DISS. ETH NO. 28168

The Empathetic Car: Detecting Emotion and Well-being of Drivers under Naturalistic Condition

A thesis submitted to attain the degree of

DOCTOR OF SCIENCE OF ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Shu LIU

Master of Science, ETH Zurich

Born on 14 February 1991

Citizen of China

accepted on the recommendation of

Prof. Dr. Elgar FLEISCH, examiner

Prof. Dr. Felix WORTMANN, co-examiner

Asst. Prof. Dr. Zimu ZHOU, co-examiner

2022

"*Think big, start small, scale fast!"*

Jim Carroll

Abstract

Vehicles are an integral part of modern life. Across the globe, the number of vehicles has witnessed an enormous increases from around 193 million in the 1970's in the last century to more than one billion today. A significant amount of the population of most developed countries, e.g., Germany and the U.S., spends up to one hour driving on a daily basis. The number of vehicles in developing countries has also experienced rapid growth in the last few decades. With their increased usage, vehicles are conceptualised as the third living space (after the home and workplace). Moreover, unlike the home and workplace, in which the introduction of new infrastructure technologies will induce potentially high costs, modern vehicles are already equipped with advanced sensors and infotainment systems that facilitate various interactions between the driver and the vehicle. As such, increasingly more researchers and practitioners view vehicles as an ideal platform into which wellness features can be engineered. In this vision, vehicles are not only a tool for transportation, but also a platform that optimises a user's psychological and physiological well-being via the monitoring and intervention of their states.

Beyond the sophisticated wellness features, one should keep in mind that road safety remains a major challenge. Globally, traffic accidents are among the top (ranked 8th) causes of mortality. In addition to the loss of human life, the economic impact is also non-negligible. It has been estimated that traffic accidents lead to a monetary cost of around 3% of the global gross domestic product (GDP). Tremendous efforts have been devoted to the development of advanced driver assistance systems (ADAS) and (semi-)autonomous driving systems. While more recently developed systems focus on the perception and interaction between the ego-vehicle and surrounding traffic participants, a relatively small portion of researchers has considered the improvement of driving safety and experience from a driver-centric perspective. Given the current development progress, autonomous driving features are mostly available only for simple scenarios on highways or under limited speed. In other words, driving safety under complex traffic conditions, such as in downtown areas or bad weather, remains heavily dependent on drivers. Therefore, the monitoring of driver's states is an essential and important approach to the improvement of driving performance, and

ultimately to the increase of driver safety.

To improve mental wellness and road safety, both academia and automobile manufacturers alike have paid increasingly more attention to the recognition of a driver's psychological state, or, be more precisely, the emotional state. Indeed, sub-optimal emotional states (e.g., anger and sadness) do not only affect a user's well-being, but are also found to be associated with risky or dangerous driving behaviours. However, the mainstream of existing approaches relies on cumbersome physiological sensors to capture data on the driver's heart rate (HR), skin conductance, and respiration rate or uses privacy-breaching cameras to record the driver's face, which reduces the acceptance of the users. To achieve a more ubiquitous and privacy-preserving emotion recognition, this thesis relies on the various advanced sensors and technology embedded in today's highly computerised vehicles. Instead of the analysis of the driver's physiological signals or driver face information, the proposed approach utilises the sensors and the technology in current vehicles that offer a comprehensive capturing of the traffic context and driving behaviours, such as heavy braking, evasive steering, and sharp turning. By leveraging machine learning techniques, the emotional states of drivers can be inferred from the traffic context and driving behaviours, thereby ultimately achieving non-intrusive and private emotion recognition.

In addition to utilising the psychological state monitoring approach that can be deployed in current vehicles, this thesis further explores the possibility of the estimation of drivers' heart rate variability (HRV) in future vehicles when driver monitoring cameras become a mandatory component. HRV and its measures are key indicators of physiological states and reflect not only a driver's health states but also whether a driver is fit for driving. Conventional methods for the measurement of HRV, such as smartwatches, electrocardiography (ECG), and photoplethysmography (PPG), suffer from various constraints, including inaccuracy and inconvenient deployment. In this thesis, a facial expression-based HRV inference approach is proposed, as facial expressions and heart activity are both controlled by the autonomic nervous system (ANS). The facial expressions of drivers can be captured via a driver monitoring camera, which is very likely to be a mandatory component of future vehicles. Driver monitoring cameras are of significant importance in level-3 (L3) and level-4 (L4) autonomous driving because the vehicle must ensure that the driver is capable of taking back control when the autonomous system fails to handle over-complicated traffic situations. The evaluation demonstrates that the proposed facial expression-based HRV inference approach is superior in the detection of the outliers of HRV and its measures as compared to conventional methods, such as smartwatches. By monitoring the physiological states of drivers, the vehicles in the future are not only capable

of improving driving safety but also open up various opportunities to facilitate the well-beings related features.

The experiments conducted in this thesis were performed based on the datasets collected from a field study consisting of a total driving distance of approximately 50,000 kilometres (by nine drivers) on public roads. This naturalistic dataset allows for an in-depth and reliable analysis of driver behaviours and facial expressions, as well as the interplay between drivers and the traffic context. The presented approaches, to the best of the author's knowledge, are the first-of-its-kind research that was evaluated under naturalistic settings and exhibit conceptual improvements over the state-of-the-art model. The promising results of the proposed approaches highlight the importance and potential of leveraging vehicle sensors to achieve a better driving experience and improved safety.

Ultimately, this thesis first utilised artificial intelligence to improve the driving experience and safety of current vehicles via the non-intrusive and privacy-preserving inference of drivers' psychological states. Furthermore, an approach for the physiological state monitoring of drivers based on driver cameras is envisioned for future vehicles. While the proposed approaches for current and future vehicles rely on different sensory settings (without and with driver cameras, respectively), they are not contradictory to each other. The transition from current vehicles to L3 or higher-level autonomous vehicles equipped with driver cameras will be a gradual process. It can be anticipated that in the upcoming future, autonomous vehicles will not be immediately available among the majority of populations, especially in those low-income regions. The proposed approaches target the challenges in the different phases of this transition, and facilitate the improvement of the driving experience during this process. In our vision, an intelligent vehicle is not only a means of transportation but is also a platform where a driver's state is optimised via various interventions, such as changes in in-vehicle lightning, temperature, and music, or mindfulness exercises. For these interventions, the recognition and the estimation of a driver's psychological and physiological states are preliminary. In line with this vision, the approaches outlined in this thesis, in essence, provide driver-centric assistance to make life safer, healthier and more comfortable.

Zusammenfassung

Das Fahrzeug ist für viele Menschen ein wesentlicher Bestandteil des modernen Lebens. Weltweit ist die Anzahl an Fahrzeugen von rund 193 Millionen in den 1970er Jahren auf zuletzt über eine Milliarde enorm gestiegen. In den meisten Industrieländern, beispielsweise in Deutschland oder den USA, verbringt ein erheblicher Teil der Bevölkerung täglich bis zu einer Stunde mit Autofahren. Darüber hinaus haben sich inzwischen auch die Menge an Fahrzeugen in den Entwicklungsländern in den letzten Jahrzehnten rasant vermehrt. Mit der zunehmenden Nutzung stellen Fahrzeuge eine Art dritten Lebensraum dar (neben Wohnort und Arbeitsplatz). Darüber hinaus sind moderne Fahrzeuge im Gegensatz zum Wohnort und dem Arbeitsplatz, wo die Einführung neuer Infrastrukturtechnologien potenziell hohe Kosten verursachen wird, bereits mit fortschrittlichen Sensoren und Infotainmentsystemen ausgestattet, die verschiedene Interaktionen zwischen Fahrenden und Fahrzeugen erleichtern. Daher sehen immer mehr Forschende und Unternehmen Fahrzeuge als ideale Plattform, in die Wellness-Funktionen integriert werden können. In dieser Vision sind Fahrzeuge nicht nur ein Transportmittel, sondern auch eine Plattform, die das psychologische und physiologische Wohlbefinden eines Benutzers durch Überwachung und Intervention seines Zustands verbessert.

Neben den ausgefeilten Wellness-Features sollte man bedenken, dass die Verkehrssicherheit eine große Herausforderung bleibt. Verkehrsunfälle gehören weltweit zu den häufigsten Todesursachen (Platz 8). Neben dem Verlust des menschlichen Lebens sind auch die wirtschaftlichen Auswirkungen nicht zu vernachlässigen. Schätzungen zufolge verursachten Verkehrsunfälle finanzielle Einbußen von rund 3 % des weltweiten Bruttoinlandsprodukts (BIP). In die Entwicklung fortschrittlicher Fahrerassistenzsysteme und (semi-)autonomer Fahrsysteme wurde ein enormer Aufwand gesteckt. Während neuer entwickelte Systeme sich auf die Wahrnehmung und Interaktion zwischen dem Ego-Fahrzeug und den umgebenden Verkehrsteilnehmern konzentrieren, haben sich ein relativ kleiner Teil der Forschenden mit der Optimierung der Fahrsicherheit und des Fahrerlebnisses aus der Fahrerperspektive beschäftigt. Beim aktuellen Entwicklungsfortschritt stehen autonome Fahrfunktionen meist nur für einfache Szenarien auf Autobahnen oder bei begrenzter Geschwindigkeit

zur Verfügung. Das heißt, die Fahrsicherheit unter komplexen Verkehrsbedingungen, wie in der Innenstadt oder bei schlechtem Wetter, bleibt stark vom Fahrer abhängig. Daher ist die Überwachung des Fahrerstatus ein wesentlicher und wichtiger Ansatz zur Verbesserung der Fahrleistung und damit zur Erhöhung der Fahrersicherheit.

Um das mentale Wohlbefinden und die Verkehrssicherheit zu verbessern, nutzen sowohl die Forschung als auch die Automobilhersteller immer stärker das Erkennen des psychischen Zustands des Fahrers, genauer gesagt des emotionalen Zustands. Tatsächlich wirken sich suboptimale emotionale Zustände (z. B. Wut und Trauer) nicht nur auf das Wohlbefinden des Benutzers aus, sondern werden auch mit riskantem oder gefährlichem Fahrverhalten in Verbindung gebracht. Der Mainstream bestehender Ansätze beruht jedoch auf umständlichen physiologischen Sensoren, die die Herzfrequenz, Hautleitfähigkeit und Atemfrequenz eines Fahrers erfassen oder Kameras nutzen, die zu sehr in die Privatsphäre des Fahrers eingreifen, was wiederum die Akzeptanz der Nutzer verringert. Um eine allgegenwärtigere und die Privatsphäre bewahrende Emotionserkennung zu erreichen, stützt sich diese Arbeit auf die verschiedenen modernen Sensoren und Technologien, die in heutigen hochcomputerisierten Fahrzeugen eingebettet sind. Anstatt physiologische Signale oder Gesichtsinformationen des Fahrers zu analysieren, nutzt der vorgeschlagene Ansatz die Sensoren und die Technologie in aktuellen Fahrzeugen, die eine umfassende Erfassung des Verkehrskontexts und des Fahrverhaltens wie starkes Bremsen, ausweichendes Lenken und scharfes Abbiegen ermöglichen. Durch den Einsatz von maschinellen Lerntechniken können die emotionalen Zustände aus dem Verkehrskontext und dem Fahrverhalten abgeleitet werden und somit eine nichtintrusive und vertrauliche Emotionserkennung erreicht werden.

Neben dem psychologischen Zustandsmonitoring-Ansatz, der in aktuellen Fahrzeugen eingesetzt werden kann, untersucht diese Arbeit die Möglichkeit der Schätzung der Herzfrequenzvariabilität (HRV) des Fahrers in zukünftigen Fahrzeugen, wenn die Fahrerüberwachungskamera eine obligatorische Komponente wird. Die HRV und ihre abgeleiteten Metriken sind Schlüsselindikatoren für physiologische Zustände und spiegeln nicht nur den Gesundheitszustand des Fahrers, sondern auch seine Fahrtüchtigkeit wider. Herkömmliche Methoden zur Messung der HRV, wie Smartwatches, Elektrokardiographie (EKG) und Photoplethysmographie (PPG), leiden unter verschiedenen Einschränkungen, einschließlich Ungenauigkeit und umständlicher

Nutzungsmöglichkeiten. In dieser Arbeit wird ein mimikbasierter HRV-Inferenzansatz vorgeschlagen, da Mimik und Herzaktivität beide vom autonomen Nervensystem (ANS) gesteuert werden. Die Mimik von Fahrern kann über eine Fahrerüberwachungskamera erfasst werden, die mit hoher Wahrscheinlichkeit ein obligatorischer Bestandteil zukünftiger Fahrzeuge sein wird. Fahrerüberwachungskameras sind im autonomen Fahren der Level 3 und Level 4 von großer Bedeutung, da das Fahrzeug sicherstellen muss, dass der Fahrer die Kontrolle übernehmen kann, wenn das autonome System zu komplizierte Verkehrssituationen nicht meistert. Unsere Evaluation zeigt, dass der mimikbasierte HRV-Inferenzansatz bei der Erkennung der Ausreißer der HRV und deren Messgrößen herkömmlichen Methoden wie Smartwatches überlegen ist. Durch die Überwachung der physiologischen Zustände des Fahrers sind die Fahrzeuge der Zukunft nicht nur in der Lage, die Fahrsicherheit zu verbessern, sondern eröffnen auch vielfältige Möglichkeiten, die Wohlfühlfunktionen zu optimieren.

Die Experimente in dieser Dissertation wurden basierend auf den Datensätzen einer Feldstudie durchgeführt, die sich aus Gesamtfahrstrecken von rund 50.000 Kilometern (neun Fahrer) auf öffentlichen Straßen zusammensetzt. Dieser naturalistische Datensatz ermöglicht eine tiefgreifende und zuverlässige Analyse des Fahrerverhaltens, der Mimik sowie des Zusammenspiels zwischen Fahrer und Verkehrskontext. Die vorgestellten Ansätze sind nach unserem besten Wissen die erste ihrer Art, die unter naturalistischen Bedingungen evaluiert wurden und den Stand der Technik konzeptionell verbessern. Die vielversprechenden Ergebnisse der vorgeschlagenen Ansätze unterstreichen die Bedeutung und das Potenzial der Nutzung der Sensoren von Fahrzeugen für ein besseres Fahrerlebnis und mehr Sicherheit.

Zusammenfassend nutzt diese Arbeit künstliche Intelligenz, um das Fahrerlebnis und die Sicherheit aktueller Fahrzeuge durch nicht-intrusive und die Privatsphäre bewahrende Rückschlüsse auf die psychischen Zustände des Fahrers zu verbessern. Darüber hinaus zeigt die Dissertation einen Ansatz, wie die Überwachung des physiologischen Zustands von Fahrern basierend auf Fahrerkameras für zukünftige Fahrzeuge aussehen kann. Die vorgeschlagenen Ansätze für aktuelle und zukünftige Fahrzeuge beruhen zwar auf unterschiedlichen Sensorikkonfigurationen (ohne bzw. mit Fahrerkamera), sie widersprechen sich aber nicht. Der Übergang von aktuellen Fahrzeugen zu autonomen Fahrzeugen des Level 3 oder höherwertigen Fahrzeugen, die mit Fahrerkameras ausgestattet sind, ist ein schrittweiser Prozess. Es ist zu erwarten, dass autonome Fahrzeuge in der kommenden Zukunft, insbesondere

in einkommensschwachen Regionen, für die Mehrheit der Bevölkerung nicht sofort verfügbar sein werden. Die vorgeschlagenen Ansätze zielen auf die Herausforderungen in den verschiedenen Phasen dieses Übergangs ab und ermöglichen die Verbesserung des Fahrerlebnisses während dieses Prozesses. In unserer Vision ist ein intelligentes Fahrzeug nicht nur ein Fortbewegungsmittel, sondern auch eine Plattform, auf der der Zustand des Fahrers durch verschiedene Eingriffe wie Beleuchtung im Fahrzeug, Temperatur, Musik oder Achtsamkeitsübungen optimiert wird. Für diese Eingriffe sind das Erkennen und die Einschätzung der psychischen und physiologischen Zustände des Fahrers notwendig. Im Einklang mit dieser Vision bieten die in dieser Arbeit skizzierten Ansätze im Wesentlichen eine fahrerzentrierte Assistenz, um das Leben sicherer, gesünder und komfortabler zu gestalten.

Disclaimer

Parts of this doctoral thesis have already been published elsewhere.

Except for a few sections and adaptations, Section 1.2, Section 2, Section 3.1 - 3.7, and Appendix A have been published at the Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies (Liu et al., 2021a). Instead of citing this doctoral thesis, we recommend to cite directly the following paper:

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". In: *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5.3, Art. no. 117:1–34.

Except for a few sections and adaptations, Section 1.2, Section 4.1 - 4.6, and Appendix B have been accepted for publication in the IEEE Internet of Things Journal (Liu et al., 2021b). Instead of citing this doctoral thesis, we recommend to cite directly the following paper:

Shu Liu, Kevin Koch, Zimu Zhou, Martin Maritsch, Xiaoxi He, Elgar Fleisch, and Felix Wortmann (2021). "Towards Non-Intrusive Camera-Based Heart Rate Variability Estimation in the Car under Naturalistic Condition". In: *IEEE Internet of Things Journal*. Accepted.

Acknowledgements

The journey of my doctoral study would not have been possible without the help and encouragement of many people. I am genuinely grateful for being a part of the Bosch IoT Lab, a collaboration between ETH Zürich, the University of St. Gallen, and the Bosch Group.

First and foremost, I would like to express my profound gratefulness to my supervisor, Prof. Dr. Elgar Fleisch, for giving me the opportunity even to start this endeavour. This is probably the most fantastic experience in my life. During this endeavour, I was given the freedom to explore edge-cutting artificial intelligence and IoT technologies and apply them to solve various human-centric challenges in the real world. In addition, Elgar has created an inspiring and rewarding work environment that allowed me to dive into timely research in an unparalleled cooperative and vibrant manner. From the discussion and the interaction with Elgar, I learned how rigorous research leads to in-negligible impact in practices and that a successful endeavour is usually guided by a insightful vision.

Next, I want to express my ultimate thanks to Prof. Dr. Felix Wortmann for his countless support during my work and personal development. I am very grateful for his encouraging character. Without his support and patience, it would be impossible for me to overcome the most difficult time during my doctoral study. His resilient optimism was a constant source of motivation during my doctorate. His out-of-the-box thinking and ideas have enabled me to conduct various first-of-its-kind research. Furthermore, Felix's precise advice about scientific writing, invaluable systematic understanding of internet-of-things and inspiring way of critical thinking have helped me structuralise my domain knowledge and build up the fundamental skills in research. Thank you Felix for everything!

In addition, I would like to thank Dr. Wolfgang Bronner and Timo Gessmann, who both supported me and provided me with valuable connections to industries. Likewise, I feel very grateful to the Robert Bosch Group for financing this research lab and all the Bosch colleagues who provided industrial insights to guide my research. Their dedication to innovation and social responsibility have been a role model not

only for my research but also for my career life in the future. I would also like to thank Prof. Dr. Zimu Zhou for inspiring me in many ways. As an expert in mobile and ubiquitous computing, he helps me open up my mind in this fantastic domain. My collaboration with Zimu further enabled me to learn and to grow. I truly appreciate for our intensive discussions and for his guidance in developing novel and fancy scientific ideas and bringing them to paper.

I am also thankful for countless help from Elisabeth Vetsch-Keller, Monica Heinz, Judith Holzheimer, Ursula Elsässer, Sonja Baumgartner, and Jörg Klaus. It is their effort and commitment that supported us doctoral students to solve various allegedly unsolvable problems.

I believe that it is of particular importance to emphasise that the work presented in this thesis results from the group effort. First, I am indebted to Kevin Koch, who accompanied me on my endeavour throughout my doctorate and was a collaborator on all papers. His reliable skills in management and leadership have ensured the successful execution of various projects that involved many uncertain factors coming from real-world experiments, communications with ethic committees, and multilateral collaborations between industries and universities. I also would like to extend my thanks to other colleagues within our connected car team. I am grateful to Simon Föll for his critical thinking and excellent academic writing that help me publish my first major publication at Bosch IoT lab. Also, I would like to thank Martin Maritsch, who helped me formulate the fundamental idea in driver physiological state estimation, which led to my second major publication at the lab. I am also thankful to Dr. Bernhard Gahr and Dr. Benjamin Ryder for introducing me to this lab. Their valuable suggestions have helped me in many ways throughout the whole doctoral study.

Finally, it was not only the fantastic projects but also various experiences during recreational time, in-depth conversation beyond academia, and team event and chair activities that made my doctoral study so memorable. I am deeply grateful for the enjoyable time I was privileged to spend with colleagues and friends at Elgar' chairs and Bosch Digital Solutions team in Zurich, including Dr. Liliane Ableitner, Caterina Bérubé, Dr. Dominik Bilgeri, George Boateng, Dr. Thomas von Bomhard, Tobias Brudermüller, Raquel Brüngger, Dr. Mathieu Chanson, David Cleres, Dr. André Dahlinger, Philipp Etschel, Dr. Remo Frey, Dr. Klaus Fuchs, Dr. Johannes Hübner, Dr. Alexander Ilic, Robert Jakob, Sven Jung, Dr. Cristina Kadar, Moritz Kaminski, Kevin Koch, Prof. Dr. David Kotz, Prof. Dr. Tobias Kowatsch, Dr. Jan-Niklas Kramer, Florian Künzler, Claudio Lamprecht, Yanick Lukic, Dr. Arne Meeuw, Prof. Dr. Varun Mishra, Dr. Daniel Müller, Kevin O'Sullivan, Dr. Dominik Rüegger,

Prabhakaran Santhanam, Dr. Sandro Schopfer, Fabian Schäfer, Dr. Iris Shih, Prof. Dr. Thorsten Staake, Dr. Funk Te, Gisbert Teepe, Prof. Dr. Verena Tiefenbeck, Dr. Peter Tinschert, Dr. Denis Vuckovac, Eva van Weenen, Fan Wu, Dr. Anselma Wörner, and Dr. Dominic Wörner.

Last but not least, I am greatly indebted to my parents Lingling Li and Ligong Liu and my marvellous sister Yi Liu for their selfless love, inspiration and guidance throughout my life. Without their support, I would not have been able to come this far. I also feel lucky to have shared several years of study in Europe with my beloved wife Yayin Su, whose encouragement, love and energy have helped me overcome all kinds of difficulties to finalise this thesis.

To my family.

Contents

Contents (detailed)

List of Figures

List of Tables

List of Abbreviations

ADS	**A**utomated **D**riving **S**ystem
ADAS	**A**dvanced **D**river-**A**ssistance **S**ystem
AI	**A**rtificial **I**ntelligence
ANS	**A**utonomic **N**ervous **S**ystem
AU	**A**ction **U**nit
CAD	**C**omputer-**A**ided **D**esign
CAN	**C**ontrol **A**rea **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
CPU	**C**entral **P**rocessing **U**nit
DF	**D**eep **F**orest
ECG	**E**lectro**c**ardio**g**raphy
EDA	**E**lectro**d**ermal **A**ctivity
EEG	**E**lectro**e**ncephalo**g**raphy
FACS	**F**acial **A**ction **C**oding **S**ystem
FPS	**F**rame **P**er **S**econd
GDP	**G**ross **D**omestic **P**roduct
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
GPS	**G**lobal **P**ositioning **S**ystem
GPU	**G**raphics **P**rocessing **U**nit
GRU	**G**ated **R**ecurrent **U**nit
HAR	**H**uman **A**ctivity **R**ecognition
HR	**H**eart **R**ate
HRV	**H**eart **R**ate **V**ariability
IBI	**I**nter-**B**eat **I**nterval
IoT	**I**nternet **o**f **T**hings
JITI	**J**ust-**I**n-**T**ime **I**ntervention
LED	**L**ight-**E**mitting **D**iode
LF/HF ratio	**L**ow **F**requency and **H**igh **F**requency **R**atio
LOSO	**L**eave-**O**ne-**S**ubject-**O**ut
MLP	**M**ulti-**L**ayer **P**erceptron
NHTSA	**N**ational **H**ighway **T**raffic **S**afety **A**dministration

NMVCCS	**N**ational **M**otor **V**ehicle **C**rash **C**ausation **S**urvey
OEM	**O**riginal **E**quipment **M**anufacturer
pNN50	**p**roportion of successive **N**ormal beat to **N**ormal beat intervals with a difference greater than **50** ms
PPG	**P**hoto**p**lethysmo**g**raphy
RMSSD	**R**oot **M**ean **S**quare of **S**uccessive **D**ifferences between inter-beat intervals
RF	**R**andom **F**orest
RNN	**R**ecurrent **N**eural **N**etwork
rPPG	**r**emote **P**hoto**p**lethysmo**g**ram
SNS	**S**ympathetic **N**ervous **S**ystem
SVM	**S**upport **V**ector **M**achine
TJP	**T**raffic **J**am **P**ilot
TPFN	**T**ree-based **P**robabilistic **F**usion **N**etwork
UWB	**U**ltra-**W**ide **B**and
V2X	**V**ehicle-to-everything
WHO	**W**orld **H**ealth **O**rganization

Chapter 1

Introduction

"Bosch's approach is to envisage the vehicle as the '3rd living space', a private space in which you feel comfortable and like to spend time. This will be realised through new control concepts and features that are not feasible in today's cars."

Robert Bosch GmbH, 2016

1.1 Context and Motivation

Driving is an inseparable part of many people's daily lives. In the U.S., for example, about 90% of the population (aged 16 or older) drove 2.5 trips daily from 2019-2020 on average, which corresponds to about one hour of driving time or a distance of approximately 50 *km* of distance per day (American Automobile Association Foundation for Traffic Safety, 2020). In Germany, approximately 60% of employees use their car for commuting, among which over 25% of them spend at least 30 minutes per direction (Statistisches Bundesamt, 2016). Therefore, the industry imagines that vehicles are being transformed into people's third living space (after the home and workplace) of people (Robert Bosch GmbH, 2016). With the help of advanced information technologies, many new features that can improve the comfort, well-being, and safety of drivers are expected to be integrated into such a transformation. Among these features, two functionalities of our particular interests are the monitoring of a driver's psychological and physiological states, as the modern life style and the ageing population have had serious negative impacts on people's wellness. In the following, the significance and the potential of psychological and physiological state monitoring in vehicles are introduced. After that, we motivate furthermore from road safety perspective that optimising driver states is a critical measure to improve safe driving.

Driver Mental Well-being. The prolonged driving time in modern life has a profound impact on people's mental states, which, in turn, influences the driving experience and road safety. First, during daily driving tasks, people often inevitably encounter frustrating situations such as traffic jams, unreliable navigation, and unfriendly traffic participants. Such events are often the triggers of negative emotions such as anger, stress, and sadness (Chatterjee et al., 2020; Legrain, Eluru, and El-Geneidy, 2015; Zepf et al., 2019). Subsequently, the accumulated negative emotions and stress adversely affect the driving experience and further cause the immediate impairment of driving performance. For example, *anger* and *sadness* can lead to risky and degraded driving performance (Chan and Singhal, 2015; Underwood et al., 1999). Therefore, the timely recognition and regulation of such negative emotions is an effective way to restore drivers into their optimal mental states and hence ultimately reduce the possibility of human errors.

From a broader perspective, in-vehicle emotion recognition and regulation have derivative benefits for improving mental health, which has escalated to a major concern of the public health sector (Kohn et al., 2004). With the accelerated pace of life and increased work stress, the number of cases of psychological illness has witnessed an astonishing increase over the last decade. For example, major depressive disorders and anxiety disorders have increased by 18.4% to 320 million cases per year and by 14.9% to 264 million cases per year, respectively (Vos et al., 2016; World Health Organization, 2017). In addition to the lowered quality of life, the economic losses due to these illnesses are uncountable due to high treatment costs and reduced productivity (Greenberg et al., 2015). Large scale conventional dedicated preventive interventions and treatment are needed to combat increasing mental illness, which, however, exceeds the available capacity of current medical systems (Ebert et al., 2017). Therefore, researchers and experts have called for light-weighted novel interventions that can be performed ubiquitously and provide the advantage of being independent of both time and place (Ebert et al., 2017; Nahum-Shani, Hekler, and Spruijt-Metz, 2015). As such, cars are considered to have a great potential in the monitoring and improvement of people's mental states because the in-vehicle setting is a unique environment where commuters spend a considerable amount of time.

Driver Physiological Health. Health and well-being are important aspects of people's daily lives. In particular, driving is a demanding task that requires full attention and optimal performance from the driver, especially in high-speed dynamic scenarios (Coughlin, Reimer, and Mehler, 2011). Empirical statistics show that inattentiveness, drowsiness, and fatigue constitute one of the main factors of traffic accidents (Choi et al., 2016). The situation is even more serious for the ageing driving population,

especially in those high income countries. Therefore, it is of vital importance to promptly recognise such sub-optimal states of drivers. Heart rate variability (HRV) and its measures are very informative physiological signals for the recognition of driver states, and have been proven to be relevant to stress, drowsiness, and inattentiveness. Indeed, a substantial amount of efforts in both academia and industry has been devoted to the monitoring of the heart activities of drivers. However, the existing solutions, such as smartwatches, remote photoplethysmogram (rPPG), the ultra-wide band (UWB) technique, and ballistics-based methods, etc., suffer from various constraints including the mediocre accuracy, inconvenient deployment, and the lack of ubiquity, which prevents their practical usage (D'Angelo et al., 2010; Stricker, Müller, and Gross, 2014; Wartzek et al., 2011; Zheng et al., 2020). Therefore, a ubiquitous and reliable HRV monitoring technique is yet to be invented to ensure optimal driving performance, reduce human errors, and improve driving safety has yet to be invented.

In addition, the monitoring of HRV measures provides a significant extra benefit from the perspective of well-being. Heart activities are not only an indicator of driving performance, but also reflect multiple psychological and physiological states. For example, occupational burnout, depression, and mood disorders, etc., can lead to sub-optimal states and manifest as changes in HRV measures changes (Holzman and Bridgett, 2017; Lo, Wei, and Hwang, 2020; Malik et al., 1996). A vehicle capable of monitoring HRV can be combined with well-being interventions that restore the sub-optimal psychological and physiological states of drivers (Koch et al., 2021; Lee, Elhaouij, and Picard, 2021) due not only to driving events, but also to other daily activities.

Road Safety. In 2019 the World Health Organisation (WHO) and the Global Health Estimates reported that road traffic accidents are the major cause of mortality of the young population aged between 15 and 29 years (World Health Organization, 2019). While various efforts, such as better infrastructures and more advanced driver assistance systems, have successfully reduced the rate of road traffic deaths by almost two-fold in the past years (from 115 deaths / 100,000 vehicles in 2006 to 64 death / 100,000 vehicles in 2016), the continuous growth of vehicle usage has neutralised the advantage of the reduced rate of road traffic deaths. In the same period from 2006 to 2016, the number of vehicles worldwide has increased from approximately 1 billion to 2.1 billion. As a result, the road traffic fatality has increased steadily, from around 1.25 million in 2013 to 1.35 million in 2016 (World Health Organization, 2015; World Health Organization, 2018). In addition to the significant humanitarian concerns of the globally increasing traffic fatalities, each year the non-fatal traffic

casualties reach about 50 million (World Health Organization, 2015). WHO has estimated that the economic cost of traffic accidents constitutes approximately 3% of the global gross domestic product (GDP) (Gahr et al., 2019; World Health Organization, 2018).

Among various factors, driver errors are the most significant critical reason for traffic crashes. Based on the National Motor Vehicle Crash Causation Survey, conducted in the U.S. from 2005 to 2007, the National Highway Traffic Safety Administration (NHTSA) summarised that approximately 94% of the traffic crashes were attributed to driver errors (National Highway Traffic Safety Administration, 2015). A rising technology to combat the consequence of human errors is the artificial intelligence facilitated semi-, and fully-autonomous driving. With the advancement of sensing and decision-making techniques, vehicles in the future are promised to overtake the driving task for humans. Nevertheless, the development of automated driving system (ADS) is a gradual process. Despite progressive achievements in autonomous driving techniques, only a few vehicles in the market have level-3 (L3) autonomy, and the current features of L3 autonomy are constrained to a very limited number of use cases. For instance, the Traffic Jam Pilott (TJP) from Audi can be operated at up to 60 km/h on highways with a longitudinally-divided traffic flows (Autovista Group, 2020). Moreover, the Drive Pilot from Mercedes is only usable in dense traffic, tailbacks, or on appropriate sections of the motorway in Germany (Patel, 2020). The NHTSA of the U.S. envisions fully automated safety features for highway autopilot only after 2025 (National Highway Traffic Safety Administration, 2020a). Traffic on highways represents a relatively simple scenario, whereas many complicated situations still require manual control. Therefore, it is justified to hypothesise that, in the near future, vehicles will still require human operation in many scenarios when, for example, the vehicle is outside predefined level-4 (L4) regions or in complicated scenarios when the system requires the driver to take control. Moreover, autonomous cars will not be directly widely available; it may take years or decades until manual vehicles are fully replaced. This progress may take even longer in developing countries, as the envisioned ADS might involve an upgrade of not only vehicles, but also infrastructures (Chen et al., 2019a). Furthermore, the barrier to the deployment of ADS originates not only from the technical perspective, but also from the legal perspective. For example, Audi postponed the release of L3 autonomous features due to legal structures and model life cycles related issues (Autovista Group, 2020); it remains unclear whether the L3 autonomous features of the Mercedes-Benz S-Class will be available in the US market at the same time as they will be in the European market (Patel, 2020). In summary, drivers will continue to play a key role in traffic; compared with autonomous driving systems that completely overtake driving tasks,

the approaches that reduce human errors can bring immediate benefit. As such, we believe that the monitoring of the psychological and physiological states of drivers is of significant importance from the road safety perspective.

1.2 Research Objective and Approach

Today's cars are highly computerised with a rich sensor set that captures the interaction between the driver and the vehicle as well as the surrounding traffic context. These sensors collect a large amount of information about the driver's states. Nevertheless, there is a gap in both academia and industry regarding the potential utilisation of such vehicle data for the inference of driver states. This motivated the two research objectives that this thesis sets out to investigate. First, it is investigated whether a driver's emotions can be inferred based on sensors that are available in current vehicles without causing inconvenience and compromising the driver's privacy. The second research question aims at whether we can further estimate a driver's physiological state in a non-intrusive manner, given the additional sensors that will be installed in future vehicles. In the following subsections, these two objectives are described and elucidated in higher granularity. Furthermore, an overview of the overall target of this thesis is provided.

1.2.1 Inference of Driver Emotions

An empathetic car that is capable of reading the driver's emotions has been envisioned by many car manufacturers. Emotion inference enables in-vehicle applications to improve driver comfort, well-being, and safety. Available emotion inference approaches use physiological, facial, and speech-related data to infer emotions during driving trips. However, existing solutions have two major limitations: First, the reliance on sensors that are not built into the vehicle restricts the inference of emotions to those people leveraging corresponding devices, for example smartwatches, electrocardiography (ECG) devices, or electroencephalography (EEG) devices. Second, the utilisation of modalities such as facial expressions and speech raises privacy concerns. By contrast, researchers in mobile health have been able to infer affective states (e.g., emotions) based on behavioural and contextual patterns decoded in available sensor streams, e.g., those obtained by smartphones (Canzian and Musolesi, 2015; LiKamWa et al., 2013; Zhang et al., 2018). In this way, no additional sensors are needed and users' privacy is better protected. In this thesis, the rationale from the mobile health domain is transferred to an in-vehicle setting by analysing the

feasibility of inferring drivers' emotions by passively interpreting the data streams of vehicles, leading to the first research question (RQ) of this thesis:

RQ 1: *Can a driver's emotions be inferred based on the control area network (CAN-bus) and the traffic context (from the front-view camera or on-board radar system)?*

To verify the idea that resided in this question, we conducted a four-month field study on public roads (around 50,000 km of driving data from nine drivers) covering a variety of uncontrolled daily driving activities. Given the naturalistic setting of the field study, the results of the experiments were generated beyond the confines of a laboratory environment and can provide more reliable insight into the feasibility of driver emotion inference in the wild.

1.2.2 Estimation of Driver Heart Rate Variability

Driver states monitoring systems will be a vital component of smart cars in the future, especially in the era when (semi-)autonomous driving vehicles will need to ensure a driver's capability to take back control of the vehicle when necessary. The heart rate (HR) is one of the most important physiological signals concerning the driver's state. To infer the HR of drivers, the mainstream of existing research has focused on capturing the subtle heartbeat-induced vibration of the torso, or has leveraged photoplethysmography (PPG) to detect cardiac cycle-related blood volume changes in the microvascular. However, existing approaches rely on dedicated sensors, which are expensive and cumbersome to be integrated, or are vulnerable to ambient noise. Moreover, their performance on the detection of HR does not guarantee a reliable computation of HRV measures, which are more applicable metrics for the inference of mental and physiological states. The accurate computation of HRV measures is based on the precise measurement of the inter-beat interval (IBI), which can only be accomplished by medical-grade devices for which electrodes are attached to the body. These existing challenges limit the utility of driver health monitoring in the real world, and hence lead to the second research question investigated in this thesis:

RQ 2: *Can a driver's HRV be reliably estimated in a non-intrusive manner (via a driver monitoring camera) in future vehicles?*

To address this question, a facial expression based HRV estimation approach is proposed. The rationale behind this approach is to establish a link between facial expressions and heartbeat, as both are controlled by the autonomic nervous system (ANS). Moreover, it should be noted that driver state monitoring systems in future cars will

most likely be realised by driver monitoring cameras. Therefore, the proposed approach does not introduce additional sensors, nor does it impose any inconvenience to drivers, and thus has a high degree of ubiquity. The experiments conducted to investigate *RQ 2* were based on the same study data used for the *RQ 1*; however, only a subset (two weeks) of the data was used, as the HRV data were not recorded throughout the entire field study. To further improve facial expression-based HRV estimation, a tree-based probabilistic fusion neural network approach is proposed and respectively compared with the conventional random forest and neural network methods and the PPG-based measurements used in smartwatches.

It should be noted that, unlike in-vehicle emotion recognition that relies only on the existing sensors in current vehicles, the proposed HRV estimation approach features a future application scenario in which driver monitoring cameras will be a prevalent or even mandatory component of vehicles. Therefore, this thesis targets different phases of the transition from current vehicles to future (semi-)autonomous driving vehicles.

1.3 Structure of the Dissertation and Contribution Statement

The remainder of this dissertation is organised around the two research questions and is structured as follows.

Chapter 2 presented the foundational dataset that is used in this work, and describes the properties of the participants of the experiments and the characteristics of the driving data. This dataset was collected by my colleague Kevin Koch.

Chapter 3 focuses on the *RQ 1*, exploring driver emotion inference via driving behaviour and traffic context. This chapter is based on our ACM IMWUT 2021 publication (Liu et al., 2021a).

Chapter 4 focuses on the *RQ 2*, investigating the possibility of non-intrusive camera-based HRV estimation in vehicles. This chapter is based on our latest manuscript, which has been accepted for publication in the IEEE Internet of Things Journal (Liu et al., 2021b).

Chapter 5 finally summarises the key discoveries, the implications for both researchers and practitioners, and the significance of the proposed monitoring approaches in the up-coming era of more intelligent vehicles. Furthermore, an outlook is provided that outlines potential work that could be performed and research topics that could be

investigated to further improve driver safety and the driving experience in the future.

As stated previously, the contributions presented in this dissertation are partially based on the work of other researchers, and primarily consist of the study design, application for the approval of the field study by the ethics committee, the replication of the state-of-the-art algorithm, and the collection of dataset. All collaborators mentioned previously are acknowledged with the co-authorship in the related publications. In Section 1.4, a comprehensive list of all other co-authors is given. These co-authors have provided intensive discussion and ingenious feedback on the study design, the argumentation of concepts, algorithm implementation, and the literature review. Apart from that, unless specifically mentioned, the contribution described in this dissertation is originally from the author.

1.4 List of Publications

The research of this dissertation have been presented as the following publications, which constitute the fundamental contribution of this dissertation.

1. **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". In: *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5.3, Art. no. 117:1–34.

2. **Shu Liu**, Kevin Koch, Zimu Zhou, Martin Maritsch, Xiaoxi He, Elgar Fleisch, and Felix Wortmann (2021). "Towards Non-Intrusive Camera-Based Heart Rate Variability Estimation in the Car under Naturalistic Condition". In: *IEEE Internet of Things Journal*. Accepted.

In addition, the publications that were part of my PhD research, but are outside the scope of this dissertation, are listed below:

3. Kevin Koch, Verena Tiefenbeck, **Shu Liu**, Thomas Berger, Elgar Fleisch, and Felix Wortmann. "Taking Mental Health & Well-Being to the Streets: An Exploratory Evaluation of In-Vehicle Interventions in the Wild". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Art. no. 539:1–15.

4. Kevin Koch, Varun Mishra, **Shu Liu**, Thomas Berger, Elgar Fleisch, David Kotz, and Felix Wortmann (2021). "When Do Drivers Interact with In-vehicle Well-being Interventions? An Exploratory Analysis of a Longitudinal Study on Public Roads". In: *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5.1, Art. no. 19:1–30.
5. Kevin Koch, **Shu Liu**, Thomas Berger, and Felix Wortmann (2020). "Towards the Healing Car: Investigating the Potential of Psychotherapeutic In-vehicle Interventions". In: *European Conference on Information Systems*. Research-in-Progress Papers, pp. 1–8.
6. **Shu Liu**, Kevin Koch, Bernhard Gahr, and Felix Wortmann (2019). "Brake Maneuver Prediction–An Inference Leveraging RNN Focus on Sensor Confidence". In: *IEEE Intelligent Transportation Systems Conference*, pp. 1415–1420.
7. Bernhard Gahr, Katherine Caves, Junhan Wen, Kevin Koch, **Shu Liu**, and Felix Wortmann (2019). "The Costs of Traffic Accident Hotspots". In: *IEEE Intelligent Transportation Systems Conference*, pp. 883–888.
8. Bernhard Gahr, **Shu Liu**, Kevin Koch, Filipe Barata, André Dahlinger, Benjamin Ryder, Elgar Fleisch, and Felix Wortmann (2019). "Driver Identification via the Steering Wheel". In: *arXiv*, arXiv:1909.03953.

Chapter 2

Foundational Materials and Methods

Our analysis is based on a four-mouth field study during which a variety of empirical sensory data were collected from participants during daily drives. The field study involved nine participants (originally 10; data from one participant were corrupted, and were thus removed), and lasted from July 4 to November 5, 2019. Data collection was approved by the university ethics committee of University of Bern prior to starting the study.

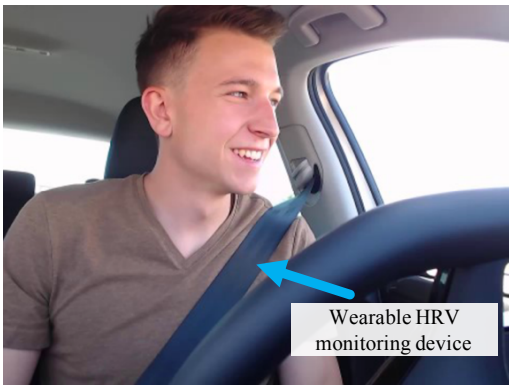
2.1 Participants

We recruited nine (four females and five males, mean age, 37 ± 8 years) participants using an internal call in an enterprise with more than 1,000 employees. Our selection followed the idea of recruiting ordinary daily commuters. They were selected to represent a large variety of people (purposive sampling). Two participants were single and eight were married. Three had children and two had pets. The preferred activities while driving included making phone calls, listening to music or the radio, and talking to other occupants of the car. We assigned each participant the same type of vehicle (with modifications for data collection as described below). The participants were supposed to use the vehicles for their daily drives, including business trips and vacations.

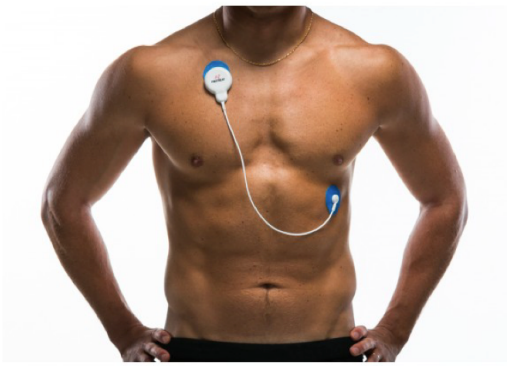
2.2 Data Collection Equipment and Protocol

Our hypothesis is that the driver's emotion can be inferred from driving behaviours and traffic contexts, which can be measured in turn by the vehicle's control area network (CAN-bus) and front-view cameras, respectively. The emotion labels based

on ground-truth facial expressions were captured by another camera mounted on the dash-board of the vehicle. Because the state-of-the-art affect recognition schemes (Healey and Picard, 2005; Nardelli et al., 2015; Schmidt et al., 2019) rely on physiological sensors, we also collected physiological data of the participants for comparison.



(A) Participant in vehicle



(B) Heart monitoring device - Firstbeat (Firstbeat Technologies Oy, 2019)

(C) CAN data collection device.

(D) Webcam deployment

FIGURE 2.1: Experimental setup

We accessed the CAN-bus data via a PCAN-USB Pro FD-Adapter (PEAK-System Technik GmbH, 2019). Two webcams (Logitech HD Pro Webcam C920) were mounted on the dashboard of the vehicle to record videos of the traffic context [1] as well as the driver's facial expressions. The CAN-bus and video data streams were controlled by an industrial-grade embedded computer (Compulab IOT-GATE-IMX7), and were stored locally in the vehicles on external hard disks. When the

[1] We used a separate camera for this, and did not rely on the CAN-bus-based radar or camera systems, to make our analysis more flexible as both systems in the car had only a limited feature set available.

vehicle was started, the computer initialised the recordings of both types of data. We collected 49 CAN signals, including those for the speed of the four wheels, accelerator position, angle of the steering wheel, and brake pedal pressure. A complete list of CAN signals is shown in Table A.1 in Appendix A. Figure 2.1c and Figure 2.1d show setups for collecting the CAN-bus data, data from the front-view camera, and data from videos of the driver's face.

In line with previous emotion recognition studies that used physiological data (Nardelli et al., 2015; Schmidt et al., 2019), we collected the heart rate (HR) and the heart rate variability (HRV) using a heart monitoring device (Firstbeat Bodyguard 2), as shown in Figures 2.1a and 2.1b. All participants were asked to wear the device for the two weeks of the field study [2]. The sampling rate of the HR and HRV of the heart monitoring device was 1000 Hz. For the sake of a more comprehensive comparison participants also wore a recent consumer smartwatch (Garmin vívoactive 3) during driving.

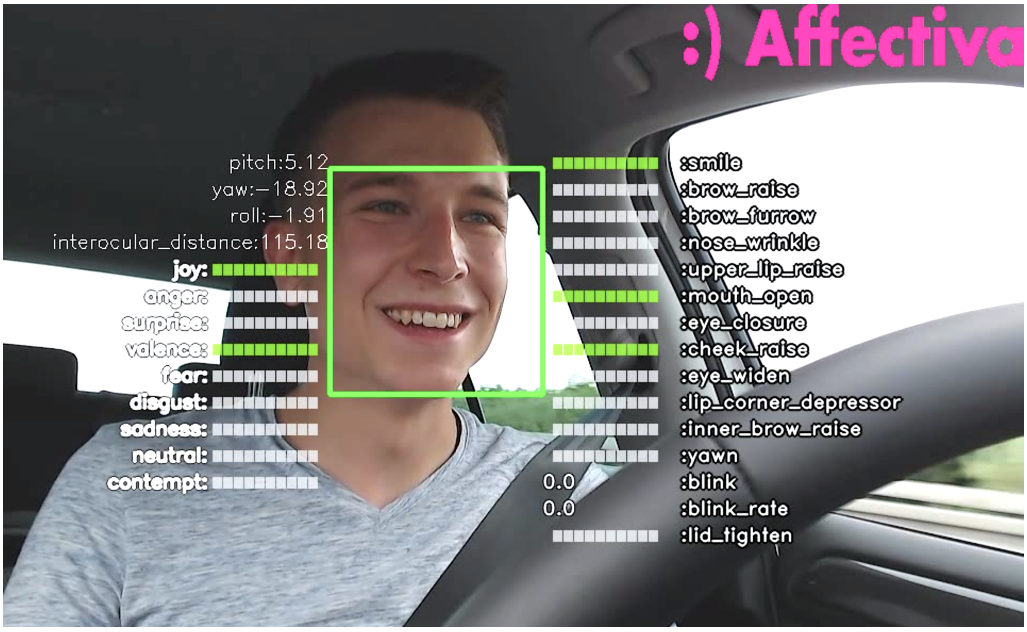


FIGURE 2.2: An example of Affectiva annotation

2.3 Characteristics of Driving Data

It was crucial for our dataset to capture representative driving situations. This subsection presents some important statistics related to our dataset.

[2] We recorded the heart rate signals of the participants only in the first two weeks because wearing the heart monitoring device for a prolonged time may cause discomfort.

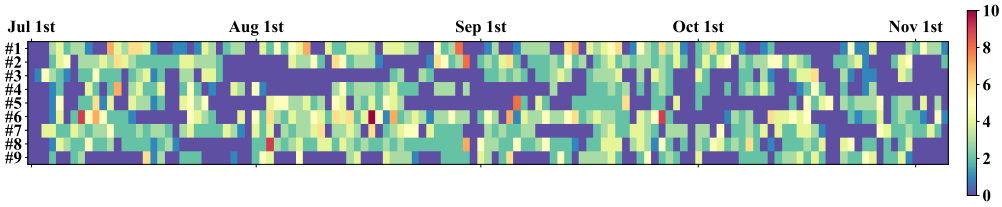
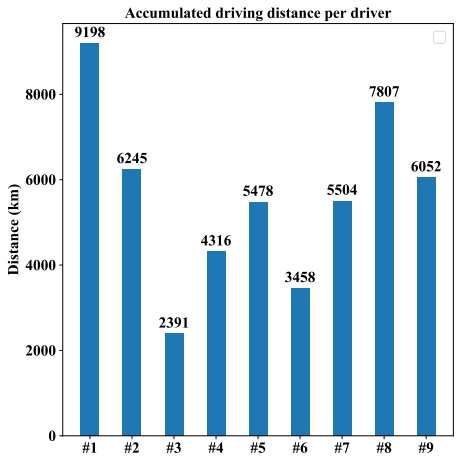
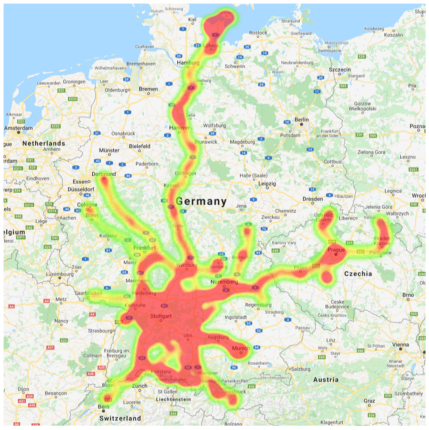


FIGURE 2.3: Number of trips per driver and per day



(A) Total driving distance per driver



(B) GPS heatmap of trips

FIGURE 2.4: Spatial coverage of driving trips

After data cleansing, we had around 675.6 hours of driving data with videos from front view camera as well as videos of the driver's face. Figure 2.3 plots the number of trips[3] per day for each participant during the field study. The overall average number of trips per day was greater than one, and the participants had up to 10 trips per day. Most participants were working during the period of our field study, except drivers #3 and #9, who had taken vacations in July and August. The total driving distances of each participant are plotted in Figure 2.4a. Most drivers drove for reasonably long distances (more than 3000 *km*) during the field study. The average trip length and duration were around 25 *km* and 23 *min*, respectively. The GPS records of the vehicles are presented as a heatmap in Figure 2.4b. As is shown there, most participants drove around the area of Stuttgart, Germany. A few long-distance trips were also taken to Prague and northern Germany.

Overall, our dataset covered a wide range of daily driving activities. All participants

[3]A trip was defined as a segment of continual driving behaviour without a pause longer than 10 minutes.

were active in terms of the number of trips per day and driving distance. The coverage and diversity of the dataset ensured that our experiments were generalisable to heterogeneous situations and drivers.

Chapter 3

Driver Emotion Inference

The previous chapters have outlined the motivation for driver state monitoring, and the potential that psychological state recognition brings to this increasingly popular field. The data collection system and detailed information about participants and driving statistics have been described and discussed. While the results of existing works have shown enormous and increasing enthusiasm in driver emotion recognition, a major disadvantage of them are their intrusive (relying on cumbersome physiological sensors) and privacy breaching (relying on driver face information) property. As such, this chapter focuses on the first research question of this thesis:

RQ 1: *Can a driver's emotion be inferred based on the control area network (CAN-bus) and the traffic context (from the front-view camera or on-board radar system)?*

The emotion recognition based on CAN-bus data stream and traffic context does not introduce inconvenience to the drivers and can better protect privacy, and hence targets the key limitation of existing works. In the following of this chapter, we start by revisiting the motivation of non-intrusive and privacy preserving driver emotion recognition. In addition, more information about the definition and the annotation of emotions, as well as recent advancements in emotion recognition is provided. The data preprocessing and transformation that are specific to this research question are described and enable deeper insight into the underlying challenge of this topic. Subsequently, the methodology and the results are presented, which are followed by a discussion and outlook that may inspire researchers in the future.

3.1 Context and Motivation

The aim of affective computing is to recognise and adapt to the affective state of the user. Examples of its implications include reducing the user's frustration through

adaptive and comfortable communication or just-in-time interventions (Sarker et al., 2014). Driving is often a cause of stress, and is associated with cognitive burden (Stutzer and Frey, 2008). Traffic, driving task, and navigation systems, etc., frequently constitute sources of negative emotions during driving (Chatterjee et al., 2020; Legrain, Eluru, and El-Geneidy, 2015; Zepf et al., 2019) and hence lead to a sub-optimal mental states while driving than other daily tasks (Chatterjee et al., 2020; Kahneman et al., 2004). The accumulated cognitive load and negative affective states do not only have negative consequences for the drivers' physiological well-being (Stutzer and Frey, 2008), but can also cause the immediate impairments of driving performance. For instance, *anger* and *sadness* are found to be associated with risky and degraded driving performance (Chan and Singhal, 2015; Underwood et al., 1999) and a *positive valence* is found to be correlated with better steering behaviours (Trick, Brandigampola, and Enns, 2012). It is of vital importance to detect negative or stress-related emotions, such as *anger*, *disgust*, *fear*, and *sadness*, as well as positive emotions such as *joy* and *valence*. In the environment of a car, technologies that can infer the driver's emotions can help improve their comfort, well-being, and safety. Thus, detecting the driver's emotional state is an important part of the vision for the car of the future. In pursuit of this vision, OEMs (e.g., BMW, KIA and Mercedes–Benz (Corby, 2017; KIA, 2019; Mcmanus, 2020b)) and start-up companies (e.g., Affectiva (Mcmanus, 2020a)) have already taken initial steps.

However, past research on in-vehicle emotion recognition has yielded two major limitations that hinder the large-scale integration of this technology into the car (Braun, Weber, and Alt, 2020). First, available solutions such as the ENERGIZING COACH, introduced by Mercedes–Benz (Corby, 2017), rely on smartwatches that incur additional cost, and hence are still not extensively used. Hence, such approaches are limited in their scalability. Second, existing work has focused on emotion recognition via facial expressions or speech, which particularly compromises drivers' privacy (Zepf et al., 2020). This substantially differs from the approach considered here because we rely on the CAN-bus and front-view camera data. The CAN-bus is a standard for communication among in-vehicle sensors, controllers, and actuators, and contains detailed information about driving behaviours such as steering, braking, and accelerating. Front-view cameras can easily be mounted on vehicles to record videos that capture the ambient traffic environment, and are already built into the latest generation of cars. Hence, approaches based on CAN-bus and front-view camera can be applied to existing vehicles without requiring expensive, special-purpose hardware.

Our approach draws on an analogy between recent advancements in generic emotion

recognition via *user behaviour* and *the contexts of application* to enable in-vehicle emotion inference. Many studies have shown the feasibility of inferring a users' emotions from the patterns of his/her smartphone usage (LiKamWa et al., 2013; Lu et al., 2019; Zhang et al., 2018). Emotion-induced behavioural patterns are highly correlated with context (e.g., dining, working, and entertainment), which can improve the accuracy of emotion recognition (Gjoreski et al., 2016; Mishra et al., 2018). Because driving is a unique activity with predefined rules and interactions that differ from what has been investigated in previous studies (LiKamWa et al., 2013; Lu et al., 2019; Zhang et al., 2018), we apply the concept of re-purposing the available sensor modalities to develop inference models targeting behavioural and context-based emotion recognition. In summary, the novelty of our approach is in applying machine learning to CAN-bus and front-view video data streams to reliably detect the emotions of drivers while minimising privacy-related concerns. The *RQ 1* is then divided into the following two sub-research questions:

- *RQ1.1: To what extent can the emotions of drivers be inferred based on (a) CAN-bus data streams, (b) the front-view camera, and (c) a combination of both (fusion)?*
- *RQ1.2: How much improvement does emotion recognition based on vehicle data offer compared with state-of-the-art methods based on physiological sensors?*

In successfully answering these research questions, the main contributions and results of this study can be summarised as follows:

- We conducted a four-month field study involving nine participants to collect various (CAN-bus, front-view camera, driver facial camera and physiological sensors) empirical sensory data during uncontrolled daily driving activities on public roads. In total, we collected valid data on 675.6 hours driving data made by nine participants covering various scenarios.
- Recording CAN-bus and video data requires pre-processing in order to use them in machine learning pipelines. We outline a comprehensive pre-processing and feature engineering pipeline for both kinds of data. We comprehensively summarise important features for time-series and video data as the basis of the classification algorithms.
- We develop an emotion classification algorithm that can process and classify CAN-bus and video data streams as well as fuse them. Based on either kind

of CAN-bus or video data, our algorithm can detect the emotions[1] based on facial expressions with an average macro F1-score of around 70% in user-dependent settings, and around 60% in user-independent settings. The results of our experiments showed that the fusion of the two modalities can further improve the performance.

- While the methods based on physiological sensors are the most prevalent among in-vehicle emotion recognition, there is a clear trend towards more ubiquitous affective state monitoring methods (Zepf et al., 2020). We demonstrate that our proposed method can accurately recognise drivers' emotions and achieve comparable performance as the medical-grade physiological sensor-based state-of-the-art baseline method (Nardelli et al., 2015). Our solution is more ubiquitous, and uses only sensors available in modern cars.

- To the best of our knowledge, this is the first study that verifies the feasibility of non-intrusive inference of driver emotions in empirical situations based on driving behaviours and traffic contexts. Based on 675 hours of driving data collected on public roads in real driving scenarios, a challenging environment compared with laboratory conditions, our results are likely to be more reliable.

The remainder of this chapter is organised as follows: The related work on the definition, the annotation of emotions and the recent advancements in emotion recognition are reviewed in Section 3.2. We present the emotion label transformation and data cleansing procedure in Section 3.3. We explain our methods for emotion recognition in Section 3.4. Section 3.5 summarises the results of verification of our method, and we discuss them further in Section 3.6. Finally, Section 3.7 provides the conclusions of our study.

3.2 Related Work

In this section, we outline common standards of emotion measurement, the current trends in emotion recognition, and the progress in research on in-vehicle emotion recognition for drivers.

[1] In this work, we focused on: anger, disgust, fear, joy, neutral, sadness, surprise, and valence.

3.2.1 Emotions

To recognise the emotional state of drivers, it is important to obtain a reliable ground truth of emotions that can be used to train and evaluate models. The affective computing community often uses several expressions interchangeably to describe emotions (Calvo and D'Mello, 2010), and there is no consensus on a general classification, even in the field of psychology (Izard, 2009). The challenge is that the emotional spectrum ranges around different origins: short and raw (affect), directed and intensely felt (emotions), or long and diffuse (moods) (Barrett, 2006; Schwarz, 1990). Researchers commonly summarise these differences in origins as experiences of feeling basic emotional states (Schwarz, 1990), and various models have been proposed to reliably measure these emotional states in a standardised way.

The common methods of measurement are discrete category models and two- (2D) or three-dimensional (3D) models (Calvo and D'Mello, 2010). Discrete category models (e.g., (Ekman et al., 1987)) allow subjects to categorise their emotional states into a set of basic emotions, such as happiness, sadness, anger, surprise, fear, and disgust. By contrast, 2D and 3D models measure emotions in a multidimensional space (Calvo and D'Mello, 2010). An example is Russel's circumplex model, in which subjects can rate their levels of arousal (i.e., degree of activeness) and valence (i.e., degree of happiness) (Russell, 1980). Combinations of the two express specific emotional states, e.g., low arousal and low valence represent sadness, whereas high arousal and high valence indicate excitement.

3.2.2 Recent Advancements in Emotion Recognition

Inferring emotions is an objective that has been addressed in many prior studies. Although the task considered here is similar, the approach differs with regard to the input used. A variety of inputs, ranging from physiological sensors (Nardelli et al., 2015; Schmidt et al., 2019) and facial images (Lopes et al., 2017) to speech (El Ayadi, Kamel, and Karray, 2011), have been used. Physiological sensors such as smartwatches allow for the continuous estimation of a subject's emotions, inferred based on the heart rate (variability), electrodermal activity, and accelerometer data. The potential of this technique has been recognised by researchers. Ubiquitous devices that record physiological data streams can be used to detect emotional states (Nardelli et al., 2015; Schmidt et al., 2019). However, monitoring physiological signals requires that subjects wear one or multiple devices, which may introduce inconvenience or discomfort for regular daily use. As an alternative, non-intrusive approaches have been developed. The most popular methods of emotion recognition are based on facial expressions and speech (El Ayadi, Kamel, and Karray, 2011; Lopes et al., 2017;

Verma and Choudhary, 2018). Nevertheless, the continuous recording of a user's visual or audio information may raise privacy-related concerns.

Inferring emotions from behaviours and contexts is a promising less-intrusive alternative. Data accumulated from a user's interactions with everyday devices contain behavioural and contextual information that can act as a proxy for the experiences or specific emotional states of the users. Researchers rely on devices such as smartphones to gather this information. The data gathered from smartphones are diverse, and contain information on app usage, screen time, accelerometer, GPS, SMS, call activity, Wi-Fi, and Bluetooth signals. These data constitute a digital representation of user behaviour and context, from which their emotional state can be deducted (Trifan, Oliveira, and Oliveira, 2019). Several studies (Buda, Khwaja, and Matic, 2021; LiKamWa et al., 2013; Reece and Danforth, 2017; Taylor et al., 2020; Zhang et al., 2018) have shown that users' emotions can be inferred from their patterns of mobile phone usage. Canzian et al. used mobility trace from a smartphone to detect a tendency toward depression (Canzian and Musolesi, 2015). The car, as an everyday device with sensor modalities, allows us to derive the driver's behaviour and context as the basis for our emotion recognition algorithms. We propose detecting the driver's emotions using driving behaviours as represented by the CAN-bus signals of the car and the context (surrounding traffic) as determined by the front-view video camera.

3.2.3 Facial Expressions and Their Annotations

Facial expressions are among the most informative source for the estimation of affective and cognitive states (Ekman, Friesen, and Ancoli, 1980). The Facial Action Coding System (FACS) is an objective and quantitative way to measure facial expression. In the FACS, *action units* describe the expressions currently active in the face at any given time, such as "brow furrow" and "eye widen". As a consequence, facial expressions can be quantified based on the combination and the level of presence of *action units* (Ekman and Rosenberg, 1997; Sayette et al., 2001). Various existing works have relied on facial expressions or action units for the estimation of psychological states (Katsis et al., 2008; Sharma et al., 2020; Zhou, Phadnis, and Olechowski, 2020) or the detection of deception (Sen et al., 2018).

However, manual FACS-coding requires profound expert knowledge and the process is laborious due to the manual labelling required. With recent advances in computer vision and machine learning, numerous studies have proposed the automated recognition of facial expressions (Baltrusaitis, Robinson, and Morency, 2016; Dhall et al.,

2011; McDuff, Kaliouby, and Picard, 2012; McDuff et al., 2016; Yang, Ciftci, and Yin, 2018).

Developments in the automated recognition of facial expressions has had a major impact on affective computing. Whereas earlier works relied on the manual annotation of facial expressions, an increasing number of researchers now detect facial action units or acquire emotion labels by using various algorithms. For example, based on automatically detected facial action units, Sen et al. analysed deceptive communication (Sen et al., 2018), and Sharam et al. focused on the assessment of cognitive performance (Sharma et al., 2020). Rostaminia et al. leveraged the the output of detection of OpenFace (Baltrusaitis, Robinson, and Morency, 2016) as ground truth labels for the unobtrusive sensing of upper facial action units (Rostaminia et al., 2019). To estimate emotional experiences during collaborative computer-aided design (CAD), Zhou et al. utilised the results of detection of facial expressions from Affectiva (McDuff et al., 2016) as the emotion labels of CAD users (Zhou, Phadnis, and Olechowski, 2020).

Compared with self-report questionnaires, the automated annotation of emotions based on facial expressions has several advantages. First, automated annotation can significantly reduce the manual labour required, thus enabling the acquisition of a large number of emotion labels at a more temporally granular level. Second, the unobtrusive emotion annotation via facial expressions means that the subject's experience is uninterrupted. The frame-by-frame annotation of facial expressions enables dynamic representations of how emotion evolve over time (McDuff, Kaliouby, and Picard, 2012). Finally, by using facial expressions, the cognitive load imposed by self-reports is avoided and the subjects' responses are less likely to be biased due to the form of the questionnaires, their context, and other irrelevant factors (Schwarz and Strack, 1999).

Given the above advantages, we acquire facial expressions-based emotion labels of drivers in this study by using a facial monitoring camera mounted on the dashboard of the vehicle. Past work (Verma and Choudhary, 2018) has shown that state-of-the-art algorithms can reliably detect the facial expressions of drivers with an accuracy of around 95% in various contexts. Thus, in this work, we rely on the automated facial expression annotation tool for the emotion label acquisition.

3.2.4 In-vehicle Emotion Recognition

As in the wider field of emotion recognition, researchers use sensors to detect the driver's emotions in cars.[2] Physiological sensors are preferred for measuring stress levels as a specific emotional response of the driver (Healey and Picard, 2005; Rigas, Goletsis, and Fotiadis, 2012; Saeed and Trajanovski, 2017; Wang, Lin, and Yang, 2013) because stress and emotions in general are highly correlated with physiological measures, such as the heart rate, the variation in heart rate, and blood pressure (Sano and Picard, 2013). Malta et al. used electrodermal activity (EDA) in combination with facial expressions, driving events, and pedal behaviours to build a Bayesian network to predict the frustration of drivers (Malta et al., 2011). To infer the comprehensive mental and physical states (concentration, tension, tiredness, relaxation) of drivers, the authors of (Rebolledo-Mendez et al., 2014) built a body sensor network to monitor signals, the such as electrocardiogram (ECG), electroencephalography (EEG), EDA, and respiration rate. Kato et al. classified emotions as positive and negative based on ECG and pulse wave measurements during traffic jams (Kato et al., 2011). Most in-vehicle emotion recognition based on physiological signals relies on numerous sensors, which are inconvenient to deploy. Data from physiological sensors as well as those on facial expressions were used by Zhang et al. to monitor a driver's emotional states and degree of fatigue (Zhang et al., 2017). Guang et al. introduced the first neuromorphic vision based distracted driving recognition system that analyses driver drowsiness, driver gaze-zone, driver hand-gesture behaviour from the generated streams of asynchronous events with a dynamic vision sensor (Chen et al., 2020). Shafaei et al. proposed a multimodal system that combines facial expressions with steering wheel usage and vehicular acceleration for emotion recognition (Shafaei, Hacizade, and Knoll, 2018). Facial expressions have been used in industry solutions in this vein. For example, Affectiva has developed an automotive software development kit that can analyse emotions using a driver-monitoring camera system (Mcmanus, 2020a). Research has also used the driver's speech to detect emotions (Tawari and Trivedi, 2010). However, the fundamental barriers of emotion recognition also apply to these methods. The use of video data for the face or data for speech analysis raises privacy concerns. Physiological sensors are also particularly intrusive.

Given that today's car already have a set of sensor modalities, they are already well prepared for emotion inference. Cars are equipped with a large number of sensors that are accessible via the CAN-bus. In the CAN-bus, the sensors and actuators

[2] We recommend a recent review by Zepf et al. (Zepf et al., 2020) that reports details of research on emotion recognition in a vehicle.

of a car transmit comprehensive information about driving-related activities and the vehicle's dynamics while the radar and camera systems interpret the environmental context. Researchers have shown that CAN-bus data can be used to detect driver behaviours to derive the relevant contextual information, for example, identifying the driver among a group of users (Enev et al., 2016), the profile of the driving style (Martinez et al., 2017), and the anticipation of the driver's intentions (Hallac et al., 2018; Liu et al., 2019). Several studies have explored the use of CAN-bus information (Dobbins and Fairclough, 2019; Paredes et al., 2018a) to detect stress using driving behaviours. As this past research indicates, the available sensor data allow for a wider interpretation beyond their intended usage (i.e., controlling the car). To the best of our knowledge, this is the first study to detect the driver's detailed emotional state by passively interpreting data streams available in today's cars.

3.3 Emotion Label Transformation

This subsection describes the statistics of the ground-truth emotion labels acquired from the Affectiva algorithm as well as the procedures for data cleansing, pre-processing, and transformation that were applied to them.

3.3.1 Emotion Annotation Tool

Affectiva spun out of MIT's Media Lab. Its emotion recognition technology uses computer vision algorithms and deep learning models to estimate emotions based on facial expressions. Unlike other solutions based on the recognition of facial expressions, Affectiva's algorithms are built on a very large foundational dataset, containing more than 9.7 million facial images of people from 90 countries, with over 5 billion facial frames and six years of video data. Affectiva's deep learning algorithms are optimised with automotive in-cabin data, including more than 20,000 hours featuring more than 4,000 unique individuals (Mcmanus, 2020a). Given these features, Affectiva's solution can reliably capture driver emotions. In our experiment, we used one of the latest stable versions (ics-2.2.1) of Affectiva for annotation.

Affectiva detects the facial expression of the subject for every frame. For all emotions except for *valence*, it outputs a score between 0 (absent) and 100 (present), indicating the *presence level* of the relevant emotions (McDuff et al., 2016). The emotions included anger, disgust, fear, joy, neutral, sadness, and surprise. The score of *valence* ranged from -100 to 100, and thus was divided into negative and positive valence (unhappy to happy).

3.3.2 Emotion Persistence

Facial expressions are often not long-lasting, with a duration between 0.5s - 4s (Ekman, 2007). An example is provided in Figure 3.1 and 3.2. To ensure that the driver's most prevalent and stable affective states were captured accurately, we applied a non-overlapping sliding window and divided the driving data into ***driving segments***. The emotion labels were defined according to their average level of presence in each ***driving segment***. Our objective was therefore to predict driver emotion using the CAN-Bus data and data from the front-view video of the same ***driving segments***. An illustration of this setting is provided in Figure 3.3. By adjusting the length of the sliding windows (and hence the length of the driving segments), driver emotion could be recognised at different granular levels. We defined the default length of the driving segments as 10 *mins* to capture emotions (directed and intensely) rather than affects (short and raw) or moods (long and diffuse) [3]. In addition, the combination of the sliding window and the temporally continuous annotation from Affectiva enabled emotion recognition at any time during drive.

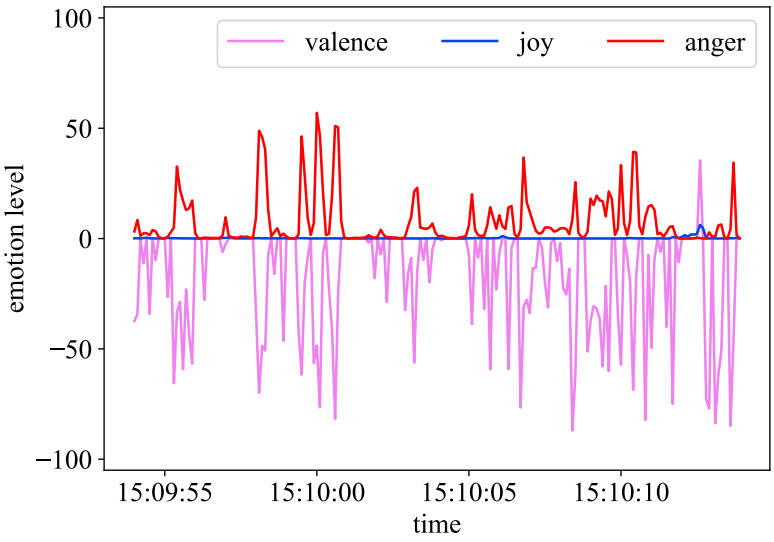


FIGURE 3.1: An example of the raw output of Affectiva results over time, dominance of anger and negative valence

Owing to inevitable occlusion (e.g., from driver turning head), and undesirable illumination, facial expressions could not be detected in every frame. Only in a subset of the frames were both the driver's face and their facial expressions detected. We refer to such frames as *valid frames*. To ensure the quality of the label, we considered only the driving segments that contained more than 70% of valid frames. The labels of

[3]The analysis of different lengths of driving segments is provided in Section 3.5.3

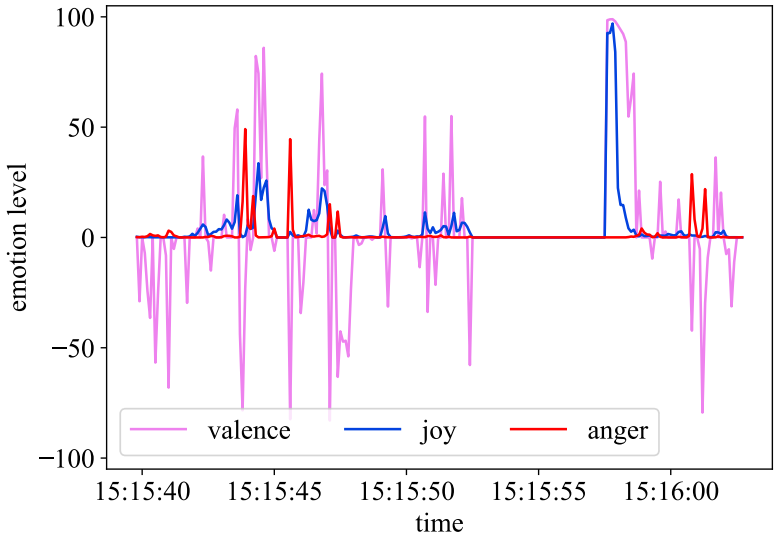


FIGURE 3.2: An example of the raw output of Affectiva results over time, dominance of joy and positive valence

the driving segments were then computed as the average level of presence of an emotion over all valid frames. Such a quality check reduced the amount of driving data being used for training and testing, as some driving segments were discarded owing to an inadequate number of valid frames. From 675.6 hours of driving data, we obtained a total of 19,885 two-minute driving segments (equivalent to 662.8 hours of data) or 3,377 10-minute driving segments (equivalent to 562.7 hours). Trips shorter than 10 *mins* were not included in the 10-minute driving segments, which led to the different number of driving hours between the types of segments.

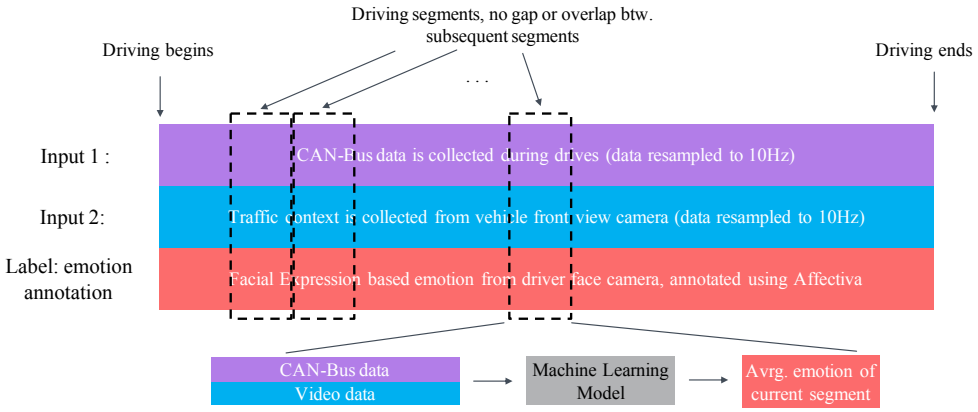


FIGURE 3.3: Data and emotion labels used for our data analysis

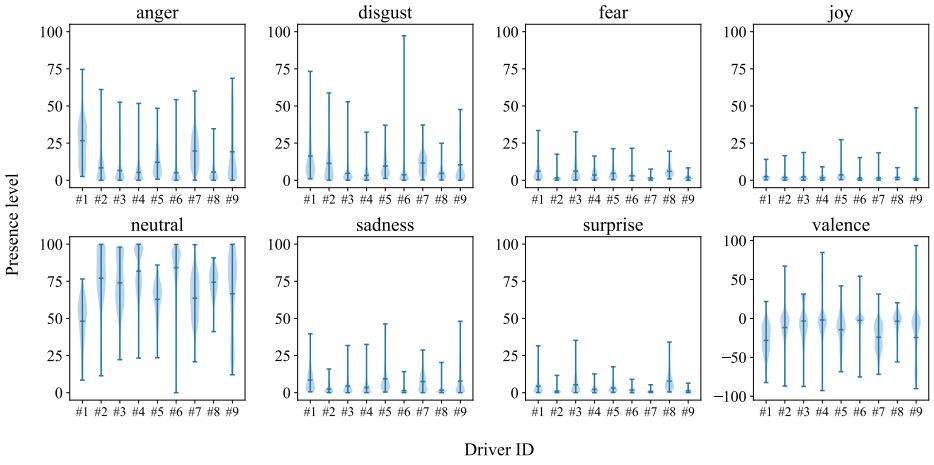


FIGURE 3.4: Violin plots of the emotion distribution of each driver

3.3.3 Emotion Distribution

It is critical to inspect the label distribution to understand how reliable we can predict the emotions and to ensure that our ground truth is valid. As described in Section 3.3.2, the emotion labels were the average presence level of emotions over driving segments of a predefined length. Figure 3.4 illustrates the distribution of emotion labels of the 10-*min* driving segments The following observations can be made from it:

- The drivers elicited more negative *valence* than positive *valence*. This can be explained by the fact that driving is a task that requires high cognitive load and induces stress (Chatterjee et al., 2020; Kahneman et al., 2004).
- Driving had a varying impact on the drivers, which was reflected in the distinct personalised mean values of each emotion across different drivers.
- Most emotions have a low presence level. This can be explained by the instantaneous nature of facial expressions: The values shown in Figure 3.4 were averaged over segments of 10 *mins* of driving; owing to the instantaneous nature of facial expressions, as illustrated in Figure 3.1 and 3.2, the values were balanced out by non-present moments, in which a given expression was absent from the subject's face.

3.3.4 Transformation of Emotion Labels and Data Cleansing

Before conducting further data analysis, we needed to address common affective computing-related issues with our data.

As discussed in Section 3.3.3, each participant had their personalised baseline (i.e. different mean values) of emotions in the context of driving because emotions are subjective, and their interpretation among people differs (Martínez, Yannakakis, and Hallam, 2014; Yang and Chen, 2011; Yannakakis, Cowie, and Busso, 2017). Such subjective factors introduced bias to the emotion recognition. Therefore, the emotion labels of each participant were calibrated following a personalised emotion label transformation, as described in Equation 3.1. This binarisation processing procedure was similar to that in (Canzian and Musolesi, 2015; Egilmez et al., 2017; Mishra et al., 2018; Sen et al., 2018).

$$emotion\ label = \begin{cases} low\ class,\ if\ presence\ level < personalised\ median \\ high\ class,\ if\ presence\ level >= personalised\ median \end{cases} \tag{3.1}$$

Such a transformation accounted for variations in individual perceptions of the driving task. Owing to the continuous values of the annotation, we have almost balanced the emotion labels for low and high classes (average proportion of majority class = $50.6\% \pm 0.7\%$) after label transformation. The objective of our final prediction was to determine whether a driver's emotion positively or negatively deviated from their personal baseline.

3.4 Methodology

This section explains our methodology to infer the driver's emotions based on their driving behaviours (CAN-bus data) and the traffic contexts (video data). We first introduce the data pre-processing and feature engineering for each sensor modality in Section 3.4.1, and then detail the inference models based on the CAN-bus and video data as well as their combination in Section 3.4.2.

3.4.1 Data Pre-processing and Feature Engineering

We briefly explain the candidate features extracted from CAN-bus, front-view video, and additional data sources. Note that we focus on interpretable features that have been proven to be effective in research on sensing the activities of the driver.

Features from CAN-bus Data

The CAN-bus data were used to capture driving behaviours and vehicular dynamics. By using 49 CAN data signals, we chose the following as candidate features because

they were the most common across different vehicles and, thus, could guarantee the capability of generalisation of our analysis. They were as follows: angle of the steering wheel, yaw rate, brake pressure, pedal position of the accelerator, speeds of the four wheels, longitudinal and lateral acceleration, and rotational speed of the motor.

Because the recording of the raw CAN-bus data was not synchronised, we re-sampled them to 10 Hz, which is suitable for CAN-bus data processing (Hallac et al., 2018). Following the common practices for such data processing (Enev et al., 2016), the re-sampled CAN-bus data streams were split into sliding windows, from which features such as statistical features, auto-correlation, etc., are derived to form a feature vector. Our tests of several sliding windows with different lengths resulted in five-second-long windows without overlap. These comparably short windows are common in CAN processing and seem to capture single driving manoeuvres (Enev et al., 2016; Liu et al., 2019). Table 3.1 lists the features derived from the CAN-bus data streams.

TABLE 3.1: Input signals and derived features from CAN-bus data. The features have been widely used for CAN-bus data processing (Enev et al., 2016), and were intended to capture driving behaviours and vehicular dynamics. The dimensions of certain features are noted in brackets.

Input Signal	Derived Features for CAN data
Steering wheel angle, Yaw rate, Brake pressure, Accelerator pedal angle, Speeds of the four wheels, Longitudinal and lateral acceleration, Motor rotational speed	min, max, mean, std. dev., median, kurtosis, skewness, quantile (25%, 75% and 95%), piece-wise approximation (14D), auto-correlation (50D), log(FFT) (26D),

Features of Front-view Video Data

Video data from the front-view camera were expected to reflect the traffic contexts. Because we hypothesise that the driver's emotions can be inferred from traffic contexts, it is reasonable to assume that these contexts are easy to perceive and interpret by humans. Following this assumption, we derived the video-based features from objects detected by using the available object detection algorithms. Deriving pixel-wise features from videos for recognising the driver's emotion is beyond the scope of our work.

We applied Yolo-v3 (Redmon and Farhadi, 2018) to detect a subset of objects related to the driving context (small vehicles, trucks, pedestrians, and cyclists), and used the location and size of each object as candidate features. The videos were re-sampled to 10 frames per second, and Yolo-v3 detection was used on each frame.

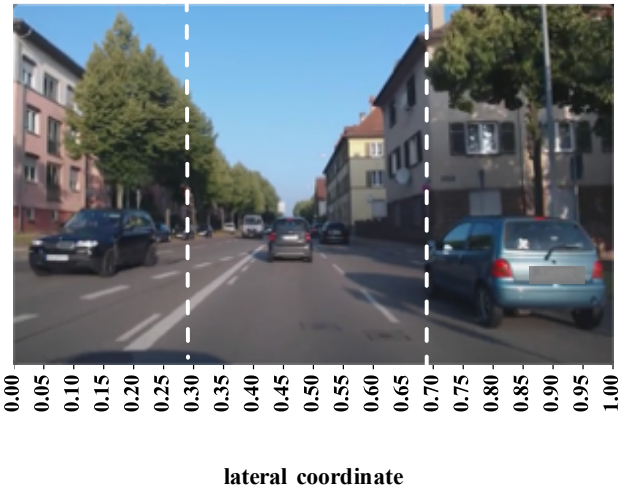


FIGURE 3.5: Simplified lane separation: front view of webcam

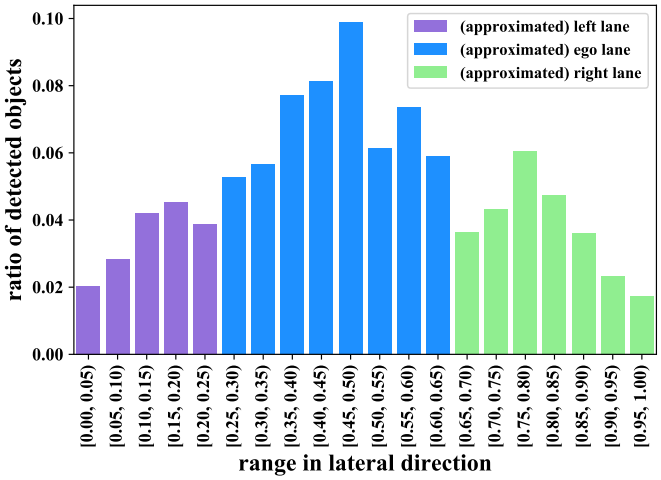


FIGURE 3.6: Simplified lane separation: lateral distribution of number of detected objects

To infer driver emotion from traffic situation, it is important to involve lane information as it can affect the cognitive load on drivers (Lee, Lee, and Boyle, 2007). With lane information, the relative position of the surrounding traffic participants can be better determined. However, state-of-the-art lane detection (Neven et al., 2018) algorithms are computationally expensive and limited in their availability to researchers. We used a simplified method to gain lane-related information. We split an image into

three columns, corresponding to [0,0.25), [0.25,0.65) and [0.65,1] along the lateral axis. These columns were approximated as the left, middle, and right lane from a drivers perspective, respectively. In addition, past research (Lee, Lee, and Boyle, 2007) has shown that drivers are more sensitive to closer vehicles than distant ones. Therefore, the trade-off between accuracy and simplicity was acceptable even though the correspondence between the split and the lanes was valid only in the near-range of the ego-vehicle. Moreover, such an approximation was supported by the distribution of the detected positions and number of vehicles captured by the front-view camera as shown in Figure 3.5 and 3.6. The lateral distribution of the detected objects was approximated by a Gaussian mixture model with three clustering centres.

For each approximated lane we computed the statistical features and auto-correlation of the number and the sum of the sizes of the detected objects in it in each sliding window (the same sliding windows as for CAN-bus data). Each approximated lane was used to compute 120 features, for a total of 360 features from the results of detection using Yolo-v3. These are summarised in Table 3.2.

TABLE 3.2: Features of traffic

Input Signal	Derived Features for Video data
Yolo detection results: - Class (vehicle, cyclist, or pedestrian) - Confidence, coordinates of bounding boxes 10 most confidently detected objects in each frame	min, max, mean, std. dev., median, kurtosis, skewness, quantile (25%, 75%, or 95%), and auto-correlation of number of objects and sum of sizes in each approximated lane in a sliding window

Features from Auxiliary Data Sources

Because emotions vary over the course of a day (Egloff et al., 1995), we considered temporal features to recognize the emotions of the driver using the following: seconds before dawn, seconds after dusk, seconds before sunrise, seconds after sunset, indicator of driving at night, current time (formatted in the 24h-scale), and the day of the week. The first four features were set to zero if driving had occurred after or before the corresponding event to ensure that there were no negative values in the temporal features. The temporal features were computed for every five seconds by using a sliding window based on the time associated with the corresponding windows. From each five-second sliding window, a temporal feature vector of seven

dimensions was computed. That is to say, for instance, from a 10-*min* driving segment, a 120-*step* ($10min \ / \ 5s \ = \ 120$) sequence of 7D temporal feature vectors was computed. We did not perform feature selection on the temporal features, and simply concatenated this temporal sequence of feature vectors to the sequences obtained from CAN data, videos, or a fusion of the two.

Summary of Features

From each sliding window, we computed 1100D, 360D, and 7D feature vectors for the CAN-bus, front-view video, and temporal modalities, respectively. Each modality therefore contained a sequence of feature vectors of the same sequence length. For example, the sequence length of a 10-*min* driving segment was 120 ($10min \ / \ 5s = 120$). For each driver, we computed the p-values associated with each dimension of the feature vector by using an ordinary least-squared regression for every emotion. We selected only the 10 dimensions with the lowest p-values as input from each modality (i.e., the CAN-bus and front-view videos) per emotion for each driver. The cumulative distribution of the p-values of the selected dimensions of the feature vectors are plotted in Figure A.1 in Appendix A. Note that for most signal sources, more than 80% of the selected features had p-values lower than 0.05. We leveraged multi-task learning approach, a method that has been proven to be useful in recognising emotions or activities (Saeed, Ozcelebi, and Lukkien, 2019; Schmidt et al., 2019), to build a neural network to predict all emotions at the same time. This meant that for every driver in the training set, 80 (10D, eight emotions = 80) CAN features, 80 video features and seven temporal features were considered. If a feature was relevant to multiple emotions, it was selected only once. If multiple drivers were in the training set, the union of the selected features was used. Therefore, the number of selected features varied depending on the drivers in the training set.

3.4.2 Driving Behaviour- and Context-based Inference Models

In this subsection, we first introduce the driving behaviour- and context-based models using data from only either CAN-bus or front-view videos as input, followed by a combination of the two (i.e., sensor fusion).

CAN-bus-only Model

Random forest approaches, as for example in (Enev et al., 2016), can be used to process CAN-bus data. However, they require hand-crafted features and are unable to capture temporal dependencies between these feature vectors. By contrast, the approaches based on recurrent neural networks (RNN) like the one in (Hallac et al.,

2018) can model the time dependence on a wider time scale even without explicit feature creation. However, such end-to-end training methods must learn the knowledge of carefully designed features and requires larger amounts of data. Our method combines the advantages of both. Our pipeline begins by computing the feature vectors of CAN-bus data using sliding windows as shown in Table 3.1 and Section 3.4.1. Then, the feature vectors are fed sequentially into a RNN.

The RNN can learn a high-level abstract summary of the data from the sequence of feature vectors. This summary is then processed by a fully connected network that outputs a probability distribution as a prediction of low and high states of a certain emotion. The detailed settings of the proposed method are as follows: We choose an RNN architecture with two layers, where each consists of 64 gated recurrent units (GRUs). This architecture is similar to (Hallac et al., 2018), where they have input dimension of 665 and two layers RNN with 256 gated recurrent units is used. We have proportionally applied the similar reduction from our input dimension to the gated recurrent units. The CAN-bus and time feature vectors are concatenated in each sliding window. Our RNN has a simple structure because manual feature extraction and feature selection are applied a priori, which significantly reduces the complexity of the input. The input to the RNN is a sequence of feature vectors with reduced dimensions. We leveraged a multi-task learning approach to build a neural network that predicted all emotions at the same time. As the labels were almost balanced for every emotion, we randomly shuffled the training batches to obtain an almost equal number of low- and high-state samples on average for each emotion in every training batch. We used the Adam (Kingma and Ba, 2014) optimiser and cross-entropy as loss functions, as described in Equation 3.2. The ReLU (Nair and Hinton, 2010) was applied as activation function to the fully connected layers. The learning rate was set to 0.005. The hyper-parameters/parameters are empirically tuned to achieve the best emotion recognition performance. We trained the neural network until its loss converges. The inverse proportion to the class ratio was assigned as weight to the loss function:

$$\mathscr{L} = \sum_{i=1}^{n} -w_i[y_i \cdot log\ \sigma(x_i) + (1-y_i) \cdot log(1-\sigma(y_i))] \tag{3.2}$$

with

$$w_i = n / \sum_{j=1}^{n} (1-y_i)(1-y_j) + y_i y_j \tag{3.3}$$

where x_i and y_i are the prediction and ground truth for the i^{th} sample, respectively, and n is the total number of samples. The framework of the proposed method is

illustrated in Figure 3.7.

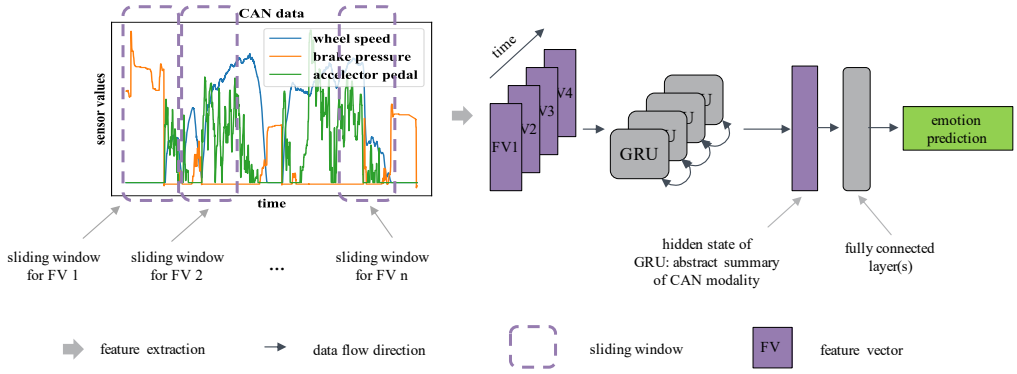


FIGURE 3.7: Recurrent neural architecture for recognising the driver's emotion: CAN-only model

Front-view Video-only Model

Apart from the modality of the input, the structure and settings for the front-view video-only model were identical to the CAN-bus-only model. The feature vectors of video replaced those of the CAN-bus as explained in Section 3.4.1. Vectors of the video and the time feature were concatenated in each sliding window. The framework of the front-view video-only model is illustrated in Figure 3.8.

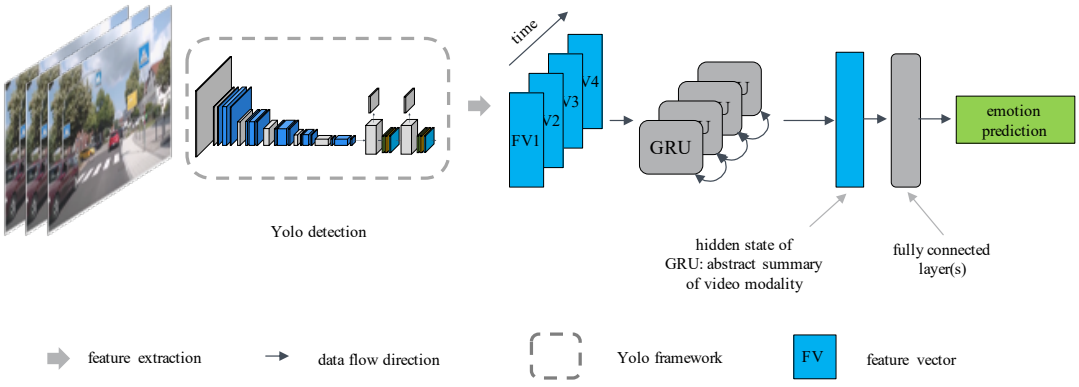


FIGURE 3.8: Recurrent neural architecture for recognising the driver's emotion: Video-only model

Fusion Model

We then integrated the two sensor modalities into a joint inference model, called the fusion model. To fuse the sensor data of the CAN-bus with the front-view video, we constructed fused feature vectors formed by concatenating the feature vectors as described in the previous two sub-sections. The rest of the network architecture

and settings constant are kept constant. The sensor fusion model is illustrated in the Figure 3.9.

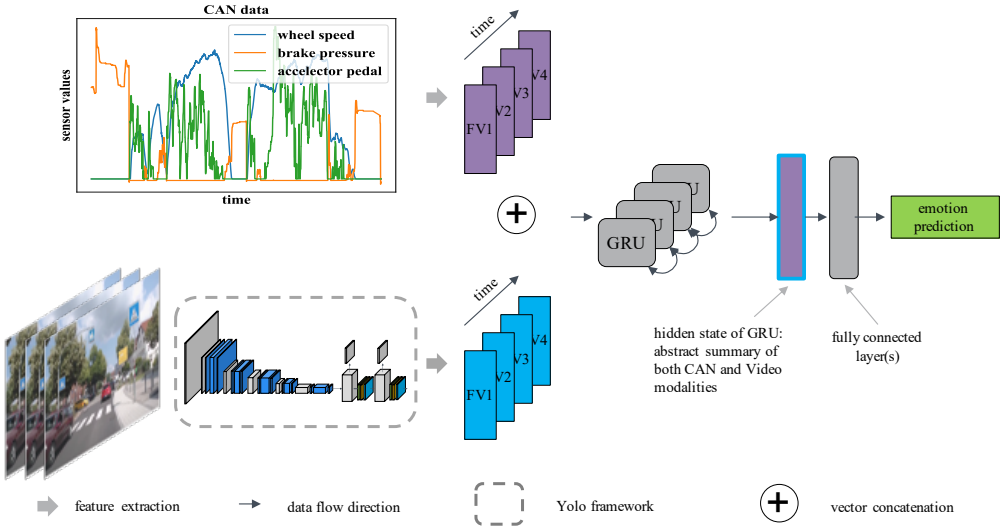


FIGURE 3.9: Multi-modal recurrent architecture for recognising the driver's emotion: fusion model

3.5 Evaluation

Our evaluation section has three parts. In Section 3.5.1, we evaluated our *driving behaviour- and context-based inference models*. In Section 3.5.2, we described the baseline methods and compared their results with that of our proposed method. Finally, in Section 3.5.3 we investigated the stability, usability, and complexity of our approach.

To avoid misinterpretation, the macro-F1-score that we used is formalised in Equations 3.4 - 3.7, where tp, fp, and fn represent the true positive, false positive and false negative, respectively, depending on the level of presence of each emotion, and subscript (c) distinguishes between low-and high-scoring classes, i.e., $(c) = (low)\ or\ (high)$. For the sake of brevity, all *F1-scores* in this paper refer to the *macro F1-score*.

As mentioned in Section 3.3.2, the *driving segments* were generated by using non-overlapping sliding windows to ensure that there was no intersection between any pairs of *driving segments*. For intra-subject evaluation, we built a personalised model for each driver by randomly dividing their driving segments into training (70%) and test (30%) datasets. For the leave-one-subject-out (LOSO) evaluation, we used the driving segments from the i^{th} driver as test data and the data from the remaining

$N-1$ drivers is used as training data. In our case, i was iterated from one to nine. All the experiments were repeated 10 times by using 10 different random seeds for both intra-subject and LOSO evaluations. The F1-scores presented are the average of the 10 repetitions of all drivers. The standard deviations of the 10 repetitions are indicated as error bars in the corresponding figures.

$$F1 = (F1_{(low)} + F1_{(high)})/2 \tag{3.4}$$

$$precision_{(c)} = tp_{(c)}/(tp_{(c)} + fp_{(c)}) \tag{3.5}$$

$$recall_{(c)} = tp_{(c)}/(tp_{(c)} + fn_{(c)}) \tag{3.6}$$

$$F1_{(c)} = \frac{2}{recall_{(c)}^{-1} + precision_{(c)}^{-1}} \tag{3.7}$$

3.5.1 RQ1.1: To what extent can the emotions of drivers be inferred based on (a) CAN-bus data streams, (b) front-view camera, and (c) the combination of both *(fusion)*?

Table 3.3 summarises the results of our proposed models for intra-subject and leave-one-subject-out (LOSO) evaluations. The best performance scores are highlighted in bold. In the intra-subject evaluation settings, the results indicate that the fusion model achieved the best performance scores on all emotions. For all emotions combined, our CAN-bus-only model achieved an F1-score of 68.8%. Our video-only and fusion models improved this score to 69.9% and 71.0%, respectively. With limited fluctuations, all emotions achieved comparable equally high scores. This suggests that a comprehensive recognition of the driver emotional state was possible in intra-subject setting. We then inspected the generalisability of the proposed models by exploring the results of the LOSO evaluation.

Compared with the setting for the intra-subject evaluation setting, the LOSO evaluation yielded lower scores. This was expected because emotion is subjective and user-independent emotion recognition remains a daunting challenge in affective computing community. The F1-scores achieved by the fusion model highlighted the varying performance across all emotions in the LOSO setting. Whereas the emotions of anger, disgust, neutral, sadness, and valence yielded relatively high scores of 63.8%,

64.5%, 62.8%, 63.5%, and 62.4%, respectively, those of fear, joy and surprise were less accurately detected, with F1-scores of 48.7%, 58.7% and 49.8%, respectively. These varying performance scores across emotions also accounted for the results of the CAN-bus-only and front-view video-only models. We conclude that some emotions were more difficult to detect than others independently of the sensor modality in the LOSO setting.

TABLE 3.3: Intra-subject and LOSO cross-validation: comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities. The best results of the three models (CAN, video, and fusion) are highlighted.

	Personalised Model			LOSO Model		
F1-score (%)	CAN	Video	Fusion	CAN	Video	Fusion
anger	69.8	71.2	***72.3***	62.9	62.3	***63.4***
disgust	69.9	71.8	***72.8***	63.7	63.7	**64.5**
fear	69.1	69.2	***70.5***	48.0	***49.3***	48.7
joy	67.8	67.5	***68.6***	***59.3***	58.3	58.7
neutral	68.3	69.6	***70.9***	61.9	60.6	***62.8***
sadness	68.1	70.2	***70.6***	62.4	61.8	***63.5***
surprise	69.5	69.6	***70.9***	49.3	49.7	***49.8***
valence	68.3	70.4	***71.1***	60.5	60.2	***62.4***
average	68.8	69.9	***71.0***	58.5	58.2	***59.2***

Furthermore, we plotted the confusion matrix of the fusion model to better visualise the results in Figure 3.10 and 3.11. For the personalised setting, the fusion model predicted the low and high classes with comparatively equal accuracy for all emotions, despite small variations. For the LOSO settings, we observed similar patterns (i.e., higher values along the diagonal of the matrix) except in case of the emotions of *fear* and *surprise*. The confusion matrix shows that our proposed solution was not biased towards the low or the high class, which reflects the stability of our results. While the LOSO evaluation in general achieved lower F1-scores than the personalised evaluation, the confusion matrix showed that *anger*, *sadness*, and *valence* could be relatively accurately recognised in the LOSO settings. The more reliable detection of these three emotions in the LOSO settings is in line with existing findings in literature providing evidence that they are closely related to driving performance and safety (Chan and Singhal, 2015; Jeon, 2016; Trick, Brandigampola, and Enns, 2012; Underwood et al., 1999). For the sake of completeness, the confusion matrices of the CAN-only and video-only models are provided in Appendix A.

In summary, the F1-scores estimated using intra-subject and LOSO cross-validation indicate that emotions could best be inferred based on a combination of CAN-bus

data and front-view camera data. The use of single modality models leads only to a minor reduction in performance.

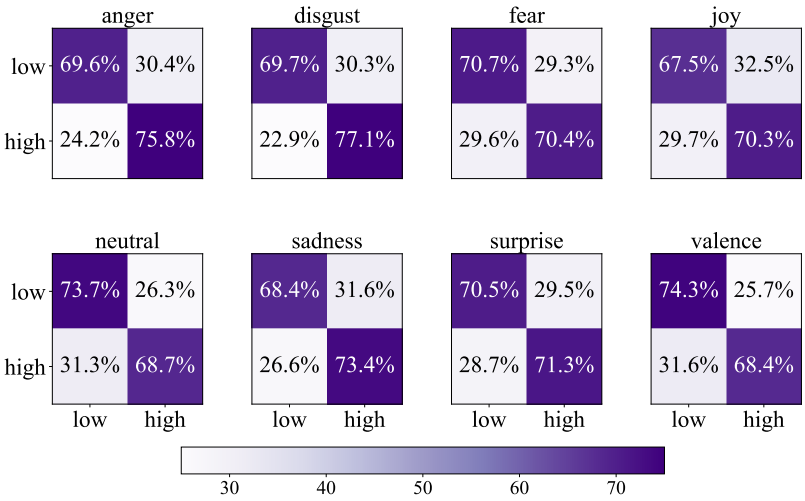


FIGURE 3.10: Confusion matrix for fusion modality under intra-subject evaluation

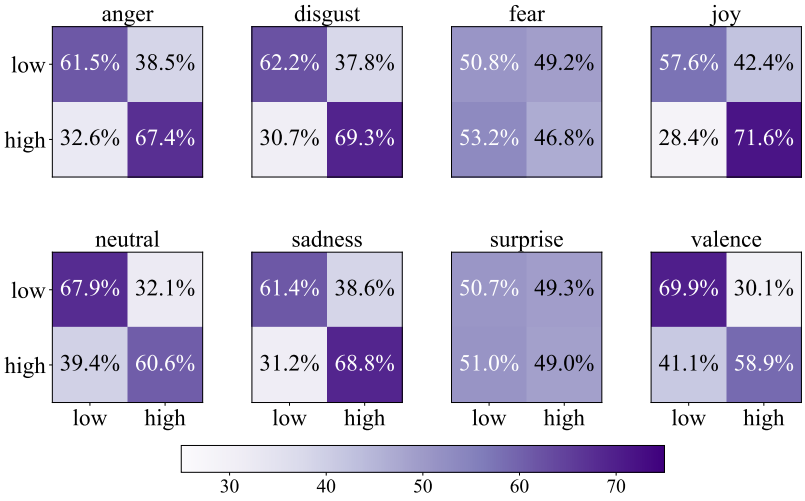


FIGURE 3.11: Confusion matrix for fusion modality under LOSO evaluation

3.5.2 RQ1.2: How much improvement does emotion recognition based on vehicle data offer compared with state-of-the-art methods based on physiological sensors?

The performance scores of our models were compared with a physiological signals-based baseline. Physiological signals are widely used for emotion recognition (Dobbins and Fairclough, 2019; Gjoreski et al., 2016; Healey and Picard, 2005; Schmidt

et al., 2019) because emotions have an influence on the autonomic nervous system (Kreibig, 2010), which is highly correlated with such physiological measures as the heart rate, heart rate variability, and blood pressure (Sano and Picard, 2013). We used the emotion recognition models proposed by Nardelli et al. (Nardelli et al., 2015) as our physiological sensor baseline. In (Nardelli et al., 2015), the emotion of subjects were induced by affective sounds in lab settings. During the experiments, the heart rate variability (HRV) data of the subjects were collected. The authors derived features of the time and frequency domains as well as non-linear features, and performed feature selection using the Friedman test and Wilcoxon signed-rank test. A quadratic discriminant classifier was then used for classification, and yielded an accuracy above 80% for *valence* and *arousal* by using the LOSO procedure. For the convenience of the readers, a comprehensive list of HRV features are provided in Table A.2.

Combinations of *valence* and *arousal* express specific emotional states according to Russel's circumplex model (Russell, 1980). For example, low arousal and low valence represent sadness, whereas high arousal and medium valence indicate happiness. Therefore, the model from (Nardelli et al., 2015), designed for valence and arousal recognition, was used as a proxy for the recognition of a variety of emotions, in preference to models that focus on detecting stress or frustration (Healey and Picard, 2005; Malta et al., 2011; Rigas, Goletsis, and Fotiadis, 2012; Saeed and Trajanovski, 2017). Furthermore, unlike other physiological sensors-based approaches (Gjoreski et al., 2016; Healey and Picard, 2005; Schmidt et al., 2019), the one in (Nardelli et al., 2015) does not require such physiological signals as the photoplethysmogram, skin temperature, and skin conductance. Therefore, this baseline approach, by relying only on HRV data represents the available commercial wearable solutions to emotion recognition (Garmin Ltd., 2019). Moreover, our in-vehicle environment resembled the laboratory settings in (Nardelli et al., 2015) because the subjects were seated and no dramatic physical movements were allowed. Therefore, (Nardelli et al., 2015) was a suitable and competitive baseline.

We replicated the method developed by Nardelli et al. as our physiological sensor baseline, and refer to it as the *baseline* method. In accordance with our evaluation scheme in Figure 3.3, we used HRV data of the driving segments to predict the emotion labels in them. The baseline was evaluated and compared with our approach on a subset of the available driving data because HRV data were collected from the participants over only a two-week period. Therefore, only around 65 of the 675 hours of the data on the trips contained the relevant HRV data, corresponding to 388 samples of 10-*mins* driving segments. Although the HRV subset constituted only 9.6% of the

entire dataset, it was still larger than the dataset used in (Nardelli et al., 2015), which consisted of only 208 samples.

Since the baseline (Nardelli et al., 2015) was conducted in lab condition, we optimised it to our context by exploring prevalent machine learning models in affective computing including Random Forest, Gradient Boosting, Support Vector Classifier, Extra Trees Classifier, Decision Trees as well as our proposed recurrent neural network. We performed grid search for optimal parameters and found that Extra Trees Classifier (the number of trees = 50, maximum depth of the tree = 20, and the minimal number of samples per split = 2) achieved the best performance. We refer to this optimised model as *baseline**.

A comparison between the proposed method and *baseline** method is provided in Table 3.4 and 3.5 [4]. Our proposed methods outperformed the *baseline** method in most of the cases. The scores in Table 3.4 and 3.5 are lower than those in Table 3.3. Such a decrease in model performance is expected, because we used a smaller dataset for comparison with the HRV baseline, and our proposed method relies on deep learning. We discuss this instance of performance and the dataset size in more detail in Section 3.5.3. The reduced performance of the baseline method compared to our approach is barely explainable due to a small dataset because our sample size (388) is much larger than the dataset (208) used in (Nardelli et al., 2015). Therefore, our proposed method is better suited for in-vehicle emotion recognition than the state-of-the-art based physiological sensor-based approach.

TABLE 3.4: Intra-subject cross-validation: comparison between the baseline and the proposed three models

	Personalised Model			
F1-score (%)	CAN	Video	Fusion	baseline*
anger	63.1	63.4	61.5	58.7
disgust	64.4	66.1	62.1	60.6
fear	62	59.5	56.8	55.8
joy	62.1	63.5	64.5	59.7
neutral	64.3	64.5	64	56.5
sadness	63.7	64.3	62.9	59.8
surprise	66.4	64.7	66.1	58.2
valence	66.7	61	62.4	62.5
average	64.1	63.4	62.5	59

[4] The comparison between the proposed method and *baseline* method is provided in Table A.3 and A.4 in the Appendix

TABLE 3.5: LOSO cross-validation: comparison between the baseline and the proposed three models

	LOSO Model			
F1-score (%)	CAN	Video	Fusion	baseline*
anger	51.4	52.9	50.8	44.4
disgust	52.4	50.8	49.5	47.9
fear	54.7	54	50.6	49.9
joy	56.6	54.8	53.3	54.3
neutral	54.6	51.6	52.4	49
sadness	48	49.8	47.4	45.3
surprise	54.3	50.7	49.3	55.4
valence	47.9	50.2	46.4	45.7
average	52.5	51.8	50	49

3.5.3 Analysis of Stability, Usability, and Complexity of Proposed Solution

In this section, the stability, usability, and complexity of the proposed solution are analysed in detail.

Stability vs. Dataset Size

Like all intelligent learning systems, the performance of our proposed method is sensitive to the amount of training data used. Hence, we analysed changes in the performance of our models with the size of the training dataset for both intra-subject and LOSO evaluations. We changed the size of our training dataset by randomly removing a certain amount of data while keeping the size of the test dataset constant: a total of 30% of the dataset (intra-subject evaluation) or data for one driver (LOSO evaluation).

Figure 3.12a depicts the results of the stability analysis of the intra-subject evaluation. In general, despite fluctuations in case of small amounts of available data, the performance patterns of different modalities stabilised once the training dataset had reached 40% of its original size. After stabilisation, our fusion model outperformed the CAN-bus-only and front-view video-only models by around one standard deviation, independently of the size ($> 40\%$) of the dataset. Overall, there is potential to further improve the performance by increasing training dataset size. However, it should be noted that the performance improvement relative to dataset size starts to show the sign of saturation after the training dataset size reaches 40% of its original

size. Therefore, the performance we showed in Table 3.3 is very close the the upper bound if more data is available.

We performed the same stability analysis on the LOSO evaluation as shown in Figure 3.12b. The performance began to saturate as the size of the dataset increased to over 40% of the original dataset. After saturation, the fusion model consistently showed slight improvements over the CAN-only or Video-only models.

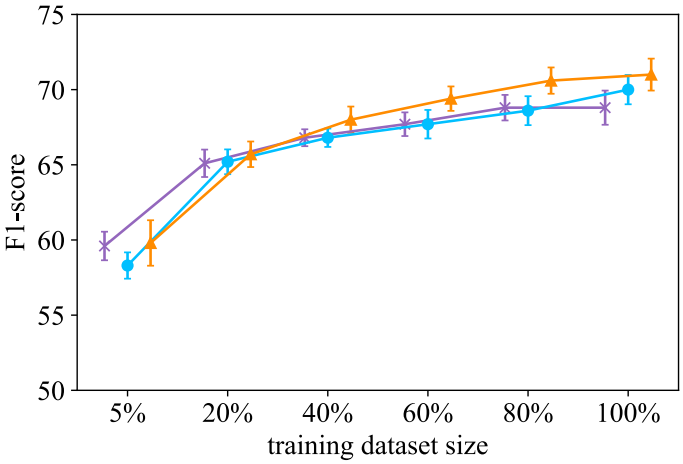
Another bottleneck in the LOSO model was the dataset diversity, i.e., the number of training subjects. The higher the diversity of the dataset was, the greater was the possibility that the model finds generalisable patterns among the drivers. To verify this hypothesis, we evaluated our LOSO model with different numbers of subjects in the training dataset. This is depicted in Figure 3.13. While the overall performance saturated once four drivers had been included in the training dataset, the standard deviation dropped drastically as more drivers were included. This indicates the importance of diversity among the training subjects for model stability in emotion recognition.

Analysis of Usability

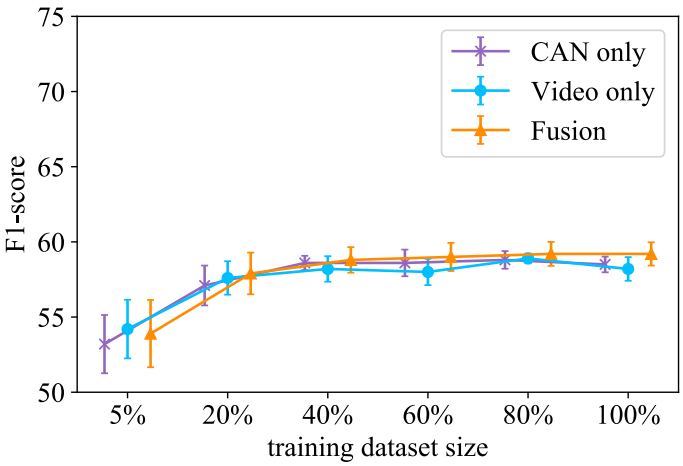
First, we analyse emotion recognition at different temporal granularities. As described in Section 3.3.4, the length of the driving segments could be adjusted to infer driver emotions at different granular levels. To inspect the performance of the model from this perspective, experiments were carried out on segments with lengths of 2-*mins* (19,884 samples), 5-*mins* (7,540 samples) and 10-*mins* (3,377 samples), and *entire segments* of trips [5] (1,773 samples). The result is illustrated in Figure 3.14.

Both the personalised and the LOSO models delivered the best performance on data on 10-*min* segments of driving. Despite having around two times more training samples than the setting for the 10-*min* driving segments, emotion recognition on the 5-*min* driving segments achieves inferior results. The performance on entire trips in terms of recognition was by some distance the poorest. There are several possible explanations for this: a) driver emotions fluctuating while driving, where the proposed method was able to better infer a driver's prevalent emotional state over a certain period (in our case, around 10 *mins*) than the instantaneous detection of emotion, b) events during driving having varying impacts on the drivers' global emotions, and c) the total number of training samples for segments of entire trips being only half

[5] A trip is defined as the duration from the driver starts driving until he/her ends driving. Therefore the trips have varying length.



(A) Personalised model



(B) LOSO model

FIGURE 3.12: Model performance vs. size of training dataset

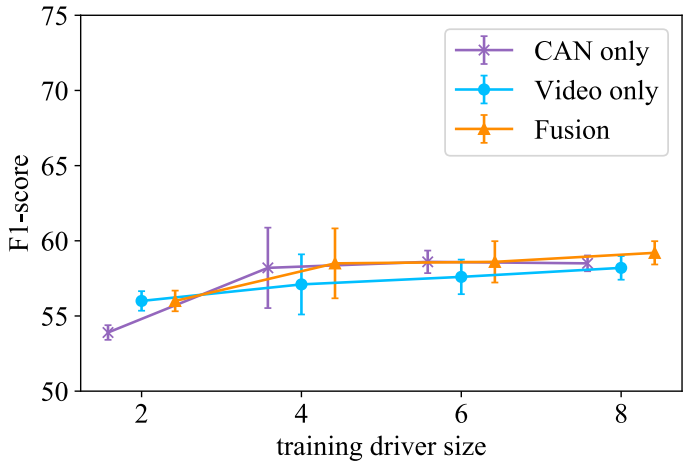


FIGURE 3.13: Model performance vs. number of drivers in training set, LOSO model

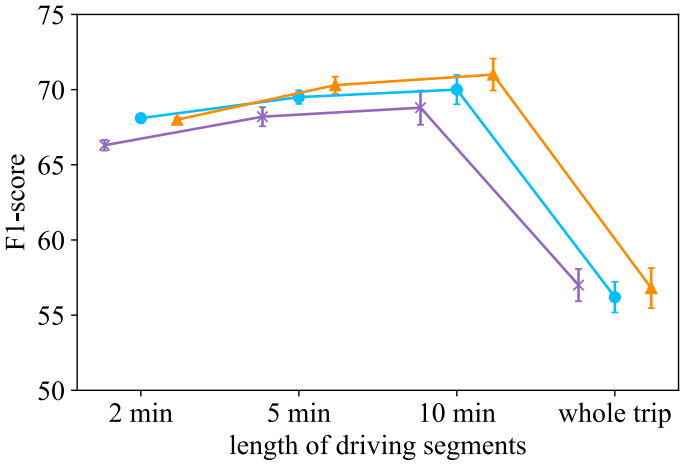
of those for the 10-*min* segments, in which case the reduced dataset had a negative impact on performance.

The proposed solution could best infer driver emotions at approximately 10-*min* intervals. Emotion recognition at a higher temporal resolution led to a decline in performance. This analysis shows that minute-level emotion recognition is possible by using the proposed solution. The high temporal resolution enables a more granular understanding of the evolution of driver's mental states evolving over time, and hence provides more opportunities for numerous applications. For example, just-in-time emotion regulation can be better applied at the appropriate time in this way.

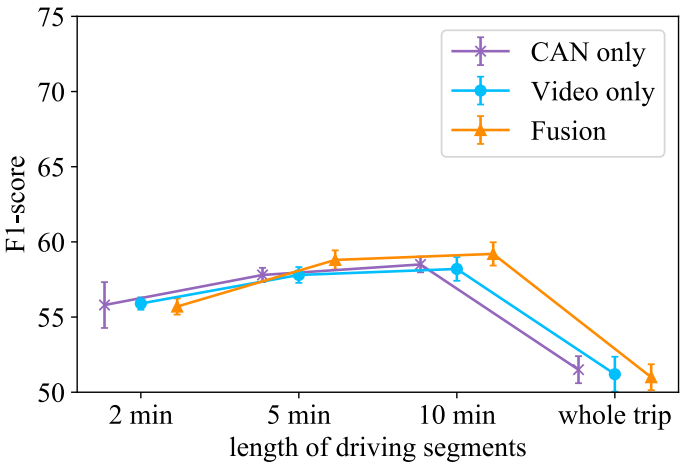
Next, the analysis of ablation study on CAN sensors is provided. We performed an ablation study by using only one, or a subset of, CAN sensor(s) to understand behaviour-based affective computing. The results are illustrated in Figure 3.15.

For personalised models, lateral acceleration (*lat. acc.*), *yaw rate*, and steering wheel angle and vehicle acceleration (*Steer. & Veh. Acc*) perform the best among all CAN sensors. That is to say, the emotion of individual person, to a certain degree, can be better traced by using vehicular dynamics (i.e., *lat. acc.* and *yaw rate*) than the interaction between the driver and the vehicle (*accelerator pedal* and *brake*). For the LOSO model, steering wheel angle (*steer. wheel*) achieved the best F1-score with a low standard deviation, which means that steering wheel-related behaviours were more generalisable than other sensors among users.

Overall, the reduced number of CAN sensors had a negative impact on the *CAN-only model*. However, the degradation in performance in terms of emotion recognition

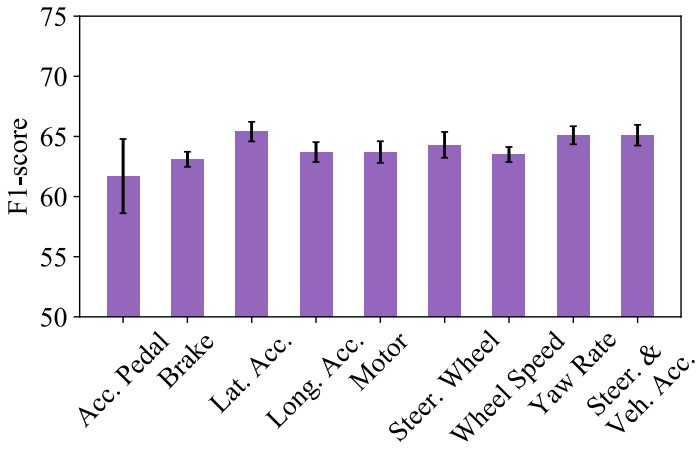


(A) Personalised model

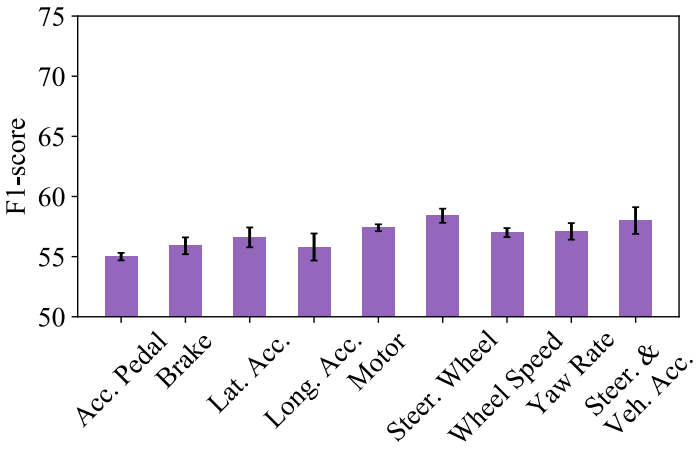


(B) LOSO model

FIGURE 3.14: Model performance vs. segment length



(A) Personalised model



(B) LOSO model

FIGURE 3.15: CAN-only model performance vs. ablation on CAN sensors: select only features from one sensor. Note: *Steer. & Veh. Acc* contains steering wheel angle, latitudinal & longitudinal acceleration and accelerator pedal; this subset replicates the settings in (Shafaei, Hacizade, and Knoll, 2018).

performance was moderate. Even with only one CAN sensor (*lat. acc.* for personalised models and *steer. wheel* for LOSO models), the *CAN-only model* achieved F1-scores that were only around $2-3\%$ lower than if all sensors were used. The ablation study thus demonstrated the flexible usability of the proposed solution in case certain CAN sensors are not available.

Model Complexity

In this section, the model complexity of the proposed solution is analysed as it is vital for mobile and ubiquitous applications to run efficiently. Our model had a varying number of input dimensions depending on the feature selection process. Therefore, the number of model parameters was not fixed. As a reference, a model with a feature vector containing 100 dimension had 79,560 parameters and a size of 180KB. Time complexity was mainly evaluated based on the CPU setup: Intel Core i5 1.4 GHz Quad-Core, 16 GB LPDDR3. To better outline the performance of current advancements in GPU-accelerated parallel computing, Yolo-v3 object detection was also evaluated on GPU setup: Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz, 196 GB LPDDR3, GeForce RTX 2080 Ti.

Table 3.6 shows the time complexity of each stage of processing of the proposed method on a 10 *min* segment. The proposed solution relied on explicit feature engineering in combination with a simple recurrent neural network. Therefore, the data pre-processing and the feature generation had higher time complexity. The main bottleneck of data pre-processing was the Yolo-v3 object detection on CPU-setup. However, potential improvements can help avoid this for ubiquitous in-vehicle emotion recognition, especially when the proposed solution is deployed on mobile devices:

- Subject to local legislative permission, the Yolo-v3 object detection can be delegated from mobiles devices to a cloud service with GPU devices. In our case, a 10-*min* segment of a video at 10FPS and a resolution of 640×360 had a size of only around 50 MB, which is acceptable in light of current mobile network capacity.
- An increasing number of mobile devices are now equipped with neural processing units that can significantly reduce the computation time and energy consumption for object detection on mobile devices.
- Our proposed video feature engineering relies only on size and location of the surrounding objects. Such information is already available in vehicles that

are equipped with a radar or visual sensing systems. Thus, integrating the proposed solution into vehicles will not impose an additional computation burden.

TABLE 3.6: Time complexity for processing one 10 *min* driving segment

procedure	Wall-time (s)
CAN resampling	10.7
Yolo-v3 detection (CPU)	8028.0
Yolo-v3 detection (GPU)	75.0
CAN feature generation	1.2
Video feature generation	1.1
Inference	0.2

In conclusion, the run-time complexity analysis demonstrates that the proposed solution has significant potential for the deployment on mobile devices (e.g., in combination with driving recorders or smartphones mounted on dashboards), or for integration into an on-board computer of vehicles.

3.6 Discussion

In this section, we discuss the results of our work. We outline the potential and limitations of the proposed non-intrusive approach with a special focus on privacy protection.

3.6.1 Non-intrusive In-Vehicle Emotion Recognition and its Contribution to the Car of the Future

In this subsection, we present the current solution offered by Original Equipment Manufacturers (OEMs) to infer emotions and highlight the possible impact of our approach on it. Braun et al. reviewed concept cars because corporate research is not often published in the technical literature (Braun, Weber, and Alt, 2020). For example, Audi presented its concept car, Elaine, which is similar to the solution proposed by Mercedes–Benz (Audi AG, 2018). Elaine is capable of detecting stress and fatigue based on body temperature and heart rate obtained through wearable devices. Actions are subsequently triggered based on the recognised emotions. Such actions can be interventions, adaptive music, ambient light, or empathetic speech. Although the actions differ among OEMs, most of them are similar in that emotions are recognised based on sensors that increase cost and might breach the user's privacy.

Unlike the above-mentioned emotion recognition methods that require physiological sensors or the facial images of users, our methods rely solely on a set of existing sensor modalities - CAN-bus and sensors that are going to be installed in future vehicles - traffic sensing systems, such as the radar, Lidar, and visual sensing systems. We used front-view cameras to capture traffic context. This paper showed that the analysis of driving behaviour and the context of an everyday task, i.e., driving, can help form a reliable estimate of the driver's emotional state. We used Affectiva to annotate the facial expressions of the drivers. Our algorithms were able to differentiate among 8 emotions to provide a detailed picture of the individually perceived emotions. In contrast to prevalent technologies that conduct frame-by-frame predictions (e.g., Affectiva (Mcmanus, 2020a)), our solution offers an assessment of the emotion of the driver over a certain period, based on their driving behaviour and the traffic context. Although, initially it seems that both approaches are ambivalent, we believe that they are complementary. The emotional spectrum ranges around different origins. While Affetiva focuses on short and raw affect, our approach is more geared towards directed and less volatile emotions and can better protect user privacy.

Our approach is fully comparable with recent advancements in behaviour-based emotion recognition in the wild (Buda, Khwaja, and Matic, 2021; Reece and Danforth, 2017; Taylor et al., 2020). Since most in-vehicle emotion recognition relied on physiological sensors, facial expression, or speech and were conducted in lab settings (Zepf et al., 2020), we benchmark our results with that of other in the wild studies based mainly on behaviours. Relying on Instagram photos, Reece and Danforth were able to detect the depression of users at an accuracy of 70% (Reece and Danforth, 2017). Taylor et al. built a personalised model to classify binary states (sad/happy) at an accuracy of 78.4%; similarly, Buda et al. (Buda, Khwaja, and Matic, 2021) predicted lower or upper outliers of users' happiness and achieved and macro F1-score at 64.7% and 59.2% [6] for user-dependent and user-independent model, respectively. In particular, (Taylor et al., 2020) excluded neutral days from both train and test, and (Buda, Khwaja, and Matic, 2021) focused on classification of anomaly. In comparison, we did not exclude any data in our approach, meaning that our method shows more robustness in predicting ambiguous emotion responses that are close to the boundary between low and high classes. As such, our competitive performance against the state-of-the-art approaches demonstrate that our in-vehicle solution can serve as a good complement of prevalent behaviour-based emotion recognition approaches.

[6] Both macro F1-scores are computed from their reported confusion matrix and data distribution.

Our proposed non-intrusive system to recognise driver emotions allows for and encourages new opportunities to exploit the available data streams to infer the emotional state of the driver. Because our data were collected on public roads in empirical driving scenarios, a challenging environment was provided compared with laboratory conditions, and so our results are likely to be more reliable.

3.6.2 Emotion Recognition with Respect to Privacy Protection

We evaluated our method in terms of protecting user privacy. CAN-bus data and the surrounding visual information are much less sensitive to intrusion or leaking than physiological and facial data. For example, physiological data can reveal sensitive health conditions. Capturing surrounding traffics via a camera can violate privacy laws in certain regions (e.g., the General Data Protection Regulation (GDPR) in Europe (European Union, 2020)). In such regions, the proposed *CAN-only model* can be applied as it achieves similar performance to that of the *Video-only* and *fusion models* while better ensuring preserve user privacy. Nevertheless, our models require only locations and distances (we used object size as a proxy) of the surrounding traffic participants. Surrounding object information can be retrieved from on-board radar systems (Sun, Fei, and Pohl, 2019). Acquiring such information does not capture sensitive information (e.g., license plates, facial images or explicit activities) from traffic participants. Hence, our approach adheres to the idea of privacy by design by minimising the data collection. It is compliant with current privacy-related laws and should be more acceptable to customers than prevalent technologies, such as driver-monitoring cameras or health-sensor-based approaches. Thus, the novelty of our non-invasive emotion recognition algorithms is in their potential for scalability and privacy-preserving capabilities.

3.6.3 Flexible Deployment

The proposed *CAN-only*, *Video-only* and *Fusion models* delivered similar performances, as described in Section 3.5.1. We thus believe that the proposed methods have great flexibility in terms of deployment depending on the available sensor modalities:

- The *CAN-only model* can be deployed in regions in which vehicular front-view videos are disallowed (by law), or in the vehicles that are not equipped with the requisite sensing systems to capture traffic context information. Furthermore, the *CAN-only model* is robust against adversarial attacks and can be used in security sensitive scenarios.

- The *Video-only model* is more flexible than *CAN-only model*. On the one hand, in the regions (e.g., China, India and USA) where the front-view video are permitted, the *Video-only model* can be used in combination with driving recorders or smartphones mounted on the vehicle dashboard. Such a setting allows for the recognition of driver emotion without access to CAN data, and hence further increases the ubiquity. On the other hand, the *Video-only model* is a natural fit in the context of increasingly intelligent vehicles. Vehicles in the future will have better sensing capability of surrounding traffic. The proposed *video-only model* relies exactly on the location and the distance of surrounding traffic participants. Such information can be reused in more and more intelligent vehicles.

- The *fusion model* further improved emotion recognition performance compared with single modality models. It can be applied to achieve the best user experience if the relevant conditions allow for it.

Our comparison should inform future research in the area on the suitability of these three models. Hence, we believe that our findings actually revealed high flexibility of the proposed method in the deployment.

3.6.4 Limitations

Despite our best efforts, this study has several limitations. First, our data collection took place in real-world traffic, a complex environment. To ensure a competitive baseline, we relied on heart rate variability measures obtained by FirstBeat - a recording device for cardiovascular activity. To record the data, electrodes needed to be attached to the chests of the subjects. Such a procedure is cumbersome, and requires additional training for the subject to correctly attach the electrodes. Moreover, our prototypical setup, its cost, and the cumbersome physiological sensors forced us to limit our sample size to nine drivers. These limitations should be, however, evaluated under the consideration that ours is the first study to explore drivers' emotions inference in a longitudinal setup (over four months). Our comprehensive sensor set that was used to collect information on driver behaviour and the environment of the car covered a wide range of influential factors.

Furthermore, we did not explicitly analyse the relation between emotion and specific driving manoeuvres or traffic events. Such a relation was modelled implicitly by using a neural network. Identifying driving manoeuvres or traffic events remains challenging. Even the state-of-the-art methods can identify only simple manoeuvres, such as *lane changing* or *turning right while accelerating* (Liu et al., 2017; Peng

et al., 2020b; Xie, Hilal, and Kulić, 2018). Future research should address this issue and a more explainable model is promised to better benefit both academia and industries.

Lastly, the LOSO model does not achieve as good performance as personalised model. Generalising from a user-dependent to a user-independent model remains a challenging topic in affective computing community. However, our results show that a generilisability across users is possible, especially for the emotions such as *anger*, *sadness*, and *valence* that are very relevant to driving safety. Further research is needed to investigate the extent to which such generalisation is achievable.

3.7 Conclusion

In-vehicle emotion recognition can enable applications of intelligent automobiles to improve comfort, well-being, and safety by adapting the car to the needs of the drivers. Current applications rely on physiological, facial, and speech-based data for emotion recognition. However, physiological sensors are cumbersome to wear during daily commute, and incur additional costs. Moreover, recording facial expressions and speech may raise privacy-related concerns. In this paper, we leverage recent advancements in generic emotion recognition through user behaviours, and used the idea for in-vehicle emotion inference. We relied solely on data streams of today's cars (i.e., CAN-bus and front-view camera data): a strong advantage that allows for a scalable and privacy-preserving implementation in cars. We collected four months of CAN-bus front-view video data from nine users under naturalistic driving settings on public roads. The in-situ emotions of the drivers' faces were recorded by a monitoring camera and the relevant videos were annotated by using a the state-of-the-art facial expression recognition software - Affectiva.

Our results can be summarised as follows: First, we evaluated our models based on intra-subject and LOSO evaluations, and compared the performance of different sensor modalities. This evaluation revealed that a fusion model that combined CAN-bus data and front-view video data achieved the best results among our models. A single modality model yielded similar performance scores to those of a fusion model, which allows for the flexible deployment of the proposed model when certain sensor modalities are unavailable. Second, we compared the results of the proposed model with an HRV baseline. This evaluation revealed that our models can achieve comparable performance as the HRV baseline. Therefore, inferring driver emotions within a vehicle based on driving behaviour and context by using data from CAN-bus and video segments yields better ubiquity than physiological sensor-based approaches.

Chapter 4

Driver Heart Rate Variability Estimation

The previous chapters have so far described how the proposed psychological (emotion) state recognition approach contributes to driver state monitoring in current vehicles. As a further improvement that incorporates new sensors available in future vehicles, the possibility of the driver's heart rate variability (HRV) monitoring is inspected. The monitoring of HRV does not only improves driving safety, but to a certain degree, is in-line with the vision of digital health that the well-being of people can be monitored and improved in ubiquitous and pervasive manner. However, existing works on driver HRV monitoring suffer from disadvantages ranging from inaccurate measurement in complicated environments to inconvenient deployment. As such, this chapter focuses on the second research question of this thesis:

RQ 2: *Can a driver's HRV be reliably estimated in a non-intrusive manner in future vehicles?*

In the following of this chapter, we start by revisiting the motivation of driver HRV monitoring. Instead of presenting existing works in an independent section, we merged the recent advancement of driver HRV monitoring techniques into the description of the context of this research question because the latest related work comprehensively depicted the current limitations and thus well motivated this topic. The experiment settings and data transformation that are specific to this research question are described and justified. After that, the methodology and the evaluation results are presented. Subsequently, the usability, reliability, and limitations of the proposed approach are discussed in detail. Finally, this chapter concludes the contributions to the community and outlines potential improvements for future research.

4.1 Context and Motivation

Daily driving is an integral part of the day for many people, a fact that is frequently demonstrated by statistics. For example, in Germany 68% of the working population uses their car for commuting and more than 25% of them commute daily more than 30 minutes per direction (Statistisches Bundesamt, 2016). However, in the U.S., about 90% of all citizens (aged 16 or older) drove 2.5 trips daily from 2019–2020 on average, which corresponds to about 1 hour of driving time or 30 miles ($\approx 48.3\ km$) of distance (American Automobile Association Foundation for Traffic Safety, 2020) per day. Moreover, the industry imagines the vehicle as the 3rd living space (after home and workplace) of people (Robert Bosch GmbH, 2016), which has a tremendous impact on people's lives. Nevertheless, driving is still a cognitively demanding task (Stutzer and Frey, 2008). The prolonged driving time induces excessive stress (Chatterjee et al., 2020; Legrain, Eluru, and El-Geneidy, 2015; Zepf et al., 2019), which has the potential to impair mental and physiological health (Stutzer and Frey, 2008). Furthermore, inattentiveness, drowsiness, and fatigue constitute one of the main factors of traffic accidents (Choi et al., 2016). The timely recognition of the driver's states can be of significant benefit to improve driver states with just-in-time intervention (JITI) (Koch et al., 2021; Lee, Elhaouij, and Picard, 2021; Sarker et al., 2014). The recognition and regulation of driver states are particularly meaningful in the era of (semi-)automated vehicles. During the transition to fully automated vehicles (L2, L3 automation), drivers need to be mentally and physically prepared to take over the driving task at any given moment (National Highway Traffic Safety Administration, 2020a). Therefore, the vision of future intelligent cars extends the idea of being a simple means of transportation toward a dedicated space where drivers' mental and physiological states are taken care of. Ultimately, identifying the states of drivers in vehicles is one important step toward safer driving and better life quality. From a broader perspective, the enhanced in-vehicle experience under the concept of ambient intelligence facilitates Internet-of-Things (IoT) enabled the transformation of a vehicle into a well-being and safety platform, where the driving performance, mental and physiological states are improved by restoring driver states in an optimal range, as illustrated in Figure 4.1 (Aarts and Ruyter, 2009; Coughlin, Reimer, and Mehler, 2011).

Heart rate variability (HRV) and its measures are the most promising physiological signals to recognise driver states. Various studies have demonstrated their relevance to infer stats like stress, drowsiness, or inattentiveness. HRV is the variation in the time interval between heartbeats (inter-beat interval, IBI), and it can be characterised by HRV measures in time and frequency domains. In the context of our work, there

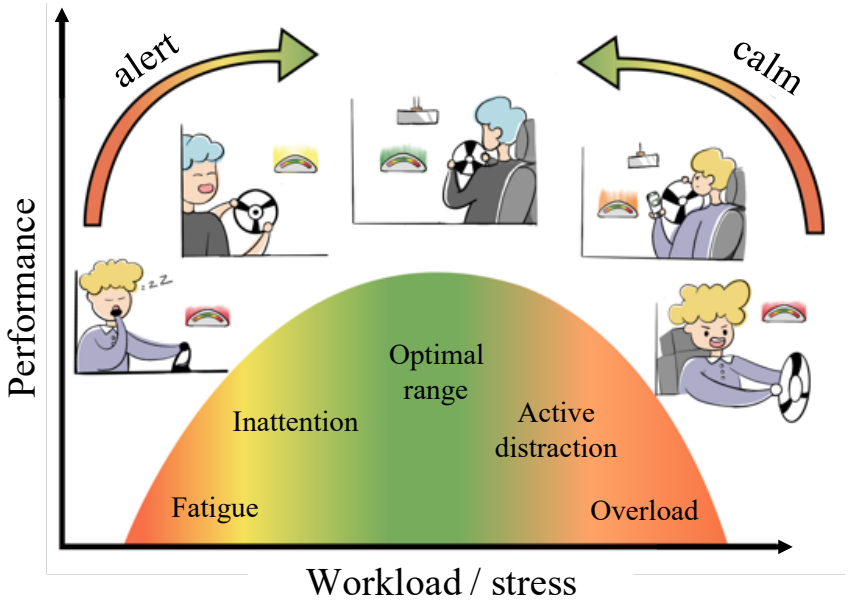


FIGURE 4.1: Driving performance vs. cognitive load according to Yerkes-Dodson law, adapted from (Coughlin, Reimer, and Mehler, 2011). In future cars, intelligent vehicle systems are envisioned to be able to regulate excessive fatigue or stressful states of drivers, in order to further improve driving experience and safety (Coughlin, Reimer, and Mehler, 2011).

are three most relevant and fundamental HRV measures in existing literature. In time domain, the root mean square of successive differences between IBIs (RMSSD) is a widely used measure. Increased RMSSD is associated with fatigue or drowsiness states, whereas stress can cause a decrease in HRV (Hernando et al., 2018; Kim et al., 2018; Lohani, Payne, and Strayer, 2019). Furthermore, Taelman et al. observed that mental tasks can significantly reduce the proportion of successive normal beat to normal beat intervals (NN-interval) with a difference greater than 50 ms (pNN50) (Taelman et al., 2009). In addition to the signal in the time domain, HRV measures in the frequency domain are also powerful indicators. Patel et al. and Vicente et al. found statistically significant evidence that low frequency (LF) and high frequency (HF) ratio (LF/HF ratio) is in alert states higher than fatigue states while driving (Patel et al., 2011; Vicente et al., 2016). To sum it up, excessively low or high states of the depicted HRV measures (RMSSD, LF/HF ratio, and pNN50) are strongly associated with the drivers' cognitive load and psychological states.

As a consequence, researchers and automobile manufacturers have taken pioneer efforts in driver heart rate detection. For example, BMW built a skin-resistance sensor into the steering wheel for heart rate monitoring (Choi et al., 2016; D'Angelo et al.,

2011). Similarly, Toyota and Denso monitored electrocardiography (ECG) and photoplethysmogram (PPG) using a steering wheel equipped with different electrodes and green light LEDs (525nm) (Choi et al., 2016; Osaka et al., 2008; Osaka, 2012). In contrast, Ford and Denso utilised the driver seat (Sakai et al., 2013; Walter et al., 2011; Wartzek et al., 2011). Although these methods seem to promise advanced and widely validated technology (as PPG used in today's smartwatches), researchers agree that these approaches cannot yet provide reliable measurements. For instance, (D'Angelo et al., 2010) evaluated the performance of steering wheel integrated sensor under lab condition and concluded that there was an average error of 6% and the maximal error could escalate to 20%. Such error is far greater than commodity smartwatches, as we evaluated in this study (see Figure 4.10a). In (Wartzek et al., 2011), Wartzek et al. found that seat-integrated sensors could not reliably detect heart rate from drivers in all situations because seat integrated sensors are, for example, vulnerable to the thickness and the material of outer clothing as well as the weight of drivers.

To overcome such drawbacks, Zheng et al. recently designed a radio frequency device and leveraged ultra-wide band (UWB) impulse. The drivers' heart rate can be inferred by analysing of the Doppler frequency shift of UWB signal induced by heartbeat, respiration and ambient noise (Zheng et al., 2020). Although the method of (Zheng et al., 2020) can accurately detect drivers' heart rate, inherent disadvantages exist. First, such a method requires a special purpose UWB device, which is not readily available. Second, due to ambient noise and physical constraint of sampling rate, IBI can only be measured with moderate accuracy (about 50% of IBI measurements have an error greater than 50 ms (Zheng et al., 2020)). Such an accuracy limitation can be tolerated, when only the average heart rate is detected since HR is computed as the inverse of the mean IBI in a certain interval. The noise in the IBI measurement is cancelled out by the mean operation. However, considering critical metrics, such as RMSSD or LF/HF ratio, the inaccuracy will be magnified because these measures are sensitive even to the small inaccuracies in the measurements. Recently, with the advancement of computer vision techniques, remote PPG (rPPG) (McDuff, Gontarek, and Picard, 2014; Wang et al., 2017) has attracted prominent attention. The fundamental principle of rPPG is as follows. Heartbeat (hence blood volume in vessels) induces subtle colour variations on the human skin surface, which can be captured by an RGB camera. Signal processing techniques are then applied to analyse such variation; thus human cardiac activities can be monitored. Although rPPG technique is appealing, many efforts are still needed before it can be applied to the real world scenario. Remote PPG is sensitive to illumination and motion artefacts. More importantly, commodity cameras record video at 30 or 60 Hz, which by

Nyquist–Shannon sampling theorem is insufficient for the accurate measurements of IBIs, of which the variation is at millisecond-level. Existing research on HR/HRV detection using rPPG was conducted in well-defined lab conditions; therefore their generalisability to real-world scenarios remains unclear (Chen et al., 2019b; Gudi et al., 2019; Stricker, Müller, and Gross, 2014; Tasli, Gudi, and Uyl, 2014). In a nutshell, the existing contact-less monitoring methods do not guarantee a reliable measurement of HRV in real world scenarios.

In light of these existing challenges, we propose an alternative way to monitor driver states through HRV. As described above, drivers's cognitive load and mental states are strongly characterised by excessively low or high HRV measures (i.e., RMSSD, LF/HF ratio, and pNN50). Therefore, instead of attempting to derive HRV measures from inherently noisy IBI measurements, we propose a facial expression-based approach to detect the onset of HRV outliers. On the basis of existing literature, we define HRV outlier as samples whose values are one standard deviation below or above the mean (Buda, Khwaja, and Matic, 2021; Smyth and Heron, 2016).

Facial expressions are strongly connected and influenced by the autonomic nervous system (ANS). On the one hand, human cardiac activity is controlled by ANS. The sympathetic nervous system (SNS) accelerate the heart rate through the discharge of epinephrine and norepinephrine while the parasympathetic nervous system (PNS) releases acetylcholine to induce deceleration (Gordan, Gwathmey, and Xie, 2015; Robinson et al., 1966). On the other hand, ANS also functions involuntarily and cope-with affective arousal in reaction to circumstance accordingly (Isaacson, 2013); To estimate HRV from facial expression, we employed the state-of-the-art machine learning scheme and developed a novel tree-based probabilistic fusion neural network approach. Compared with existing contact-less and non-intrusive UWB or rPPG based methods (Wang et al., 2017; Zheng et al., 2020), the advantage of our facial expression-based method and our contribution can be summarised as follows.

- Our approach relies on commodity RGB camera working at 30 FPS, which is very likely to be integrated in future vehicles as a part of driver monitoring systems (European Commission, 2020; Official Journal of the European Union, 2019). Thus, no additional UWB devices are needed.
- We verified our approach based on around $3,400$ km (68.6 hours) of driving data collected from a two-week field study, involving nine participants during uncontrolled daily driving activities on public roads.

- A novel tree-based probabilistic fusion neural network approach is developed to optimise HRV estimation performance. The proposed tree-based probabilistic fusion framework outperformed conventional convolutional or recurrent neural networks and classic tree based machine learning models by up to 6.9% in balanced accuracy.
- We benchmark our method against consumer smartwatch measurement. Smartwatch can be seen as a proxy of the upper bound of rPPG since its close contact with the skin mitigates a large portion of noise due to illumination and motion artefacts. Our evaluation shows that the proposed approach can even outperform high-end consumable smartwatches by a large margin.
- To the best of our knowledge, this is the first study that verifies the feasibility of facial expression-based HRV outlier detection based on driving data collected from public roads in real driving scenarios. Since the overall environment is challenging compared with laboratory conditions, our results are likely to be more reliable.

The remainder of the chapter is organised as follows: We present our field study in Section 4.2. We introduce our methods for HRV estimation in Section 4.3. Section 4.4 summarises the results of our methods. The implication of the method and discovery is discussed in Section 4.5. Finally, Section 4.6 presents the conclusion.

4.2 Experiment Settings

We conducted a two-week field study with nine daily commuters (originally then; one participant's data were removed due to corruption) during their normal driving routine on public roads. A variety of sensory data, including HRV, facial expression, and smartwatch records, is collected from the daily the driving activity of participants in naturalistic condition. The participants were supposed to use the vehicles for their daily drives, including business trips and vacations.

4.2.1 Characteristics of Driving Activity

It was crucial for our dataset to capture representative driving situations. This subsection presents some important statistics related to the subset of the whole dataset where the precise HRV recording (recorded by Firstbeat Bodyguard 2) was available.

As described in Section 2.2, all the drivers wore the Firstbeat Bodyguard 2 for a period of two weeks. After data cleansing, we had about 68.6 hours of video data with associated HRV measurements during driving. The total driving distances of each participant are plotted in Figure 4.2. Most drivers drove for reasonably long distances (more than 300 *km*) during the field study. The GPS records of the vehicles are presented as a heatmap in Figure 4.3. As shown in this figure, most participants drove around the area of Stuttgart, Germany. An additional heatmap covering a driver's vacation activities is also provided in Figure B.1. Overall, our dataset covered a wide range of daily driving activities like commuting, shopping trips, and leisure activities at the weekend.

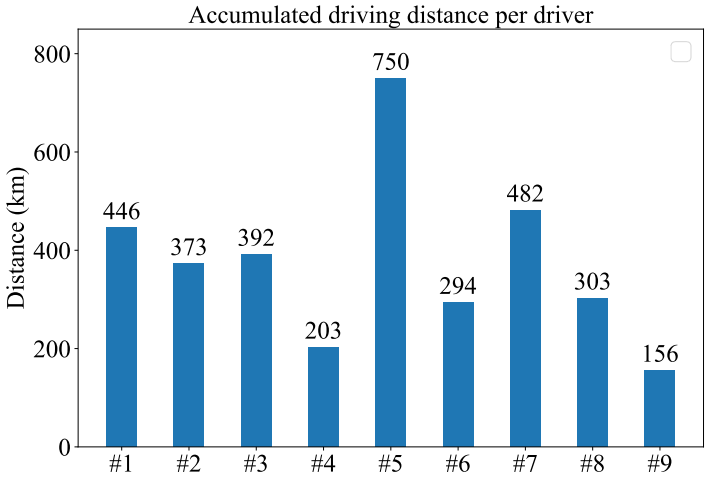


FIGURE 4.2: Accumulated driving distance

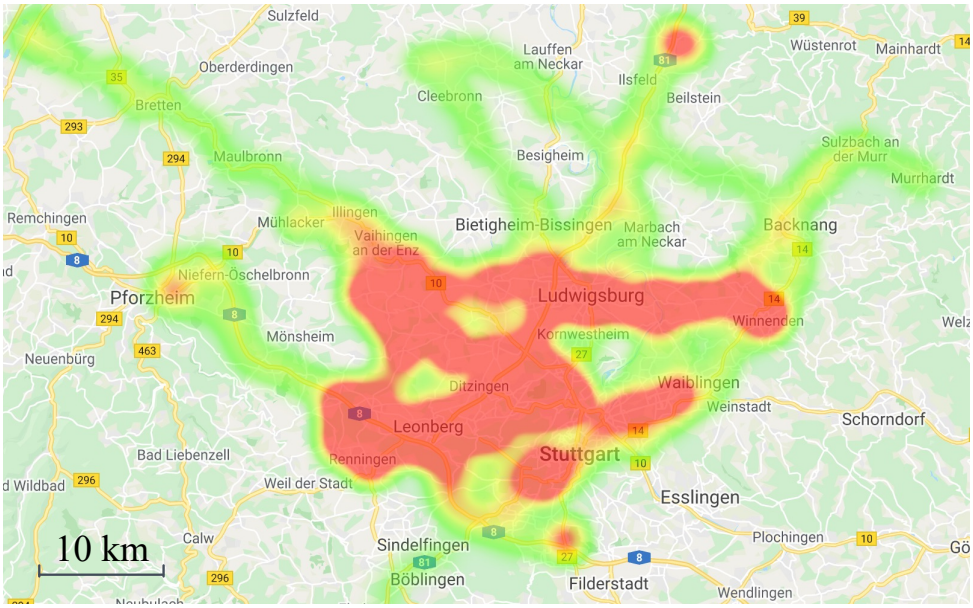


FIGURE 4.3: GPS heatmap of the most active area

4.2.2 Characteristics of Heart Rate Variability Measure

HRV is measured over a period of time. We applied an overlapping sliding window with a length of 5 *min* and a step size of 30 *s* to compute HRV measures (i.e., RMSSD, LF/HF ratio, and pNN50). The choice of 5 *min* follows the convention of short term HRV measurement (Malik et al., 1996; Shaffer and Ginsberg, 2017). We refer to such 5-*min* segments as ***HRV segments***. The entire facial expression-based HRV detection framework is shown in Figure 4.4. We take facial expressions in ***HRV segments*** as input data to estimate the HRV measures associated with the corresponding segments. From the recording of Firstbeat Bodyguard 2, the ground truth of HRV measures is computed based on a standardised wearable data processing toolkit (Föll et al., 2021). These HRV segments are randomly shuffled for training and testing. To avoid the intersection between training and test datasets caused by overlapping sliding windows, we discard HRV segments that intersect both datasets for each shuffle, as illustrated in Figure B.2.

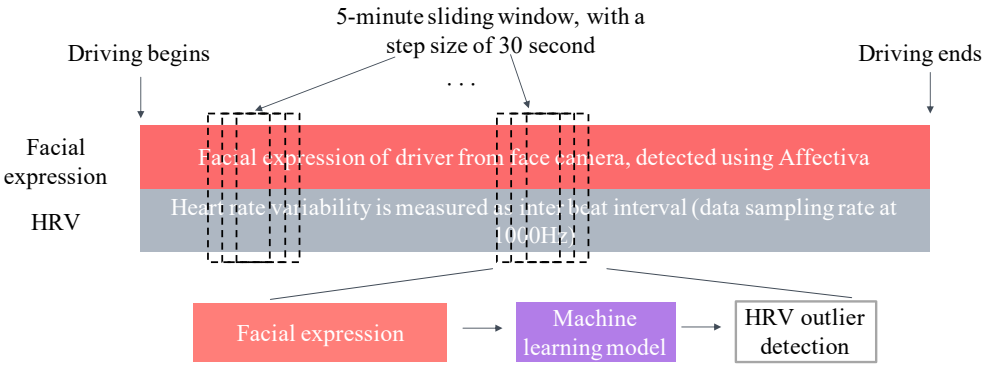


FIGURE 4.4: HRV outlier detection framework

Next, we inspect the distribution of HRV, which is illustrated in Figure 4.5 to 4.7. Owing to the influence of age and gender, there is significant difference among participants in terms of the median and range (Malik et al., 1996; Voss et al., 2015). To account for such individual factors, we define HRV outlier detection as a binary classification problem and predict whether a given driver's HRV measures are excessively low or high with respect to his/her personal empirical distribution. We distinguished between low and high HRV outlier detections, as formulated in Equation 4.1 and 4.2, respectively. Such definition is similar to (Buda, Khwaja, and Matic, 2021; Smyth and Heron, 2016), in where they defined outliers for stress or mental states estimation as one standard deviation below or above the mean. Consequently, HRV measures within one standard deviation of the personal mean are considered normal. It means that we develop two machine learning models, one for the detection of low outliers of HRV measures and the other for high outliers.

$$low\ detector = \begin{cases} low\ outlier, < per.mean - per.std \\ rest, >= per.mean - per.std \end{cases} \tag{4.1}$$

$$high\ detector = \begin{cases} rest, <= per.mean + per.std \\ high\ outlier, > per.mean + per.std \end{cases} \tag{4.2}$$

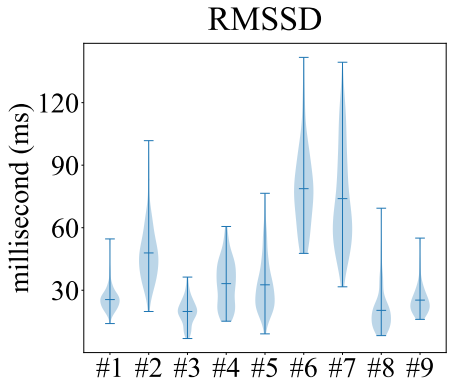


FIGURE 4.5: Distribution of RMSSD

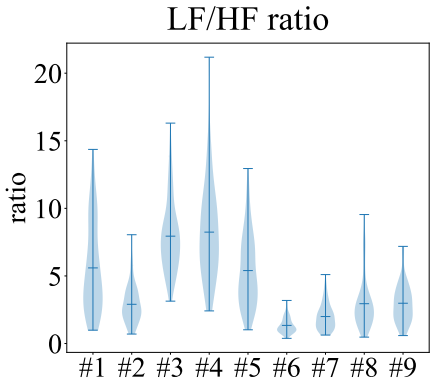


FIGURE 4.6: Distribution of LF/HF ratio

We performed data cleansing and removed IBI artefacts (< 250 *ms* or > 2000 *ms*). We removed HRV segments where the driver faces appeared in less than 70% of video frames as well as HRV segments with no valid IBI signal. Finally, we obtained in total 3876 HRV segments, the distribution of low and high outliers, and normal samples are given in Table 4.1.

4.3 Methodology

This section explains our approach to infer HRV outliers from facial expressions.

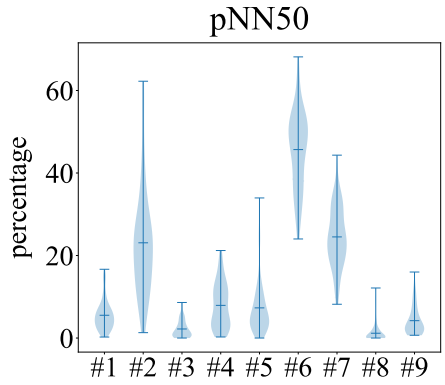


FIGURE 4.7: Distribution of pNN50

TABLE 4.1: Sample Counts for Different Category

	RMSSD	LF/HF ratio	pNN50
low	513	584	423
normal	2840	2636	2858
high	523	656	595

4.3.1 Data Preprocessing

Detection of Facial Action Units. The facial action units (AUs) is used in the facial action coding system (FACS) to describe the muscle movement currently active in the face, such as "nose wrinkle" or "cheek raise" (Ekman and Rosenberg, 1997) Based on the active level and the combination of AUs, facial expressions such as anger, fear, and joy can be quantitatively determined.

The manual coding process of FACS requires profound expert knowledge and is laborious. To alleviate this problem, we leveraged the automatic FACS coding algorithm from Affectiva, a spin-off of MIT's Media Lab. Affectiva's facial expression recognition technology uses computer vision and deep learning techniques to first detect the active level of AUs, based on which another mapping function is established between facial expression and AUs (Mcmanus, 2020a). The Affectiva's major advantage is that it is built on a very large foundational dataset, consisting of more than 9.7 million facial images of people, with more than 5 billion facial frames (Mcmanus, 2020a). Additionally, based on the in-vehicle data of more than 20,000 hours featuring more than 4,000 unique individuals, Affectiva is well optimised to automotive in-cabin environment (Mcmanus, 2020a). Given these features, Affectiva's solution can reliably capture facial movements. In this study, we used one of the latest stable versions (ics-2.2.1).

Feature Engineering. Affectiva detects AUs for each frame and presents the results

as the activation level of each AU in the range of 0 to 100. The entire list of detected AU is the following:“browRaise”,“browFurrow”, “noseWrinkle”, “upperLipRaise”, “mouthOpen”, “eyeClosure”, “cheekRaise”, “eyeWiden”, “innerBrowRaise”, “yawn”, “blink”, “blinkRate”, “lipCornerDepressor”, “lidTighten” and “smile” (in total 15 AUs). We build feature vectors (FVs) for AUs through a sliding window, with both the length and the step size equal to five seconds. In each sliding window, we compute mean, min, max, median, standard deviation, quantile-25%, quantile-75%, kurtosis and skewness for each AU. That is to say, for every 5 seconds, a 135-dimension (15 AUs $\times$ 9 features) FV is generated. From a 5-min HRV segment, a sequence of 60 FVs ($5min$ / $5s$ = 60) is generated.

In addition to AUs, HRV is heavily influenced by the time of day. We incorporate this prior information by including time features defined as current time (formatted in the 24 h-scale), day of the week, an indicator of driving at night, seconds before dawn, seconds after dusk, seconds before sunrise, and seconds after sunset. The last four features were set to zero if driving had occurred after or before the corresponding event. By merging the time features to each FV, the final input to the machine learning models has the shape of $60\,steps \times 142\,dim$.

4.3.2 Machine Learning Approaches

Standard Pipeline. We first verify the feasibility of the HRV outlier detection in the wild by exploring a random forest model. Despite the simplicity of tree-based models, they often outperform more complicated models such as neural networks or support vector machines (SVM) (Fernández-Delgado et al., 2014). This is especially the case with a lack of prior insight about underlying data property or domain knowledge (Wang, Aggarwal, and Liu, 2017).

Our random forest based pipeline is depicted in Figure 4.8. In the training phase, we assign all FVs in an HRV segment the same label as the HRV segment, meaning that the input instance to the random forest is each FV. In the test phase, we perform prediction on all FVs in each HRV segment. The final prediction for one HRV segment is aggregated from prediction results of all FVs in that HRV segment. In this study, we use the majority vote as the aggregation function.

The input FVs are sequences of time-series data. Therefore, to further explore the possibility of other machine learning models, the choice of random forest can be replaced by prevalent (1D) convolutional neural network (CNN), recurrent neural network (RNN), and multilayer perceptron (MLP), etc. We used random forest as

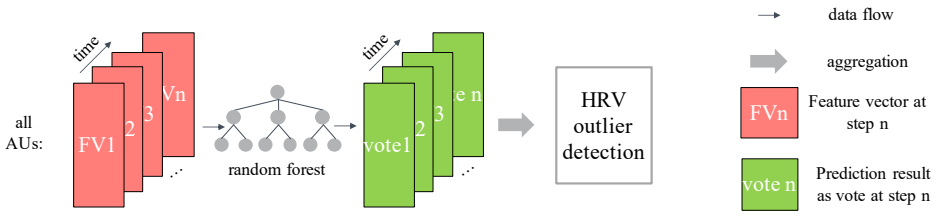


FIGURE 4.8: HRV outlier detection using standard random forest pipeline

well as various neural networks as baseline method and the evaluation is presented in Section 4.4.

Tree-based Probabilistic Fusion Network (TPFN). As we will show in Table 4.4 and 4.5 in Section 4.4, when tree-based models are applied in the standard pipeline, they usually outperform neutral network models. The tree-based model often perform better than other models in practise (Fernández-Delgado et al., 2014; Wang, Aggarwal, and Liu, 2017). The reason is that the hierarchy decision stage of tree-based models does not impose restrictions on the distribution of inputs; the merit of the ensemble mechanism of random forest makes it extremely robust to unseen data (Fernández-Delgado et al., 2014; Wang, Aggarwal, and Liu, 2017; Zhou and Feng, 2017). Unlike neural networks whose architecture is sensitive to specific data distribution and requires profound domain knowledge, recent work suggested that random forests can help discover the underlying structure of data (Fernández-Delgado et al., 2014; Wang, Aggarwal, and Liu, 2017). As such, we develop a hybrid model that uses a tree-based model to create a probabilistic embedding from data, which is further fused and processed by a neural network. The details of our proposed model are explained as follows.

We first compute the probability embedding of each AU. This is performed by building 15 random forests for the 15 feature subsets of all AUs. Each feature subset contains not only features of the corresponding AU, but also the prior mentioned time features. Therefore, each random forest takes FVs of 16 dim. (nine statistical features from AUs and seven time features) as input. We train these 15 random forests similar to those in Figure 4.8. After that, instead of aggregating the predictions of the random forests, we take the prediction probability (with closer to 0 being more likely to be class 0, and vice versa), which is again a time-series sequence of form 60 step $\times$ 15 AUs, as input to a neural network. The neural network take the fusion of the probability from the random forests and further predicts the HRV outlier for the entire sequence. In this study, we used a multilayer perceptron (two layers, each with 16 neurons and sigmoid as activation) to classify on every step the fused probability

and then with a final classification that is aggregated (by majority vote) from the 60 votes.

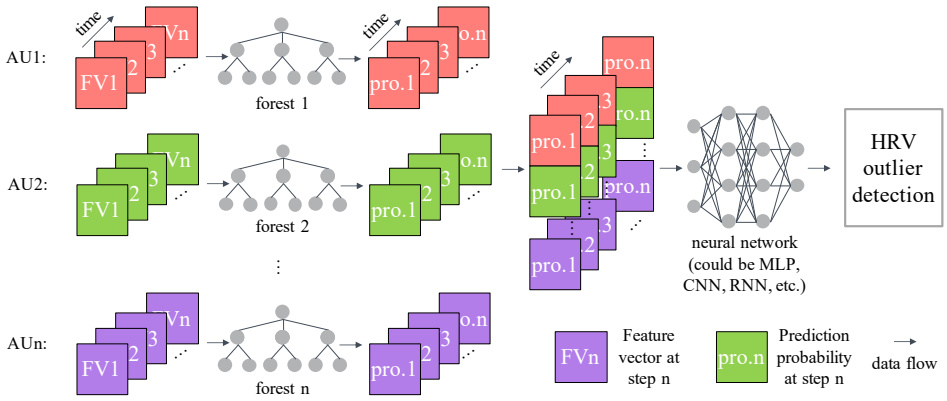


FIGURE 4.9: Tree based probabilistic fusion model: feature vectors are transformed into probabilistic embedding before fed into neural networks

4.4 Evaluation

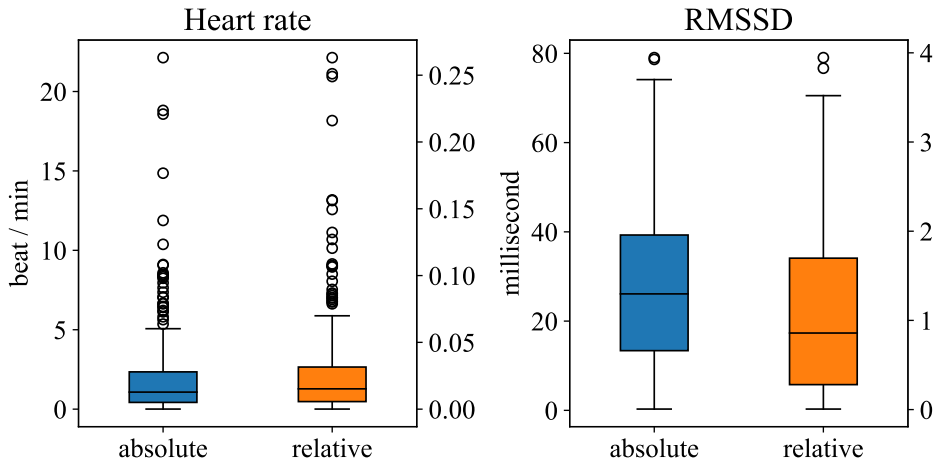
In this section, we evaluate the proposed method against state-of-the-art machine learning models. The evaluation is performed by constructing a *general model* for all drivers.

In the following, we will provide an insight into the HRV measurement accuracy of the current high-end commodity smartwatch compared with medical-grade heart rate monitor (Firstbeat Bodyguard 2). Next, we comprehensively compare the proposed approach with various baseline methods.

4.4.1 Measurement Accuracy of Smartwatch

Smartwatches and other wearable devices are becoming popular in people's daily life. The low cost and ubiquitous property make them an ideal tool for health monitoring. Therefore, we should first inspect if their measurement accuracy meets the requirement of HRV detection in the wild.

For this purpose, we use the measurement of the Firstbeat as the gold standard to compute the errors of smartwatches. The absolute and relative errors of mean heart rate and RMSSD of HRV segments are illustrated in Figure 4.10. In this study, we have an overlap of 21.25 hours of Firstbeat data with smartwatch measurements. Due to the in-the-wild property of the experiment, drivers sometimes did not wear the smartwatch.



(A) Heart rate (B) RMSSD

FIGURE 4.10: Absolute and relative errors of high-end smartwatch compared with Firstbeat bodyguard 2 (Firstbeat Technologies Oy, 2019)

It it obvious in Figure 4.10a that the smartwatch can very accurately measure the average heart rate. The mean value of the absolute error is only around 1 beat per minute. This magnitude of error agrees with the latest systematic evaluation of smartwatches (Hernando et al., 2018; Shcherbina et al., 2017). However, the errors become significant when using the measured IBI from a smartwatch to compute RMSSD, as illustrated in Figure 4.10b. The mean of the relative errors is almost 100%. More comparisons between smartwatch and Firstbeat measurements are given in Figure B.5. Although the sensors of smartwatches tremendously improved and will continue (e.g., ECG monitoring is now available in certain smartwatches, the prerequisite of its usage is that the users must sit still without arm movements; thus, limited applicability while driving (Hernando et al., 2018)), the current high-end smartwatch that measures the accurate mean heart rate does not provide reliable HRV measurements while driving.

4.4.2 Comparison with Baseline Methods

In this subsection, we present the baseline methods to be compared and analyse the results quantitatively.

Baseline Methods. This part describes the baseline methods in detail. On the one hand, the chosen baselines, such as smartwatch and time models, are used to demonstrate that our proposed facial expression based approach is a good and necessary complement of currently prevalent heart rate monitoring methods; on the other hand,

the comparison with the tree-based and neural network models can demonstrate that the proposed tree-based probabilistic fusion is an efficient way to learn data representation.

- **Smartwatch Model.** Although smartwatches exhibit unreliable measurements of HRV, as described in Section 4.4.1, it is still meaningful to evaluate whether the noise in the smartwatch is consistent. This means, for example, if the noise adds a consistent offset to the HRV measurements, HRV outlier detection can still be accurately performed since we are interested in whether HRV is lower or higher than the personal baseline level. We refer to the smartwatch model as **SM**.

- **Time Model.** It is well known that the time of day has a strong impact on HRV (Tsuji et al., 1996). For example, HRV tends to be higher during working hours than at night because the body must react to the accumulated stress and cognitive load. We demonstrate the time-dependent variation of RMSSD in Figure 4.11. More examples of LF/HF ratio and pNN50 are given in Figure B.3 and B.4 in the Appendix B.

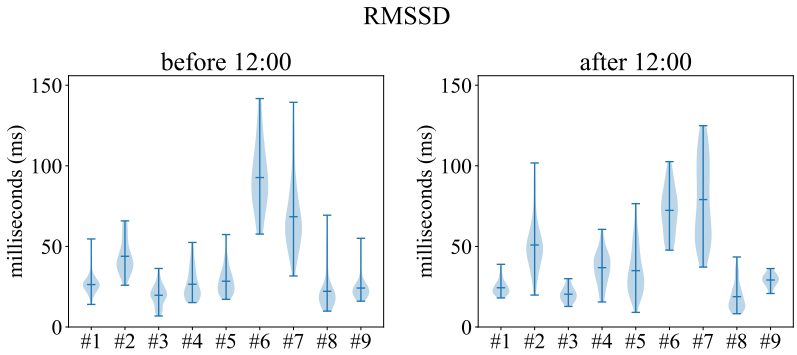


FIGURE 4.11: RMSSD of the nine drivers in different time interval

Therefore, it is crucial to inspect the possibility of inferring HRV outlier purely based on time. To this end, instead of defining a rule based model, we build a Time Model (referred as **TM**) by constructing a random forest using only time features (7D). The TM resembles the settings in Figure 4.8 except for the input FVs.

- **Tree Based Models.** As described in Section 4.3.2, random forest (refered as **RF**) can be used as the machine learning back-end in the pipeline. To explore the feasibility of other tree-based models, we further replace random forest with one of the latest tree-based models, the Deep Forest (referred as **DF**) (Zhou and Feng, 2017). For RF, DF and tree-part of TPFN, a grid-search of

parameters is performed. The candidate parameters are described in Table B.1. The optimal parameters for all tree-based models are determined as depth of tree = None (i.e., unlimited depth), number of trees = 200, min samples split = 2, min samples leaf = 1.

- **Neural Network Models.** Over the last decade, neural network techniques have experienced tremendous improvement. Therefore, it is meaningful to benchmark our proposed tree-based probabilistic fusion approach with them. We implemented 1D convolutional neural network (referred as **CNN**), Multilayer-perceptron (referred as **MLP**) as well as recurrent neural network (referred as **RNN**) to the time-series FVs. To be more precise, the **CNN**, as depicted in Figure 4.12, consists of two cascaded 1D convolutional filters (kernel size = 3, filter size = 64, dropout rate = 0.5, and activation = sigmoid) followed by a linear fully connected (FC) layer with 16 neurons and a Softmax operation that reduces the flatted convolutional output to two dimensions, corresponding to the binary classification.

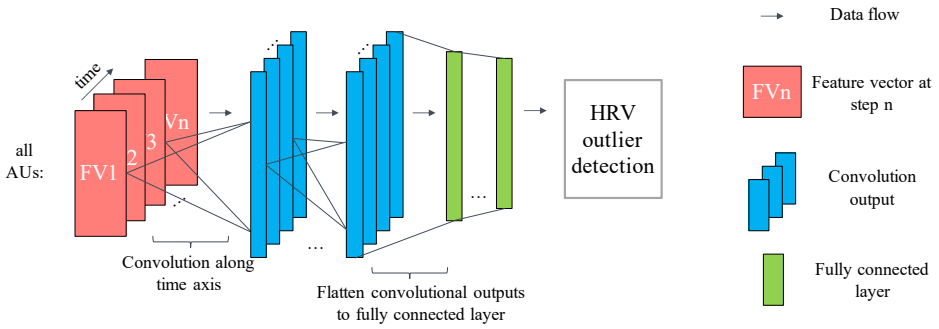


FIGURE 4.12: CNN baseline model

RNN, as shown in Figure 4.13, uses two recurrent layers (dropout rate = 0.5) with 16 gated recurrent units (GRU) followed by a linear FC layer with 16 neurons and a Softmax operation that reduces the hidden states of GRUs to two dimensions, similar to CNN.

MLP resembles the pipeline in Figure 4.8, where random forest is replaced by a two-layer multilayer perceptron (activation = sigmoid, dropout rate = 0.5) with 32 units in each layer. The classification is performed on each FV and the final prediction is the aggregation (majority vote) of all FVs in an HRV segment.

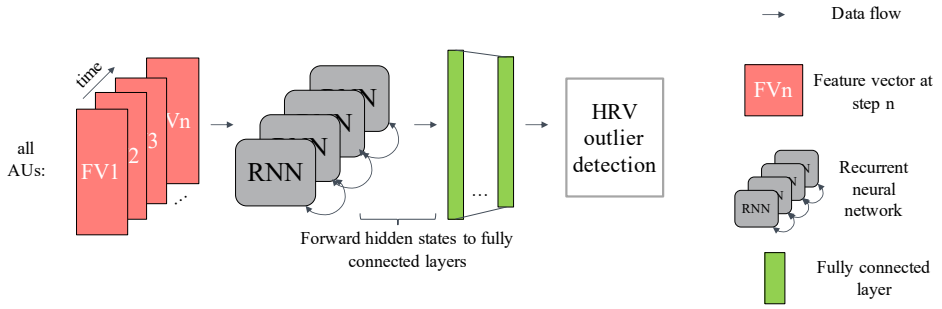


FIGURE 4.13: RNN baseline model

The chosen architectures for CNN, RNN and MLP are similar to the ones used in (Schmidt et al., 2019; Taylor et al., 2020), which have been proven to be effective in predicting various physiological and psychological states. Additionally, the optimal parameter settings of the above mentioned neural networks were determined using grid search. This is done by applying a 5-fold cross validation to the training dataset. Subsequently, each network is re-trained on a total training dataset with the optimal parameters. This procedure of parameter searching is similar to that of (Taylor et al., 2020). The candidate parameters for the grid search are described in Table B.2 - B.4. Furthermore, for all networks, we further grid searched on optimiser (SGD and ADAM), normalisation schemes (Z-score normalisation, min-max normalisation, and logarithmic transformation[1]), and gradient clipping schemes (norm type = 2-norm, the option for max. norm was iterated over 1, 10, 100). Finally, all neural network models (including the neural network part of TPFN) are trained by ADAM with a learning rate of 0.005; Z-score normalisation and gradient clipping with max. norm = 10 are applied. The loss function is defined as cross-entropy.

Numeric Results. We perform the binary classification on an unbalanced dataset ($majority : minority \approx 82\% : 18\%$). Therefore *balanced accuracy* is used as the metric, which is an unweighted mean of accuracy over all classes. Thus, this metric is not biased towards the majority class and can provide a more accurate evaluation of the overall performance. As relevant HRV metrics, we selected RMSSD and evaluated LF/HF ratio and pNN50 since they are closely related to mental states, as explained in Section 4.1.

We first compare the proposed approach with HRV outlier detection based on smartwatch measurements. That is to say, for SM we computed HRV measures from smartwatch-measured IBI and use the computed HRV measures to detect HRV outliers. Smartwatch measurements are available for 21.25 hours of 68.58 hours. To

[1] to avoid numerical issues, logarithmic transform is applied as $A = log(|A| + 1)$

ensure a fair comparison, the proposed TPFN is trained for the remaining 47.3 hours. After that, TPFN and SM were validated on the same dataset where smartwatch measurements are available. The result is described in Table 4.2 and 4.3. The proposed TPFN method outperforms SM in all cases, with an improvement ranging from 3.6% to 13.1%. The evaluation demonstrated that the IBI measured by smartwatches could not precisely compute HRV measures despite accurate heart rate measurement. The measurement noise does not constitute a constant offset, making HRV outlier detection based on smartwatches imprecise.

TABLE 4.2: Balanced accuracy of low HRV outlier detection, smartwatch (SM) vs. proposed solution (TPFN)

Model	RMSSD	LF/HF ratio	pNN50
SM	68.2	52.7	55.1
TPFN	**73.3**	**60.3**	**68.2**

TABLE 4.3: Balanced accuracy of high HRV outlier detection, smartwatch (SM) vs. proposed solution (TPFN)

Model	RMSSD	LF/HF ratio	pNN50
SM	64.7	58.6	61.0
TPFN	**68.3**	**70.6**	**71.9**

Next, we compare the proposed TPFN with prevalent machine learning models. We randomly split the dataset into train (70%) and test (30%) sets. The final results are presented as the average of 10 repeated experiments using 10 different random seeds. The standard deviations of the 10 repetitions are indicated in brackets in corresponding tables.

The results are presented in Table 4.4 and 4.5. TM performs by distance the worst despite the strong correlation between HRV and time (Malik et al., 1996). Neural network approaches (CNN, RNN and MLP), despite their higher complexity, achieve worse results than tree-based models (RF and DF). Finally, the best performance is achieved by the proposed hybrid TPFN model that combines the merits of both tree based model and neural networks. The TPFN model outperforms other best performing baseline models by an average of 3.4% and up to 6.9% in balanced accuracy.

To better visualise the performance of the outlier detection, the confusion matrices of TPFN are plotted in Figure 4.14 - 4.16. The confusion matrices show that the proposed TPFN is not particularly biased towards the majority class, except for the high outlier detection of LF/HF ratio in Figure 4.15. Meanwhile, the model maintains

TABLE 4.4: Balanced accuracy of low HRV outlier detection

Model	RMSSD	LF/HF ratio	pNN50
TM	60.3 (1.7)	58.0 (2.2)	61.5 (2.5)
RF	62.4 (2.1)	61.6 (2.7)	66.8 (2.2)
DF	62.8 (2.8)	61.8 (2.7)	66.6 (2.2)
CNN	59.1 (3.3)	57.9 (2.9)	55.3 (3.7)
RNN	58.4 (3.2)	56.1 (3.1)	57.3 (3.3)
MLP	60.5 (2.5)	55.5 (5.3)	56.3 (4.7)
TPFN	**69.7 (2.2)**	**65.3 (2.7)**	**71.4 (2.1)**

TABLE 4.5: Balanced accuracy of high HRV outlier detection

Model	RMSSD	LF/HF ratio	pNN50
TM	56.0 (2.4)	61.4 (2.8)	59.2 (2.6)
RF	65.5 (2.3)	64.5 (2.7)	65.5 (2.4)
DF	66.8 (2.4)	64.8 (2.7)	65.8 (2.3)
CNN	56.0 (3.5)	60.2 (3.4)	60.3 (4.7)
RNN	60.5 (2.3)	61.1 (2.7)	62.2 (4.2)
MLP	59.7 (3.8)	58.1 (3.8)	60.4 (4.5)
TPFN	**68.3 (2.1)**	**65.7 (2.7)**	**69.2 (2.2)**

low false negative and false positive rates, as illustrated in the sub-diagonals of the confusion matrices.

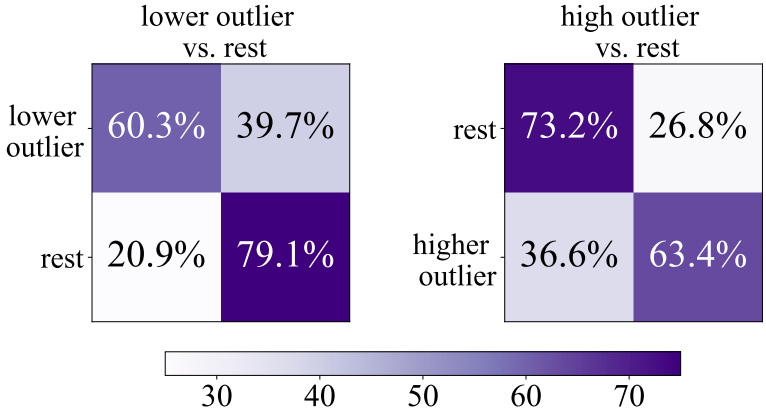


FIGURE 4.14: Confusion matrix of RMSSD outlier detection

4.5 Discussion

In this section, we discuss the proposed approach in terms of prediction usability, reliability and potential limitations.

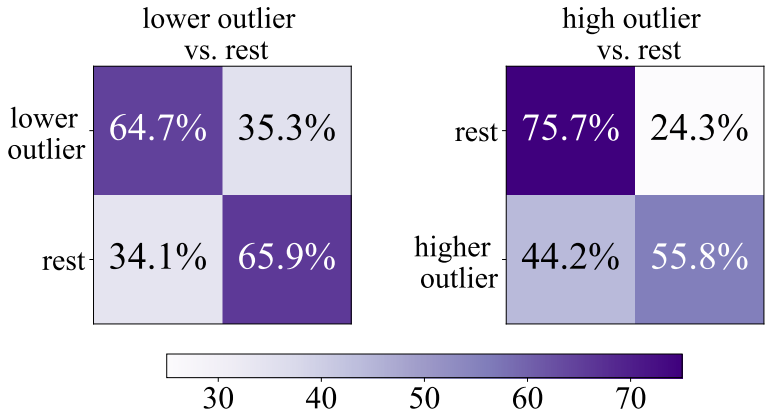


FIGURE 4.15: Confusion matrix of LF/HF ratio outlier detection

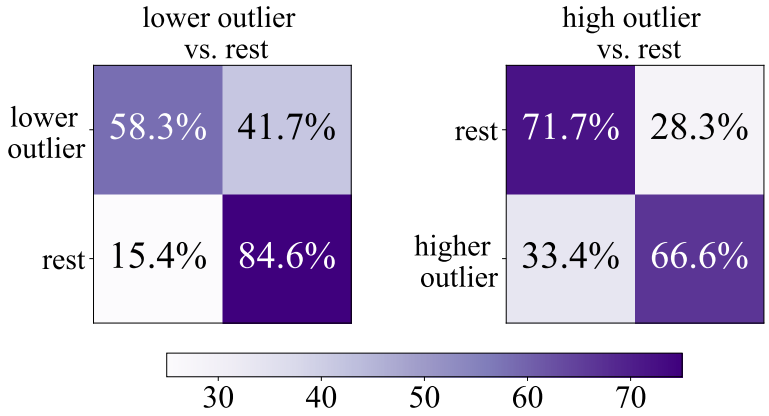


FIGURE 4.16: Confusion matrix of pNN50 outlier detection

4.5.1 Usability

The proposed approach relies on a driver monitoring camera. Although such a camera is not widely installed in current vehicles, it is becoming an integral and essential component of future cars. The reason is that a driver monitoring camera is an essential safety feature that prevents inattention or drowsiness while driving. European Union (EU) is the pioneer in pushing forwards this safety feature. In 2019, a general safety regulation was passed by the EU Council of Ministers. The safety regulation requires that all new vehicles on the EU market must install advanced safety systems to prevent distraction and drowsiness. Such an advanced safety system is very likely to be implemented through a driver monitoring camera (European Commission, 2020; Lyrheden, 2020; Official Journal of the European Union, 2019). Starting in 2022, all new type-approved vehicles with a certain level of autonomy must fulfil this requirement. By 2026, this law will cover all newly produced cars regardless of their level of automation (European New Car Assessment Programme, 2021; Lyrheden, 2020). In the United States, two safety-related traffic bills have been introduced or passed (H.R.2 - Moving Forward Act and S.4123 - SAFE Act of 2020). This may lead to the requirement that driver monitoring camera becomes mandatory in new vehicles (Lyrheden, 2020). In China, the regulations requiring long-distance trucks to use driver monitoring have already been implemented in certain regions, in particular for vehicles transporting hazardous goods. More similar regulations are expected to follow (Lyrheden, 2020). We can anticipate that driver monitoring cameras will become essential and mandatory in many regions of the world in the future. The proposed solution can be well integrated into future cars without any additional hardware cost.

The proposed solution plays an important complementary role to the emerging driver monitoring solutions such as activity recognition and gaze detection (Vicente et al., 2015; Xing et al., 2019). While driver activity recognition and gaze detection algorithms can infer whether a driver's behaviours are allowed during driving, these algorithms do not guarantee if a driver's mental states is favourable. Various studies suggested that incremental cognitive load impact drivers' visual behaviour and their gaze is therefore focused on the central road region (Sodhi, Reimer, and Llamazares, 2002; Victor, Harbluk, and Engström, 2005). Even worse is that increased cognitive load reduces drivers' awareness of incidents occurring within the restricted visual field, namely "look but fail to see" (Coughlin, Reimer, and Mehler, 2011). Such shortcomings of driver activity or gaze monitoring techniques can be well compensated by our proposed algorithm, which assesses the cognitive load of drivers through HRV estimation.

Finally, the monitoring of HRV measures provides a significant derivative benefit from the well-being perspective. Driving is not the only source of stress and mental load in daily lives. Occupational burn-out, sentimental relation between couples, and mood disorders, etc., can lead to sub-optimal states and could manifest themselves in the changes of HRV measures (Holzman and Bridgett, 2017; Lo, Wei, and Hwang, 2020; Malik et al., 1996). The HRV monitoring technique in combination with well-being interventions that regulate drivers' psychological states (Koch et al., 2021; Lee, Elhaouij, and Picard, 2021), in essence, does not only reduces stress and cognitive load from driving, but also from other daily events (Coughlin, Reimer, and Mehler, 2011). The smart vehicles in the future should not only be a tool for transportation, but also an intelligent 3rd living space integrated with a wellness platform (Coughlin, Reimer, and Mehler, 2011).

4.5.2 Reliability

The proposed HRV estimation solution provides supportive service to improve the user experience. Upon the detection of excessively low or high HRV measures, intelligent vehicle systems can deliver corresponding interventions to regulate the sub-optimal states of drivers. State-of-the-art driver stress or mental regulation strategies mainly consist of music or mindfulness intervention, breath exercise, control of ambient auditory, lighting or aero (wind) feedback, and odour stimulation, etc. (Balters et al., 2020; Dmitrenko et al., 2020; Koch et al., 2021; Lee, Elhaouij, and Picard, 2021; Paredes et al., 2018b).

Unlike obstacle avoidance or pedestrian detection systems that have almost zero tolerance for false detection, the HRV estimation in our context can engage in ambiguity when the system is uncertain about its estimation. This is in line with the Guidelines for Human-AI Interaction proposed by Amershi et al. (Amershi et al., 2019). There is a certain grey zone that tolerates ambiguous decisions. In the case of uncertainty, the reliability of the system can be further improved by adopting the interaction between system and users, for instance, through verbal communication (Rudovic et al., 2019), an inquiry of the necessity of intervention (Koch et al., 2021), or adjusting intensity/option of intervention (Nahum-Shani, Hekler, and Spruijt-Metz, 2015), etc. On the other hand, interventions yield a stronger effect, especially if users are in a sub-optimal state (and hence a state of high "vulnerability"), because more potential for improvement exists. That being said, wrongly applied interventions (i.e., user in the optimal state) to regulate low HRV measures are unlikely to move the user from optimal state to a state of high HRV measures (Nahum-Shani, Hekler, and Spruijt-Metz, 2015). If interventions are provided based on wrong HRV estimation,

the consequence is not as dangerous as, for instance, miss-detection of lane marks or pedestrians.

4.5.3 Numerical Issue and Revisit of Tree-based Models

Our intensive experiments showed that tree-based models (RF, DF, and TPFN), despite their simplicity, outperform prevalent neural network models. This is not surprising: In a recent research (Fernández-Delgado et al., 2014), Fernandez et al. compared random forest against other prevalent models like SVM and various type neural network models, and concluded that random forest was statistically significantly better than other machine learning models.

To further explore the merit of tree-based models in our context, we visualise the mean and maximal absolute values of the input feature vectors in Figure 4.17 and 4.18. As can be seen, while the average input values are very small, the maximal values could scale up to 100, which is around two orders of magnitude larger than the mean. This high variant property of the input makes it difficult for neural networks to minimise loss. Indeed, both neural networks and tree-based models can approximate arbitrary decision boundaries. They differ in that neural networks learn a continuous decision boundary according to Universal Approximation Theorem (Hornik, 1991). In comparison, a tree-based model can create a decision boundary based on the minimisation of Gini-index and therefore is not constrained on continuity (Breiman, 2001), which means the decision boundary of tree-based model can jump between very large and very small values. Due to this reason, tree-based models perform better on our high variant input data.

While various schemes (Z-score normalisation, min-max normalisation, logarithmic transformation, and gradient clipping) have been proposed to handle the high variance of input data, our evaluation demonstrated their limited capability. In comparison, our proposed TPFN first embeds input data into a probability space via a tree-based model and after that, the fusion of the probability intermediate outputs is further processed by a neural network. Therefore, the proposed method avoids numerical issues due to the high variance of input data. Our evaluation showed that the proposed TPFN outperforms pure neural network models and can further boost tree-based models.

Though deep learning techniques are becoming dominating in the machine learning domain, our discovery should inspire researchers to revisit the merit of tree-based

models, especially in practice when it comes to the area where few domain knowledge is available or/and the underlying data distribution could cause unstable behaviour of neural networks (Wang, Aggarwal, and Liu, 2017).

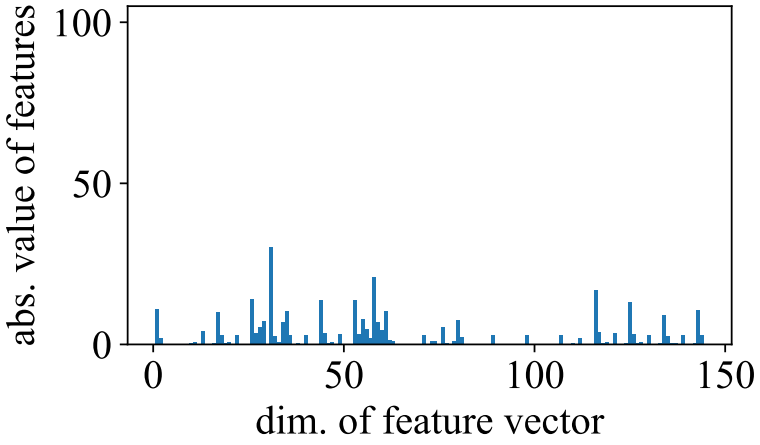


FIGURE 4.17: Mean value of each dimension

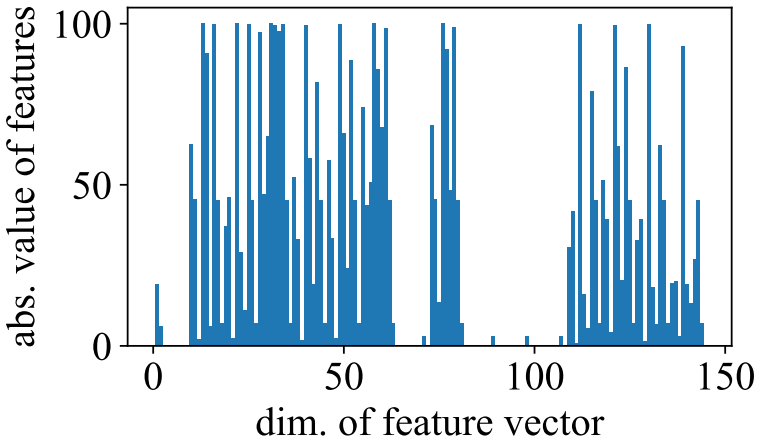


FIGURE 4.18: Maximal value of each dimension

4.5.4 Limitations

This research should be assessed considering its limitations. Even though our experiments were performed under naturalistic conditions, a very challenging setting, the proposed approach does not generalise to the leave-one-subject-out setting. This drawback could be attributed to the fact that we have only nine drivers in our dataset. The limited sample size is not diverse enough for a machine learning model to learn a generalisable pattern among different subjects. We expect that a large scale field study with a greater number of the subject could be conducted to further explore the generalisability of facial expression-based HRV estimation.

In addition, it is worth noting that the estimation of LF/HF ratio is generally inferior than the other HRV measures. We believe that this difference can be explained by the fact that subject respiration heavily influences the frequency components of HRV (Brown et al., 1993; Poh, McDuff, and Picard, 2010). More specifically, both respiration and autonomic nerve activities contribute to the deviation in LF/HF ratio, whereas only the latter factor can be reliably interpreted by the proposed facial expression-based inference model. For future work, we see potential in integrating a respiration detection module and thus fusing the information of breath to further improve the HRV estimation, which shall be one of our focuses in future research.

Furthermore, although the excessively low or high HRV measures are strong indicators of certain physiological and psychological states, an exact measurement of HRV measures could bring more insight into a user's health states (e.g., monitoring of hypertension or other cardiovascular diseases), which is not accomplished in this study. The exact measurement of HRV relies on precise capturing of IBI, which can be achieved by rPPG under well-defined lab condition. The fundamental mechanism of rPPG that detects the heartbeat induced peak of blood volume in a vessel is a more straightforward approach for measuring the exact value of HRV. However, rPPG is not as robust as our approach and is vulnerable to ambient noise due to illumination and motion artefacts (Niu et al., 2019). With the positive results demonstrated in this study, researchers in the future could focus on a fusion approach that leverages the robustness of our approach to reduce noise in rPPG; thus, achieving a reliable HRV measurement in the wild. At the same time, future work could extend our work by investigating the feasibility of applying the proposed facial expression-based HRV estimation outside the vehicle. For example, a potential use case could be patients with cardiovascular diseases who need low-cost monitoring of their current condition. The mandatory step to validate our approach would be the collection of medical data from affected patients and subsequent experiments on this data.

4.6 Conclusion

Several studies and surveys pointed out that the sub-optimal state of drivers is the main cause of traffic accidents (Choi et al., 2016). The National Highway Traffic Safety Administration (NHTSA) suggested that 94% of accidents resulted from human errors (National Highway Traffic Safety Administration, 2020b). Therefore, a strategy for monitoring drivers' states and driving performance becomes crucial in the reduction of the number of accidents. Such a driver monitoring system is particularly meaningful in the upcoming era of ever automated vehicles, where driver

states needs to be maintained to ensure a seamless takeover of the control of cars. Although several HRV estimation approaches have been proposed, the mediocre accuracy, inconvenient deployment and the lack of ubiquity prevent them from becoming a practical and prevalent solution.

To address the existing challenges and embrace future technologies, we proposed a facial expression-based approach for HRV measure outlier detection. The reason is that empirical research showed that excessively low or high HRV measures are strongly correlated with various sub-optimal mental and psychological states of people (Hernando et al., 2018; Kim et al., 2018; Patel et al., 2011; Vicente et al., 2016). The merit of the proposed approach is three-fold. First, HRV estimation is a meaningful and even necessary complement to visual human activity recognition (HAR) based driver monitoring. While HAR captures drivers' physical behaviours, HRV estimation evaluates their mental states. Second, driver monitoring cameras will become a mandatory component of future vehicles in many regions. Therefore the proposed approach does not induce any extra hardware cost, providing a higher degree of ubiquity than smartwatches and UWB based technologies. Our evaluation demonstrates that the proposed tree-based probabilistic fusion network approach outperforms a consumer smartwatch in HRV measure outlier detection by up to 13.1% in terms of balanced accuracy. The positive results and the ubiquity of the proposed approach demonstrated its great potential in improving driving experience and safety. Finally, the proposed tree-based probabilistic fusion network approach outperforms other prevalent pure tree-based or neural network based methods by an average of 3.4% in balanced accuracy. The idea of the tree-based probabilistic embedding should inspire researchers to consider the possibility of hybrid models that leverages the merits of the tree-based models, especially when no rich prior domain knowledge is available.

The concept of facial expression-based estimation of HRV measures proposed in this work could further facilitate various IoT-based services and applications. For example, in mobile crowd sensing (Wang et al., 2018a; Wang et al., 2020a), car ridesharing companies (Uber, Didi, etc.) could determine whether a driver is an optimal state based on the proposed HRV estimation approach. After that, task allocation can be optimised by assigning more demanding tasks to the drivers of better states or enforcing mandatory pause to the drivers who are temporally not fit for working. Thus, the quality of service will be improved. Another example is smartphone-based mobile sensing of user physiological and psychological states. One major limitation of smartphone-based sensing is the lack of accurate physiological data (Buda, Khwaja, and Matic, 2021; Taylor et al., 2020; Wang et al., 2018b). With the help

of the proposed method, users' HRV estimation can be shared via data link between smartphones and the devices that capture facial expressions (e.g., intelligent vehicles, webcam of laptops, and surveillance cameras). In this way, smartphone-based mobile sensing can achieve a more comprehensive understanding of users' states. The method proposed in this work, in essence, conceptualises a more robust and more accurate way of pervasive monitoring of users' mental states. The concept targets "IoT Data Analytical Services", one of the ten main challenges in developing an IoT service outlined by Bouguettaya et al. (Bouguettaya et al., 2021). The purpose of IoT data analytics is to distil heterogeneous IoT data in order to provide domain-specific actionable knowledge of adequate quality (Bouguettaya et al., 2021). In our vision, the facial expression-based HRV estimation of users should not only be limited to drivers, but can also be generalised to broader applications where users' mental state should be considered. As such, we expect to see interdisciplinary research from psychology, neuroscience, and computer science could benefit from our idea and further push forward the pervasive sensing of user states.

Chapter 5

Conclusion and Outlook

The previous chapters of this thesis have so far introduced the context and the motivation for driver states monitoring and presented the relevant existing works that can help readers gain a deeper understanding of current development in this domain. Moreover, the experimental settings, the methodology, and evaluation results for the two research questions regarding the monitoring of the psychological and physiological states of drivers were also presented in the corresponding chapters. In this chapter, the overall context and motivation for the detection of the emotions and well-being of drivers are reintroduced. Combined with the key findings and the concepts of the newly proposed approaches, the reintroduction provides a better understanding of the contribution of this thesis to the community. The two research questions are subsequently revisited and the core results of this thesis are summarised. Furthermore, the implications for researchers and practitioners are provided, along with limitations that should be addressed in future research. Finally, this chapter concludes the thesis with closing remarks.

5.1 Summary of Contributions

The vision of a safer and more comfortable driving experience is being intensively researched by both academia and industry. Among various techniques and concepts, driver-centric monitoring and assistance techniques have been identified as opportunities to realise this goal (Coughlin, Reimer, and Mehler, 2011; Zepf et al., 2020). The traditional concept of driver monitoring and assistance typically improved road safety by measures such as the promotion of the warning of upcoming hazards, the prediction of locations with a higher risk of traffic accidents, and the anticipation of a driver's next manoeuvres (Gahr et al., 2019; Hallac et al., 2018; Ryder et al., 2017). In contrast to this concept, an increasing amount of research has explored alternative

possibilities via the improvement of the mental and physical states of drivers, in order to reduce the risk of human errors before it becomes unmanageable (Dobbins and Fairclough, 2019; Paredes et al., 2018a; Zepf et al., 2020; Zheng et al., 2020). Following this vein, the thesis at hand focused on the monitoring of the psychological and physiological states of drivers, and thus contributes to the new concept of driver monitoring systems.

Furthermore, with the growing concerns of people's psychological and physiological health, there is an emerging market that requires the more ubiquitous and persuasive measurement of users' bio-markers (Koch et al., 2021; Nahum-Shani, Hekler, and Spruijt-Metz, 2015). Given the important role that vehicles play in modern life and the sophisticated sensors and technologies embedded in current and future vehicles, increasingly more researchers and practitioners have identified the great potential of health monitoring and intervention in in-vehicle environments. However, many existing techniques involved cumbersome physiological sensors or driver monitoring cameras that could compromise user privacy. Such limitations prevent the prevalent usage of these techniques. With the progressive advancement of computer science and engineering technologies, the real-time collection and processing of the wealth of information from dynamic sources of naturalistic driving data becomes realistic in ordinary vehicles. Yet, to date, only few existing works have leveraged these data to perform driver states inference in a non-intrusive manner. When considering the great potential and progressive achievement of user states inference in the mobile computing domain (Canzian and Musolesi, 2015; LiKamWa et al., 2013; Lu et al., 2019; Zhang et al., 2018), for instance user emotion recognition based on smartphone usage pattern, there is a research gap regarding driving data based user states inference, which was addressed in this thesis.

Moreover, previous studies on vehicle data based user states inference were mostly conducted under laboratory settings or controlled naturalistic conditions, which are characterised by a limited amount of data or the restricted diversity of experimental subjects (Zepf et al., 2020). Therefore, the community lacks a more reliable and robust evaluation of such vehicle data-based user states inference. To deliver stronger empirical evidence of the utility of this new concept, all the experiments conducted in the present work were based on a longitudinal field study of four months involving nine participants driving on public roads.

Overall, a novel concept was adopted in this thesis to improve existing driver monitoring systems via the inference and estimation of a driver's psychological and physiological states. The proposed approaches targets the limitation of existing works by emphasising the ubiquity of the solution while better protecting the privacy of users.

The real world assessment of the proposed methods enables the provision of reliable answers to the two research questions, namely (*RQ 1*) can a driver's emotion be inferred based on the control area network (CAN-bus) and the traffic context (from the front-view camera or on-board radar system), and (*RQ 2*) can a driver's heart rate variability (HRV) be reliably estimated in a non-intrusive manner (via a driver monitoring camera) in future vehicles? The analysis of the key findings targeting these two questions are respectively presented in the subsequent sub-sections, which summarise the implications for researchers and practitioners in the future.

5.1.1 Driver Emotion Inference

To address the challenge depicted in the first research question, a longitudinal field study of four months, involving nine participants driving on public roads, was conducted. The relevant data, including CAN-Bus data (driving behaviour), front-view video streams (traffic context), driver facial video streams (emotion labels), and physiological signals, were collected. In contrast to other existing research on driving behaviour-based emotion recognition, in which experiments were performed on the data collected from a driving simulator or under a controlled naturalistic environment, the evaluation and analysis in this work were based on the wealth of naturalistic data from the field study, which cover a wide range of driving areas and distances. Given the in-the-wild nature of the data, the following research question was thoroughly investigated:

RQ 1: *Can a driver's emotion be inferred based on the driving behaviour (from CAN-bus) and the traffic context (from the front-view camera)?*

A series of systematic experiments were performed to investigate this research question, which included the investigation of the usability of various sensor modalities, comparisons with state-of-the-art physiological sensor-based methods, and the flexibility and ubiquity of the proposed method in the real world. Based on these comprehensive evaluations, the following key findings, and their implications for both researchers and practitioners, are subsequently highlighted.

Driver emotions can be inferred from either the driving behaviour or traffic context, whereas the fusion of both modalities can further improve the emotion recognition performance. Driving behaviour and traffic context information are acquired from very different sensor modalities. Although driving behaviour information (based on CAN-Bus data) information is available in almost all modern vehicles, it has certain limitations including: (a) restricted access: the data are typically only available to Original Equipment Manufacturers (OEMs); (b) a lack of generalisation

to different types of vehicles due to individual definitions of the format of CAN data by OEMs; and (c) the decrease of the interaction between drivers and vehicles due to the increased automation of smart vehicles. Given these drawbacks, the feasibility of emotion recognition based on the traffic context, which is typically acquired via front-view cameras, was also investigated. The advantages of using the traffic context can be summarised as follows: (a) front-view cameras are becoming prevalent in current and future vehicles, (b) the machine learning models based on video streams are generalisable to different types of vehicles, and (c) if a vehicle is not yet equipped with a front-view camera, front-view video streams can be acquired by a smartphone mounted on the dashboard of the vehicle. However, the privacy regulations in certain regions might restrict the processing of front-view video streams; moreover, video based models could be vulnerable to adversarial attacks.

To address these concerns, we specifically built driver emotion recognition models based on either driving behaviour (CAN-only model), traffic context (video-only model), or the combination of both (fusion model). Overall, the CAN-only and video-only model achieved comparable performance, reaching a macro-F1 score of approximately 69% for intra-subject evaluation and approximately 59% for leave-one-subject-out (LOSO) evaluation. The fusion model exhibited further improvement in these scores by around 1.2%. These numbers are comparable with recent advancements in behaviour-based emotion recognition in the wild. Furthermore, it should be noted that the model under LOSO evaluation achieved an F1-score that was almost 10% higher than the random guess baseline (50%). LOSO evaluation is the most challenging scenario in affective computing and is even not possible in many behaviour-based emotion recognition models (LiKamWa et al., 2013; Zhang et al., 2018). The experiments conducted in the present work demonstrated that the proposed models are able to capture generalisable patterns across different subjects, which should inspire future researchers regarding how to perform feature engineering and construct the architecture of machine learning models.

Overall, this is a positive result for the field of driver emotion recognition. While there is still room for the improvement of the performance, the analysis suggests that the current bottleneck lies in the limited size and diversity of training data. This limitation should, however, be assessed with consideration that ours is the first study to explore drivers' emotions inference in a longitudinal setup (over four months). It could be anticipated that when a larger amount of data and more diverse subjects are involved, a better driver emotion recognition performance can be achieved.

The driving behaviour and traffic context based emotion recognition method outperforms the physiological sensor-based method. To date, the mainstream of

user emotion recognition has been based on physiological signals such as HRV, electrocardiography (ECG), or electroencephalography (EEG). While such methods have achieved impressive performance, especially under laboratory conditions, their generalisabiliy in the real world remains a challenging. On the one hand, not only emotions, but also other factors such as physical exercises, diseases, and the environment have impact on physiological signals; on the other hand, despite the progress in electronic technologies, most of physiological sensors are still cumbersome in practices. Even the most high-end smartwatch can only provide reliable HR measurements but not HRV measurements.

To address this issue, the HRV data of subjects were also collected during the field study using a medical-grade device. The state-of-the-art HRV-based emotion recognition algorithm was replicated (Nardelli et al., 2015) and its performance was compared with that of the proposed driving behaviour and traffic context-based methods. The comparison revealed that the proposed solution achieved higher performance than the HRV-based method in terms of the F1-score.

The result of the comparison is significant. Without the hurdles of cumbersome physiological sensors, the proposed emotion recognition solution has great potential for practical use. More importantly, this concept can be extended beyond the environment of cars. For example, manufacturers of mobile devices and smart infrastructure could leverage this idea and embed mental state monitoring functions in their products, thereby passively interpreting the behaviours of users. This idea is in line with the recently proposed idea of the Artificial Internet of Things (AIoT), which is the fusion of artificial intelligence (AI) technologies and Internet of Things (IoT) equipment to achieve more powerful and efficient data management and analysis, and ultimately to facilitate human-machine interactions.

The lightweight property and the flexible deployment requirement facilitate the ubiquitous usage of the proposed non-intrusive driver emotion recognition approaches. As described above, our evaluation demonstrated that driver emotion can be recognised, even when only one sensor modality (CAN-bus data or front-view video) is available. This revealed the great ubiquity and flexibility in the deployment of the proposed driver emotion recognition model. Automobile manufacturers, health service providers, and third-party infotainment systems can choose appropriate settings for sensor modalities to infer a driver's mental states and ultimately improve the driving experience of users.

For example, the CAN-only model can be deployed in regions where the recording of front-view videos is forbidden, or in the vehicles that are not yet equipped with

traffic context sensing systems. Moreover, the CAN-only model is less vulnerable to adversarial attacks. The video-only model has a higher degree of flexibility. In the regions where the front-view video recording is allowed, the video-only model can be deployed in combination with driving recorders or smartphones mounted on the vehicle dashboard. This setting can be directly deployed in current vehicles without the retrofitting of the hardware of vehicles to gain the access to CAN data. Furthermore, intelligent vehicles in the future will most likely be equipped with advanced sensing systems that capture surrounding traffic information. In this case, the video-only model can reuse such information to distil the knowledge about the traffic context. In general, the fusion model achieved the best performance among the three modality settings, and can be applied when relevant regulations and conditions permit.

In addition to flexible deployment, the proposed solution has a lightweight property. For the CAN-only model, the prevalent CPU configuration can satisfy real-time computation, including data pre-processing, feature generation, and machine learning inference. The main bottleneck lies in the processing of context information from video streams, as image processing relies heavily on deep learning models. The real-time processing of video streams can only be achieved by a GPU device. However, there are many ways to bypass this limitation. First, when local legislative regulation allows for it, the video processing task can be delegated to a cloud service with GPUs. The transmission of video streams is acceptable given the current mobile network capability. Second, with the advancement of neural processing units, it is anticipated that the computation time and energy consumption for video processing on mobile devices will be greatly reduced in the future. Third, in the case of future smart vehicles that are equipped with radar or visual perception systems, the traffic context information will be directly available from vehicles, and therefore no additional video processing will be required for the video-only model.

Overall, the thorough evaluation and analysis demonstrated the flexible and lightweight properties of the proposed non-intrusive emotion recognition model. This is of important significance, as it indicates that not only automobile manufacturers, but also other service providers, such as healthcare or mobile social network service providers, can leverage the pervasive sensing of driver emotions and deliver more customised services for users.

5.1.2 Driver Heart Rate Variability Estimation

The monitoring of a driver's physiological state has been a vision of intelligent vehicles for a long amount of time. On the one hand, driver well-being has a direct impact on road safety, as fatigue, stress, and inattentiveness all contribute to traffic

accidents; on the other hand, with the increasing trend towards a healthier lifestyle and the more precise recording of health information, many researchers consider vehicles as an ideal platform that can improve user well-being. Both in industry and academia, HRV is widely used as a proxy for the estimation of physiological states (D'Angelo and Lüth, 2011; D'Angelo et al., 2011; Wartzek et al., 2011; Zheng et al., 2020). However, the existing approaches for the monitoring of the HRV of drivers suffer from various constraints ranging from intrusive deployment to mediocre accuracy. Given the current drawbacks and the ultimate vision of intelligent vehicles, the following research question was raised:

RQ 2: *Can a driver's HRV be reliably estimated in a non-intrusive manner (via a driver monitoring camera) in future vehicles?*

To address this research question, the data collected for the investigation of *RQ 1* were reused, and the possibility of establishing the link between heart activities and facial expressions via the autonomic nervous systems (ANS) was explored. It should be noted that, unlike existing approaches that focus on the computation of the exact values of the HR or HRV measures, the proposed approach aims at the estimation of the outliers (i.e., one standard deviation below or above the mean) of HRV measures. This alternative health monitoring approach is supported by existing literature, which has found that a user's stress level, cognitive load, fatigue, and alert states can manifest themselves in excessively low or high states of certain HRV measures (RMSSD, LF/HF ratio, and pNN50) (Kim et al., 2018; Lohani, Payne, and Strayer, 2019; Patel et al., 2011; Taelman et al., 2009; Vicente et al., 2016). Ultimately, the extensive experiments and evaluation verified that via the link of the ANS, the outliers of HRV measures can be reliably inferred from the facial expressions of drivers, and can even outperform prevalent solutions such as smartwatches. Moreover, the opportunity to improve machine learning models was explored by integrating the decision tree model into neural networks. A series of systematic experiments were performed and led to the following key findings.

The implicit relationship between heart activity and facial expressions can be leveraged by camera technology in conjunction with a tree-based probabilistic fusion network model. The ANS is the part of the human nervous system that controls bodily reactions at the unconscious level. Functions such as the HR, pupillary response, facial expressions, and sexual arousal are regulated via the ANS (Schmidt and Thews, 1989). Furthermore, the ANS consists of the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS emerges from the spinal cord and stimulates the HR through the discharge of epinephrine and norepinephrine, whereas the PNS releases acetylcholine to decelerate the HR (Gordan,

Gwathmey, and Xie, 2015; Robinson et al., 1966). Despite the close relationship between heart activity and facial expressions, to date, there exists no explicit knowledge about how they influence each other, even in the medical domain. Given this, tree-based machine learning models were leveraged in this work, as they often perform well when there is a lack of prior insight about the underlying data structure or domain knowledge.

To be more specific, the performance of HRV outlier estimation was compared by using various deep learning architectures, such as the convolutional neural network, recurrent neural network, and multi-layer perceptron, as well as conventional machine learning models, such as the random forest and deep forest models. The comparison revealed a consistent advantage of the random forest and deep forest models. Given this finding, the merits of both tree-based models and deep learning architectures were further leveraged, and a hybrid combination of the two, namely the tree-based probabilistic fusion network (TPFN), was proposed. The TPFN first uses a deep forest to create a probabilistic embedding from the features of facial action units (AUs). The embedding is then further processed by a deep neural network. The advantages of this arrangement are twofold, namely that (a) deep forests are good at handling the heterogeneous distribution of AU features, and (b) after AU features are converted into probabilistic embedding by a deep forest, a neural network can learn a better mapping from facial cues to HRV measures due to the increased numerical stability. The results of extensive experiments demonstrated that the proposed TPFN model outperformed prevalent neural architectures and tree-based models. It is worth noting that this is not the first attempt to combine a tree-based model with neural networks; various researchers have exploited similar ideas and verified their advantages in different contexts (Kong and Yu, 2018; Kontschieder et al., 2015; Wang, Aggarwal, and Liu, 2017). Overall, a novel TPFN model that can reliably estimate the outliers of HRV measures, and hence provide the reliable well-being monitoring of drivers, was proposed in this thesis. Furthermore, the idea of the hybrid usage of a tree-based model and neural network should inspire future researchers in the development of machine learning models, especially when no prior domain knowledge is available.

HRV measurements from consumer smartwatches are conditionally unreliable. Wearable devices such as smartwatches are becoming popular in people's daily lives. Thanks to the progressive advancements in electronic technology and signal processing techniques, nowadays these devices provide an accurate assessment of the average HR in various contexts. Compared to other existing driver HR monitoring solutions, such as methods based on ballistocardiograph, ultra-wide band (UWB), or

remote photoplethysmography (rPPG) (McDuff, Gontarek, and Picard, 2014; Sakai et al., 2013; Walter et al., 2011; Zheng et al., 2020), and smartwatches have exhibited a great leap in terms of measurement accuracy. In the presented assessment, the mean absolute error between high-end smartwatches and medical-grade heart monitoring devices (Firstbeat) of drivers is only as large as one beat per minute, corresponding to a relative error of less than 2%. This magnitude of errors is in line with the latest systematic assessment of smartwatches (Hernando et al., 2018; Shcherbina et al., 2017). However, the accurate HR monitoring does not imply reliable computation of derived HRV measures. For example, the mean absolute error of RMSSD between smartwatches and medical-grade devices is approximately 30 milliseconds, corresponding to a relative error of around 90%. Such a substantial drop in accuracy can be attributed to the fact that driving involves frequent hand and arm movements, while the monitoring of HRV measures in smartwatches works best if users sit still and avoid arm movements (Hernando et al., 2018). Given this, smartwatches are not a suitable device for the direct monitoring of the physiological conditions of drivers, when precise HRV measures are required.

Furthermore, the requirement was relaxed and smartwatches were used to perform the outlier estimation of HRV measures, as was done for the proposed TPFN model. This evaluation was necessary to verify if the measurement error of smartwatches constitutes consistent offsets. If so, smartwatches can still serve as a reliable tool for outlier estimation of HRV measures because an offset error does not affect outlier detection. The evaluation showed that TPFN based on facial expressions outperformed smartwatches in all HRV measures with an improvement ranging from 3.6% to 13.1% in terms of balanced accuracy. As such, the comparison results highlight the advantages of facial expression-based HRV estimation, which not only provides higher estimation accuracy, but also facilitates easier application because no wearable device is needed.

5.2 Limitations

The encouraging results and implications of this thesis should be assessed in the light of the limitations of the research and evaluation settings, as discussed in Section 3.6 and Section 4.5. This section revisits the previously mentioned limitations and challenges while additionally emphasising the common issues shared by the analysis of both *RQ 1* and *RQ 2*. Discussions on potential improvements that can be made to address these limitations are also provided to advance the research in the domain of ubiquitous and mobile computing.

Regarding driver emotion recognition, a critical obstacle is the accuracy of emotion labels. Emotions are subjective feelings and can be induced by various inner and external factors. For example, a driver might get upset due to the aggressive actions of other traffic participants. Meanwhile, he or she might also be experiencing anger while driving due to a conflict with family members at home. Therefore, the emotion labels of drivers inherently contain a large amount of noise due to the diverse sources of emotions. Similar driving behaviours and traffic contexts might be mapped to contradictory emotional states. For this reason, the upper bound of the emotion recognition accuracy achieved in this study was approximately 70% in terms of the F1-score. In fact, label noise has been a long-standing problem in the machine learning community. One potential approach by which to improve emotion recognition in the current context is to adopt uncertainty suppression mechanisms (Peng et al., 2020a; Wang et al., 2020b). Briefly explained, this mechanism creates a score for each training sample, indicating the reliability of the corresponding label. By doing so, the machine learning model is able to adjust its loss function in order to put more emphasis on more reliable samples and hence increase the overall performance.

Furthermore, the usability of the outlier estimation of HRV measures should be thoroughly inspected and improved. First, unlike smartwatch-, rPPG-, or UWB-based approaches, the proposed facial expression-based outlier estimation method does not provide exact values of HRV measures. While the outliers can reflect certain extreme physiological and psychological states of users, the exact measurement of HRV measures could provide a much more comprehensive assessment of a user's health condition. This goal can be achieved only by exploring beyond pure learning-based methods, as the exact HRV measures rely on the precise detection of the IBI. A promising direction could be the combination of rPPG and the facial expression-based method, as their underlying mechanisms complement each other. The rPPG-based method is able to detect exact peaks of IBI when the light conditions are ideal and the subject remains still. The major limitation of rPPG is its vulnerability to fluctuating illumination and motion artefacts. This issue could potentially be overcome by fusing the facial cues into the detection, as the detection of facial expressions is robust against ambient noise, and the present evaluation already demonstrated the strong correlation between facial expressions and HRV measures. Second, the estimation of the LF/HF ratio is generally inferior to other HRV measures, which can be attributed to the fact that the frequency components of HRV are strongly influenced by respiration. Unlike heart activities, respiration does not induce obvious deviations in facial expressions, and hence cannot be reliably detected by the proposed approach. As such, the integration of a visual respiration detection module into the existing pipeline is posited

to further improve the HRV estimation. Finally, the generalisability of the proposed approach to patients with cardiovascular diseases remains unclear; all participants in the field study were healthy, and none reported any relevant complications. While the proposed concept of HRV estimation via facial videos has the potential to be deployed in hospitals or smart homes to facilitate seamless health states monitoring, it is critical to inspect whether the connection between facial expressions and heart activities via the ANS is affected by certain diseases. Such an investigation will require the collection of the medical data of inpatients via experiments and long-term observation. As such, future research in this direction is anticipated.

Finally, the common limitations shared by the two investigations should be discussed. The greatest challenge is the generalisability to different subjects. Regarding the LOSO evaluation of driver emotion recognition, the proposed method achieved an average F1-score that was approximately 10% lower than that of the intra-subject evaluation. Regarding the estimation of HRV measures, this problem is even more severe, as the proposed solution did not work for LOSO evaluation. This limitation can be attributed to two major factors. First, the experimental scale was limited in terms of the dataset size and the subject diversity. The total amounts of data available for driver emotion recognition and HRV outlier estimation were only approximately 675.6 and 68.6 hours, respectively. In addition, there were only nine drivers represented in the dataset. These numbers are insignificant as compared to other in-the-wild machine learning projects. For example, the dataset used by Affectiva consists of more than 20,000 hours of driving data from more than 4,000 unique individuals (Mcmanus, 2020a). Moreover, due to the limited experimental scale, it would be difficult for a machine learning model to learn a powerful representation that can be sufficiently generalised to different subjects. Second, emotions are subjective feelings, and each individual will have different or even contradictory interpretations of the same event; this is an inherent problem of emotion recognition. Therefore, when an emotion recognition model is applied to a test dataset of a previous unseen subject, the performance of the model will be upper-bounded by the number of common patterns shared between the training and test dataset. Empirically, this number is very limited as compared to that in other machine learning tasks, such as image recognition or object detection.

Another common limitation shared by the two investigations stems from the lack of explainability of the machine learning models. While the presented emotion recognition and HRV estimation solutions demonstrated promising and robust prediction performance, their underlying mechanisms have not yet been fully explored. From

the practical perspective, knowing the decision rules of an AI system can help facilitate the trust between users and machines (Amershi et al., 2019; Ignatiev, 2020; Norkute et al., 2021). From the research perspective, the intransparency of AI systems has been a long-standing problem, and the increase of the explainability of machine learning models has attracted increasingly more attention. Further exploration of the decision rules of the proposed solutions would provide insights for AI scientists to better understand how learning algorithms distil knowledge about emotion and heart activities from user behaviours. More importantly, both medical researchers and psychologists can benefit from such explainability, and can gain novel knowledge in their respective domains (Došilović, Brčić, and Hlupić, 2018; Tjoa and Guan, 2021).

5.3 Conclusion

Improving the driving experience has been a popular research topic ever since the invention of cars (Zepf et al., 2020). With the continuous development and innovation of engineers, scientists, policy makers, and automotive manufacturers, the forms and functions of the human-machine interfaces of cars have been improved over the years. Nevertheless, the emerging technologies and the evolution of modern lifestyles have created new requirements and opportunities for safer and more comfortable in-vehicle experiences. Among the various challenges, we identify the monitoring of driver psychological and physiological states are one of the most valuable functions for current and future vehicles. The most fundamental significance of the driver state monitoring is improved road safety because the psychological and physiological states directly indicate whether a driver is fit for driving. While the advent of self-driving vehicles may be the ultimate solution to safer traffic, the progress of such technology is uncertain and it might take decades before self-driving vehicles become prevalent, especially in those low-income countries. In other words, drivers will continue to play a central role in traffic. Furthermore, the monitoring of a driver's psychological and physiological states has derivative benefits. People nowadays have more demands regarding health-related services. For example, the proportion of the ageing population in the developed countries continues to grow, and this population suffers from various chronic diseases. Moreover, the threat of mental diseases, such as depression and burn-out, is increasing for the working population in the most part of the world due to the accelerated pace of life, work pressures, etc. Researchers and public-health related authorities are calling for more innovative approaches to handle the massive need for health monitoring and treatments, among which the ubiquitous and the pervasive monitoring of psychological and physiological states plays a vital

role (Ebert et al., 2017; Nahum-Shani, Hekler, and Spruijt-Metz, 2015). With the advanced sensors embedded in modern vehicles, the explosive progress of machine learning techniques and the growing usage of cars, an increasing amount of research suggests that cars have the potential to become a wellness platform, in which the users' health states are monitored and improved.

Nevertheless, there exists only limited research targeted at such aspects. First, the mainstream of driver state monitoring focuses on behaviour recognition. While behaviour recognition can protect drivers from distractions like making phone calls or over-engagement in conversation, this solution is unable to detect a driver's reduced awareness of incidents occurring even within the visual field, namely the problem of "look but fail to see". Furthermore, driver behaviour recognition does not provide health information about drivers that can be utilised to improve the driving experience and well-being of users. Finally, many of the latest emotion recognition and physiological signal monitoring approaches were evaluated under laboratory conditions or controlled naturalistic settings. Despite their positive results in well-defined settings, the generalisability of these approaches to the in-the-wild environment is not guaranteed. As such, this thesis demonstrated the evaluation from the field assessment of the estimation of the psychological and physiological states of drivers in a non-intrusive manner while relying only on existing or soon-to-come sensors in cars.

Overall, the insights of this thesis reveal that the exploration of driver behaviours and the traffic context has the potential to overcome the long-standing limitations of driver emotion recognition that relies on cumbersome physiological sensors or privacy-breaching facial videos. In addition, the results demonstrate that the HRV measures of drivers can be reliably estimated from their facial expressions. The presented approaches for the non-intrusive monitoring of the psychological and physiological states of drivers provide the opportunity for a ubiquitous, cost-effective, and scalable solution to promote road safety and the driving experience. The idea of the pervasive inference of driver states should encourage the collaboration between health service providers, automotive manufacturers, and even players in the smart infrastructure industry. Together, they can facilitate a safer and smoother transition from current to semi- and fully-autonomous vehicles in the future, while concurrently providing health-related services to improve the driving experience and eventually benefit the whole of society.

Bibliography

Aarts, E. and Ruyter, B. de (2009). "New research perspectives on ambient intelligence". In: *Journal of Ambient Intelligence and Smart Environments* 1.1, pp. 5–14.

American Automobile Association Foundation for Traffic Safety (2020). *New American Driving Survey: Updated Methodology and Results from July 2019 to June 2020*. Accessed: 2021-08-09. URL: https://aaafoundation.org/new-american-driving-survey-updated-methodology-and-results-from-\#july-2019-to-june-2020/.

Amershi, S. et al. (2019). "Guidelines for human-AI interaction". In: *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Audi AG (2018). *Audi Elaine*. Web Page. Accessed: 2021-04-23. URL: https://www.audi.com/en/experience-audi/models-and-technology/concept-cars/audi-elaine.html.

Autovista Group (2020). *Audi A8 will not feature level 3 autonomy*. Accessed: 2021-04-23.

Balters, S. et al. (2020). "Calm Commute: Guided Slow Breathing for Daily Stress Management in Drivers". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.1, Art. no. 38:1–19.

Baltrusaitis, T., Robinson, P., and Morency, L. (2016). "OpenFace: An open source facial behavior analysis toolkit". In: *IEEE Winter Conference on Applications of Computer Vision*, pp. 1–10.

Barrett, L. F. (2006). "Are Emotions Natural Kinds?" In: *Perspectives on Psychological Science* 1.1, pp. 28–58.

Bouguettaya, A. et al. (2021). "An Internet of Things Service Roadmap". In: *Communications of the ACM* 64.9, pp. 86–95.

Braun, M., Weber, F., and Alt, F. (2020). *Affective Automotive User Interfaces – Reviewing the State of Emotion Regulation in the Car*. arXiv: 2003.13731 [cs.HC].

Breiman, L. (2001). "Random forests". In: *Machine Learning* 45.1, pp. 5–32.

Brown, T. E. et al. (1993). "Important influence of respiration on human R-R interval power spectra is largely ignored". In: *Journal of Applied Physiology* 75.5, pp. 2310–2317.

Buda, T. S., Khwaja, M., and Matic, A. (2021). "Outliers in smartphone sensor data reveal outliers in daily happiness". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.1, Art. no. 5:1–19.

Calvo, R. A. and D'Mello, S. (2010). "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications". In: *IEEE Transactions on Affective Computing* 1.1, pp. 18–37.

Canzian, L. and Musolesi, M. (2015). "Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis". In: *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1293–1304.

Chan, M. and Singhal, A. (2015). "Emotion matters: Implications for distracted driving". In: *Safety Science* 72, pp. 302–309.

Chatterjee, K. et al. (2020). "Commuting and wellbeing: a critical overview of the literature with implications for policy and future research". In: *Transport Reviews* 40.1, pp. 5–34.

Chen, G. et al. (2019a). "Deep anticipation: lightweight intelligent mobile sensing for unmanned vehicles in IoT by recurrent architecture". In: *IET Intelligent Transport Systems* 13.10, pp. 1468–1474.

Chen, G. et al. (2020). "NeuroIV: Neuromorphic Vision Meets Intelligent Vehicle Towards Safe Driving With a New Database and Baseline Evaluations". In: *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13.

Chen, X. et al. (2019b). "Video-based heart rate measurement: Recent advances and future prospects". In: *IEEE Transactions on Instrumentation and Measurement* 68.10, pp. 3600–3615.

Choi, Y. et al. (2016). "Driver Status Monitoring Systems for Smart Vehicles Using Physiological Sensors: A safety enhancement system from automobile manufacturers". In: *IEEE Signal Processing Magazine* 33.6, pp. 22–34.

Corby, S. (2017). *Mercedes-Benz Vitality Coach revealed*. Web Page. Accessed: 2020-05-10. URL: https://www.drive.com.au/motor-news/mercedes-wants-to-make-you-fitter--even-while-you-re-sitting-still-gvkm25.

Coughlin, J. F., Reimer, B., and Mehler, B. (2011). "Monitoring, managing, and motivating driver safety and well-being". In: *IEEE Pervasive Computing* 10.3, pp. 14–21.

D'Angelo, L. T. and Lüth, T. (2011). "Integrated systems for distraction-free vital signs measurement in vehicles". In: *ATZ worldwide eMagazine* 113, pp. 52–56.

D'Angelo, L. T. et al. (2010). "A system for unobtrusive in-car vital parameter acquisition and processing". In: *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare*, pp. 1–7.

D'Angelo, L. T. et al. (2011). "Unobtrusive in-car vital parameter acquisition and processing". In: *Ambient Assisted Living*, pp. 257–271.

Dhall, A. et al. (2011). "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark". In: *IEEE International Conference on Computer Vision Workshops*, pp. 2106–2112.

Dmitrenko, D. et al. (2020). "CARoma Therapy: Pleasant Scents Promote Safer Driving, Better Mood, and Improved Well-Being in Angry Drivers". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Dobbins, C. and Fairclough, S. (2019). "Signal Processing of Multimodal Mobile Lifelogging Data towards Detecting Stress in Real-world Driving". In: *IEEE Transactions on Mobile Computing* 18.3, pp. 632–644.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). "Explainable artificial intelligence: A survey". In: *IEEE International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215.

Ebert, D. D. et al. (2017). "Prevention of mental health disorders using internet-and mobile-based interventions: a narrative review and recommendations for future research". In: *Frontiers in psychiatry* 8, Art. no. 116:1–16.

Egilmez, B. et al. (2017). "UStress: Understanding college student subjective stress using wrist-based passive sensing". In: *IEEE International Conference on Pervasive Computing and Communications Workshops*, pp. 673–678.

Egloff, B. et al. (1995). "Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure". In: *Motivation and Emotion* 19.2, pp. 99–110.

Ekman, P. (2007). *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Holt Paperbacks.

Ekman, P., Friesen, W. V., and Ancoli, S. (1980). "Facial signs of emotional experience". In: *Journal of Personality and Social Psychology* 39, pp. 1125–1134.

Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press.

Ekman, P. et al. (1987). "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion". In: *Journal of Personality and Social Psychology* 53, pp. 712–717.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases". In: *Pattern Recognition* 44.3, pp. 572–587.

Enev, M. et al. (2016). "Automobile Driver Fingerprinting". In: *Proceedings on Privacy Enhancing Technologies* 1, pp. 34–50.

European Commission (2020). *Report on Advanced Driver Distraction Warning Systems*. Web Page. Accessed: 2021-08-09. URL: https://ec.europa.eu/docsroom/documents/45901?locale=en.

European New Car Assessment Programme (2021). *Euro NCAP 2025 Roadmap*. Accessed: 2021-07-19. URL: https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf.

European Union (2020). *General Data Protection Regulation (GDPR) Compliance Guidelines*. Web Page. Accessed: 2020-09-24. URL: https://gdpr.eu/.

Fernández-Delgado, M. et al. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *The Journal of Machine Learning Research* 15.1, pp. 3133–3181.

Firstbeat Technologies Oy (2019). *Firstbeat Bodyguard 2*. Web Page. Accessed: 2020-05-10. URL: https://international-shop.firstbeat.com/product/bodyguard-2/.

Föll, S. et al. (2021). "FLIRT: A Feature Generation Toolkit for Wearable Data". In: *Computer Methods and Programs in Biomedicine* 212.106461, pp. 1–11.

Gahr, B. et al. (2019). "The Costs of Traffic Accident Hotspots". In: *IEEE Intelligent Transportation Systems Conference*, pp. 883–888.

Garmin Ltd. (2019). *What Is the Stress Level Feature on My Garmin Watch?* Web Page. Accessed: 2021-04-23. URL: https://support.garmin.com/en-US/?faq=WT9BmhjacO4ZpxbCc0EKn9.

Gjoreski, M. et al. (2016). "Continuous Stress Detection using a Wrist Device: in Laboratory and Real Life". In: *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 1185–1193.

Gordan, R., Gwathmey, J. K., and Xie, L.-H. (2015). "Autonomic and endocrine control of cardiovascular function". In: *World Journal of Cardiology* 7.4, pp. 204–214.

Greenberg, P. E. et al. (2015). "The economic burden of adults with major depressive disorder in the United States (2005 and 2010)". In: *The Journal of clinical psychiatry* 76.2, pp. 155–162.

Gudi, A. et al. (2019). "Efficient real-time camera based estimation of heart rate and its variability". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pp. 1570–1579.

Hallac, D et al. (2018). "Drive2Vec: Multiscale State-Space Embedding of Vehicular Sensor Data". In: *IEEE International Conference on Intelligent Transportation Systems*, pp. 3233–3238.

Healey, J. A. and Picard, R. W. (2005). "Detecting Stress During Real-world Driving Tasks using Physiological Sensors". In: *IEEE Transactions on Intelligent Transportation Systems* 6.2, pp. 156–166.

Hernando, D. et al. (2018). "Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects". In: *Sensors* 18.8, Art. no. 2619:1–11.

Holzman, J. and Bridgett, D. (2017). "Heart rate variability indices as bio-markers of top-down self-regulatory mechanisms: a meta-analytic review". In: *Neuroscience & Biobehavioral Reviews* 74, pp. 235–255.

Hornik, K. (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2, pp. 251–257.

Ignatiev, A. (2020). "Towards Trustable Explainable AI". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Ed. by C. Bessiere, pp. 5154–5158.

Isaacson, R. (2013). *The limbic system*. Springer Science & Business Media.

Izard, C. E. (2009). "Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues". In: *Annual Review of Psychology* 60.1, pp. 1–25.

Jeon, M. (2016). "Don't Cry While You're Driving: Sad Driving Is as Bad as Angry Driving". In: *International Journal of Human–Computer Interaction* 32.10, pp. 777–790.

Kahneman, D. et al. (2004). "A survey method for characterizing daily life experience: the day reconstruction method". In: *Science* 306(5702), pp. 1776–1780.

Kato, T. et al. (2011). "Classification of positive and negative emotion evoked by traffic jam based on electrocardiogram (ECG) and Pulse wave". In: *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1217–1222.

Katsis, C. D. et al. (2008). "Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38.3, pp. 502–512.

KIA (2019). *Amplify Your Joy with Emotive Driving*. Web Page. Accessed: 2020-09-01. URL: https://pr.kia.com/en/future/future/emotive-driving-ces.do.

Kim, H.-G. et al. (2018). "Stress and heart rate variability: A meta-analysis and review of the literature". In: *Psychiatry Investigation* 15.3, pp. 235–245.

Kingma, D. P. and Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. URL: http://arxiv.org/abs/1412.6980.

Koch, K. et al. (2021). "Taking mental health & well-being to the streets: An exploratory evaluation of in-vehicle interventions in the wild". In: *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 539:1–15.

Kohn, R. et al. (2004). "The treatment gap in mental health care". In: *Bulletin of the World health Organization* 82.11, pp. 858–866.

Kong, Y. and Yu, T. (2018). "A deep neural network model using random forest to extract feature representation for gene expression data classification". In: *Scientific Reports* 8.1, pp. 1–9.

Kontschieder, P. et al. (2015). "Deep neural decision forests". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1467–1475.

Kreibig, S. D. (2010). "Autonomic nervous system activity in emotion: A review". In: *Biological Psychology* 84.3, pp. 394–421.

Lee, J., Elhaouij, N., and Picard, R. (2021). "AmbientBreath: Unobtrusive just-in-time breathing intervention using multi-sensory stimulation and its evaluation in a car simulator". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2, Art. no. 71:1–30.

Lee, Y.-C., Lee, J. D., and Boyle, L. N. (2007). "Visual Attention in Driving: The Effects of Cognitive Load and Visual Disruption". In: *Human Factors* 49.4, pp. 721–733.

Legrain, A., Eluru, N., and El-Geneidy, A. M. (2015). "Am stressed, must travel: The relationship between mode choice and commuting stress". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 34, pp. 141–151.

LiKamWa, R. et al. (2013). "MoodScope: Building a Mood Sensor from Smartphone Usage Patterns". In: *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services*, pp. 389–402.

Liu, H. et al. (2017). "Visualization of Driving Behavior Based on Hidden Feature Extraction by Using Deep Learning". In: *IEEE Transactions on Intelligent Transportation Systems* 18.9, pp. 2477–2489.

Liu, S. et al. (2019). "Brake Maneuver Prediction – An Inference Leveraging RNN Focus on Sensor Confidence". In: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 3249–3255.

Liu, S. et al. (2021a). "The Empathetic Car: Exploring Emotion Inference via Driver Behaviour and Traffic Context". In: *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5.3, Art. no. 117:1–34.

Liu, S. et al. (2021b). "Towards Non-Intrusive Camera-Based Heart Rate Variability Estimation in the Car under Naturalistic Condition". In: *IEEE Internet of Things Journal*. Accepted.

Lo, E., Wei, Y., and Hwang, B. (2020). "Association between occupational burnout and heart rate variability: a pilot study in a high-tech company in Taiwan". In: *Medicine* 99.2, pp. 1–11.

Lohani, M., Payne, B. R., and Strayer, D. L. (2019). "A review of psychophysiological measures to assess cognitive states in real-world driving". In: *Frontiers in Human Neuroscience* 13, Art. no. 57:1–27.

Lopes, A. T. et al. (2017). "Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order". In: *Pattern Recognition* 61, pp. 610–628.

Lu, L. et al. (2019). "I3: Sensing Scrolling Human-computer Interactions for Intelligent Interest Inference on Smartphones". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.3, Art. no. 97:1–:22.

Lyrheden, F. (2020). *Driver Monitoring (DMS) on its way to becoming mandatory in vehicles around the world*. Web Page. Accessed: 2021-07-19. URL: https://smarteye.se/blogs.

Malik, M. et al. (1996). "Heart rate variability. Standards of measurement, physiological interpretation, and clinical use". In: *Circulation* 93.5, pp. 1043–1065.

Malta, L. et al. (2011). "Analysis of Real-World Driver's Frustration". In: *IEEE Transactions on Intelligent Transportation Systems* 12.1, pp. 109–118.

Martinez, C. M. et al. (2017). "Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey". In: *IEEE Transactions on Intelligent Transportation Systems* 19.3, pp. 666–676.

Martínez, H. P., Yannakakis, G. N., and Hallam, J. (2014). "Don't Classify Ratings of Affect; Rank Them!" In: *IEEE Transactions on Affective Computing* 5.3, pp. 314–326.

McDuff, D., Kaliouby, R. E., and Picard, R. W. (2012). "Crowdsourcing Facial Responses to Online Videos". In: *IEEE Transactions on Affective Computing* 3.4, pp. 456–468.

McDuff, D., Gontarek, S., and Picard, R. (2014). "Remote measurement of cognitive stress via heart rate variability". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2957–2960.

McDuff, D. et al. (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*, pp. 3723–3726.

Mcmanus, A. (2020a). *Affectiva Automotive AI*. Web Page. Accessed: 2020-10-28. URL: https://go.affectiva.com/auto.

Mcmanus, A. (2020b). *BMW: How In-Cabin Sensing Helps Build the Ultimate In-Vehicle Experience*. Web Page. Accessed: 2020-10-28. URL: https://blog.affectiva.com/bmw-how-in-cabin-sensing-helps-build-the-ultimate-in-vehicle-experience.

Mishra, V. et al. (2018). "Investigating the Role of Context in Perceived Stress Detection in the Wild". In: *Proceedings of ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pp. 1708–1716.

Nahum-Shani, I., Hekler, E. B., and Spruijt-Metz, D. (2015). "Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework". In: *Health Psychology* 34.0, pp. 1–20.

Nair, V. and Hinton, G. E. (2010). "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *International Conference on International Conference on Machine Learning*, pp. 807–814.

Nardelli, M. et al. (2015). "Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability". In: *IEEE Transactions on Affective Computing* 6.4, pp. 385–394.

National Highway Traffic Safety Administration (2015). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. Web Page. Accessed: 2021-09-07. URL: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115.

National Highway Traffic Safety Administration (2020a). *Automated Vehicles for Safety*. Accessed: 2021-04-23. URL: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety\#topic-road-self-driving.

National Highway Traffic Safety Administration (2020b). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*. Accessed: 2021-07-19. URL: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812506.

Neven, D. et al. (2018). "Towards End-to-End Lane Detection: an Instance Segmentation Approach". In: *IEEE Intelligent Vehicles Symposium (IV)*, pp. 286–291.

Niu, X. et al. (2019). "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation". In: *IEEE Transactions on Image Processing* 29, pp. 2409–2423.

Norkute, M. et al. (2021). "Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization". In:

Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Art. no. 53:1–7.

Official Journal of the European Union (2019). *REGULATION (EU) 2019/2144 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. Accessed: 2021-08-09. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R2144&from=EN\#d1e32-24-1.

Osaka, M. et al. (2008). "Application of heart rate variability analysis to electrocardiogram recorded outside the driver's awareness from an automobile steering wheel". In: *Circulation Journal* 72, pp. 1867–1873.

Osaka, M. (2012). *Customized heart check system by using integrated information of electrocardiogram and plethysmogram outside the driver's awareness from an automobile steering wheel*. Accessed: 2021-06-04. URL: http://cdn.intechopen.com/pdfs-wm/27021.pdf.

Paredes, P. E. et al. (2018a). "Fast & Furious: Detecting Stress with a Car Steering Wheel". In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 665–677.

Paredes, P. E. et al. (2018b). "Just Breathe: In-Car Interventions for Guided Slow Breathing". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1, Art. no. 28:1–23.

Patel, M. et al. (2011). "Applying neural network analysis on heart rate variability data to assess driver fatigue". In: *Expert systems with Applications* 38.6, pp. 7235–7242.

Patel, V. J. (2020). *2021 Mercedes-Benz S-Class Comes With Built-In Level 3 Autonomy*. Accessed: 2021-04-23.

PEAK-System Technik GmbH (2019). *PCAN-USB Pro FD-Adapter*. Web Page. Accessed: 2020-05-10. URL: https://www.peak-system.com/PCAN-USB-Pro-FD.366.0.html.

Peng, X. et al. (2020a). "Suppressing Mislabeled Data via Grouping and Self-Attention". In: *European Conference on Computer Vision*. Springer, pp. 786–802.

Peng, X. et al. (2020b). "Driving maneuver early detection via sequence learning from vehicle signals and video images". In: *Pattern Recognition* 103, pp. 1265–1270.

Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2010). "Advancements in noncontact, multiparameter physiological measurements using a webcam". In: *IEEE Transactions on Biomedical Engineering* 58.1, pp. 7–11.

Rebolledo-Mendez, G. et al. (2014). "Developing a Body Sensor Network to Detect Emotions During Driving". In: *IEEE Transactions on Intelligent Transportation Systems* 15.4, pp. 1850–1854.

Redmon, J. and Farhadi, A. (2018). *Yolov3: An incremental improvement*. arXiv: 1804.02767.

Reece, A. G. and Danforth, C. M. (2017). "Instagram photos reveal predictive markers of depression". In: *EPJ Data Science* 6, Art. no. 15:1–12.

Rigas, G., Goletsis, Y., and Fotiadis, D. I. (2012). "Real-Time Driver's Stress Event Detection". In: *IEEE Transactions on Intelligent Transportation Systems* 13.1, pp. 221–234.

Robert Bosch GmbH (2016). *The Car as 3rd living space*. Accessed: 2021-06-04. URL: https://www.bosch.com/stories/the-car-as-3rd-living-space/.

Robinson, B. F. et al. (1966). "Control of heart rate by the autonomic nervous system. Studies in man on the interrelation between baroreceptor mechanisms and exercise". In: *Circulation Research* 19.2, pp. 400–411.

Rostaminia, S. et al. (2019). "W!NCE: Unobtrusive Sensing of Upper Facial Action Units with EOG-Based Eyewear". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.1, Art. no. 23:1–26.

Rudovic, O. et al. (2019). "Personalized estimation of engagement from videos using active learning with deep reinforcement learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 217–226.

Russell, J. A. (1980). "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39.6, pp. 1161–1178.

Ryder, B. et al. (2017). "Preventing traffic accidents with in-vehicle decision support systems-The impact of accident hotspot warnings on driver behaviour". In: *Decision support systems* 99, pp. 64–74.

Saeed, A., Ozcelebi, T., and Lukkien, J. (2019). "Multi-Task Self-Supervised Learning for Human Activity Detection". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2, Art. no. 61:1–30.

Saeed, A. and Trajanovski, S. (2017). "Personalized driver stress detection with multi-task neural networks using physiological signals". In: *Machine Learning for Health Workshop at 31st Conference on Neural Information Processing Systems*.

Sakai, K. et al. (2013). "Design of seat mounted ecg sensor system for vehicle application". In: *SAE International Journal of Passenger Cars-Electronic and Electrical Systems* 6.1, pp. 342–348.

Sano, A. and Picard, R. W. (2013). "Stress Recognition using Wearable Sensors and Mobile Phones". In: *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 671–676.

Sarker, H. et al. (2014). “Assessing the availability of users to engage in just-in-time intervention in the natural environment”. In: *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 909–920.

Sayette, M. A. et al. (2001). “A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression”. In: *Journal of Nonverbal Behaviour* 25.3, pp. 167–185.

Schmidt, A and Thews, G (1989). “Autonomic Nervous System”. In: *Human Physiology (2 ed)*, pp. 333–370.

Schmidt, P. et al. (2019). “Multi-target Affect Detection in the Wild: an Exploratory Study”. In: *Proceedings of the ACM International Symposium on Wearable Computers*, pp. 211–219.

Schwarz, N. and Strack, F. (1999). “Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications”. In: *Well-Being: The Foundations of Hedonic Psychology*, pp. 61–84.

Schwarz, N. (1990). “Feelings as information: Informational and motivational functions of affective states”. In: *Handbook of motivation and cognition: Foundations of social behavior*. Ed. by E. T. Higgins and R. Sorrentino, pp. 527–561.

Sen, T. et al. (2018). “Automated Dyadic Data Recorder (ADDR) Framework and Analysis of Facial Cues in Deceptive Communication”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4, Art. no. 163:1–22.

Shafaei, S., Hacizade, T., and Knoll, A. (2018). “Integration of Driver Behavior into Emotion Recognition Systems: A Preliminary Study on Steering Wheel and Vehicle Acceleration”. In: *Asian Conference on Computer Vision Workshop*, pp. 386–401.

Shaffer, F. and Ginsberg, J. (2017). “An overview of heart rate variability metrics and norms”. In: *Frontiers in Public Health* 5, Art. no. 258:1–17.

Sharma, K. et al. (2020). “Assessing Cognitive Performance Using Physiological and Facial Features: Generalizing across Contexts”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3, Art. no. 95:1–41.

Shcherbina, A. et al. (2017). “Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort”. In: *Journal of Personalized Medicine* 7.2, pp. 1–12.

Smyth, J. M. and Heron, K. E. (2016). “Is providing mobile interventions "just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management”. In: *Proceedings of the IEEE Wireless Health*, pp. 1–7.

Sodhi, M., Reimer, B., and Llamazares, I. (2002). "Glance analysis of driver eye movements to evaluate distraction". In: *Behavior Research Methods, Instruments, & Computers* 34.4, pp. 529–538.

Statistisches Bundesamt (2016). *Ergebnisse des Mikrozensus 2016. Report Report EVAS-Nr.12211*. Accessed: 2020-05-10. URL: https://www.destatis.de.

Stricker, R., Müller, S., and Gross, H.-M. (2014). "Non-contact video-based pulse rate measurement on a mobile service robot". In: *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062.

Stutzer, A. and Frey, B. S. (2008). "Stress that Doesn't Pay: The Commuting Paradox". In: *The Scandinavian Journal of Economics* 110.2, pp. 339–366.

Sun, Y., Fei, T., and Pohl, N. (2019). "A High-Resolution Framework for Range-Doppler Frequency Estimation in Automotive Radar Systems". In: *IEEE Sensors Journal* 19.23, pp. 11346–11358.

Taelman, J. et al. (2009). "Influence of mental stress on heart rate and heart rate variability". In: *Proceedings of the European Conference of the International Federation for Medical and Biological Engineering*, pp. 1366–1369.

Tasli, H. E., Gudi, A., and Uyl, M. den (2014). "Remote PPG based vital sign measurement using adaptive facial regions". In: *Proceedings of the IEEE International Conference on Image Processing*, pp. 1410–1414.

Tawari, A. and Trivedi, M. M. (2010). "Speech Emotion Analysis in Noisy Real-World Environment". In: *20th International Conference on Pattern Recognition*, pp. 4605–4608.

Taylor, S. et al. (2020). "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health". In: *IEEE Transactions on Affective Computing* 11.2, pp. 200–213.

Tjoa, E. and Guan, C. (2021). "A survey on explainable artificial intelligence (XAI): toward medical XAI". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11, pp. 4793–4813.

Trick, L. M., Brandigampola, S., and Enns, J. T. (2012). "How fleeting emotions affect hazard perception and steering while driving: The impact of image arousal and valence". In: *Accident Analysis & Prevention* 45, pp. 222–229.

Trifan, A., Oliveira, M., and Oliveira, J. L. (2019). "Passive Sensing of Health Outcomes Through Smartphones: Systematic Review of Current Solutions and Possible Limitations". In: *JMIR Mhealth Uhealth* 7.8, Art. no. e12649.

Tsuji, H. et al. (1996). "Determinants of heart rate variability". In: *Journal of the American College of Cardiology* 28.6, pp. 1539–1546.

Underwood, G. et al. (1999). "Anger while driving". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 2.1, pp. 55–68.

Verma, B. and Choudhary, A. (2018). "A Framework for Driver Emotion Recognition using Deep Learning and Grassmann Manifolds". In: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 1421–1426.

Vicente, F. et al. (2015). "Driver gaze tracking and eyes off the road detection system". In: *IEEE Transactions on Intelligent Transportation Systems* 16.4, pp. 2014–2027.

Vicente, J. et al. (2016). "Drowsiness detection using heart rate variability". In: *Medical & Biological Engineering & Computing* 54.6, pp. 927–937.

Victor, T. W., Harbluk, J. L., and Engström, J. A. (2005). "Sensitivity of eye-movement measures to in-vehicle task difficulty". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 8.2, pp. 167–190.

Vos, T. et al. (2016). "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015". In: *The lancet* 388.10053, pp. 1545–1602.

Voss, A. et al. (2015). "Short-term heart rate variability—influence of gender and age in healthy subjects". In: *PloS ONE* 10.3, Art. no. e0118308.

Walter, M. et al. (2011). "The smart car seat: personalized monitoring of vital signs in automotive applications". In: *Personal and Ubiquitous Computing* 15.7, pp. 707–715.

Wang, J.-S., Lin, C.-W., and Yang, Y.-T. C. (2013). "A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition". In: *Neurocomputing* 116, pp. 136–143.

Wang, J. et al. (2018a). "Task Allocation in Mobile Crowd Sensing: State-of-the-Art and Future Opportunities". In: *IEEE Internet of Things Journal* 5.5, pp. 3747–3757.

Wang, J. et al. (2020a). "HyTasker: Hybrid Task Allocation in Mobile Crowd Sensing". In: *IEEE Transactions on Mobile Computing* 19.3, pp. 598–611.

Wang, K. et al. (2020b). "Suppressing uncertainties for large-scale facial expression recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897–6906.

Wang, R. et al. (2018b). "Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1, Art. no. 43:1–26.

Wang, S., Aggarwal, C., and Liu, H. (2017). "Using a random forest to inspire a neural network and improving on it". In: *Proceedings of the SIAM International Conference on Data Mining*, pp. 1–9.

Wang, W. et al. (2017). "Algorithmic Principles of Remote PPG". In: *IEEE Transactions on Biomedical Engineering* 64.7, pp. 1479–1491.

Wartzek, T. et al. (2011). "ECG on the Road: Robust and Unobtrusive Estimation of Heart Rate". In: *IEEE Transactions on Biomedical Engineering* 58.11, pp. 3112–3120.

World Health Organization (2015). *Global status report on road safety 2015*. Web Page. Accessed: 2021-09-07. URL: https://www.who.int/violence_injury_prevention/road_safety_status/2015/GSRRS2015_Summary_EN_final.pdf.

World Health Organization (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. Web Page. Accessed: 2020-03-05. URL: https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/.

World Health Organization (2018). *Global status report on road safety 2018*. Web Page. Accessed: 2021-09-07. URL: http://apps.who.int/iris/bitstream/handle/10665/277370/WHO-NMH-NVI-18.20-eng.pdf?ua=1.

World Health Organization (2019). *Global Health Estimates: Life expectancy and leading causes of death and disability*. Web Page. Accessed: 2021-09-07. URL: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates.

Xie, J., Hilal, A. R., and Kulić, D. (2018). "Driving Maneuver Classification: A Comparison of Feature Extraction Methods". In: *IEEE Sensors Journal* 18.12, pp. 4777–4784.

Xing, Y. et al. (2019). "Driver activity recognition for intelligent vehicles: A deep learning approach". In: *IEEE transactions on Vehicular Technology* 68.6, pp. 5379–5390.

Yang, H., Ciftci, U., and Yin, L. (2018). "Facial Expression Recognition by De-Expression Residue Learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2168–2177.

Yang, Y. and Chen, H. H. (2011). "Ranking-Based Emotion Recognition for Music Organization and Retrieval". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4, pp. 762–774.

Yannakakis, G. N., Cowie, R., and Busso, C. (2017). "The ordinal nature of emotions". In: *International Conference on Affective Computing and Intelligent Interaction*, pp. 248–255.

Zepf, S. et al. (2019). “Towards empathetic car interfaces: emotional triggers while driving”. In: *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Art. no. LBW0129: 1–6.

Zepf, S. et al. (2020). “Driver Emotion Recognition for Intelligent Vehicles: A Survey”. In: *ACM Computer Survey* 53.3, Art. no. 64: 1–30.

Zhang, X. et al. (2018). “MoodExplorer: towards Compound Emotion Detection via Smartphone Sensing”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4, Art. no. 176:1–30.

Zhang, Y. et al. (2017). “SOVCAN: Safety-Oriented Vehicular Controller Area Network”. In: *IEEE Communications Magazine* 55.8, pp. 94–99.

Zheng, T. et al. (2020). “V2iFi: In-vehicle vital sign monitoring via compact rf sensing”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.2, Art. no. 70:1–27.

Zhou, J. J., Phadnis, V., and Olechowski, A. (2020). “Analysis of Designer Emotions in Collaborative and Traditional Computer-Aided Design”. In: *Journal of Mechanical Design* 143.2, Art. no. 021401.

Zhou, Z.-H. and Feng, J. (2017). “Deep forest: Towards an alternative to deep neural networks”. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3553–3559.

Appendix A

Inference of Driver Emotion via Driving Behaviour and Traffic Context

TABLE A.1: All signals of CAN data

Signal ID	Description	Signal ID	Description
1	Accelerator pedal position	26	Left turn indicator
2	Belt buckle indicator 1	27	Longitudinal acceleration
3	Belt buckle indicator 2	28	Motor rotational speed
4	Belt buckle indicator 3	29	Odometer
5	Belt buckle indicator 4	30	Parking light indicator
6	Belt buckle indicator 5	31	Rear fog light
7	Brake indicator	32	Right turn indicator
8	Brake pressure	33	Steering wheel angle
9	Clutch switch	34	Steering wheel direction
10	Daytime running lamp	35	Steering wheel velocity
11	Dimmed headlights indicator	36	Steering wheel velocity direction
12	Electronic stability control	37	Tank level percent
13	External temperature sensor 1	38	Temperature sensor
14	External temperature sensor 2	39	Time
15	Flasher	40	Wheel direction (back left)
16	Fog light indicator	41	Wheel direction (back right)
17	Front wiper	42	Wheel direction (front left)
18	Gear position	43	Wheel direction (front right)
19	GPS altitude coordinate	44	Wheel speed (back left)
20	GPS latitude coordinate	45	Wheel speed (back right)
21	GPS longitude coordinate	46	Wheel speed (front left)
22	High beam	47	Wheel speed (front right)
23	High beam indicator	48	Yaw rate
24	Humidity	49	Yaw rate direction
25	Lateral acceleration		

TABLE A.2: HRV features of the baseline method (Nardelli et al., 2015)

Feature Index	Description
	Time Domain Measures
1	the mean value (RR mean: R refers to the peak of the electrocardiography wave; RR is the interval between successive Rs)
2	the standard deviation (RR std)
3	the standard deviation of Normal-to-Normal (NN) intervals (SDNN)
4	the square root of the mean of the sum of the squares of differences between subsequent NN intervals (RMSSD)
5	the number of successive differences of intervals which differ by more than 50 ms, expressed as a percentage of the total number of heartbeats analysed (pNN50)
6	the integral of the probability density distribution divided by the maximum of the probability density distribution (HRV tiangular index)
7	the triangular interpolation of NN interval histogram (TINN)
	Frequency Domain Measures
8-10	the power calculated within the very low frequency (VLF), low frequency (LF), and high frequency (HF) bands.
11-13	the frequencies containing maximum magnitude (VLF peak, LF peak, and HF peak).
14-16	the power expressed as percentage of the total power (VLF power %, LF power %, and HF power %).
17-18	the power normalised to the sum of the LF and HF power (LF power nu and HF power nu)
19	the LF/HF power ratio
	Nonlinear HRV Measures
20	Approximate Entropy
21-22	Detrended Fluctuation Analysis: short-term fluctuations (α_1) and long-term flunctuations (α_2) Lagged Poincaré Plots: SD1, SD2, SD12, S, SDRR (details below)
23	SD1: the standard deviation related to the points that are perpendicular to the line-of-identity
24	SD2: the standard deviation that describes the long-term dynamics and measures the dispersion of the points along the identity line.
25	SD12 (SD1/SD2): the ratio between SD1 and SD2.
26	S (πSD1SD2): the area of an imaginary ellipse with axes SD1 and SD2
27	SDRR: an approximate relation indicating the variance of the whole HRV series

TABLE A.3: Intra-subject cross-validation: comparison between the baseline and our driving behaviour- and context-based inference

	Personalised Model			
F1-score (%)	CAN	Video	Fusion	baseline
anger	63.1	63.4	61.5	54.0
disgust	64.4	66.1	62.1	55.7
fear	62	59.5	56.8	55.2
joy	62.1	63.5	64.5	61.4
neutral	64.3	64.5	64	58.3
sadness	63.7	64.3	62.9	54.1
surprise	66.4	64.7	66.1	53.9
valence	66.7	61	62.4	56.5
average	64.1	63.4	62.5	56.1

TABLE A.4: LOSO cross-validation: comparison between the baseline and our driving behaviour- and context-based inference

	LOSO Model			
F1-score (%)	CAN	Video	Fusion	baseline
anger	51.4	52.9	50.8	43.6
disgust	52.4	50.8	49.5	44.3
fear	54.7	54	50.6	53.3
joy	56.6	54.8	53.3	54.5
neutral	54.6	51.6	52.4	44.9
sadness	48	49.8	47.4	44.4
surprise	54.3	50.7	49.3	49.2
valence	47.9	50.2	46.4	48.7
average	52.5	51.8	50	47.9

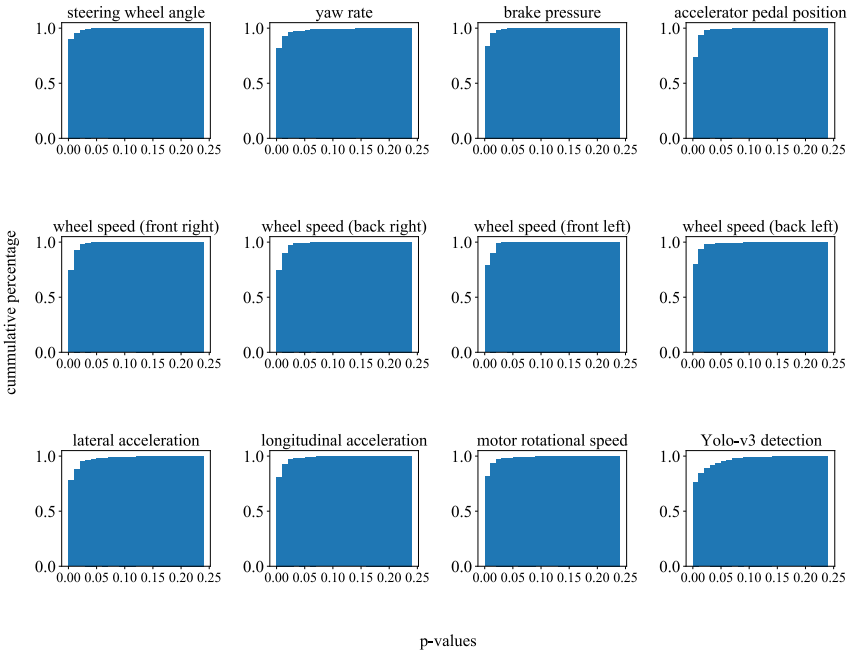


FIGURE A.1: Cumulative distribution of p-values of the selected features according to the source of the signal

TABLE A.5: Results of precision for low class: Intra-subject and LOSO cross-validation, comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities

low class	Personalised Model			LOSO Model		
precision (%)	CAN	Video	Fusion	CAN	Video	Fusion
anger	71.6	73.5	76.6	64.1	64.3	66.7
disgust	71.4	73.7	74.1	64.7	65.3	65.0
fear	67.9	68.7	69.0	45.9	47.2	45.6
joy	68.7	67.4	71.0	60.2	58.0	61.1
neutral	64.0	65.5	68.5	53.9	53.2	58.6
sadness	71.4	73.7	74.0	65.6	64.8	66.8
surprise	70.2	68.6	70.2	49.4	47.6	47.9
valence	65.4	65.2	69.1	52.1	51.0	57.2
average	68.8	69.5	71.6	57.0	56.4	58.6

TABLE A.6: Results of precision for high class: Intra-subject and LOSO cross-validation, comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities

high class	Personalised Model			LOSO Model		
precision (%)	CAN	Video	Fusion	CAN	Video	Fusion
anger	67.5	68.7	68.1	58.3	56.8	58.8
disgust	67.5	69.5	70.7	58.8	57.8	62.0
fear	70.3	69.7	71.9	50.5	51.1	52.3
joy	66.2	67	65.8	48.5	51.5	48.8
neutral	72.6	73.4	73.0	66.3	63.7	64.9
sadness	64.3	66.4	67.0	55.1	55.5	58.0
surprise	68.8	70.2	71.6	49.1	51.4	51.8
valence	71.3	76.1	73.0	63.9	65.8	65.3
average	68.6	70.1	70.1	56.3	56.7	57.7

TABLE A.7: Results of recall for low class: Intra-subject and LOSO cross-validation, comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities

low class	Personalised Model			LOSO Model		
recall (%)	CAN	Video	Fusion	CAN	Video	Fusion
anger	66.4	68.3	69.2	60.0	59.7	61.1
disgust	66.6	68.9	69.8	60.8	61.0	62.3
fear	69.5	70.3	71.2	49.4	52.3	51.3
joy	67.3	65.8	67.7	58.8	56.8	57.8
neutral	71.6	72.9	73.5	70.8	68.8	67.7
sadness	66.4	67.9	69.2	60.8	60.0	62.2
surprise	69.8	70.2	71.2	50.2	51.5	51.4
valence	72.3	75.3	74.7	73.2	71.9	70.3
average	68.7	70.0	70.8	60.5	60.3	60.5

TABLE A.8: Results of recall for high class: Intra-subject and LOSO cross-validation, comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities

high class	Personalised Model			LOSO Model		
recall (%)	CAN	Video	Fusion	CAN	Video	Fusion
anger	74.2	74.7	76.2	70.5	69.9	67.8
disgust	74.6	75.5	77.0	71.9	72.5	69.2
fear	68.7	68.1	69.9	46.5	47.1	46.3
joy	69.1	69.9	70.1	75.6	69.6	71.4
neutral	65.7	67.3	68.9	59.3	58.9	60.8
sadness	70.8	73.3	72.6	70.1	68.4	68.0
surprise	69.2	69.4	70.6	48.5	48.5	48.3
valence	64.8	66.4	68.0	56.7	56.3	58.5
average	69.6	70.6	71.7	62.4	61.4	61.3

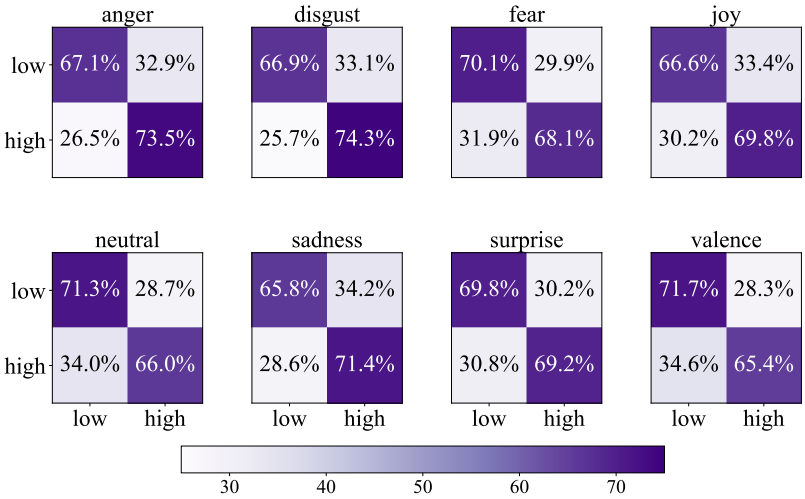


FIGURE A.2: Confusion matrix for CAN-only modality under intra-subject evaluation

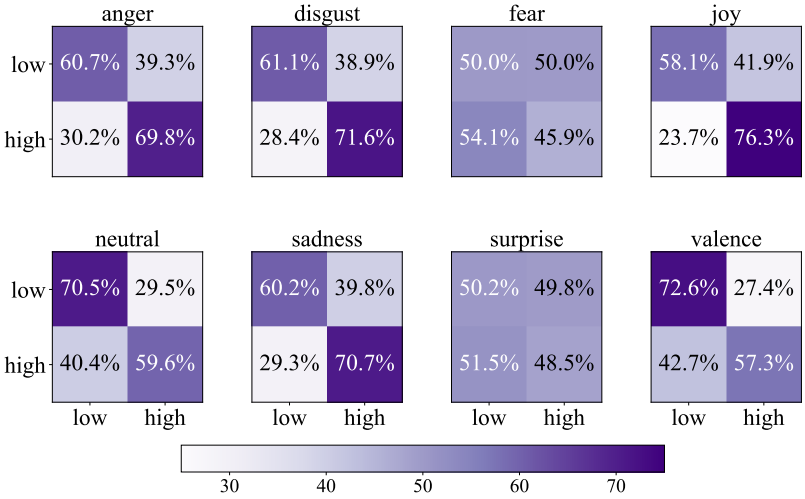


FIGURE A.3: Confusion matrix for CAN-only modality under LOSO evaluation

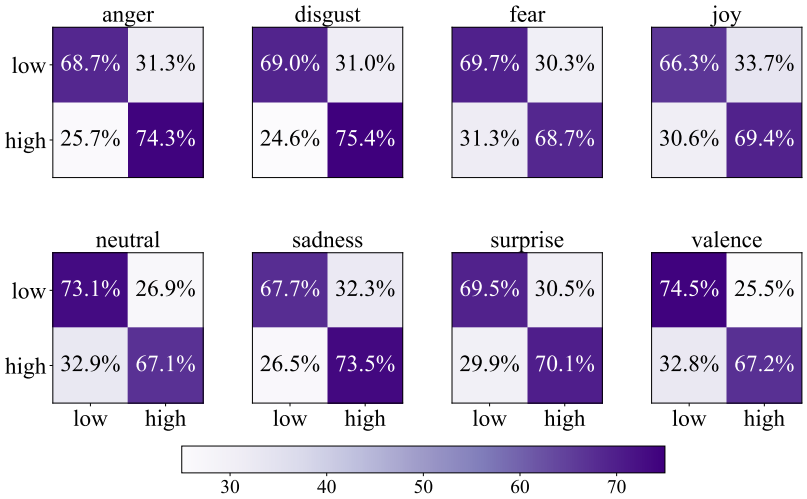


FIGURE A.4: Confusion matrix for Video-only modality under intra-subject evaluation

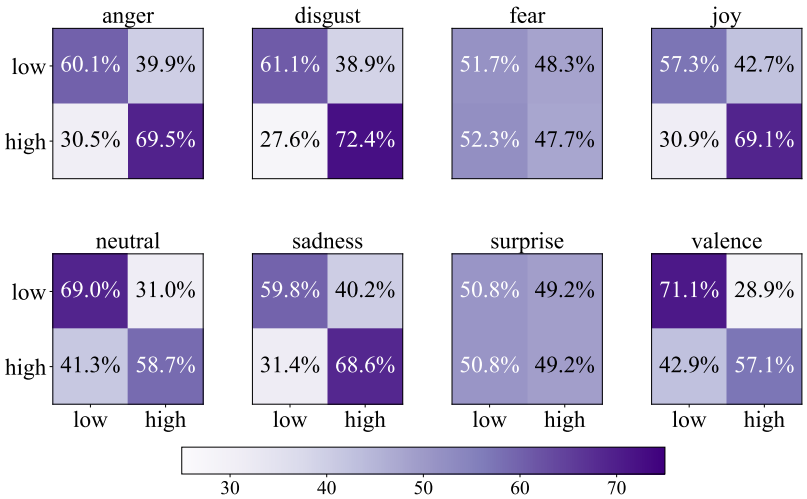


FIGURE A.5: Confusion matrix for Video-only modality under LOSO evaluation

Appendix B

Estimation of Driver Heart Rate Variability

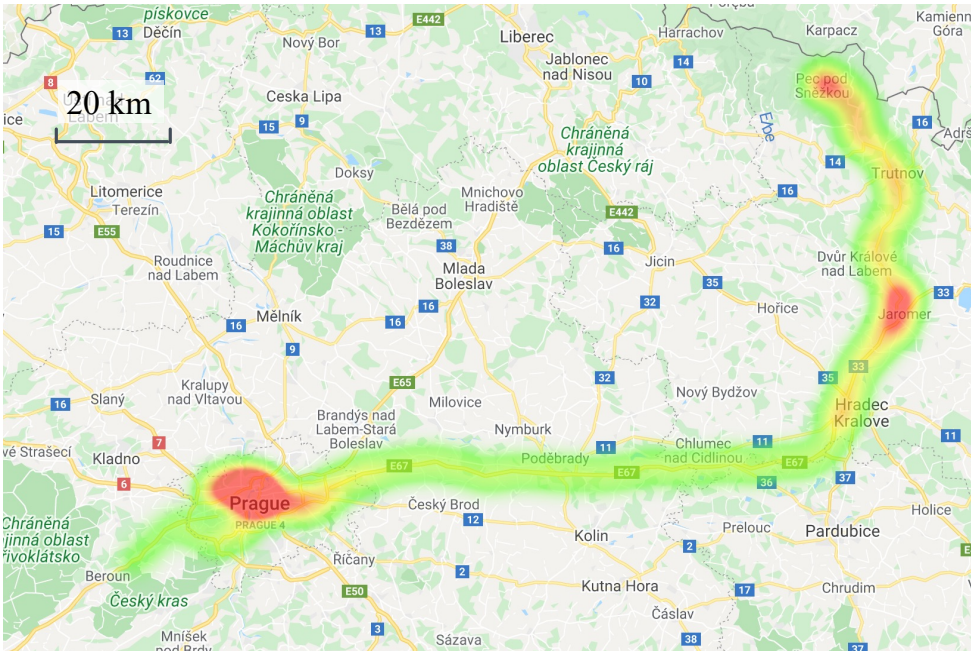


FIGURE B.1: GPS heatmap of the additional active area

TABLE B.1: Candidate parameters for grid search for tree-based models

RF, DF, TFPN	
depth of tree	10, 20, 50, None
number of trees	50, 100, 200, 300
min samples split	2, 5, 10
min samples leaf	1, 2, 5

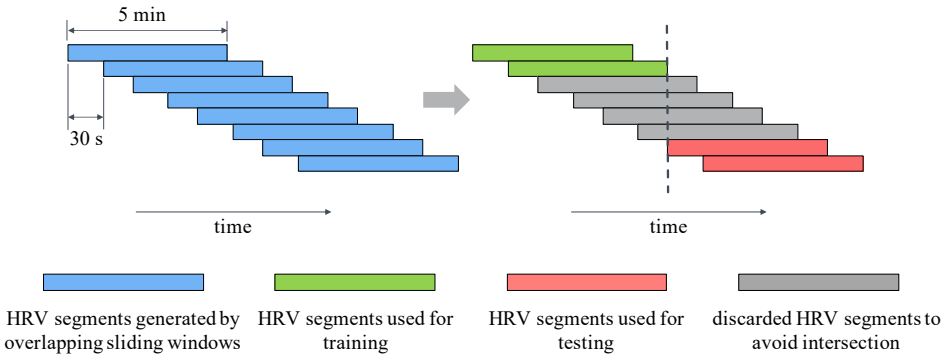


FIGURE B.2: Processing of HRV segments to avoid intersection between training and test data

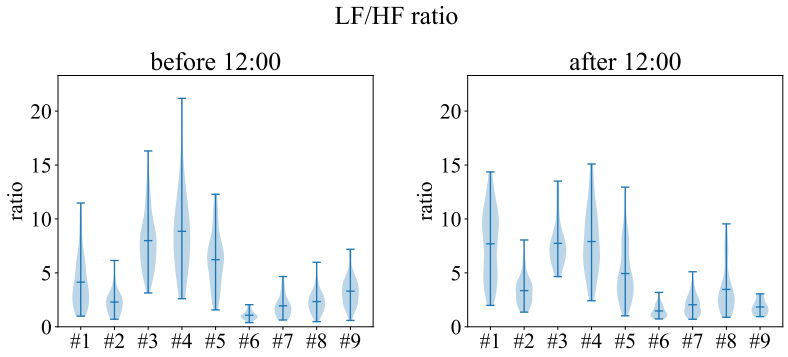


FIGURE B.3: LF/HF ratio of the nine drivers in different time interval

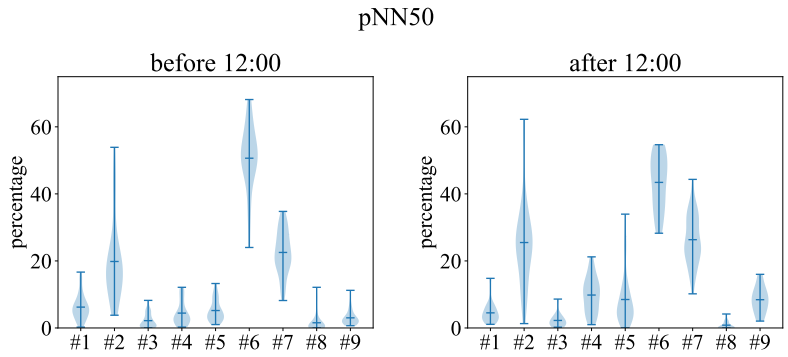
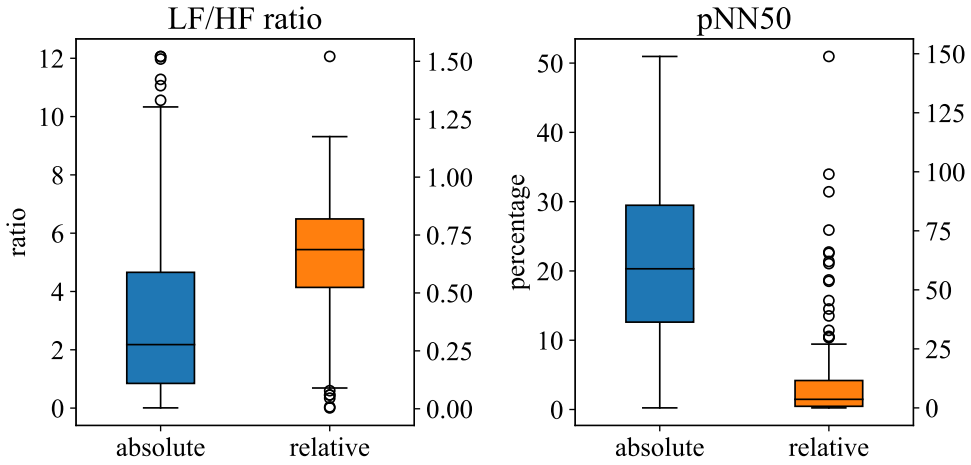


FIGURE B.4: pNN50 of the nine drivers in different time interval



(A) LF/HF ratio (B) pNN50

FIGURE B.5: Absolute and relative errors of high-end smartwatch compared with Firstbeat bodyguard 2 (Firstbeat Technologies Oy, 2019)

TABLE B.2: Candidate parameters for grid search for CNN

CNN	
# conv. layers	2, 4, 8
# filter per layer	8, 16, 32
kernel size	3, 5, 7
dropout rate	0.3, 0.5, 0.7
learning rate	0.01, 0.05, 0.005
activation	sigmoid, Relu
FC layer settings (neurons in each layer)	[16], [32], [64], [16,16], [32,32], [64,64]

TABLE B.3: Candidate parameters for grid search for RNN

RNN	
# layers	1, 2, 4
# hidden units	8, 16, 32
dropout rate	0.3, 0.5, 0.7
learning rate	0.01, 0.05, 0.005
activation	sigmoid, Relu
FC layer settings (neurons in each layer)	[16], [32], [64], [16,16], [32,32], [64,64]

TABLE B.4: Candidate parameters for grid search for MLP

MLP	
# layers	2, 4, 8
# neurons in layer	8, 16, 32
dropout rate	0.3, 0.5, 0.7
learning rate	0.01, 0.05, 0.005
activation	sigmoid, Relu