Engineering Applications of Artificial Intelligence 115 (2022) 105309

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Survey paper

Detecting and recognizing driver distraction through various data modality using machine learning: A review, recent advances, simplified framework and open challenges (2014–2021)[☆]

Hong Vin Koay [a], Joon Huang Chuah [a,*], Chee-Onn Chow [a], Yang-Lang Chang [b]

[a] Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, 50603, Malaysia
[b] Department of Electrical Engineering, National Taipei University of Technology, Taipei, 10608, Taiwan

ARTICLE INFO

Keywords:
Driver distraction
Intelligent Transportation System (ITS)
Machine learning
Deep learning

ABSTRACT

Driver distraction is one of the main causes of fatal traffic accidents. Therefore, the ability to detect driver inattention is essential in building a safe yet intelligent transportation system. Currently, the available driver distraction detection systems are not widely available or limited to specific class actions. Various research efforts have approached the problem through different techniques, including the usage of intrusive sensors, which are not feasible for mass production. Most of the work in early 2010s used traditional machine learning approaches to perform the detection task. With the emergence of deep learning algorithms, many research has been conducted to perform distraction detection using neural networks. Furthermore, most of the work in the field is conducted under simulation or lab environment, and did not validate the proposed system under naturalistic scenario. Most importantly, the research efforts in the field could be further subdivided into many subtasks. Thus, this paper aims to provide a comprehensive review of approaches used to detect driving distractions through various methods. We review all recent papers from 2014–2021 and categorized them according to the sensors used. Based on the reviewed articles, a simplified framework to visualize the detection flow, starting from the used sensors, collected data, measured data, computed events, inferred behaviour, and finally its inferred distraction type is proposed. Besides providing an in-depth review and concise summary of various published works, the practicality and relevancy of driver distraction detection towards increasing vehicle automation are discussed. Further, several open research challenges and provide suggestions for future research directions are provided. We believe that this review will remain helpful despite the development towards a higher level of vehicle automation.

1. Introduction

Intelligent Transportation Systems (ITS) has become part of our life today. They are meant to enhance transportation safety, efficiency and sustainability. At the same time, sustainable transportation systems are vital elements of sustainable cities, and they align with Goal 3 (Good health and well-being), Goal 9 (Industry, innovation and infrastructure), and Goal 11 (Sustainable cities and communities) of Sustainable Development Goals (SDGs) (General Assembly, 2015). Specifically, the global rate of mortality from road traffic injuries fell by 8.3 per cent, from 18.1 deaths per 100,000 population in 2010 to 16.7 deaths per 100,000 in 2019 (United Nations Economic and Social Council, 2021). The drop in road traffic mortality is largely contributed by the smarter vehicles and advanced driver assistance system (ADAS), which was introduced to enhance the safety of road users further. A survey reveals that ADAS can prevent up to 10,000 fatal crashes each year (McDonald et al., 2018).

Thanks to the continuous development and improvement of ADAS, some of the vehicles are now equipped with driver fatigue warning systems (Sikander and Anwar, 2018). Besides, a few automated systems are being used, such as Forward Collision-Avoidance Assist (FCA), Lane Keeping Assist (LKA), Intelligent Speed Limit Assist (ISLA), etc. (Hyundai Motor Group Tech, 2021). All these systems allow drivers to disengage driving tasks temporarily and can assist the driver in handling some dangerous situations. In reality, this assistant becomes incompetent in detecting dangerous driving situations. This is because ADAS depends on sensors and in adverse situations, and sensors have limitations and wear-offs. For example, LKA uses sensors to register lane markings on the road and might fail to activate if the road is not well marked. However, most ADAS focuses on understanding the

☆ This work is financially supported by the Faculty Research Grant (FRG) of University of Malaya with the grant no. GPF055B-2020.
* Corresponding author.
E-mail address: jhchuah@um.edu.my (J.H. Chuah).

https://doi.org/10.1016/j.engappai.2022.105309
Received 19 April 2022; Received in revised form 7 July 2022; Accepted 1 August 2022
Available online 13 August 2022
0952-1976/

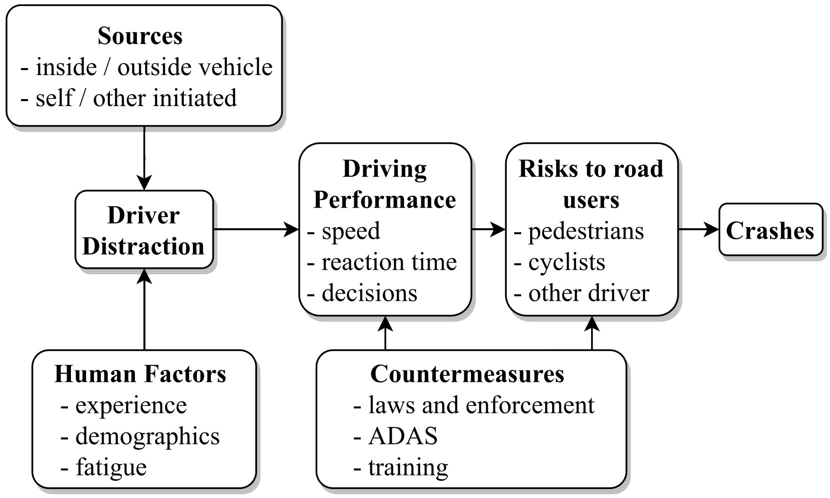

Fig. 1. Outline of road safety aspects with driver distraction.

surroundings, and little effort is applied to monitor the driver. While we are looking forward to fully automated vehicles yet accident-free vehicles on the road, there are still gaps in technological perspective. Moreover, the main obstacle in achieving a safe transportation system is driver distraction.

Driver distraction is one of the main contributors to the number of road accident death (Olson et al., 2009; National Safety Council, 2010). World Health Organization (WHO) reported that nearly 1.25 million fatalities each year, averaging about 3287 deaths a day (World Health Organization, 2018). It is estimated that this figure will continue to climb over the years. In the USA, between 15% to 18% of road traffic accidents were caused by driver distraction, as reported by the National Highway Traffic Safety Administration (NHTSA) (National Highway Traffic Safety Administration, 2019, 2020). Another report suggested that young drivers were distracted in 58% of the analysed crashes (Carney et al., 2015). In Australia, 14% of all crashes involve a distracted driver (DriveRisk, 2020); in Canada, 27% of all crashes involve a distracted driver (Marija, 2022). From an economic perspective, it is estimated that road injuries will cost the world economy as much as US$1.8 trillion from 2015 to 2030. This amount is equivalent to an annual tax of 0.12% on the global gross domestic product (GDP) (Chen et al., 2019).

Driving requires the driver's full attention to control a vehicle safely and respond to events happening on the road. It is a skill that involves constant yet complex coordination between mind and body. Distraction occurs when there are events preventing drivers from operating a vehicle safely. Driver distraction can be categorized into four visual categories (eyes off the road), manual (hands off the wheel), auditory (listening to phone) and cognitive (mind off the task) (Strayer et al., 2013; European Commission, 2015).

A survey conducted in Australia found that the most common distracting activities during driving were the usage of mobile phones, which contributed to the lack of concentration while driving (McEvoy et al., 2006). It is estimated that about 85% of American drivers use a cell phone while driving (Goodman et al., 1997) and about 5% of cars on US roadways are driven by people on phone calls during daylight hours (Pickrell et al., 2015). The research found that a driver who uses his phone while driving suffers from reduced detection abilities and a slower reaction time (Caird et al., 2008; Lamble et al., 1999).

A general outline of road safety aspects is illustrated in Fig. 1. We can observe how driver distraction is propagated along with driving performance, putting risks on other road users and ultimately involving crashes. Ultimately, there exists countermeasures to prevent crashes to happen, such as enforcing stricter rules, usage of ADAS or driver education. Note that the outline did not include many other aspects, such as extreme weather condition, poor road condition and hazards due to other road users. From the figure, we can observe that driver distraction is one of the key element in achieving safer ITS and efforts should be made to prevent deadly crashes. As for the countermeasures, we identify the three common ways to prevent crashes caused by driver distraction, which are laws and enforcement, ADAS, and re-educating the driver. However, we believe that ADAS is the easiest way to prevent a deadly crash due to distractions. Specifically, ADAS includes driver distraction system that could identify if a driver is distracted and provide warnings to the driver.

Distraction detection is a trending topic, which encapsulates several challenges, such as background noises, and illumination variations. The methods for detecting driver distraction can be grouped into three subgroups: mathematical models, rule-based models and models that are based on machine learning (ML) algorithms (Sikander and Anwar, 2018). Besides categorizing into the methods used, we could also categorized them based on the sensors used to collect driver or vehicle data. Specifically, we can divide the sensors used into three categories, which are physiological, visual and external sensors. From these three different categories of sensors, we provide an in-depth overview of works done in the respective category. Specifically, we observe that many research works tackle specific part of problem, and most of them are done in a simulated or lab environment, which prohibits them to apply in naturalistic driving scenario. For example, physiological sensors are intrusive and not feasible to deploy in real-world scenario, however they are capable of accessing driver's body condition and sensitive to minor changes. Visual sensors are the most feasible choices since they are cheap and easy to setup, however the process of collecting data is invading driver's privacy. Some recent works also look into the possibility of combining different category of sensors to achieve higher accuracy and sensitivity. Thus, to further understand how far driver distraction detection has developed, we present this paper to investigate and analyse all existing methods of detecting distraction using ML models. Various methods that have been proposed throughout the years are studied, and a general flow to tackle this problem is developed. We hope that this paper can provide fundamental knowledge for future researchers and accelerate the building of novel systems.

Acknowledging the fact that the field of ML moving in a rapid phase, there is a need to have a summary of the work done in the field of driver distraction detection. Coupled with the easy-to-use ML library and state-of-the-art pretrained models, many overlapping works have been done and the progress in the field has been blurred out. Besides, with the introduction of many publicly available datasets, we noticed two streams of works, where most of the work utilizes the advance deep learning models in performing distraction detection,

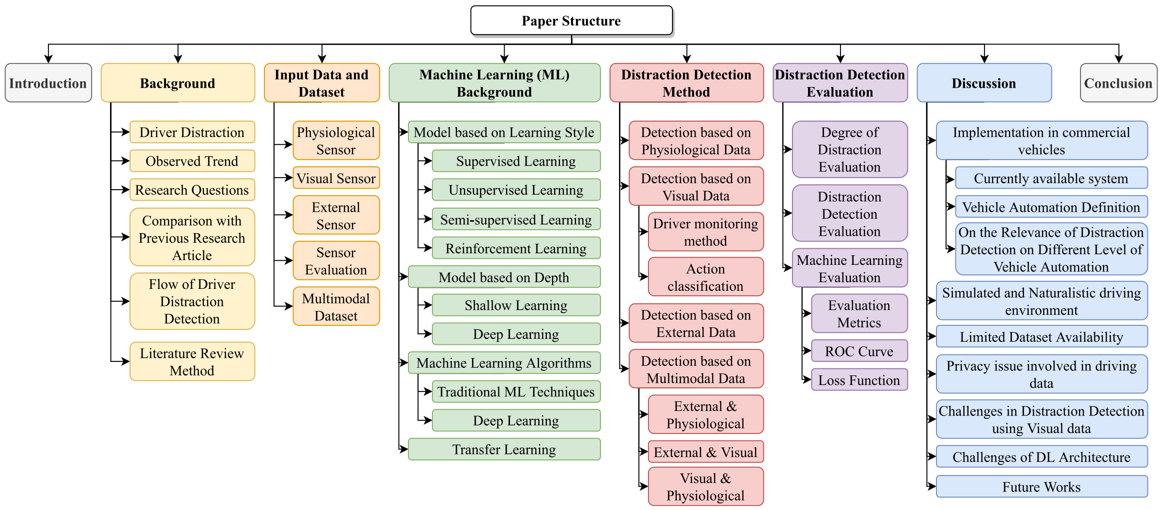

Fig. 2. Visualization of this paper structure.

while the other still uses traditional ML methods to indicate if a driver is being distracted. Both of the streams usually shared the same keyword of "driver distraction detection", but ultimately the contribution is different. Further, even though the level of vehicle automation keeps on increasing.

This paper's primary work is to analyse algorithms that have been proposed over the past decade, and then the advantages and disadvantages of each algorithm are further discussed. Given the papers involved in the field is growing in an exponential speed, there is a need of research article to summarize all the works done, preventing repetitive works while focusing the current limitation and improving it. The main contribution is twofold, where we first define the research questions to guide the flow of this review and then address them accordingly in the following sections. Specifically, the contribution of this work can be summarized as follows.

1. To provide a methodological review of recent trends in driver distraction detection based on various input data.
2. To define and understand the actions and sources of distraction while driving.
3. To present an overview of various sensors used to collect the raw input data, including the usage of multi-view, multimodal and multi-spectral data to detect distractions.
4. To explore the available datasets used in the field of driver distraction detection.
5. To conduct a comparison of different ML algorithms used to populate unimodal and multimodal features.
6. To provide an introduction to all traditional ML, deep learning (DL) models, and the concept of transfer learning.
7. To put together a review of articles on detecting driver distraction from 2014–2021 using only ML algorithm.
8. To present an overview of common evaluation metrics used to evaluate the proposed algorithms.
9. To discuss the outlook and future directions for the development of driver distraction detection system.

The paper structure is illustrated in Fig. 2. The remainder of this paper is organized as follows. In Section 2, we include the background of distracted driving and the observed trend. We explain the research protocol, including research questions, selection of papers and comparison to previous survey articles. In Section 3, we describe the type of input sensors used to collect raw data and datasets available for driver distraction detection. In Section 4, we discuss the classification of the ML algorithm. We provide a brief overview of various traditional ML and DL models used in the field. The concept of transfer learning is discussed in the same section. Then, we include the current studies on distracted driving and discussed its limitation in Section 5. The merits and demerits of each method are discussed, alongside the results obtained from the proposed methods. In addition, this section also presents the current limitation and future directions in this domain. Section 6 provides an overview of various evaluation methods used to examine the proposed methods. The general discussion, requirements, challenges and future directions of this field are discussed in Section 7. Section 8 concludes this paper.

2. Background

2.1. Driver distraction

In this subsection, we compile the well-known definitions of driver distraction by various parties.

- "Any activity that diverts attention from driving, including talking or texting on your phone, eating and drinking, talking to people in your vehicle, fiddling with the stereo, entertainment or navigation system – anything that takes your attention away from the task of safe driving" (National Highway Traffic Safety Administration, 2021).
- "A diversion of attention from driving, because the driver is temporarily focusing on an object, person, task or event not related to driving, which reduces the driver's awareness, decision-making ability and/or performance, leading to an increased risk of corrective actions, near-crashes, or crashes" (Hedlund et al., 2006).
- "A diversion of attention away from activities critical for safe driving toward a competing activity" (Lee et al., 2008).
- "Insufficient or no attention to activities critical for safe driving" (Regan et al., 2011).
- "A triggering event induces an attentional shift away from the task" (Horberry et al., 2006).
- "Delayed in recognition of information needed to accomplish the driving task safely, because of some event, activity, object, or person within or outside the vehicle compels or induces the driver's shifting attention away from the driving task" (Stutts et al., 2001).

From the above definitions, we extract the key element of driver distraction as follows:

- There is a shift of attention away from driving.

Table 1
Common distraction actions and their mapping to type of distraction.

Activity	Self-initiated	Location	Distractions
Using Phone	✓	Inside vehicle	Auditory, Cognitive, Visual, Manual
Eat, Drink	✓	Inside vehicle	Visual, Physical
Looking advertisement	✗	Outside vehicle	Visual, Cognitive
Listening music	✓	Inside vehicle	Auditory, Cognitive
Daydreaming	✓/✗	Inside vehicle	Cognitive

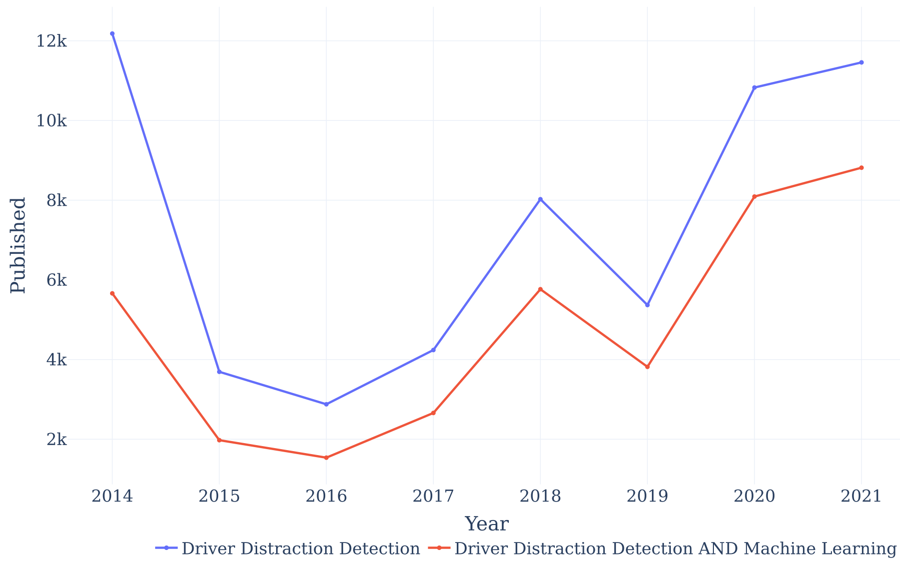

Fig. 3. Paper published from year 2014–2021 with the keyword “driver distraction detection” and “machine learning”.

• Safe driving is not achieved.

However, the definition of driver inattention and driver distraction could be different. There are two views of the relationship between inattention and distraction. Some researcher recognizes driver distraction as a form of inattention while others did not. We refer the reader to Section 4 of the paper by Regan et al. (2011) for the differences between driver distraction and driver inattention. In this work, we recognize that distraction is a subset of inattention.

Distractions come in many forms. As discussed earlier, we can categorize distraction into four forms, which are cognitive, visual, manual and auditory (Strayer et al., 2013; European Commission, 2015). The driver tends to shift their attention away from the driving task to non-driving secondary tasks by taking their hands (manual distraction), eyes (visual distraction), mind (cognitive distraction), and/or ear (auditory distraction) off the driving task. Some activities can involve two or more distractions. For example, the usage of cell phones involves all four distractions.

• Manual distraction because it interferes with controlling the vehicle.
• Visual distraction because the user watches the device instead of the road.
• Cognitive distraction because interpreting information from the device directs attention away from driving.
• Auditory distraction because the sound diverts attention away from driving.

We include some of the examples of distraction action and its mapping to the types of distraction in Table 1.

2.2. Observed trend

The number of paper published from the year 2014–2021 is shown in Fig. 3. The keywords used are “driver distraction detection” and “machine learning” to generate the figure.[1] It is observed that there is an increasing trend of published papers in the field, signalling that this field is highly active in development.

An overview of technological evolution in the development of driver distraction detection is shown in Fig. 4. We can observe that the trend of using DL techniques started as early as 2014. Undeniably, there are still research works that utilize the standard ML approaches due to their simplicity and lightweight properties. Since the introduction of vision transformer (Dosovitskiy et al., 2020) in 2021 for computer vision tasks, we believe that the trend of adopting vision transformer in this field will begin to populate fast. In terms of data modality, we observe that multi-modality (cross sensor category (e.g. visual and physiological), where RGB and IR are not considered multimodal in this case) started in 2017 when researchers found that more data could churn accurate detection. Multisensor, such as GPS and EEG, in detecting distraction started as early as 2014. The last viewpoint is the inferencing devices. We notice that the use of IoT devices started to populate since its introduction in the early 2010s. As for edge inference, or cloud inference, we observe that such technology started only at the end-2020. We believe the trend of embedded devices and edge inferencing is the future of distraction detection.

2.3. Research questions

We provide the research questions in Table 2 as the starting point of this review article. We include the queries related to the field of driver distraction detection.

RQ1–RQ3 are answered in this section, RQ4–RQ6 are answered in Section 3, RA7–RQ9 are answered in Section 4, RQ10–RQ11 are answered in Section 5, RQ 12 is answered in Section 6, and RQ13–RQ14 are answered in Section 7.

[1] The website used to collect the number is https://dimensions.ai with the search of keywords in full text.

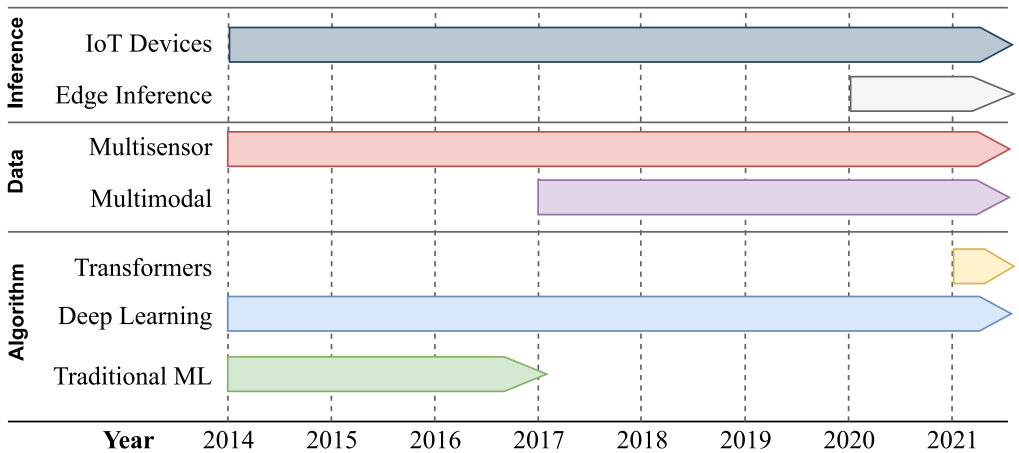

Fig. 4. Trend and technology evolution of driver distraction detection in terms of algorithm, data modality and inferencing.

Table 2
Summary of RQs and motivations involved in this study.

RQ	Research Question	Motivation
RQ1	What type of research is being done on driving distraction detection?	To understand studies done in this domain.
RQ2	How do we define driver distraction?	To gather various definition of driver distraction in the field and summarize them.
RQ3	What actions done by drivers are considered as distraction while driving?	To know the boundary of actions that are considered as distractions, which may bring deadly accidents.
RQ4	What sensors are used to collect data for detecting distractions?	To understand different sensors used to collect the input driver distraction data.
RQ5	Which types of input data is most suitable and commonly used in the field?	To evaluate the feasibility and effectiveness of the collected data in detecting distracted drivers.
RQ6	What dataset of driving distraction are available?	To understand types of dataset is collected (inside or outside vehicle; videos or images) in this domain, and identify the availability of dataset for training and testing.
RQ7	What are the types of ML algorithm used from the past to present?	To identify the advancement of ML algorithms used in this domain.
RQ8	What is the role of DL architecture in this domain?	To study the latest advancement in ML and understand if DL would benefit the system.
RQ9	Does transfer learning (TL) could effectively used in detecting the distractions?	To study the usage TL in detecting distractions.
RQ10	What is the typical flow of detecting driver distraction?	To understand the framework used to model and detect driver distractions.
RQ11	Does multi-view and multi-modal could benefits the distraction detection system?	To understand if more features collected would benefit the system at a whole.
RQ12	What are the evaluation metrics used to differentiate between studies?	To understand the common metrics used in studies to evaluate the proposed models.
RQ13	What are the strengths and weaknesses of different of ML models?	To identify the strengths and weaknesses of different ML models.
RQ14	What are the current limitations and the future direction of this research areas?	To identify the research gap of this domain.

2.4. Comparison with previous research article

The authors in Sigari et al. (2014) presented a review on driver face monitoring systems for fatigue and distraction detection. They focused on various techniques to detect distraction and fatigue through the use of driver's face images.

The authors in Kaplan et al. (2015) reviewed several techniques of driver monitoring, mainly for drowsiness and distraction. Besides, they explored the use of smartphones and wearable devices for driver monitoring. However, this article mainly focuses on detection techniques based on head pose and gaze. There is a lack of modern neural network methods in the survey article since they only review works from the year 2008–2014.

The authors in Oviedo-Trespalacios et al. (2016) covered the mechanisms involved with mobile phone distractions and the driving task and subsequent outcomes. This article analysed distraction assessment methods in detail, comprising articles from the year 2005–2014. However, the study mainly focuses on the distractions caused by mobile phones. This article's main contribution is the introduction of the human–machine framework to isolate the components, and various interactions associated with mobile phone distracted driving.

The author in Chhabra et al. (2017) provided overviews of various driver behaviour monitoring, such as driving style, fatigue, inattentiveness, drunk and aggressive driving. The author split the driver behaviour detection technique into real-time and non-realtime. However, this article provided a less detailed overview of the detection mechanism and most of the articles reviewed uses traditional ML algorithms, such as fuzzy logic and principal component analysis (PCA).

The author in Khan and Lee (2019) summarized the work in the field of distraction and fatigue detection and driving style analysis. The author summarized the driving monitoring system through the three main components of driving tasks: driver, vehicle, and driving environment. Despite the broad topics, the authors provide a clear and deep overview of driver distraction detection through physiological sensors, such as electroencephalogram (EEG) and electrocardiogram (ECG). In addition, they also included the modelling and recognition of driving style behaviour and the models and systems developed to avoid collisions between vehicles. Overall, this research mainly focuses on using physiological sensors only to monitor the driver intrusively, without discussing the use of other non-intrusive sensors to collect input data.

The author in Sikander and Anwar (2018) reviewed the recent advancement in driver fatigue detection. They first reviewed various

Table 3
Comparison with previous review articles.

Ref	Year	Reviewed years	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7	RQ8	RQ9	RQ10	RQ11	RQ12	RQ13	RQ14
Sigari et al. (2014)	2014	1998–2013	✓			✓			✓			✓				✓
Kaplan et al. (2015)	2015	2008–2014	✓				✓	✓		✓			✓			✓
Oviedo-Trespalacios et al. (2016)	2016	2005–2014	✓										✓			✓
Chhabra et al. (2017)	2017	2004–2013	✓			✓		✓								
Khan and Lee (2019)	2019	1968–2019	✓	✓	✓	✓			✓			✓			✓	✓
Sikander and Anwar (2018)	2019	1982–2019	✓			✓			✓	✓						
El Khatib et al. (2019)	2019	1990–2019	✓		✓											✓
Abou Elassad et al. (2020)	2019	2009–2019	✓						✓			✓		✓	✓	✓
Alkinani et al. (2020)	2020	2018–2020	✓	✓	✓				✓	✓		✓			✓	✓
Zhang and Eskandarian (2020)	2020	1983–2020	✓			✓	✓		✓	✓			✓		✓	
Kashevnik et al. (2021)	2021	1967–2021	✓	✓	✓							✓	✓	✓		✓
Abbas and Alsheddy (2021)	2021	2009–2021	✓			✓		✓	✓	✓	✓		✓	✓	✓	✓
Moslemi et al. (2021)	2021	2005–2020	✓	✓	✓			✓								✓
Ours	2022	2014–2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

commercial systems in detecting driver fatigue. Then, they categorize novel driver fatigue detection methods into three categories: mathematical model, rule-based, and ML. The authors also look into the feature-based fatigue classification methods, which they categorized into five groups: subjective reporting, driver biological features, driver physical features, vehicular features while driving, and hybrid features. However, this paper only looks into fatigue detection without considering other sources of distractions.

The authors in El Khatib et al. (2019) reviewed various forms of distractions, including visual, cognitive, and manual distraction, and various features and algorithms used in both the research literature and in industrial production. They contextualized distraction detection systems within the SAE's framework of automation. They explored the transfer of control of the vehicle from an automated system to a human in the same work.

The authors in Abou Elassad et al. (2020) investigated and summarized the driver behaviour concept in various articles. They proposed a framework to conceptualize different facets of driver behaviour analysis and a scheme for future development and implementation of driver behaviour assessment strategies. An overview of various ML and non-ML techniques is used in the same work for driver behaviour analysis.

The author in Alkinani et al. (2020) classified human driver inattentive driving behaviour into distraction and fatigue. Moreover, they discussed the causes and effects of aggressive human driving behaviour. A brief introduction to DL algorithms used in detecting inattentive driving behaviour was included in the study. This article ends with the open challenges in the field. However, they reviewed very limited articles in the field.

The author in Zhang and Eskandarian (2020) provides a meta-analysis on EEG-based brain monitoring for driver state analysis. This article only considers EEG sensors as input data to monitor the driver's state. The authors explained the commonly used EEG system setup for driver state studies, followed by different methods for signal preprocessing, feature extraction, and classification.

The author in Kashevnik et al. (2021) reviewed various driver distraction detection methods and integrated the identified approaches into a framework. Their work included reviews of distractions, distraction evaluation methods, and outlooks for the automated driving era. A distraction detection framework is proposed. However, the framework lacked some aspects, such as more data can be inferred from the processed raw data. This work has almost a similar structure to this study, but lesser articles are considered.

The author in Abbas and Alsheddy (2021) reviewed methods for multi-stage hypovigilance detection. They highlighted the recent trends of hypovigilance detection systems through various input data. This paper is quite complete with the review of various ML and DL algorithms used for hypovigilance detection and the usage of multimodal data for performance improvement. However, this work is more towards hypovigilance/fatigue detection, while we focus more on distraction detection.

The author in Moslemi et al. (2021) reviewed various approaches to recognition of driver distraction actions using computer vision and discussed the types of its approaches. However, this article did not perform a deeper review on the field, with limited papers considered in the article.

We observed a lack of completeness in reviewing various techniques in detecting driver distraction from all previous review articles stated above. Many review articles only focus on certain type of distraction (Oviedo-Trespalacios et al., 2016), fatigue or drowsiness detection (Kaplan et al., 2015; Sikander and Anwar, 2018; Abbas and Alsheddy, 2021), driver monitoring (Sigari et al., 2014; Chhabra et al., 2017; Khan and Lee, 2019; Abou Elassad et al., 2020), or certain types of input data only (Zhang and Eskandarian, 2020). We wish to bridge the gap by including all sensors, methodology, and ML algorithms used to detect driver distractions.

We summarize all previous research articles discussed above in Table 3 according to our predefined research questions.

2.5. Literature review method

We limit our search from the year 2014–2021 and only consider articles utilizing ML algorithms. We started this review article by conducting a keyword search through Google Scholar with "driver distraction", "driver monitoring system", "distraction detection", and "driving distraction". To further broaden the scope, we searched the references of identified papers.

2.6. Flow of driver distraction detection

We identify the main components of driver distraction detection techniques. There are four main components, which are the input sensors used, collected data, feature extraction methods, and the ML algorithm, as shown in Fig. 5. We will discuss every component in this study.

3. Input data and dataset

The first step of detecting driver distraction is the input data from sensors. Many types of input data can be collected from sensors to determine if a driver is distracted. We categorize the sensors into three main categories, which are physiological, visual and external, as illustrated in Fig. 6. Physiological input data are collected from physiological or biophysical sensors related to the driver's body. Visual input data are usually collected with visual sensors to monitor the driver, such as imaging devices. As for external sensors, we regard all other sensors that collect vehicle data and driving data.

The sensors are further divided into two categories, intrusive (solid line boxes in Fig. 6) and non-intrusive (dotted line boxes in Fig. 6). The intrusiveness of sensors is defined as such that the sensor may interfere with the drivers and affect their driving behaviour. Intrusive sensors

Fig. 5. The flow of driver distraction detection.

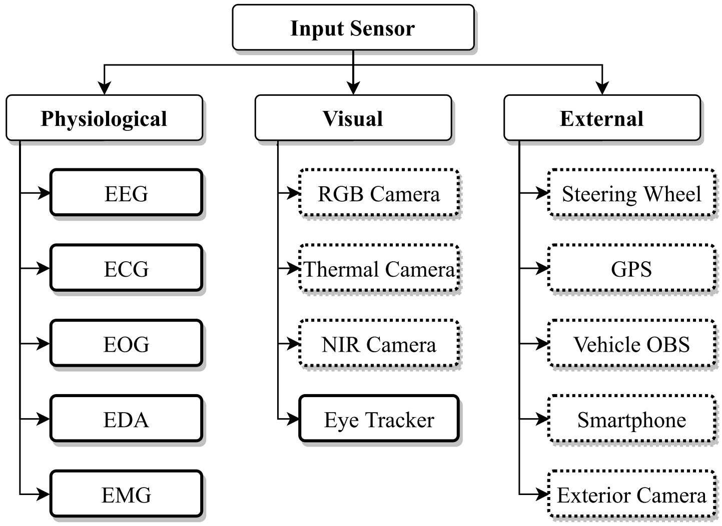

Fig. 6. Type of input sensors used to detect driver distraction. Solid and dotted boxes represent intrusive and non-intrusive sensor, respectively.

are more valuable since they directly interact with the driver to collect more accurate data. On the other side, non-intrusive sensors do not distract the driver and are more favourable in the field. These sensors are currently being built into commercial vehicles to allow several tasks to be automated, such as lane-keeping. Ideally, we utilize sensors to collect data and analyse if the driver is distracted without feeling increased effort or reduced performance. Therefore, non-intrusive sensors are preferable, such as a vision-based system that does not interfere with the driver.

The main objective of the input sensor is to collect raw data to determine if a driver is distracted. Raw data come in all forms, including numerical, text or image. In this work, we only consider two primary data types: numerical and image data. We regard all data, not in image form and are presented in numerical form as numerical data. This data is usually collected from sensors other than camera and depth sensors.

3.1. Physiological sensors

Five primary physiological sensors are commonly used to monitor the state of the driver, which are electroencephalogram (EEG), electrocardiogram (ECG), electrooculography (EOG), electro-dermal activity (EDA), and electromyography (EMG).

EEG is mainly used for brain activity research. The frequency domain, time domain, and individual component analysis of EEG data derive useful information. Frequency domain features include mean frequency, energy contents of α, β, θ and σ bands (Thorslund, 2004). These features can then be used to detect if a driver is undergoing fatigue. Time-domain features, such as standard deviation and average value, provide information on brain activity. EEG can effectively produce brain activity and differentiate between awake and asleep. The main limitation of the EEG sensor is the placement of electrodes on the driver's head, which may affect the driver's concentration on the road. Besides, the complex arrangement of the sensor makes it unpractical in a real-life scenario. Moreover, EEG data always suffers from artefacts, and noise (Taherisadr et al., 2017), therefore a filter is always deployed to preprocess the collected data. We refer the reader to a survey article by Zhang and Eskandarian (2020) for a detailed explanation of data collection techniques and signal processing.

ECG provides the electrical activity of the driver's heart voltage versus time. The collected ECG signals can then calculate heart rate, heart rate variability (HRV), and respiration rate. These inferred data can provide valuable information related to driver's fatigue. Features such as root mean square standard deviation (RMSSD), the standard deviation of normal RR-interval (SDNN), HRV, triangular index, spatial filling index, central tendency method, correlation dimension, approximate entropy, the proportion of NN20 (pNN20), heart rate RR-interval QRS complex, R peak, pulse arrival time (PAT), respiratory rate, sample entropy, complexity of ECG are extracted by researchers to identify the internal states (Murugan et al., 2020). However, distraction detection through ECG is subjective and person-dependent; besides, it came with motion artefacts (Deshmukh and Dehzangi, 2019).

EOG provides a measure of the corneo-retinal standing potential between the front and back of the human eye (Noor and Mustafa, 2016). EOG is used to measure the eye movement of the driver. From EOG, we can post-process to obtain blink duration, blink frequency, blink amplitude, percentage of eyelid closure over the pupil over time (PERCLOS), lid reopening delay, and eyeball movement. EOG data has significant variations since different subjects react to drowsiness and distractions differently. Thus, detection based on the subject's eye is not practical yet added distraction to the driver.

EDA, also known as galvanic skin response (GSR), provides a measure of skin conductance that changes due to the secretion of the sweat gland. The data is collected by applying a low, undetectable, and constant voltage to the skin and then measuring how the skin conductance varies (Benedek and Kaernbach, 2010). EDA has been closely linked to autonomic emotional and cognitive processing, and EDA is widely used as a sensitive index of emotional processing, and sympathetic activity (Braithwaite et al., 2013). EDA measures such as skin conductance level (SCL) and skin conductance response (SCR) have been reported to be sensitive to arousal, and mental workload in driving as well (Solovey et al., 2014).

EMG is a technique for evaluating and recording the electrical signal generated from muscle contraction. Frequently used time-domain features from EMG data are mean absolute value, zero crossings,

Table 4
List of public physiological-based driver monitoring datasets used to identify driver distraction.

Ref	Sensor	Year	Subjects[a]	EEG channels	Usage
Min et al. (2017)	EEG	2017	12 (12/0)	40	Fatigue detection
Zheng and Lu (2017)	EEG, EOG	2017	23 (11/12)	12	Vigilance estimation
Cattan et al. (2018)	EEG	2018	20 (13/7)	16	Fatigue detection
Cao et al. (2019)	EEG	2019	27	32	Fatigue detection

[a]Number of subjects (Male/Female).

Table 5
Summary of advantages and disadvantages of psychological sensors.

Sensor	Advantages	Disadvantages
EEG	Cheap, fast and safe to access brain activity. It has good temporal resolution.	Limited spatial resolution. Subject to electrical and physiological artifacts. Complex to set up and analyse the data.
ECG	Mainly used to monitor drowsiness since it directly collects heart data.	Does not provide sufficient insights for detecting driver distractions, and is person-dependent. Suffer from artifacts.
EOG	Easy to use. Accessing to driver's eye directly.	The placement of electrode will affect the data collected. Detection based on blinking behaviour is person-dependent.
EDA	Provide insight into the level of emotional arousal.	Highly sensitive to atmospheric humidity and temperature.
EMG	Records single muscle activity and provides access to deep musculature.	Detection area does not represent the whole muscle.

slope sign changes, waveform length, Willison amplitude, v-order, log-detector, EMG histogram, autoregression coefficient, and Cepstrum coefficients (Tkach et al., 2010).

There is some publicly available dataset collected from physiological sensors, as shown in Table 4. However, many of the studies did not disclose their data publicly; besides, most of the intended usage is to detect fatigue and vigilance.

We summarize the advantages and disadvantages of each physiological sensor in Table 5. There are still many physiological sensors available, but we only consider the main five, as discussed before.

3.2. Visual sensors

Visual sensors are used to collect visual data in RGB, grayscale, depth, and IR images. Multiple vision sensors can be used to collect different modalities of images, including monocular cameras, colour cameras, and depth cameras. These sensors collect drivers' naturalistic driving data and analyse drivers' interaction behaviour with the surrounding.

Many features could be inferred from a single image. Besides, many works started to detect, localize and track drivers' body parts such as faces, hands, feet, heads, and gazes, to detect distractions (Jegham et al., 2020a). For example, a camera pointing toward the driver's face could collect features such as eye closure, mouth opening, an object near the ears, and head pose. These collected features can be used to determine if a driver is experiencing fatigue or distractions. This refers to the basics of the driving task, where the driver must keep their hands on the steering wheel and their eyes on the road. On the other hand, if the camera is located inside the cabin, the whole pose of the driver can be collected, as well as head movement, pose estimation, hands-on wheel, and objects inside the cabin. These angles are usually used to detect and classify actions that cause distraction, such as drinking or using a mobile phone.

Usually, colour images are preferred in almost all conditions since it has proven their robustness, especially in controlled environments. However, in naturalistic driving settings, captured colour images are impacted by various weather and illumination conditions, which influence the image quality. Therefore, researchers tend to replace the monochrome and colour images, which are located in the visible spectrum, with images from other modalities to improve the system's performance, allowing it to work in any conditions. Infrared cameras can produce clear images under any conditions, and they are designed to be used in poor lighting and weather conditions.

In naturalistic driving settings, the accuracy of vision-based driver distraction detection methods remains limited due to many challenges present in this domain, such as dynamic background, occlusion, and bad visibility. Therefore, the availability of a public dataset is crucial to benchmark the suggested novel algorithm on the same dataset effectively. Some of the publicly available datasets are summarized in Table 6. Note that we include only the dataset used in the article and introduced from 2014–2021. However, many of them are not publicly available (as shown in Table 11), and therefore other researchers could not validate or enhance the proposed algorithm on the same dataset.

As shown in Table 6, we summarize the usage into four main categories, which are hand detection, body pose estimation, face detection, and distraction/drowsiness. Hand and face detection and body pose estimation are used to infer if the driver's hand is on the steering wheel, the head is on the road, and the body pose is facing towards the road. These are then used to classify if a driver is being distracted or not. For distraction detection, usually, such a dataset comes with a set of predefined actions recognized as a distraction. These datasets are usually used to perform binary or multi-class classification of action.

There is another type of sensor in the visual sensor category that can detect driver distraction — eye tracker. An eye tracker is a device for measuring eye positions and eye movement. There are two categories of eye-tracking systems based on the location of the eye tracker placed, which are head-mounted wearable and remote (Ojsteršek, 2019). Head-mounted wearables usually came in the form of glasses with a helmet, while remote eye-tracking systems came in the form of cameras that record eye movement data.

3.3. External sensors

We consider all types of sensors that did not collect direct data from drivers as external sensors. Therefore, we identify five primary sensors in this category, which are the steering wheel, a GPS, vehicle on-board diagnostics (OBS), smartphone, and exterior camera (camera that is not inside the cabin).

A variety of steering wheel metrics have been studied to detect abnormal patterns. This includes the standard deviation of steering wheel angle, steering wheel reversal rate, high-frequency steering, and steering entropy. The data collected from the steering wheel can relate to its driving behaviour and thus its attention on the road.

GPS is a satellite-based radio navigation system used to capture location. GPS devices can be deployed to collect the location of vehicles to study driving behaviour such as speeding, acceleration, and deceleration/braking. These data could then be used as a different modality to infer that distraction, such as speeding, could be resulted from distraction.

A vehicle OBS logger is a device used to collect information from the vehicle. Information, such as speed, acceleration, braking, pedal force,

Table 6
List of public vision-based driver monitoring datasets.

Ref.	Dataset	Year	Subjects[a]	Act	Views[b]	Size[c]	Streams	Ground truth	Occ	Scenario	Usage
Ohn-Bar and Trivedi (2014)	CVRR-Hands	2014	8 (7/1)	–	1	7207	RGB, Depth	Hands, actions	✓	Naturalistic driving	Hand detection
Abtahi et al. (2014)	YawDD	2014	107 (57/50)	3	1	342 videos	RGB	Actions	✗	Naturalistic driving	Drowsiness, distraction
Jain et al. (2015)	Brain4Cars	2015	10	5	2	2M	RGB	Road, head, face, action	✗	Naturalistic driving	Distraction, autonomous driving
Das et al. (2015)	VIVA Hand Detection	2015	–	–	7	54 videos	RGB	Hands	✗	Naturalistic driving	Hand detection
Martin et al. (2016)	VIVA Face Dataset	2016	–	–	1	39 videos	RGB	Head	✗	Naturalistic driving	Face detection, head pose
Diaz-Chito et al. (2016)	DrivFace	2016	4 (2/2)	–	1	606	RGB	Head	✗	Naturalistic driving	Head pose
Weng et al. (2016)	NTHU-DDD	2016	36 (18/18)	8	1	360 videos	RGB, IR	Actions	✓	Simulated settings	Drowsiness
StateFarm (2016)	StateFarm	2016	26	10	1	102k	RGB	Actions	✗	Naturalistic driving	Distraction
Borghi et al. (2017)	Pandora	2017	22 (10/12)	–	1	250k	RGB, Depth	Head, body	✓	Simulated settings	Body pose
Schwarz et al. (2017)	DriveAHead	2017	20 (16/4)	–	1	10.5h	IR, Depth	Head, objects	✓	Naturalistic driving	Body pose
Abouelnaga et al. (2017)	AUC-DDD V1	2017	31 (22/9)	10	1	17308	RGB	Actions	✗	Parked vehicle	Distraction
Billah et al. (2018)	EBDD	2018	13	5	1	0.7h	RGB	Actions	✓	Naturalistic driving	Distraction
Borghi et al. (2018)	Turms	2018	7 (5/2)	–	1	14k	IR	Hands	✗	Naturalistic driving	Hand detection
Roth and Gavrila (2019)	DD-Pose	2019	27 (21/6)	–	2	660k	RGB, Depth, IR	Head, object	✓	Naturalistic driving	Body pose
Eraqi et al. (2019)	AUC-DDD V2	2019	44 (29/15)	10	1	14478	RGB	Actions	✗	Parked vehicle	Distraction
Martin et al. (2019)	Drive&Act	2019	15 (11/4)	6	6	12h	RGB, Depth, IR	Head, body, actions, objects	✗	Naturalistic driving	Distraction, autonomous driving
Jegham et al. (2019)	MDAD	2019	50 (38/12)	16	2	6h	RGB, Depth, IR	Actions	✗	Naturalistic driving	Distraction, hand detection, face detection
Ortega et al. (2020)	DMD	2020	37 (27/10)	13	3	41h	RGB, Depth, IR	Head, body, actions, objects	✓	Naturalistic driving	Distraction, drowsiness
Jegham et al. (2020a)	3MDAD	2020	69 (49/20)	16	2	574k	RGB, Depth, IR	Actions, head, body	✓	Naturalistic driving	Distraction, Hand detection, face detection

Legend: Occ: Occlusions.
[a]Number of subjects (Male/Female), - represents the data not available since the data are scraped from multiple sources.
[b]Simultaneous views of scene.
[c]Number of images available (in whole number) or number of hours in recorded video (in h).

torque, fuel consumption, engine RPM, steering wheel angle, seat belt usage, headlight status, windshield wiper status, and the indicator used are recorded in the device. These devices collect data from the vehicle's engine control unit (ECU) via the vehicle controller area network (CAN) bus. OBD data are much more reliable than GPS data since they can log a steady stream of data. One could also collect vehicle data from the OBD sensor to determine the driver's eco-driving behaviour.

Researchers have used smartphones to collect accurate driving data. All smartphones are equipped with accelerometers, magnetometers, gyroscopes, barometers, microphones, GPS, etc. Also, studies have shown that data collected via smartphone sensors are highly comparable with popular devices used for data collection. This has made more research in the area more accessible without having expensive sensors and instruments. Besides, smartphones are usually equipped with a camera and can collect images and videos inside or outside the vehicle. Some of the famous visual datasets, such as AUC-DDD (Abouelnaga et al., 2017; Eraqi et al., 2019) are collected with a smartphone.

Exterior cameras, such as dashcams and smartphones mounted on the windscreen, could be a potential data source to determine distraction. Data such as distance to the front vehicle could be used to infer a safe following distance. This can be used to analyse a driver's driving behaviour and aggressiveness. Besides, the data could perform on-road detection, such as lane detection and signboard detection, which could aid in self-driving vehicles.

Table 7
List of public driver monitoring datasets using external sensor.

Ref	Dataset	Year	Sensors	Raw data	Subjects[a]	Size	Behaviour	Scenario	Usage
Romera et al. (2016)	UAH-DriveSet	2016	Inertial sensors, GPS, Camera	GPS–Timestamp, Speed, Latitude coordinate, Longitude coordinate, Altitude, Vertical accuracy, Horizontal accuracy, Course, Difcourse: course variation; Inertial sensors–Timestamp, Boolean of system activated (1 if > 50 km/h), Acceleration in X, Acceleration in Y, Acceleration in X filtered by KF, Acceleration in Y filtered by KF, Acceleration in Z filtered by KF, Roll, Pitch, Yaw; Camera–Timestamp, Distance to ahead vehicle in current lane, Time of impact to ahead vehicle, Number of detected vehicles in this frame	6 (5/1)	500 min	Normal, drowsy, aggressive driving	Naturalistic	Driver Behaviour
figshare (2019)	Reading while driving	2019	Smartphone	PS, accelerometer, gyroscope and magnetometer sensors	18	11.13 MB	Reading while driving	Naturalistic	Distraction
Torres et al. (2019)	Texting while driving	2019	Smartphone	Mean and standard deviation for time between touches, speed, acceleration and gyroscope	13 (8/5)	823 set	Texting while driving	Naturalistic	Distraction

[a]Number of subjects (Male/Female),–represents the data not available since the data are scraped from multiple sources.

Table 8
Evaluation of various sensors.

Sensors	Precision	Robustness	Timeliness	Unobtrusiveness	Feasibility
Physiological	⊙	○	⊙	○	○
Camera	⊙	⊙	⊙	●	●
Eye Tracker	⊙	○	⊙	●	○
Smartphone	⊙	⊙	●	⊙	–
Exterior Camera	⊙	⊙	⊙	●	●
Vehicle IMU	⊙	●	⊙	●	●
GPS	●	⊙	●	●	⊙
Steering Wheel	⊙	●	⊙	●	●

Legend: ○ Poor ⊙ Moderate ● Good — Not Applicable.

A list of the publicly available datasets for the external sensor is summarized in Table 7.

3.4. Sensor evaluation

Among all the sensors commonly used in detecting driver distraction, there are several indicators to evaluate if the sensor is effective. The main indicators used are precision, robustness, timeliness, unobtrusiveness, and feasibility. We provide the analysis of each sensor discussed above in Table 8. We follow the definition by NHTSA (Lee et al., 2013) as follows:

- Precision indicates the degree to which the sensor identifies and differentiates distraction.
- Robustness indicates the degree to which the sensor provides reliable data.
- Timeliness indicates the degree to which the sensor supports real-time estimates of distraction.
- Unobtrusiveness indicates the degree to which the sensor does not interfere with driving.
- Feasibility indicates the degree to which the sensor could be included in a production vehicle.

3.5. Multimodal dataset

There exist several datasets which leverage multiple types of sensors. We only consider the dataset to be multimodal if they use a mix of multiple types of sensors. For example, RGB, IR, and depth images are not considered multimodal since they came from the same sensor group. The list of publicly available datasets is shown in Table 9.

4. Machine learning background

ML allows computers to solve a specific task and make predictions based on previous observations. ML has come a long way, and many algorithms have been proposed. Here, we only include the ML algorithms commonly used in detecting driver distraction.

4.1. Machine learning algorithms

The ML algorithms used in detecting and classifying distracted drivers can be categorized into traditional ML and DL. Note that we did not classify the type of algorithm used based on the depth of the model since there are confusing since different scholars have

Table 9
List of public multimodal driver monitoring datasets.

Ref	Dataset	Year	Features	Subjects[a]	Views[b]	Size[c]	Streams	Occlusions	Scenario	Usage
Barnard et al. (2016)	UDRIVE	2016	Video streams, CAN, GPS, audio	–	5–8	87,871 h	–	✓	Naturalistic driving	Driver behaviour
Massoz et al. (2016)	DROZY	2016	Karolinska Sleepiness Scale (KSS), psychomotor vigilance test (PVT), EEG, EOG, ECG, EMG, NIR images	14 (3/11)	1	7 h	IR	✗	Simulated settings	Drowsiness
Taamneh et al. (2017)	OSF	2017	Drives key response variables (speed, acceleration, brake force, steering angle, and lane position), palm EDA, heart rate, breathing rate, facial expression signals, biographical and psychometric covariates, eye tracking data	68	1	1.7 TB	RGB, Thermal	✗	Simulated settings	Distraction
Fridman et al. (2019)	MIT-AVT	2019	In-cabin camera, CAN, IMU (Acceleration, Gyroscope), GPS, Audio	122	1	511,638 miles	RGB	✓	Naturalistic driving	Driving behaviour
Tavakoli et al. (2021b)	HAR-MONY	2019	Camera inside and outside cabin, GPS, Smartwatch (heart rate, hand acceleration, audio amplitude, light intensity, location, gravity, Compass, Altitude, Magnetometer, gyroscope), weather data, speed limit data, music log	21	2	1 month data	RGB	✓	Naturalistic driving	Driving behaviour

[a]Number of subjects (Male/Female), - represents the data not available since the data are scraped from multiple sources.
[b]Simultaneous views of scene.
[c]Number of images available (in whole number) or number of hours in recorded video (in h).

definitions of "shallow" and "deep" models. However, we consider all non-DL algorithms as traditional ML algorithms. There are numerous algorithms available now; thus, we only introduce the most commonly used algorithms in this subsection.

The main difference between traditional ML and DL is how features are extracted. Traditional ML uses handcrafted features by applying several feature extraction algorithms and then applying the learning algorithms. In DL, the features are learned automatically and are represented hierarchically on multiple levels.

4.1.1. Traditional machine learning techniques

The traditional ML algorithms include simple classifiers, such as linear discriminant analysis (LDA), support vector machine (SVM), naïve Bayes (NB), k-nearest neighbour (kNN), decision tree (DT), and random forest (RF). We identify the used algorithms in all considered articles and briefly introduced them in this section. Usually, these techniques are applied to numerical data, such as data collected from physiological sensors. Note that there are articles that utilize traditional ML for image data, too (usual papers published earlier than 2014).

LDA. LDA is a discriminant analysis algorithm, one of the simplest yet effective classifiers in real-world applications. LDA is commonly used as a dimensionality reduction technique. LDA transforms data linearly to a hyperplane to maximize the distance among each class of data. It calculates the transformation matrix by solving and locating the maximum eigenvalue from the between-class, and with-in-class metrics (Balakrishnama and Ganapathiraju, 1998).

SVM. SVM is a supervised ML algorithm, which works almost similarly to LDA, except SVM only considers the data points near the classification boundaries. The objective of SVM is to obtain a hyperplane in an N-dimensional space (where N is the number of features) that can distinctly classify the data points using the kernel trick. A kernel function is used to improve its classification accuracy through operations performed in the input space rather than the potentially high dimensional feature space (Jakkula, 2006).

NB. NB is a probabilistic-based classifier for multidimensional data classification. It applies Bayes' theorem with strong (naïve) independence assumptions between the features. The NB classifier assumes that every input feature is independent of each other and that the contributions among every feature are equal.

KNN. KNN is a non-parametric algorithm, which means it does not make any assumptions on underlying data. KNN is calculated based on distance estimation and majority vote. KNN first calculates the distance between known and unknown data, and then ranks the distances from low to high. The highest vote determines the results of the unknown data class in the first kth closest distance from the known data. Commonly used distance functions include Euclidean and Manhattan distance (Cunningham and Delany, 2020a). KNN stands out as it does not require any training period. Instead, it calculates the distance between testing and training data directly. We could consider KNN as a local optimization algorithm.

DT. DT has a flowchart-like structure with three parts: the root node, branch, and leaf node. DT is constructed from nodes representing circles, and the branches are represented by the segments connecting the nodes. DT starts from the root, moves downward, and generally is drawn from left to right (Ali et al., 2012). The node from where the tree starts is called a root node. The node where the chain ends is known as the leaf node. A node represents a specific characteristic, while the branches represent a range of values.

RF. Random Forest (Breiman, 2001) is a group of unpruned classification or regression trees made from the random selection of samples of the training data. Features are selected randomly in the induction process while prediction is made by aggregating the predictions of the ensemble.

4.1.2. Deep learning

For the driver distraction detection field, most of the work is done in the form of deep supervised learning, in which the models are given with a set of labelled data. Popular deep supervised learning algorithms include convolutional neural network (CNN), recurrent neural network (RNN), long–short term memory (LSTM), and gated recurrent unit (GRU).

CNN. Fukushima first proposed CNN in 1988 (Fukushima, 1988). However, due to computational resources limitation, it was not widely adopted. Until the 1990s, LeCun et al. applied a gradient-based learning algorithm to CNNs and showed promising results in handwritten digit classification (LeCun et al., 1998). Today, CNN is widely adopted in all sorts of tasks, including but not limited to image classification, speech recognition, video classification, and action recognition. CNNs are more like a human visual processing system and are highly optimized for processing images. Thus, it is effective at learning and extracting abstractions of 2D features. Besides, CNNs are trained with a gradient-based learning algorithm, restricting the model to encounter varnishing gradient problems (Alom et al., 2019).

A typical CNN is made up of two parts, which are the feature extractor and the classifier. Each layer in the feature extractor section obtains its input from the previous layer output and passes it to the next layer as input. The features are extracted through a series of convolutional and pooling layers. The higher-level features are generated from lower-level features propagated through the network. Therefore, the number of feature maps usually increases before entering the classification layer. In the classification layer, the extracted features are treated as inputs to the dimension of the weight matrix of the final neural network. The score of the respective class is then calculated in the last classification layer through a softmax layer. The highest score class is then treated as the prediction given by the network.

The main component of a CNN model is a convolutional and pooling layer. Kernels (also known as filters) are applied to the original image or the intermediate feature maps in a deep CNN in the convolutional layer. The output of the kernels will go through an activation function to form the output feature maps. Pooling layers reduce the spatial size of the representation by decreasing the required amount of computation and weights through dimensionality reduction. For example, max-pooling takes the maximum value in a specific filter region, while average pooling takes the average value in a filter region. In the classification section, the output of the previous layer (final layer of the feature map) is "flattened" into a single vector to be an input for the next stage. Then, the fully connected layer computes the score of each class from the extracted features. The final probabilities for each label are available at the fully connected output layer.

Several popular state-of-the-art CNN architectures includes LeNet (LeCun et al., 1998), AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), Inception (Szegedy et al., 2015, 2016, 2017), DenseNet (Huang et al., 2017) and EfficientNet (Tan and Le, 2019, 2021).

RNN. RNN is a neural network with time-varying behaviour, including the notion of dynamic change over time (Lipton et al., 2015). It supports sequential data processing through its looping mechanism, enabling the information to flow from one step to the next. RNN consists of three layers: the input layer, the hidden layer, and the output layer. The input layer takes in a sequential input, loops through the input values, and produces the output, while the hidden layer retains the memory from previous iterations for prediction. RNNs are helpful when time-dependent data are used, where the data must be handled in a sequential manner or processes change over time. However, RNN suffers from a vanishing gradient problem.

LSTM. LSTM was proposed as a solution to the vanishing gradient problem in RNN (Hochreiter and Schmidhuber, 1997). LSTM unit comprises different memory blocks named cells and three "regulators" named gates (input gate, output gate, and forget gate). The cell remembers values over arbitrary time intervals and decides which data to store and when to read, write and erase. The gates are responsible for regulating the flow of information inside the LSTM unit and outside the cell. The input gate controls the addition of information to the cell state. The output gate selects valuable information from the current cell state. The forget gate removes information from the cell state that is no longer required.

GRUs. GRUs are lighter versions of standard LSTM in terms of topology, computation cost, and complexity (Cho et al., 2014). GRU omits the output gate and has fewer parameters than LSTM. Thus, GRU writes the contents from its memory cell to the larger net at each time step. GRU consists of only one hidden state to hold both long-term and short-term dependencies at the same time. GRU has only two gates, where the update gate controls the information that flows into memory, and the reset gate controls the information that flows out of memory. These gates are trained to filter out irrelevant information. To propose a new hidden state, the reset gate decides which portions of the previous hidden state will combine with the current input, while the update gate determines how much the previous hidden state to retain and what portion of the new proposed hidden state derives from the reset gate to include in the final hidden state.

4.2. Transfer learning

Transfer learning (TL) is an ML technique where a model is trained and developed for one task and is then reused on another related task. It refers to the situation where what has been learned in one setting is exploited to improve optimization in another setting (Gao and Mosalam, 2018). TL tries to transfer the knowledge from the source domain to the target domain by relaxing the assumption that the training data and the test data must be independent and identically distributed (Tan et al., 2018). TL is commonly used when the new dataset is smaller than the original dataset used to train the model.

There are four techniques in TL as pointed in Tan et al. (2018), which are instance-based, mapping-based, network-based, and adversarial-based. In the field of image classification, network-based TL is commonly used. Network-based TL reuses the partial network that is pretrained in the source domain, including its network structure and connection parameters, and transfers it to a part of the deep neural network used in the target domain. The commonly used large dataset to train the state-of-the-art CNN model is ImageNet (Deng et al., 2009). It is found that when models trained on ImageNet are used as fixed feature extractors or fine-tuned for another downstream task; there is a strong correlation between ImageNet accuracy and transfer accuracy (Kornblith et al., 2019).

Usually, we strip off the classification layer of the pretrained models since it has 1000 classes (ImageNet has 1000 classes) and replaced it with a randomly initialized classification layer. Then, the newly added layer is trained with the target dataset while keeping the weights of the pretrained network frozen. One way to further improve TL performance is to "fine-tune" the weights of the top layers of the pretrained model with the added classifier. The training process will force the weights to be tuned from generic feature maps to features explicitly associated with the dataset. Fine-tuning is usually adopted in driver distraction recognition or classification tasks.

5. Distraction detection method

We review articles based on the input sensors used to collect the input data, as shown in Fig. 6. We believe that this provides a more straightforward overview of detecting distraction based on the used sensors.

5.1. Detection based on physiological data

The physiological measurements provide important clues about the driver's state, such as stress levels and drowsiness. The commonly used sensors in EEG, since many features can be collected through EEG. Wang et al. (2015a) utilized SVM with radial basis function (RBF) kernel to distinguish cognitive distraction from the focus of drivers in a dual-task experiment of lane-keeping and solving math problems using EEG sensor. They achieved 84.6% and 86.2% classification accuracy in detecting distraction versus math solving and driving, respectively.

Similar studies were carried out with different ML algorithms and features to classify distracted driver (Alizadeh and Dehzangi, 2016; Murugan et al., 2020; Schneiders et al., 2020). Recently, Li et al. (2021a) proposed a temporal–spatial information network, a combination of CNN and GRU to extract both spatial and temporal features from EEG signals. Specifically, CNN was used to recognize the spatial patterns in the constructed EEG "image", while GRU was used to recognize the temporal patterns in EEG sequences. They treat the collected EEG data as a video sequence input to the network. Their proposed model can achieve binary classification (distraction versus non-distraction) accuracy of 92% and task-specific distraction detection accuracy of 88%.

Besides EEG, the next commonly used sensor is ECG. Lee et al. (2015) measured ECG signals via electrodes that were placed on the steering wheel. Additionally, they derived the respiratory rate and heart rate variability from the ECG signals and used PPG measured from the driver's finger. They used if-then rules and applied kernel fuzzy C-mean (KFCM) to detect driving distractions. Similar works are done by using different ML approaches, including LDA (Vicente et al., 2016), ANN (Tjolleng et al., 2017), and KNN (Deshmukh and Dehzangi, 2019).

GSR data could be collected through a non-intrusive manner. This is proposed by Rajendra and Dehzangi (2017), where GSR data is collected through a wrist band wearable. They used SVM to classify several actions and demonstrated high accuracy under subject dependents scenarios. Similarly, EDA sensor data can be collected with wearable devices too. Dehzangi and Rajendra (2019) used the time and frequency domain of the EDA signal to extract features to capture the patterns of the distracted driver at the physiological level. They then employed linear and kernel-based SVM and ten-fold cross-validation to generate identification results. Their base accuracy with the SVM RBF kernel is 85.19%, while their weighting mechanism can achieve 88.91%.

While it is possible to only use a single type of sensors to infer distraction, some other works proposed to use multiple physiological sensor data to further boost the accuracy of the proposed methodology. Sonnleitner et al. (2014) used EEG and EMG as input sensors on 20 subjects. They recorded the alpha spindles rate, frequency, amplitude, and duration through EEG and were used to classify distracted drivers. EMG, on the other hand, was used to determine the brake reaction time. Regularized LDA with shrinkage of the covariance matrix was used as the classifier. This article mainly focuses on auditory distraction detection, and with the use of EEG and LDA classifiers, they achieved a classification error of 8%.

Sahayadhas et al. (2015) detect inattention using ECG and EMG signals. The inattention features were extracted from the preprocessed signals using conventional statistical, higher-order statistical, and higher-order spectral features. The features were classified using KNN, LDA, and quadratic discriminant analysis (QDA). They concluded that ECG and EMG signals could be explored further to develop a robust and reliable inattention detection system.

Huang et al. (2022) conducted a simulated driving experiment by collecting 15 drivers' biological signals, including EEG (2 channels), ECG, EDA, RSP, and HR. They used four methods in classifying driver's mental workload: feature extraction with XGBoost, CNN, ConvLSTM, and a combination of CNN and LSTM. The network model that combined CNN and LSTM has the best performance in recognition with the highest accuracy of 97.8% when using 3-second samples, and higher than that of the CNN model in all cases.

We include a summary of all articles discussed above in Table 10.

Discussion. Generally, all of the studies considered achieve relatively good accuracy with little loss. Specifically, it is observed that the usage of LSTM and neural network did not have significant improve over traditional ML methods. From the articles reviewed, most of the recent work still favour traditional ML methods over DL methods. As observed in some of the literature which uses RNN/LSTM model, they seem to predict continuous sequential data better than traditional ML methods. However, if the signal is further converted into 2D data and feed into a CNN network, the performance is on-par with RNN/LSTM model. In terms of feasibility and speed, traditional ML still strikes the balance between speed and accuracy.

It is observed that all of the works discussed above only involved a small group of participants. Some of the works only involve as few as 6 participants, which is far below the minimum threshold to support the proposed model. Besides, the metrics used to score the proposed method is non-uniform. Specifically, most of the articles only include the accuracy of the proposed system without including its sensitivity. For example, the usage of ECG with LDA can achieved high accuracy (96%), but low sensitivity (59%) (Vicente et al., 2016). High accuracy with low sensitivity system should be avoided since they might produced many false positive predictions. Therefore, we suggest that future work that involves the usage of physiological data just use traditional ML methods to perform predictions, since they are simple and efficient.

Furthermore, the raw data collected from physiological sensors usually undergo preprocessing to obtain the features before passing into the ML model. The obtained features usually suffer from large variation since the raw data is susceptible to noise. Therefore, most of the works discussed above will first remove the anomaly in the obtained features, which introduces another issue — information loss, since the anomaly in data might carry important information towards the prediction. Even though physiological data is enough to infer if a driver is distracted, it is impractical in the real-world driving scenario.

5.2. Detection based on visual data

There are multiple approaches and arguments to detect if a driver is distracted. We subdivide them into driver monitoring methods and action classification. In the driver monitoring method, driver distractions are determined through the driver's body parts, such as eye gaze, head pose, hand position, and body pose. For example, if the eyelids are closed, we could then induce the driver is experiencing fatigue and, therefore, be distracted. The same goes for detecting the hand position, where if the driver's both hands are off the steering wheel, we can infer the driver is carrying out activities unrelated to the driving task. As for action classification techniques, a set of predefined distracting actions, such as reaching behind, drinking, and eating, are determined. From these defined actions, multiple images and/or videos are collected. The collected images or videos are then fed into the ML models to classify the distraction actions. Usually, action detection methods are carried out with publicly available datasets, such as StateFarm dataset (StateFarm, 2016), and AUC-DDD dataset (Abouelnaga et al., 2017; Eraqi et al., 2019).

5.2.1. Driver monitoring method

In driver monitoring method, most earlier work only identify the usage of cellphone while driving as distractions. The image is collected using a frontal camera and preprocessed by cropping the region of interest, followed by isolating the driver skin pixels to locate the hand near the face region (Berri et al., 2014). Other works use different face parts (head, eye, mouth, lips and eyebrows) to infer distraction (Azman et al., 2014; Chuang et al., 2014; Seshadri et al., 2015; Fridman et al., 2016; Xiao and Feng, 2016; Huang and Zhang, 2018; Azim et al., 2014; Choi et al., 2016; Ali and Hassan, 2018). The usage of head pose estimation is also used in some works to boost the prediction accuracy (Chuang et al., 2014; Vicente et al., 2015; Braunagel et al., 2015; Billah and Rahman, 2016).

Ali and Hassan (2018) proposed to collect features based on facial points, especially the features computed using motion vectors and interpolation to detect the type of driver distraction, which is the driver's head rotation due to a change in yaw angle. These facial points are detected by the Active Shape Model (ASM) and Boosted Regression with Markov Networks (BoRMaN). They trained on various classifiers and showed that the approach that uses the motion vectors and interpolation outperforms other approaches in detecting driver's

Table 10
Summary of studies utilizing physiological data as input.

Ref	Year	Input sensor	Collected features	Tasks	Subject (M/F)	Distraction type	ML algorithm	Effectiveness[a]
Sonnleitner et al. (2014)	2014	EEG, EMG	EEG: Alpha spindles rate, frequency, amplitude, duration; EMG: break reaction time	Distraction classification, break reaction time	20 (15/5)	Auditory	LDA	CE: 8%
Sahayadhas et al. (2015)	2015	ECG, EMG	SD, mean, median, energy, skewness, Kurtosis, sum of the logarithmic amplitudes of the bispectrum (H1, H2, H3)	Driver distraction	15 (15/0)	Visual, cognitive	KNN, LDA, QDA	Acc: 90.97–98.12%
Wang et al. (2015a)	2015	EEG	Spectral magnitudes of alpha, low beta, delta, an theta	Cognitive distraction detection	10 (10/0)	Cognitive	SVM	Acc: 84.6–86.2%
Lee et al. (2015)	2015	EEG, PPG	EEG: Power spectrum density, centre of gravity frequency, and frequency variability of Heart Rate Variability, Peak, Median, and Mean Frequency of respiratory rate variability (RRV), Respiratory Irregularity of RRV; PPG: Power spectrum density, centre of gravity frequency, and frequency variability of Pulse Rate Variability	Driver vigilance detection	12 (8/4)	Cognitive	KFCM	Acc: 97.28%
Alizadeh and Dehzangi (2016)	2016	EEG	Short Time Fourier Transform, band power, discrete wavelet transform, fractal dimensions, auto-regressive, entropy	Distraction classification	6	Cognitive, visual, auditory, manual	DT, RF, KNN, NB	Acc: 98.99%
Vicente et al. (2016)	2016	ECG	HRF, Respiratory frequency	Drowsiness detection	30	Cognitive	LDA	Acc: 0.96
Tjolleng et al. (2017)	2016	ECG	Mean inter-beat interval (IBI), SD of IBIs, root mean squared difference of adjacent IBIs, power in low frequency, power in high frequency, ratio of power in low and high frequencies	Distraction classification	15 (15/0)	Cognitive	ANN	Acc: 82 %
Rajendra and Dehzangi (2017)	2017	GSR	Mean, variance, accGSR [i], avgGSR [i], maximum Value, and power [i]	Distraction detection	6	Manual, cognitive, visual, auditory	SVM	Acc: 89.45–91.33%
Deshmukh and Dehzangi (2019)	2019	ECG	Average heart rate, mean R-R, NN50, pNN50, SD of HR and R-R interval, RMSSD, sample entropy, power spectral entropy	Distraction classification	6	Cognitive, visual, auditory, manual	DT, RF, KNN, SVM, NB	Acc: 65.60–83.04%
Dehzangi and Rajendra (2019)	2019	EDA	Mean, variance, accumulated GSR, average GSR, maximum value of GSR, Short term Fourier transforms, Fractal dimensions, auto regressive coefficients	Distraction detection	7 (7/0)	Visual, manual, cognitive	LDA, SVM	Acc: 88.91%
Murugan et al. (2020)	2020	ECG	Mean, median, SD, Q1, Q2, Q3, IQR, maximum, minimum, variance, skewness, kurtosis, RMS, power, energy, Approximate Entropy, Central Tendency Measure (Nanmean, Trimmean, Harmonic Mean), Hurst exponent	Normal vs. drowsy, visual inattention, fatigue, cognitive inattention	10 (9/1)	Visual, cognitive	SVM, KNN, Ensemble	Acc: 58.3%
Schneiders et al. (2020)	2020	EEG	Higuchi Fractal Dimension, Petrosian Fractal Dimension, Band Power Ratio, and Discrete Wavelet Transform	Distraction classification	8 (5/3)	Cognitive	RF	Acc: 76.77–97.99%

(*continued on next page*)

head rotation. They were able to achieve an accuracy of 98.45% using NN.

All of the works above use traditional ML methods, and did not achieve high accuracy since they are incapable of capturing important features. Therefore, the usage of CNN and DL is used in recent works. Choi et al. (2016) proposed a DL approach in real-time distraction detection based on eye gaze. The driver's face is recorded, and a Haar feature-based face detector is combined with a correlation filter-based, minimizing the output sum of the square error tracker for face tracking. Then AlexNet model is used to classify the nine gaze zones. Moreover, DL techniques can also estimate the driver's head pose, valuable for distraction detection.

Hoang Ngan Le et al. (2016) identified two actions, which are cell phone usage and hands on the wheel, to infer if the driver is distracted. They proposed MultiScale Faster R-CNN (MS-FRCNN). Their approach achieves higher accuracy with less processing time compared to the FRCNN (Ren et al., 2015).

Vora et al. (2017) used CNNs to classify driver's gaze into seven zones. Since drivers' gaze direction has been previously shown as an important clue in understanding distraction, they collected their dataset and annotated the gaze zone. They pre-process the image using a face

Table 10 (*continued*).

Ref	Year	Input sensor	Collected features	Tasks	Subject (M/F)	Distraction type	ML algorithm	Effectiveness[a]
Li et al. (2021a)	2021	EEG	Spatial and temporal EEG data	Distraction detection	24 (24/0)	Manual, visual, auditory, cognitive	CNN, GRU	Acc: 88%–92%
Huang et al. (2022)	2021	EEG, ECG, GSR, RESP, RR	EEG: Mean, SD, Wavelet Mean and SD; ECG: Mean, SD, Wavelet Mean and SD; GDR: Mean, SD, Wavelet Mean and SD; RR: Mean, SD, Wavelet Mean and SD, LF and HF power sum, ratio of LF/HF, LF/AF, HF/AF; RESP: Mean, SD, Wavelet Mean and SD, Band power	Driver state classification	18 (18/0)	Cognitive	XGBoost, CNN, ConvLSTM, CNN+LSTM	0.7440–0.9780

CE: Classification Error; Acc: Accuracy
*–represent data not stated in the article.
SD: standard deviation; Q1: First Quartile; Q2: Second quartile; Q3: Third quartile; IQR: Interquartile Range.
[a]Due to different metrics used across different study, we named this column as "effectiveness".

detector and pass the detected face into VGG16, AlexNet, and RF. They found that passing half of the face to the classifier yielded higher accuracy than passing the whole face.

Instead of still RGB image, Kinect camera is used is several studies to collect RGB and depth images. The earliest study by Martin et al. (2014a) presented a vision-based analysis framework to detect in-vehicle activities using two Kinect cameras pointed towards the hand and head of the driver. They use SVM trained on RGB descriptors to detect head and hand features and achieve an accuracy of 91%. The same procedure is adopted by Ohn-Bar et al. (2014) with the addition of eye cues to achieve higher accuracy of 94%. However, these work did not utilize depth information to the fullest. Craye and Karray (2015) improved previous method by extracting features from face and body using both RGB and depth images. Their system comprises four sub-modules: eye behaviour (gaze and blinking), arm position, head orientation, and facial expressions. The information from these modules is then merged using the Adaboost classifier and Hidden Markov Model (HMM). Five distractive tasks were recorded and manually labelled for training and evaluation. They found that HMM outperforms Adaboost for both distraction recognition and classification.

Other works explored the usage of semi-supervised learning method, such as Liu et al. (2015) used two graph-based semi-supervised learning methods and compared them with supervised learning methods. Laplacian SVM (LapSVM), which is an extension of SVMs to SSL under manifold regularization framework (Belkin et al., 2006), and SS-ELM (Huang et al., 2014) were compared with three supervised learning methods (static BN with Supervised Clustering (SBN-SC) (Liang et al., 2007; Liang and Lee, 2014), Extreme Learning Machine (ELM) and SVM) and one low-density-separation-based method (Transductive SVM (TSVM) Joachims et al., 1999). They collected on-road experiment data to capture the realistic eye and head movement patterns. They concluded that using unlabelled data, the graph-based semi-supervised methods reduced the labelling cost and improved the detection accuracy. SS-ELM achieved the highest accuracy of 97.2%. The benefits of using semi-supervised learning methods increased with the size of the unlabelled data set, showing that by exploring the data structure without actually labelling them, extra information to improve models' performance can be obtained.

Beside still images, continuous stream of video of distracted driver is used too. Chui et al. (2019) proposed a distraction detection module that processes video stream and compute motion coefficient to reinforce identification of distraction conditions of drivers. The motion-coefficient of the video frames is computed follows by the spike detection via statistical filtering. The authors collected a dataset consisting of nine actions from 65 volunteers. The accuracy of the head motion analyzer on the dataset is 98.6%.

Deo and Trivedi (2019) proposed an LSTM model for continuous estimation of the driver's takeover readiness index (defined by subjective ratings of human observers viewing the feed from in-vehicle vision sensors), using features from gaze, hand, pose, and foot activity to represent the driver's states. Their best results achieve a mean absolute error (MAE) of 0.449 on a 5 point scale of the takeover readiness index.

We summarized all of the articles in Table 11. The features used in all articles are summarized in Table 12.

Discussion. The usage of ML in driver monitoring presents good performance. Most of the works focus on facial features (head, eye, nose and mouth) and hands to infer if a driver is being distracted. It was done in a two-stage manner, where the first stage is to extract features, as shown in Table 12, and a classification head is added to classify based on the extracted features. The feature extraction can be done using mathematical approaches, traditional ML approaches, or even using neural networks. Classification head can be as simple as using a fully connected layers or traditional ML algorithms.

Besides, the collected dataset is not varied, whereby most of the work captures their data in a vehicle simulator or during daytime. Therefore, the obtained accuracy could not perfectly describe real-world scenario accuracy. Further, most of the work fine-tuned the model to only perform detection on a certain region to infer if a driver is being distracted. This is not ideal if the driver moves out of the region, then the returned result is no longer accurate. Moreover, it is not ideal to model human behaviour based on a limited number of participants in the dataset. For example, the blink duration for everyone is different, and if the training dataset is not highly varied, the model might misclassify normal driving as distracted driving.

5.2.2. Action classification

Besides detecting distraction from body parts, which are highly driver-dependent yet limited to specific distractions, a more common way of recognizing and detecting distraction is through action classification. In this area, researchers tend to collect images of various modalities (RGB, IR, depth, or thermal images) from the various viewpoints (frontal or side view) with a specific set of actions (such as drinking, smoking, and reaching behind) deemed to be causing distractions. We further break it down into two sections based on the availability of the dataset. Private datasets are usually collected privately without the intention to share them publicly. Most of the work treated this task as an image classification task, with various DL models, such as ResNet and AlexNet.

Private dataset. Most of the works collect image data using single camera pointed towards the driver's face and then label the image into specific number of actions (Koesdwiady et al., 2017; Tran et al., 2018; Xing et al., 2019; Kapoor et al., 2020; Wang et al., 2021; Hu

Table 11
Summary of articles that detect driver distraction through driver monitoring method.

Ref	Year	Streams	Body parts	Goal detail	ML Algo	Accuracy
Berri et al. (2014)	2014	RGB	Head, hand	Cell phone usage	SVM	0.8906–0.9157
Azman et al. (2014)	2014	RGB	Lips, Eyebrow	Distraction detection	LR	0.7585
					SBN	0.7757
					SVM	0.7958
					DBN	0.9362
Martin et al. (2014a)	2014	RGB	Head, hand	Hand on wheel	SVM	0.91
Ohn-Bar et al. (2014)	2014	RGB	Head, hand, eye	Hand on wheel	SVM	0.94
Azim et al. (2014)	2014	NIR	Face, eye	Fatigue detection	FL	1
Chuang et al. (2014)	2014	RGB	Eye, mouth, nose	Gaze position	SVM	0.8640–0.9740
Craye and Karray (2015)	2015	RGB, Depth	Head, Eye, Arm	Gaze estimation, Arm position, Head position	Adaboost	0.8984
					HMM	0.8964
Vicente et al. (2015)	2015	RGB, IR	Head, eye	Gaze estimation	SDM	0.9449–0.9849
Seshadri et al. (2015)	2015	Grayscale	Face (SHRP-2 The National Academies of Sciences, Engineering, and Medicine, 2021)	Cell phone usage	AdaBoost	0.8440–0.9390
					SVM	0.7870–0.8420
					RF	0.8010–0.9270
Liu et al. (2015)	2015	RGB	Eye, head	Distraction detection	SBN-SC	0.8380
					SVM	0.9500
					TSVM	0.9550
					ELM	0.9560
					SS-ELM	0.9720
					LapSVM	0.9730
Braunagel et al. (2015)	2015	Eye Tracker	Eye, Head	Distraction detection	SVM	0.77
Hoang Ngan Le et al. (2016)	2016	RGB, Grayscale	Face (SHRP-2 The National Academies of Sciences, Engineering, and Medicine, 2021), Hand (VIVA Das et al., 2015)	Hand on wheel, Cell phone usage	FRCNN	0.9240
					MS-FRCNN	0.9420
Xiao and Feng (2016)	2016	RGB	Face, eye	Distraction detection	SVM	0.9170
Billah and Rahman (2016)	2016	RGB	Hand, lips, forehead	Distraction detection	KNN	0.8150–0.8167
Choi et al. (2016)	2016	RGB	Face, Eye	Gaze location	AlexNet	0.95
Fridman et al. (2016)	2016	Grayscale	Face, eye	Gaze location	RF	0.4410–0.9250
Vora et al. (2017)	2017	RGB	Eye	Gaze location	RF	0.6876
					AlexNet	0.8891
					VGG16	0.9336
Huang and Zhang (2018)	2018	–	Lips	Distraction detection	FV+NN	0.9450–0.9850
Ali and Hassan (2018)	2018	RGB	Face	Distraction detection	SVM	0.8962
					NB	0.9002
					AdaBoost	0.9262
					DT	0.9278
					J48	0.9295
					Nnge	0.9298
					NN	0.9348
Chui et al. (2019)	2019	RGB	Head	Head position	SVM	Coefficient: 0.986
Deo and Trivedi (2019)	2019	Eye, head, body, foot	RGB, IR, Depth	Observable readiness index	SVM	MAE: 0.599
					LSTM	MAE: 0.467
					LSTM key-frame weighting	MAE: 0.449

et al., 2019; Li et al., 2020; Hu et al., 2020; Li et al., 2021b; Wu et al., 2021). Some of the work further included second camera to record the hand activity (Ohn-Bar et al., 2014) or Internet images to increase the number of images (Ou et al., 2018).

RGB camera is not robust under low-light condition, while IR camera can record the driver's face during these condition. Therefore, some of the works recorded the dataset using RGB camera during daytime and IR camera during night time (Yan et al., 2016; Ou et al., 2019)

Table 12
Summary of features collected from visual sensor through driver monitoring method.

Body part	Feature group	Feature
Head	Head Position	x, y, z coordinate (Liu et al., 2015; Braunagel et al., 2015; Liu et al., 2015)
	Head rotation	Yaw, Pitch, Roll (Liu et al., 2015; Vicente et al., 2015; Braunagel et al., 2015)
Face	Face	Face detection (Hoang Ngan Le et al., 2016; Seshadri et al., 2015; Martin et al., 2014a; Azim et al., 2014; Xiao and Feng, 2016; Choi et al., 2016; Fridman et al., 2016), Face tracking (Craye and Karray, 2015; Martin et al., 2014a; Azim et al., 2014; Choi et al., 2016), Face pose (Martin et al., 2014a), Ratio of facial components (Ali and Hassan, 2018), Face landmark (Fridman et al., 2016)
	Hand near face	Hand detection (Hoang Ngan Le et al., 2016; Seshadri et al., 2015; Berri et al., 2014)
	Mouth	Lips activity (Azman et al., 2014; Azim et al., 2014; Huang and Zhang, 2018), Mouth location (Chuang et al., 2014)
	Eyebrow	Eyebrow activity (Azman et al., 2014)
	Nose	Nose location (Chuang et al., 2014)
Eye	Gaze location	Left, Right; Yaw, Pitch (Liu et al., 2015; Xiao and Feng, 2016), Gaze Estimation (Vicente et al., 2015; Deo and Trivedi, 2019; Vora et al., 2017; Chuang et al., 2014; Choi et al., 2016), Iris Location (Craye and Karray, 2015; Xiao and Feng, 2016), Pupil Detection (Azim et al., 2014), Pupil Tracking (Azim et al., 2014)
	Gaze temporal	Saccade (Liu et al., 2015; Braunagel et al., 2015), Blink (Liu et al., 2015; Braunagel et al., 2015), Blink Frequency (Liu et al., 2015), Blink Duration (Liu et al., 2015), PERCLOS (Liu et al., 2015), Fixation (Braunagel et al., 2015)
	Eye blocking	Sunglass detection (Vicente et al., 2015)
Hand	Hand location	Hand on wheel (Hoang Ngan Le et al., 2016; Martin et al., 2014a; Deo and Trivedi, 2019), Object on hand (Deo and Trivedi, 2019)
Arm	Arm location	Arm position (Craye and Karray, 2015)
Body	Body pose	Body pose (Deo and Trivedi, 2019)
Foot	Foot location	Foot activity (Deo and Trivedi, 2019)

or just collected IR images only since they are capable under both conditions (Wagner et al., 2021). Beside IR image, depth information could be used to model the driver's head (Xing et al., 2017).

Other works used multiple streams of input to do detection. Yan et al. used motion history image (MHI) (Yan et al., 2014) to benefit from temporal information, and PHOG features were used for training the classifier. Ragab et al. (2014) utilized IR and Kinect cameras to extract five visual cues: arm position, eye closure, eye gaze, facial expressions, and head orientation.

Kavi et al. (2016) used CNN-LSTM architecture for multi-view distraction recognition. They collected a triple-view dataset and were passed it into individual CNN-LSTM. The individual CNN-LSTM score is then undergone feature and score fusion to determine the final predictions. The feature fusion techniques comprised the features collected by CNN-LSTM and an SVM to classify the feature, while the score fusion techniques used a softmax layer to fuse all the scores from individual CNN-LSTM. They found that score fusion achieved higher accuracy and multi-view classification and provided higher classification accuracy than single-view classification.

Zhang et al. (2020) proposed a dedicated Interwoven Deep CNN (InterCNN) architecture that utilized four-dimensional (time, height, width, RGB channels) multi-stream inputs, including side video streams, side optical flows, front video streams, and front optical flows to classify accurately of driver behaviours in real-time. They used a simulated environment to emulate self-driving car conditions and record the body movements, and facial expressions side and front-facing cameras were deployed. The hierarchical InterCNN has two components involving two simpler architectures: the plain CNN, which uses only the side video stream as input, and the two-stream CNN (TS-CNN), which takes the side video stream and the side optical flow as input.

In terms of ML algorithm used in these works, many of them treated the task as a classification problem. To boost the classification accuracy, some works turn into using ensemble of various models (Swetha et al., 2019). On the other hand, to reduce the complexity of the model, Kapoor et al. (2020) used lightweight models, such as MobileNet which has the balance between accuracy and speed. Li et al. (2021b) proposed a lightweight CNN with an octave-like convolution mixed block, called OLCMNet. While directly treating the problem as a classification task can achieve good result, Li et al. (2020) proposed to first detect the bounding boxes of the driver's right hand and right ear using YOLO, and then take the bounding boxes as the input to predict the distraction type using MLP. Similarly, we can split the problem into multiple subtasks, as proposed by Wu et al. (2021), where a multi-feature fusion network is proposed based on hand detection, pose estimation, and general classification for image-based distracted driving detection.

A summary of all works discussed above is tabulated in Table 13.

Public dataset. In April 2016, StateFarm's distracted driver detection competition on Kaggle defined ten postures to be detected (Safe driving and nine distracted behaviours) (StateFarm, 2016). Based on the State Farm dataset (StateFarm, 2016), Okon and Meng (2017) used AlexNet to classify the distraction action and found that using triplet loss has a slight improvement over generic log loss. Hssayeni et al. (2017) compared the performance of DL approaches versus a traditional classification algorithm such as SVM with a HoG/SIFT features. However, CNNs proved to be the most effective techniques achieving high accuracy. Jegham et al. (2018) first segmented driver actions using the SURF key-points, then they extracted HoG features candidate regions, and finally used KNN for classification. They finally achieved 70% recognition rate (This study is excluded in Table 14). Lu et al. (2020) used the improved deformable and dilated faster RCNN (DD-RCNN) structure to obtain an accuracy of 92.2%. Nel and Ngxande (2021) proposed to use 3D-ResNet because 3D kernels can extract spatio-temporal features from videos. They combined both AUC-DDD dataset and StateFarm dataset and achieve higher accuracy then predicting on individual dataset. Majdi et al. (2018) proposed Drive-Net, which is a combination of a CNN and a random decision forest. Other works explored the usage of 3D-CNN (Valeriano et al., 2018; Moslemi et al., 2019; Lemley et al., 2017; Nel and Ngxande, 2021) while some fine-tuned on existing CNN models, such as AlexNet (Okon and Meng, 2017; Kumar et al., 2021), MobileNet (Li et al., 2021b; Ugli et al., 2021), Inception (Li et al., 2021b; Alotaibi and Alotaibi, 2019), ShuffleNet (Li et al., 2021b), ResNet (Li et al., 2021b; Alotaibi and Alotaibi, 2019; Ugli et al., 2021), DenseNet (Li et al., 2021b), VGGNet (Hssayeni et al., 2017; Ugli et al., 2021; Masood et al., 2020; Gumaei et al., 2020) and HRNN (Alotaibi and Alotaibi, 2019). All of the studies are summarized in Table 14.

AUC-DDD (Abouelnaga et al., 2017; Eraqi et al., 2019) is one of the famous driver distraction datasets, which follows the same action as StateFarm's. Eraqi et al. (2019) trained five unique CNNs (VGG-16, ResNet50, InceptionV3) using the original image, face image, hand image, face with hand image, and skin segmented image. Finally, the results obtained from individual CNNs were weighted with the help of the genetic algorithm, and the action was classified. Baheti et al. (2018) applied a modified VGG-16 with various regularization

Table 13
Summary of articles that perform action classification through self-collected dataset.

Ref	Year	Dataset information							Results											
		Sub	Act	Views	Size	Streams	Occ	Scenario	ML Algo	PT	TE	BS	LR	WD	Mom	Opt	DO	BN	DA	Acc
Ohn-Bar et al. (2014)	2014	4 (3/1)	3	2	11,147	RGB	✗	Naturalistic	SVM	–	–	–	–	–	–	–	–	–	–	0.94
Yan et al. (2014)	2014	20 (10/10)	5	1	20 vid	RGB	✗	Naturalistic	(MHI+PHOG)+KNN	–	–	–	–	–	–	–	–	–	–	0.8801
									(MHI+PHOG)+MLP	–	–	–	–	–	–	–	–	–	–	0.9093
									(MHI+PHOG)+SVM	–	–	–	–	–	–	–	–	–	–	0.9443
									(MHI+PHOG)+RF	–	–	–	–	–	–	–	–	–	–	0.9656
Ragab et al. (2014)	2014	6	5	1	240 min	IR, Depth	✗	Simulated	CRF	–	–	–	–	–	–	–	–	–	–	0.6757
									NN	–	–	–	–	–	–	–	–	–	–	0.6838
									HMM	–	–	–	–	–	–	–	–	–	–	0.7950
									RF	–	–	–	–	–	–	–	–	–	–	0.8278
									AdaBoost	–	–	–	–	–	–	–	–	–	–	0.8297
									CRF (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.8255
									HMM (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.8815
									NN (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.8832
									AdaBoost (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.8890
									RF (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.9047
Kavi et al. (2016)	2016	3	7	3	90 min	RGB	✗	Simulated	CNN-LSTM	–	–	32	$1e^{-5}$	–	–	RMSProp	✓	✗	✗	> 0.90
Yan et al. (2016)	2016	20 (10/10)	4	1	29,410	IR	✗	Naturalistic	CNN	–	250	128	0.01	$5e^{-4}$	0.6	–	✓	✗	✓	0.9930
					17,730	RGB	✗	Naturalistic	CNN	–	250	128	0.01	$5e^{-4}$	0.6	–	✓	✗	✓	0.9577
Koesdwiady et al. (2017)	2017	10	6	1	–	RGB	✗	Naturalistic	XGBoost	–	–	–	–	–	–	–	–	–	✗	0.7504
									VGG19	✓	–	–	–	–	–	–	✓	✗	✗	0.8026
Xing et al. (2017)	2017	5	7	1	–	RGB, Depth	✗	Naturalistic	KNN	–	–	–	–	–	–	–	–	–	–	0.623
									RF	–	–	–	–	–	–	–	–	–	–	0.736
									SVM	–	–	–	–	–	–	–	–	–	–	0.747
									NB	–	–	–	–	–	–	–	–	–	–	0.767
									FFNN	–	–	–	–	–	–	–	–	–	–	0.824
Kim et al. (2017)	2017	6	2	1	4000	RGB	✗	Simulated	Inception-ResNetV2	✓	4000	–	–	–	–	RMSProp	–	–	–	0.9290
									MobileNetV1	✓	4000	–	–	–	–	Adam	–	–	–	0.9410
Tran et al. (2018)	2018	10	10	1	35,000	RGB	✗	Simulated	VGG16	✓	40	–	$5e^{-5}$	$5e^{-5}$	0.9	SGD	✓	✓	✓	0.79
									AlexNet	✓	40	–	$9e^{-4}$	$1e^{-5}$	0.9	SGD	✓	✓	✓	0.81
									GoogLeNet	✓	40	–	$1e^{-4}$	$1e^{-5}$	0.9	SGD	✓	✓	✓	0.83
									ResNet50	✓	40	–	$1e^{-3}$	$5e^{-5}$	0.95	SGD	✓	✓	✓	0.88
Ou et al. (2018)	2018	–	3	1	1214	RGB	✗	Internet	ResNet50	✗	10	–	–	–	–	–	✓	✗	✓	0.8958–0.9560
		23	3	1	22,800	RGB	✗	Simulated	ResNet50	✗	10	–	–	–	–	–	✓	✗	✓	0.9616–0.9658
		–	3	1	24,014	RGB	✗	Mixed	ResNet50	✗	10	–	–	–	–	–	✓	✗	✓	0.9550–0.9830
Ou et al. (2019)	2019	14	4	1	16,713	RGB	✗	Simulated	SqueezeNet	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.8305
									ResNet18	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.9237
									ResNet50	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.9443
									ResNet34	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.9598
					16,770	IR	✗	Simulated	SqueezeNet	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.7721
									ResNet34	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.9033
									ResNet50	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.9036
									ResNet18	✓	–	32	$1e^{-4}$	–	–	Adam	✓	✗	✗	0.9224
Hu et al. (2019)	2018	–	6	1	33,162	RGB	✗	Simulated	BoVW-SVM	–	–	–	–	–	–	–	–	–	✓	0.6900
									PHOG-MLP	–	–	–	–	–	–	–	–	–	✓	0.7460
									AlexNet	–	–	–	–	–	–	–	–	–	✓	0.8340
									VGG19	✓	–	–	–	–	–	–	–	–	✓	0.8980
									GoogLeNet	✓	–	–	–	–	–	–	–	–	✓	0.9010
									CNN	✓	–	80	0.001	–	–	–	–	–	✓	0.8670–0.8870
									Fusion of CNNs	✓	–	32	$4e^{-4}$	–	–	–	–	–	✓	0.8990–0.9320
Xing et al. (2019)	2019	10	7	1	34,000	RGB	✗	Naturalistic	ResNet50	–	–	–	–	–	–	–	–	–	–	0.749
									GoogLeNet	–	–	–	–	–	–	–	–	–	–	0.786
									AlexNet	–	–	–	–	–	–	–	–	–	–	0.816
									ResNet50 (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.830
									GoogLeNet (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.875
									AlexNet (Cls)[a]	–	–	–	–	–	–	–	–	–	–	0.914
Swetha et al. (2019)	2019	11 (5/6)	6	1	10 h	RGB	✓	Naturalistic	AlexNet	✓	–	–	–	–	–	–	–	–	–	0.6083
									VGG16	✓	–	–	–	–	–	–	–	–	–	0.6472
									LeNet	✓	–	–	–	–	–	–	–	–	–	0.6722
									AlexNet+VGG16+LeNet	✓	–	–	–	–	–	–	–	–	–	0.6945
									AlexNet+ VGG16+ LeNet+ LR	✓	–	–	–	–	–	–	–	–	–	0.8350
Kapoor et al. (2020)	2020	–	4	1	21,625	RGB	✗	Naturalistic	MobileNetV1	✓	–	–	–	–	–	–	–	–	–	0.9911
Li et al. (2020)	2020	20 (12/8)	6	1	106,677	RGB	✗	Simulated	YOLO+MLP	✓	–	–	–	–	–	–	–	–	–	0.9200
Omerustaoglu et al. (2020)	2020	–	10	1	137,093	RGB	✗	Naturalistic	LSTM	–	–	–	–	–	–	–	–	–	–	0.76
Zhang et al. (2020)	2020	50 (31/19)	9	2	60 h	RGB	✗	Simulated	CNN	✗	136	100	–	–	–	Adam	–	–	–	0.6601–0.6986
									MobileNet	✗	136	100	–	–	–	Adam	–	–	–	0.6236–0.7081
									MobileNetV2	✗	136	100	–	–	–	Adam	–	–	–	0.6750–0.6871
									ShuffleNet	✗	136	100	–	–	–	Adam	–	–	–	0.6623–0.6942
									ShuffleNetV2	✗	136	100	–	–	–	Adam	–	–	–	0.6668–0.6998
									I3D	✗	136	100	–	–	–	Adam	–	–	–	0.6675
Hu et al. (2020)	2020	–	6	1	42,816	RGB	✗	Parked	PHOG-MLP	–	–	–	–	–	–	–	–	–	–	0.710
									PAV-SVM	–	–	–	–	–	–	–	–	–	–	0.753
									AlexNet	–	–	–	–	–	–	–	–	–	✓	0.809
									VGG19	–	–	–	–	–	–	–	–	–	✓	0.846
									PAV-Hint CNN	–	–	–	–	–	–	–	–	–	✓	0.864
									Multi-stream CNN	–	–	–	–	–	–	–	–	–	✓	0.874
									MSA-CNN	✗	–	80	0.001	–	–	–	–	✓	✓	0.940
Wu et al. (2021)	2021	–	4	1	65,902	RGB	✗	Naturalistic	VGG16	✓	50	64	0.001	$1e^{-6}$	0.9	SGD	✗	✗	✗	0.7335
									ResNet50	✓	50	64	0.001	$1e^{-6}$	0.9	SGD	✗	✗	✗	0.7570
									InceptionV3	✓	50	64	0.001	$1e^{-6}$	0.9	SGD	✗	✗	✗	0.7728
									VGG16 (Fusion)	✓	50	64	0.001	$1e^{-6}$	0.9	SGD	✗	✗	✗	0.9109
									ResNet50 (Fusion)	✓	50	64	0.001	$1e^{-6}$	0.9	SGD	✗	✗	✗	0.9258
									InceptionV3 (Fusion)	✓	50	64	0.001	$1e^{-6}$	0.9	SGD	✗	✗	✗	0.9575
Wang et al. (2021)	2021	14	10	1	2200	RGB	✓	Simulated	Xception	Train on AUCD3 dataset, test on self-collected dataset										0.8394

(*continued on next page*)

Table 13 (*continued*).

Ref	Year	Dataset information							Results											
		Sub	Act	Views	Size	Streams	Occ	Scenario	ML Algo	PT	TE	BS	LR	WD	Mom	Opt	DO	BN	DA	Acc
Li et al. (2021b)	2021	2648	6	1	267,378	RGB	✓	Naturalistic	MobileNetV3 Small	✗	20	64	$3e^{-3}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9010
									ShuffleNetV2	✗	20	64	$1e^{-3}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9149
									MSCNN	✗	20	64	$1e^{-4}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9255
									MobileNetV3 Large	✗	20	64	$3e^{-3}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9290
									ResNet50	✗	20	64	$1e^{-4}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9305
									InceptionV4	✗	20	64	$6e^{-4}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9395
									DenseNet40	✗	20	64	$1e^{-3}$	$1e^{-5}$	0.9	Adam	✗	✗	✗	0.9588
									OLCMNet	✗	20	64	0.1	$1e^{-5}$	0.9	SGD	✓	✓	✗	0.9598
Wagner et al. (2021)	2021	16	5	1	12,716	Grayscale	✗	Naturalistic	VGG16	✓	50	8	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.8735
									VGG19	✓	50	8	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.8874
									ResNeXt50	✓	50	16	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.8989
									ResNeXt34	✓	50	16	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.9288
				1	19,427	Grayscale	✗	Naturalistic	VGG16	✓	50	8	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.3809
									VGG19	✓	50	8	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.5482
									ResNeXt34	✓	50	16	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.8184
									ResNeXt50	✓	50	16	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.9036
				2	32,143	Grayscale	✗	Naturalistic	ResNeXt50	✓	50	16	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.8973
									ResNeXt34+ResNeXt50	✓	50	16	–	$2e^{-4}$	–	Adam	✓	✗	✗	0.9254

Legend: Sub : Subject (Male/Female) Act: Actions Size: Size of dataset (number of images) Occ: Occlusion PT: Pretrained TE: Training epoch BS: Batch size
LR: Learning rate WD: Weight Decay Mom: Momentum DO: Dropout Opt: Optimizer BN: Batch Normalization
DA: Data augmentation (only techniques that increases size of dataset)
* – represent data not stated in article.

[a](Cls) represent binary classification (distracted vs. non-distracted).

Table 14
Summary of articles that using SD3 to classify distraction actions (sorted accuracy in ascending order).

Ref	Year	ML Algo	PT	TE	LR	WD	BS	Mom	Opt	DO	BN	DA	Acc
Hssayeni et al. (2017)	2017	AlexNet	✓	30	$(5e^{-4}, 9e^{-4}, 1e^{-3})$	$1e^{-5}$	50	0.9	SGD	✓	✓	✗	0.7260
Lemley et al. (2017)	2017	Video (16-frame seq): C3D	✓	–	$1e^{-4}$	–	–	0.95	GD	✗	✗	✗	0.7335
Hu et al. (2019)	2018	CNN	✓	–	0.001	–	80	–	SGD	–	–	✓	0.7920
Li et al. (2021b)	2021	MobileNetV3 Small	✗	20	$3e^{-3}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.7938
Li et al. (2021b)	2021	InceptionV4	✗	20	$6e^{-4}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.8169
Majdi et al. (2018)	2018	MLP	✗	–	–	–	–	–	–	–	–	–	0.8200
Hu et al. (2019)	2018	GoogLeNet	✓	–	–	–	–	–	–	–	–	✓	0.8230
Valeriano et al. (2018)	2018	ResNet101	✓	–	–	–	–	–	–	–	–	–	0.8238
Hssayeni et al. (2017)	2017	VGG16	✓	55	$(5e^{-6}, 3e^{-5}, 1e^{-5})$	$5e^{-5}$	2	0.9	SGD	✓	✓	✗	0.8250
Li et al. (2021b)	2021	ShuffleNetV2	✗	20	$1e^{-3}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.8364
Li et al. (2021b)	2021	MSCNN	✗	20	$1e^{-4}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.8411
Li et al. (2021b)	2021	MobileNetV3 Large	✗	20	$3e^{-3}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.8482
Hssayeni et al. (2017)	2017	ResNet152	✓	30	$(1e^{-3}, 5e^{-4}, 1e^{-4}, 5e^{-5})$	$5e^{-5}$	2	0.9	SGD	✓	✓	✗	0.8500
Hu et al. (2019)	2018	ResNet34	✓	–	–	–	–	–	–	–	–	✓	0.8530
Koay et al. (2021b)	2021	ResNet34	✓	30	$(4e^{-6}, 1e^{-4})$	–	32	(0.85, 0.95)	Adam	✗	✗	✗	0.8544
Li et al. (2021b)	2021	ResNet50	✗	20	$1e^{-4}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.8565
Nel and Ngxande (2021)	2021	Video (16-frame): 3D-ResNet18	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.8570
Lu et al. (2020)	2020	DD-RCNN	✓	80k steps	$(1e^{-3}, 1e^{-4})$	$5e^{-4}$	–	0.9	SGD	✗	✗	✓	0.8600
Li et al. (2021b)	2021	DenseNet40	✗	20	$1e^{-3}, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	Adam	✓	✗	✓	0.8656
Hu et al. (2019)	2018	Fusion of 3 CNNs	✓	–	$4e^{-4}$	–	32	–	SGD	–	–	✓	0.8660
Moslemi et al. (2019)	2019	Video (Optical Flow): I3D	✓	30	0.001	$1e^{-7}$	32	–	Adam	✓	✓	✗	0.8920
Li et al. (2021b)	2021	OLCMNet	✗	20	$0.1, \alpha = 1e^{-5}$	$1e^{-5}$	64	0.9	SGD	✓	✓	✓	0.8953
Balamurugan and Kalaiarasi (2021)	2021	VGG19	✓	10	–	–	–	–	–	–	–	–	0.8969
Nel and Ngxande (2021)	2021	Video (16-frame): 3D-ResNet34	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9000
Nel and Ngxande (2021)	2021	Video (16-frame): 3D-ResNet50	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9110
Behera et al. (2018)	2018	Video (30-frame seq): M-LSTM	✓	50	$5e^{-5}$	–	32	–	RMSProp	✓	✗	✗	0.9125
Valeriano et al. (2018)	2018	Video (64-frame seq): ResNext101	✓	–	–	–	–	–	–	–	–	–	0.9125
Behera et al. (2020)	2020	MDFN	✓	–	–	–	–	–	–	–	–	–	0.9139
Majdi et al. (2018)	2018	RNN	✗	–	–	–	–	–	–	–	–	–	0.9170
Alotaibi and Alotaibi (2019)	2019	ResNet+HRNN	✓	30	0.001	$\beta_1 = 0.9,\ \beta_2 = 0.999$	80	–	Adam	✗	✗	✗	0.9172
Moslemi et al. (2019)	2019	Video (RGB): I3D	✓	30	0.001	$1e^{-7}$	32	–	Adam	✓	✓	✗	0.9180
Nel and Ngxande (2021)	2021	Video (16-frame): 3D-ResNet101	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9210
Nel and Ngxande (2021)	2021	Video (16-frame): 3D-ResNet152	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9250
Valeriano et al. (2018)	2018	Video (Optical Flow): I3D	✓	–	–	–	–	–	–	–	–	–	0.9274
Nel and Ngxande (2021)	2021	Video (16-frame): 3D-ResNet200	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9290
Balamurugan and Kalaiarasi (2021)	2021	DenseNet121	✓	10	–	–	–	–	–	–	–	–	0.9304
Balamurugan and Kalaiarasi (2021)	2021	ResNet101	✓	10	–	–	–	–	–	–	–	–	0.9410
Moslemi et al. (2019)	2019	Video (Optical Flow+RGB): I3D	✓	30	0.001	$1e^{-7}$	32	–	Adam	✓	✓	✗	0.9440
Majdi et al. (2018)	2018	Drive-Net (CNN, RF)	✗	50	0.001	$\beta_1 = 0.9, \beta_2 = 0.99$	128	–	Adam	✓	✗	✗	0.9500
Alotaibi and Alotaibi (2019)	2019	ResNet152	✓	30	0.001	$\beta_1 = 0.9,\ \beta_2 = 0.999$	80	–	Adam	✗	✗	✗	0.9531
Leekha et al. (2019)	2019	GrabCut + VGG16	✓	20	0.001	$\beta_1 = 0.9,\ \beta_2 = 0.999$	32	–	Adam	✗	✗	✗	0.9532
Balamurugan and Kalaiarasi (2021)	2021	InceptionV3	✓	10	–	–	–	–	–	–	–	–	0.9542
Alotaibi and Alotaibi (2019)	2019	ResNet+HRNN+Inception	✓	30	0.001	$\beta_1 = 0.9,\ \beta_2 = 0.999$	80	–	Adam	✗	✗	✗	0.9623
Valeriano et al. (2018)	2018	Video (Optical Flow+RGB): I3D	✓	–	–	–	–	–	–	–	–	–	0.9667
Okon and Meng (2017)	2017	AlexNet	✓	10	–	–	50	–	–	✗	✗	✗	0.9680
Kumar et al. (2021)	2021	GWE-(AlexNet+VGG16+ EffNetB0+CNN+DN201+InV3- BiLSTM)	✓	–	–	–	–	–	–	✗	✗	✗	0.9716
Balamurugan and Kalaiarasi (2021)	2021	Reg-DenseNet	✓	10	–	–	–	–	–	–	–	–	0.9806
Balamurugan and Kalaiarasi (2021)	2021	ResNeXt101	✓	10	–	–	–	–	–	–	–	–	0.9810
Alotaibi and Alotaibi (2019)	2019	HRNN	✓	30	0.001	$\beta_1 = 0.9,\ \beta_2 = 0.999$	80	–	Adam	✗	✗	✗	0.9834
Ugli et al. (2021)	2021	ResNet50	✓	–	–	–	–	–	–	–	–	–	0.9840
Leekha et al. (2019)	2019	GrabCut + CNN	✗	15	0.001	$\beta_1 = 0.9,\ \beta_2 = 0.999$	32	–	Adam	✓	✗	✗	0.9848
Ugli et al. (2021)	2021	VGG16	✓	–	–	–	–	–	–	–	–	–	0.9860
Okon and Meng (2017)	2017	AlexNet (Triplet Loss)	✓	10	–	–	50	–	–	✗	✗	✗	0.9870
Masood et al. (2020)	2020	VGG19	✓	10	$1e^{-3}$	$10e^{-6}$	32	0.9	SGD	✗	✗	✓	0.9898
Ugli et al. (2021)	2021	MobileNet	✓	–	–	–	–	–	–	–	–	–	0.9900
Gumaei et al. (2020)	2020	CDCNN	✗	10	–	–	–	–	RMSProp	✓	✗	✗	0.9936
Masood et al. (2020)	2020	VGG16	✗	50	$1e^{-3}$	$10e^{-6}$	32	0.9	SGD	✗	✗	✓	0.9939
Masood et al. (2020)	2020	VGG16	✗	50	$1e^{-3}$	$10e^{-6}$	32	0.9	SGD	✗	✗	✓	0.9943
Masood et al. (2020)	2020	VGG16	✓	10	$1e^{-3}$	$10e^{-6}$	32	0.9	SGD	✗	✗	✓	0.9957
Gumaei et al. (2020)	2020	VGG16	✓	100	–	–	–	–	RMSProp	✓	✗	✗	0.9957
Baheti et al. (2020)	2021	MobileVGG	✗	500	$1e^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$	$1e^{-6}$	32	–	Adam	✓	✓	✓	0.9975

Legends: PT: Pretrained LR: Learning Rate TE: Training Epochs WD: Weight Decay BS: Batch Size Mom: Momentum Opt: Optimizer
BN: Batch Normalization DO: Dropout DA: Data Augmentation
* Only data augmentation that increase the number of training size are considered. Resizing and normalization are not consider as DA.
** Numbers in bracket for TE represents early stopping patience epoch. Column with – represent data not stated in the article. LR with multiple values represent learning rate scheduling was used.

techniques to prevent over-fitting, and finally, they achieved 96% accuracy on the distracted driver detection task. Mase et al. (2020) benchmarked various CNN models, including ResNet50, InceptionV3, and fusion of InceptionV3 with RNN with AUC-DDD. The best performing model was InceptionV3-BiGRU. Wang et al. (2021) performed data augmentation to increase the training dataset by first using Faster RCNN to extract the driving operation areas. The extracted driving operation area images are then coupled with the original dataset to perform classification model training. They trained the dataset on AlexNet, InceptionV4, and Xception and constantly reported that the augmented dataset achieved higher accuracy. They also performed predictions on their self-collected dataset and concluded that their proposed model could be generalized well. Koay et al. (2021b) proposed a framework named Optimally-Weighted Image-Pose Approach (OWIPA), which used the global, hand pose and body pose feature to perform distraction classification. They found that fusing ResNet101 trained with global feature and ResNet50 trained with body pose feature achieve higher results than just classifying through global features. Other works fine-tuned famous state-of-the-art models, such as InceptionV3 (Mafeni Mase et al., 2020; Kumar et al., 2021), HRNN (Alotaibi and Alotaibi, 2019), ResNet (Alotaibi and Alotaibi, 2019; Koay et al., 2021b), AlexNet (Kumar et al., 2021), VGG (Kumar et al., 2021), EfficientNet (Kumar et al., 2021) and DenseNet (Kumar et al., 2021) with different training techniques to achieve a relatively high accuracy in classifying the distraction actions. Besides, there is some work demonstrating the usage of 3D-CNN (Nel and Ngxande, 2021) in classifying distraction actions. Another technique is to use ensemble of various model to boost the final classification accuracy (Alotaibi and Alotaibi, 2019; Kumar et al., 2021; Koay et al., 2021b). All of the studies are summarized in Table 15.

Since AUC-DDD and StateFarm datasets are the most common datasets, most studies used both datasets to perform benchmarks on their proposed methods. Given that both datasets have several differences, such as recording angle and variability of the subject, testing on both datasets and achieving relatively high accuracy signals a good algorithm.

Behera et al. (2018) proposed a novel MultiStream LSTM (M-LSTM) for recognizing fine-grained driver distracted activities that are difficult to distinguish. M-LSTM combines the concepts of LSTM and CNN for recognizing driver distraction activities. The proposed network is evaluated from one stream up to four streams on both AUC-DDD and StateFarm datasets. The proposed architecture has three components: the transferable deep CNN features (from VGG16), contextual cues involving body pose and body-object interaction, and the proposed M-LSTM for sequence modelling and activity recognition. They proposed contextual descriptors to represent high-level knowledge from the human pose, human–object interactions, and relationships between body parts and objects. Body-pose descriptor translates the body parts configuration to a feature vector by encoding relationships between various body parts. Body-object descriptor captures the pairwise relationship between the body joints and involved objects, encoding the relative position of an object with respect to a given joint in a scene. Their proposed method performs well on the StateFarm dataset, but not AUC-DDD, with their four-stream network achieving 91.25% on the StateFarm dataset. The same concept is used in Behera et al. (2020), where a novel Multistream Deep Fusion Network (MDFN) and classifier-level fusion for combining the CNN features with the proposed descriptor was proposed. They explored the configuration of body parts and the interaction between body parts and objects to perform the prediction. Their results on both datasets are promising. Leekha et al. (2019) proposed a simple and robust architecture using foreground extraction, GrabCut (Rother et al., 2004) and a small CNN to detect distraction. Their smaller CNN with GrabCut outperforms fine-tuned VGG-16 in both datasets, achieving 98.48% and 95.64% on StateFarm and AUC-DDD, respectively. Baheti et al. (2020) proposed MobileVGG, which is modified from VGG architecture with fewer parameters, and benchmarked on both datasets, obtaining accuracies of 95.24% and 99.75% on AUC-DDD and StateFarm, respectively.

DMD (Ortega et al., 2020) came in the form of videos. Therefore, one can classify the distraction actions through video or image classification. The author in Cañas et al. (2021) tested both approaches and found that image classification, where turning video frames into image dataset, performs better than video classification. They achieved an image classification accuracy of 99.5% with MobileNetV1, while a video classification accuracy of 97.3% with MobileNetV1 + LSTM through 30-frame video clips. Interestingly, longer frame video clips did not perform better, where 70-frame video clips with the same architecture only achieved 95.7%. The results are tabulated in Table 16.

MDAD (Jegham et al., 2019) is a multimodal and multiview dataset. In the dataset article, the author extracted Histograms of Gradients (HOG), and Histograms of Optic Flow (HOF) features from Spatio Temporal Interest Points (STIPs). Then, they quantized these features into 60 codebooks using k-means to represent each sequence as a histogram of codebooks. The classification process is done via the SVM with a polynomial kernel. They achieve 37.45–42.45% when detecting through multiview data (concatenate the feature vectors across views) and 33.64–39.41% when detecting through multimodal data. Jegham et al. (2020b) proposed a novel soft spatial attention-based network named the Depth-based Spatial Attention network (DSA) and tested it on MDAD. DSA consists of a hybrid deep network and LSTM for distraction analysis. The soft spatial attention is shown to improve the capability of the CNN by selectively highlighting relevant frame regions. The authors claim to achieve up to 75% of accuracy when fusing both front and side views, while achieving 69.79% and 65.63% when detecting on the side and front view only, respectively.

All articles utilizing public dataset are summarized in Tables 14– 16. However, these tables have several pitfalls, as summarized below.

- AUC-DDD dataset has two versions, with version two having more drivers and images. Besides, there are two types of splits, split by driver or randomly split for training and testing datasets. However, most of the articles did not specify the version and splitting method used in the study.
- StateFarm dataset is mainly used for competition back in 2016 hosted on Kaggle. The labelled dataset is only applicable for the training dataset, while the testing dataset is unlabelled. Almost all of the studies used part of the training dataset as the testing dataset, while some manually labelled the testing dataset (Li et al., 2021b).
- Not all studies explicitly stated if the model is trained from scratch or pretrained model is used to fine-tune to the dataset. We assume all works that did not state the method of training the model as using a pretrained model. The assumption is made because lower training epochs are used, whereby training from scratch will not warrant such high accuracy.
- Some papers take validation accuracy as the final classification accuracy, which are the images used to validate the training process. While we eliminated several articles which deemed to misunderstand the meaning of the test dataset, which is supposed to be "unseen" by the model, there is some article that which the author did not specify the process of obtaining the final classification accuracy.
- The training platform and framework are not summarized in the table.

Drive&Act dataset (Martin et al., 2019) is a large-scale video dataset comprising of various driver activities. The dataset contains annotations for 12 classes (coarse tasks) of top-level activities (e.g., eating and drinking), 34 categories of fine-grained activities (e.g., opening bottle, preparing food, etc.), and 372 classes of atomic action units involving triplet of action, object, and location. There are five types of actions, 17 object classes, 14 location annotations, and 372 (All)

Table 15
Summary of articles that using AUCD3 V2 dataset to classify distraction actions (sorted accuracy in ascending order).

Ref	Year	ML Algo	PT	TE	LR	WD	BS	Mom	Opt	DO	BN	DA	Acc
Mase et al. (2020)	2020	AlexNet	✓	50 (5)	$1e^{-4}$	–	32	–	Adam	✓	✗	✗	0.7380
Eraqi et al. (2019)	2019	GWE-VGG-16	✗	30	$(5e^{-3}, 1e^{-2})$	–	50	–	GD	✗	✗	✗	0.7613
Wharton et al. (2021)	2021	VGG16	–	–	–	–	–	–	–	–	–	–	0.7613
Eraqi et al. (2019)	2019	GWE-Resnet50	✗	30	$(5e^{-3}, 1e^{-2})$	–	50	–	GD	✗	✗	✗	0.8169
Wharton et al. (2021)	2021	Resnet50	–	–	–	–	–	–	–	–	–	–	0.8170
Mase et al. (2020)	2020	VGG-19	✓	50 (5)	$1e^{-4}$	–	32	–	Adam	✓	✗	✗	0.8330
Alotaibi and Alotaibi (2019)	2019	HRNN	✓	30	0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$	–	80	–	Adam	✗	✗	✗	0.8485
Mase et al. (2020)	2020	ResNet50	✓	50 (5)	$1e^{-4}$	–	32	–	Adam	✓	✗	✗	0.8770
Mase et al. (2020)	2020	InceptionV3	✓	50 (5)	$1e^{-4}$	–	32	–	Adam	✓	✗	✗	0.8840
Mafeni Mase et al. (2020)	2020	InceptionV3	✓	50	$1e^{-4}$	–	32	–	SGD	✗	✗	✗	0.8441
Nel and Ngxande (2021)	2021	Video (16-frame seq): 3D-ResNet18	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.8760
Mase et al. (2020)	2020	InceptionV3-RNN	✓	50 (5)	$1e^{-4}$	–	16	–	Adam	✗	✗	✗	0.8840
Alotaibi and Alotaibi (2019)	2019	ResNet152	✓	30	0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$	–	80	–	Adam	✗	✗	✗	0.8852
Mase et al. (2020)	2021	Densenet-201	✓	50 (5)	$1e^{-4}$	–	32	–	Adam	✓	✗	✗	0.8900
Mafeni Mase et al. (2020)	2020	InceptionV3-LSTM	✓	50	$1e^{-4}$	–	32	–	SGD	✗	✗	✗	0.8982
Eraqi et al. (2019)	2019	GWE-InceptionV3	✗	30	$(5e^{-3}, 1e^{-2})$	–	50	–	GD	✗	✗	✗	0.9006
Wharton et al. (2021)	2021	InceptionV3	–	–	–	–	–	–	–	–	–	–	0.9007
Mase et al. (2020)	2020	InceptionV3-LSTM	✓	50 (5)	$1e^{-4}$	–	16	–	Adam	✗	✗	✗	0.9020
Mase et al. (2020)	2020	InceptionV3-GRU	✓	50 (5)	$1e^{-4}$	–	16	–	Adam	✗	✗	✗	0.9030
Wu et al. (2021)	2021	InceptionV3 (Global, Hand, Pose)	✓	50	0.001	$1e^{-6}$	64	0.9	SGD	✗	✗	✗	0.9038
Mase et al. (2020)	2020	InceptionV3-BiLSTM	✓	50 (5)	$1e^{-4}$	–	8	–	Adam	✗	✗	✗	0.9170
Mase et al. (2020)	2020	InceptionV3-BiGRU	✓	50 (5)	$1e^{-4}$	–	8	–	Adam	✗	✗	✗	0.9170
Nel and Ngxande (2021)	2021	Video (16-frame seq): 3D-ResNet34	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9190
Alotaibi and Alotaibi (2019)	2019	ResNet+HRNN+Inception	✓	30	0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$	–	80	–	Adam	✗	✗	✗	0.9236
Wharton et al. (2021)	2021	Video: CTA-Net	✗	25	0.001, $\beta_1 = 0.9, \beta_2 = 0.999$	–	4	–	Adam	–	–	–	0.9250
Mafeni Mase et al. (2020)	2020	InceptionV3-BiLSTM (C-SLSTM)	✗	50	$1e^{-4}$	–	32	–	SGD	✗	✗	✗	0.9270
Wang et al. (2021)	2021	AlexNet	✓	100	0.001	–	32	–	–	✗	✗	✓	0.9314
Nel and Ngxande (2021)	2021	Video (16-frame seq): 3D-ResNet101	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9340
Nel and Ngxande (2021)	2021	Video (16-frame seq): 3D-ResNet152	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9370
Nel and Ngxande (2021)	2021	Video (16-frame seq): 3D-ResNet50	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9390
Nel and Ngxande (2021)	2021	Video (16-frame seq): 3D-ResNet200	✓	200	$1e^{-3}$	$1e^{-3}$	–	0.9	SGD	✗	✓	✓	0.9400
Leekha et al. (2019)	2019	GrabCut + CNN	✓	50	0.001, $\beta_1 = 0.9, \beta_2 = 0.999$	–	32	–	Adam	✓	✗	✗	0.9437
Kumar et al. (2021)	2021	DenseNet201	✓	–	–	–	–	–	–	✗	✗	✗	0.9442
Baheti et al. (2018)	2018	VGG-16	✓	100	$1e^{-4}$	$1e^{-6}$	64	0.9	SGD	✗	✗	✗	0.9444
Wang et al. (2021)	2021	InceptionV4	✓	100	0.001	–	32	–	–	✗	✗	✓	0.9506
Baheti et al. (2020)	2021	MobileVGG	✗	500	$1e^{-4}, \beta_1 = 0.9, \beta_2 = 0.999$	$1e^{-6}$	32	–	Adam	✓	✓	✓	0.9524
Kumar et al. (2021)	2021	InceptionV3-BiLSTM	✓	30	–	–	64	–	RMSProp	✗	✗	✗	0.9528
Kumar et al. (2021)	2021	VGG16	✓	50	$1e^{-4}$	–	64	–	RMSProp	✗	✗	✗	0.9530
Kumar et al. (2021)	2021	EfficientNet-B0	✓	30	–	–	64	–	RMSProp	✗	✗	✗	0.9530
Wang et al. (2021)	2021	Xception	✓	100	0.001	–	32	–	–	✗	✗	✓	0.9531
Baheti et al. (2018)	2018	VGG-16 (Replace FC with Conv layers)	✓	100	$1e^{-4}$	$1e^{-6}$	64	0.9	SGD	✓	✓	✗	0.9554
Behera et al. (2020)	2020	MDFN	✓	–	–	–	–	–	–	–	–	–	0.9557
Leekha et al. (2019)	2019	GrabCut + CNN	✗	50	0.001, $\beta_1 = 0.9, \beta_2 = 0.999$	–	32	–	Adam	✓	✗	✗	0.9564
Kumar et al. (2021)	2021	CNN	✓	–	–	–	–	–	–	✗	✗	✗	0.9576
Kumar et al. (2021)	2021	AlexNet	✗	50	–	–	32	–	RMSProp	✗	✓	✗	0.9624
Baheti et al. (2018)	2018	VGG-16 + Regularization	✓	100	$1e^{-4}$	$1e^{-6}$	64	0.9	SGD	✓	✓	✗	0.9631
Kumar et al. (2021)	2021	GWE-(AlexNet+VGG16+EffNetB0+CNN +DN201+InV3-BiLSTM)	✓	–	–	–	–	–	–	✗	✗	✗	0.9637
Wang et al. (2021)	2021	Faster-RCNN + Xception	✓	100	0.001	–	32	–	–	✗	✗	✓	0.9697

Legends: LR: Learning Rate TE: Training Epochs WD: Weight Decay BS: Batch Size Mom: Momentum Opt: Optimizer BN: Batch Normalization DA: Data Augmentation
* Only consider data augmentation that increase the number of training size. Resizing and normalization are not consider as DA.
** Numbers in bracket for TE represents early stopping patience epoch. Column with – represent data not stated in the article. LR with multiple values represent learning rate scheduling was used.

Table 16
Summary of articles that using DMD dataset to classify distraction actions (sorted accuracy in ascending order).

Ref	Year	DL Algo	PT	TE	LR	WD	BS	Mom	Opt	Dropout	BN	DA	Acc
Cañas et al. (2021)	2021	MobileNetV1	✓	–	$1e^{-3}$	–	32	–	Adam	✓	✗	✗	0.9950
Cañas et al. (2021)	2021	InceptionV3	✓	–	$1e^{-3}$	–	32	–	Adam	✓	✗	✗	0.9930
Cañas et al. (2021)	2021	Video (30-frame seq): MobileNetV1 + LSTM	✓	–	$1e^{-3}$	–	–	–	Adam	✓	✗	✗	0.9730
Cañas et al. (2021)	2021	Video (30-frame seq): Conv3D	✓	–	$1e^{-3}$	–	–	–	Adam	✓	✗	✗	0.9720
Cañas et al. (2021)	2021	Video (30-frame seq): Conv2DLSTM	✓	–	$1e^{-3}$	–	–	–	Adam	✓	✗	✗	0.9580

Legends: LR: Learning Rate TE: Training Epochs WD: Weight Decay BS: Batch Size Mom: Momentum Opt: Optimizer BN: Batch Normalization DA: Data Augmentation PT: Pretrained
* We only consider data augmentation that increase the number of training size. Resizing and normalization are not consider as DA.
** Numbers in bracket for TE represents early stopping patience epoch. Column with – represent data not stated in the article. LR with multiple values represent learning rate scheduling was used.

possible combinations. Differ from all the datasets discussed above, it is one of the largest datasets and has multiple tasks.

Liu et al. (2021) proposed a triple-wise multi-task learning (TML) framework to improve the accuracy of distracted driver recognition tasks. Their framework first generates positive and negative samples of the given inputs. Then, the framework is trained on the network backbone with different tasks, including recognizing the raw input and positive sample and pulling closer the distance between the features obtained from the raw input and positive sample while pushing away the distance between the features obtained from the raw input and negative sample. They tested on Drive&Act, StateFarm, and AUC-DDD datasets. They found that their proposed framework could learn more accurate clues from the videos.

Ren et al. (2021) proposed an end-to-end model which tackled accurate classification of similar driver actions in real-time, known as MSRNet. The proposed architecture comprises two branches, known as the action detection network and the object detection network, which are used to extract spatiotemporal and key-object features, respectively. A confidence fusion mechanism is introduced to aggregate the predictions from both branches based on the semantic relationships between actions and critical objects. They regrouped the fine-grained dataset from 34 to 11 categories. Their model can achieve test accuracy of 64.18% and 20 FPS inference time on an 8-frame input clip.

In CTA-Net (Wharton et al., 2021), coarse temporal branches are introduced in a trainable glimpse network. The introduction of the branch is to allow the glimpse to capture high-level temporal relationships by focusing on a specific part of a video. These branches also respect the topology of the temporal dynamics in the video, ensuring that different branches learn meaningful spatial and temporal changes. The model also uses the attention mechanism to generate high-level action-specific contextual information for activity recognition by exploring the hidden states of an LSTM. The attention mechanism helps to decide the importance of each hidden state for recognition tasks by weighing them when constructing a representation of the video. They evaluated the model on AUC-DDD, StateFarm, and Drive&Act datasets.

Open Set Drive&Act dataset is formulated from the Drive&Act dataset for open set recognition of driver monitoring (Roitberg et al., 2020). The proposed model shall be able to identify behaviours previously unseen by the classifier. The authors combined closed set models with different sets of strategies for novelty detection adopted from general action classification (Roitberg et al., 2018) in a generic open set driver behaviour recognition framework. They deployed I3D (Carreira and Zisserman, 2017) architecture extended with modules for assessing its uncertainty via Monte-Carlo dropout. They show the benefits of uncertainty-sensitive models while leveraging the output neurons' uncertainty in a voting-like fashion leads to the best recognition results.

Similarly, the ZS-Drive&Act dataset (Reiß et al., 2020) is introduced by modifying the Drive&Act dataset to become a zero-shot activity classification benchmark dataset. The dataset is designed to recognize previously unseen driver behaviours and is unique since it is focused on fine-grained activities and the presence of activity-driven attributes, which are automatically derived from a hierarchical annotation scheme. The authors evaluated multiple zero-shot learning methods on the dataset. They proposed to extend the feature generating Wasserstein GANs (f-WGAN) (Mandal et al., 2019) with a fusion strategy for linking semantic attributes and word vectors representing the behaviour labels. Their experiments demonstrated the effectiveness of leveraging both semantic spaces simultaneously, improving the recognition rate by 2.79%.

All of the studies using the Drive&Act dataset are summarized in Table 17.

Note that we only cherry-picked several datasets to discuss in details. This is because we believe some of the dataset are irrelevant, since they have smaller data size (Ohn-Bar and Trivedi, 2014; Weng et al., 2016; Billah et al., 2018), less actions considered (Abtahi et al., 2014; Jain et al., 2015; Billah et al., 2018), and only capture certain part of body (Martin et al., 2016; Diaz-Chito et al., 2016; Weng et al., 2016; Borghi et al., 2017; Schwarz et al., 2017; Borghi et al., 2018; Roth and Gavrila, 2019).

Discussion. In terms of classification, many works prefer the usage of the state-of-the-art DL model due to its out-of-the-box accuracy. Together with the models pretrained on large-scale image datasets (e.g ImageNet) open-sourced, many classification tasks can achieve a great result through transfer learning on the targeted dataset. However, to achieve better performance, many tricks are introduced to enhance the prediction. Among them, the usage of data augmentation, tuning of hyperparameters, and scheduling the learning rate are commonly performed according to the targeted dataset. This presents a pitfall, whereby these tricks might only work well on one dataset, but not for the others or even real-world scenarios. Besides, most of the models are huge in number of parameters, which is not needed when the specific downstream task contains only 10 different actions to be classified. We hypothesize that some of the fine-tuned model might still be over-parameterized and is not efficient by all means. Therefore, the usage of model compression techniques could be applied to produce lighter models. Further, the splitting of dataset plays an important role in determining the robustness of the model. Ideally, we should not "leak" the training dataset during the training process. For example, random splitting a given dataset would therefore include all subjects during training phase, and present a risk for the model to "memorize" and thus overfit the model. In AUC-DDD, the original split is done by splitting based on the human subjects, whereby the driver in test set is not included in training or validation split. Such split is not available for Statefarm, and the accuracy reported might not reflect the real performance of the models.

5.3. Detection based on external data

With the general concept that driver distraction will have a negative and observable effect on driving performance, many researchers have proposed and experimented with driving performance as an indicator of driver distraction. Typically, minimal fluctuation in vehicle signals can be used to detect distraction. For example, the steering angle and average steering speed variance were the most distinguishing signals between normal and distracted driving (Im et al., 2014). On the other hand, the usage of pedal, driving speed, steering wheel position and lane offset are the most generally used indicator of driving performance (Bingham et al., 2016; Son and Park, 2016; Li et al., 2017; Ye et al., 2017; Iranmanesh et al., 2018). These data are usually collected through vehicle's CAN bus. It is also possible to collect these data through smartphone sensors, such as GPS and IMU, to estimate vehicle kinematics, and from there, identify distracted driving patterns (Xie et al., 2018). Beside collecting data from vehicle or smartphone, Li et al. (2017) obtained driving performance data from the Integrated Vehicle Based Safety System (IVBSS) (Sayer et al., 2011). Anomaly in those indicators would therefore inferred as distracted driving.

On the other hand, Aksjonov et al. (2018, 2017) presented a method for detecting the driver's distraction by monitoring lane maintenance and speed performance on specified road segments.

Beside vehicle signals, the sensors of smartphones can be used to detect micro-movements of drivers. For example, Bo et al. (2016) proposed TEXIVE, which used the inertial sensors of the smartphones to detect if a driver is entering a vehicle or not, inferring which side of the vehicle he/she is entering, and determining whether a user is sitting in the front or rear seats. Similar work is done by Torres et al. (2019) to identify text while driving. Another line of work collects driving performance through wearable devices, equipped with accelerometer and gyroscope (Goel et al., 2018; Xie et al., 2021; Sun et al., 2021). An interesting work by Sun et al. (2021) where they collected four common gestures made while driving and then perform gesture recognition to detect distractions.

Since there are publicly available datasets collected using external sensors, some of the work chosen to use them rather than collecting own data. Tavakoli et al. (2021a) utilized the smartwatch data from HARMONY dataset (Tavakoli et al., 2021b). They analysed drivers' activities and behaviours through smartwatches in naturalistic conditions using an RF classifier. They defined eight distracting actions and used the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) to ensure the even distribution of class samples. Echanobe et al. (2021) proposed a multi-objective genetic algorithm (MOGA) that searches for the minimum number of features extracted from the driving task yet is still able to provide high recognition rates. The system has been evaluated through a real-life data collection of different subjects performing the driving task with occasionally induced distractions from the UYANIK dataset (Abut et al., 2009).

Instead of having the driver actively engaged with a certain devices (smartphone and wearable sensors) or accessing vehicle telemetric

Table 17
Accuracy of various approach on Drive&Act dataset.

Ref	Year	Model	Fine-grained		Coarse task		Action		Object		Location		All	
			Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
Martin et al. (2019)	2019	Pose	53.17	44.36	37.18	32.96	57.62	47.74	51.45	41.72	53.31	52.64	9.18	7.07
Martin et al. (2019)	2019	Interior	45.23	40.30	35.76	29.75	54.23	49.03	49.90	40.73	53.76	53.33	8.76	6.85
Martin et al. (2019)	2019	2-Stream (Wang and Wang, 2017)	53.76	45.39	39.37	34.81	57.86	48.83	52.72	42.79	53.99	54.73	10.31	7.11
Martin et al. (2019)	2019	3-Stream (Martin et al., 2018)	55.67	46.95	41.70	35.45	59.29	50.65	55.59	45.25	59.54	56.50	11.57	8.09
Martin et al. (2019)	2019	C3D (Tran et al., 2015)	49.54	43.41	–	–	–	–	–	–	–	–	–	–
Martin et al. (2019)	2019	P3D ResNet (Qiu et al., 2017)	55.04	45.32	–	–	–	–	–	–	–	–	–	–
Martin et al. (2019)	2019	I3D Net (Carreira and Zisserman, 2017)	69.57	63.64	44.66	31.80	62.81	56.07	61.81	56.15	47.70	51.12	15.56	12.12
Wharton et al. (2021)	2021	CTA-Net	72.42	65.25	62.82	52.31	57.59	56.41	63.37	59.19	56.41	63.01	46.44	49.41
Liu et al. (2021)	2021	TML	–	66.90	–	–	–	–	–	–	–	–	–	–
Ren et al. (2021)	2021	YOWO (Köpüklü et al., 2019)	67.71[a]	61.02[a]	–	–	–	–	–	–	–	–	–	–
Ren et al. (2021)	2021	MSRNet	72.36[a]	64.18[a]	–	–	–	–	–	–	–	–	–	–
Open Set Drive&Act (Roitberg et al., 2020)														
Roitberg et al. (2020)	2020	SVM	–	54.78[b]	–	–	–	–	–	–	–	–	–	–
Roitberg et al. (2020)	2020	GMM	–	62.39[b]	–	–	–	–	–	–	–	–	–	–
Roitberg et al. (2020)	2020	CNN	–	74.83[b]	–	–	–	–	–	–	–	–	–	–
Roitberg et al. (2020)	2020	Bayesian I3D – Selective Voting	–	77.62[b]	–	–	–	–	–	–	–	–	–	–
Zero-Shot Drive&Act (Reiß et al., 2020)														
Reiß et al. (2020)	2020	ConSE (Norouzi et al., 2013)	40.65	30.01	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	DeViSE (Frome et al., 2013)	40.65	28.01	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	DAP (Lampert et al., 2013)	39.44	27.61	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	f-WGAN (Mandal et al., 2019) (Word2Vec)	45.28	28.93	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	f-WGAN (Mandal et al., 2019) (DAwA)	40.96	29.96	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	f-WGAN (Mandal et al., 2019) (Word2Vec, DAwA)	46.37	32.80	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	Early Semantic Fusion	46.37	32.80	–	–	–	–	–	–	–	–	–	–
Reiß et al. (2020)	2020	Late Semantic Fusion	44.78	32.55	–	–	–	–	–	–	–	–	–	–

[a] Represents the dataset is being reconstructed from 34 categories into 11 categories.

[b] Represents the dataset is recategorized into open set, where the dataset is split into 24 seen and 10 unseen categories, of which 5 unknown classes are used for validation and 5 for testing.

data, the usage of road camera can be used to detect distracted driver too. Artan et al. (2014) utilized the near-infrared (NIR) camera system that is mounted above the road. This has posed a challenge, where the windshield could be reflective, and the camera angle causes the search of the driver's location in a larger search space. They solve this by localizing the driver using a Deformable Part Model (DPM), which reduces the search space. The face detection algorithm can then assume the most face-like region to be a face, even in partially occluded images. This is crucial as the sun visors occlude many faces. The face region, the region of interest (ROI), is then analysed through image classification to identify driver cell phone usage. First, the Scale-Invariant Feature Transform (SIFT) features are generated. Then, Bag of Visual Words (BoVW), Vector of Locally Aggregated Descriptors (VLAD), and Fisher Vectors (FV) are used to calculate the image descriptors locally. The three vectors are then classified by a linear Support Vector Machine (SVM). They run the classification with full-face images and half-face images. It is shown that classifying the right and left sides of the driver's head separately yields a better result. FV outperformed with the highest accuracy. This project utilizes vector signatures instead of the complex algorithm, yet archiving quite a good result. It is proven to be useful by only locating the driver on the right side of the windshield. The same is being explored by Xu and Loce (2015) using DPM and FV representation to classify the driver's side of the windshield for cell phone usage.

In terms of ML algorithms, since the collected data are usually in time-series, many used LSTM (Saleh et al., 2017; Xie et al., 2021), ANN (Im et al., 2014; Ye et al., 2017), GMM (Im et al., 2014; Bingham et al., 2016), and HMM (Bingham et al., 2016; Sun et al., 2021) to detect distractions. While other works turned into specially-designed algorithm (Li et al., 2017; Echanobe et al., 2021), or traditional ML methods (Goel et al., 2018; Aksjonov et al., 2018, 2017; Iranmanesh et al., 2018; Xie et al., 2018; Torres et al., 2019; Tavakoli et al., 2021a).

All of the works are summarized in Table 18.

Discussion. Generally, the performance is slightly worse than using physiological sensors or visual sensors. Besides, since the collected data are usually in form of numerical data, many traditional ML methods are used in this field of study. Therefore, the performance is somehow capped by the capability of the selected ML methods in processing the data.

The usage of external sensors in driver distraction detection is known for its non-intrusive properties. However, compared with the visual sensor, the external sensor collected external parameters (e.g vehicle state and external environmental state) to perform guesses on driver distraction is inconclusive. One cannot simply label distraction by looking at the vehicle speed or GPS location, instead, they should be used as an enhancing parameter to eliminate false positives.

5.4. Detection based on multimodal data

5.4.1. External and physiological data

Solovey et al. (2014) used five classification methods (KNN, NB, DT, LR, and MLP) to classify the extent of workload into the two levels and reported an accuracy of 69.4%–75.7%. Although several classification methods have been applied to classify the extent of cognitive workload level, their accuracies are low because they do not consider the individual differences in heart response to the cognitive workload in developing a classification model.

de Salis et al. (2019) assessed the impact on the estimation of the driver state without access to the driver's physiological data. CNN, KNN, and RF are implemented in the study. F1-Score of 0.63, 0.66, and 0.9711 are achieved for KNN, RF, and CNN, respectively, when using only driving data. Using both driving and physiological data, all three models achieved higher F1-Score at 0.85, 0.92, and 0.9848 for KNN, RF, and CNN, respectively.

5.4.2. External and visual data

Li and Busso (2014) proposed to use features extracted from the CAN-bus signal, a microphone array, and two video cameras facing the road and the driver. Du et al. (2018) collected the facial, audio and simulation driving data through MultiSense sensor (Stratou and Morency, 2017). Similar works include combining the features from both driving performance and eye movements (Liang and Lee, 2014; Yang et al., 2015; Liao et al., 2016) as well as driving performance and driver face (Li and Busso, 2015; Omerustaoglu et al., 2020). In this type of settings (camera and CAN bus), the dataset usually contains driver images and sensor data collected from naturalistic settings. In terms of the proposed algorithm, a two-stage detection system is proposed to classify the distracted behaviours. In the first stage, a vision-based CNN

Table 18
Summary of studies utilizing external sensor data as input.

Ref	Year	Input sensor	Collected features	Subject (M/F)	Distraction Type	ML algorithm	Effectiveness[a]
Im et al. (2014)	2014	In-vehicle signal	Acceleration pedal value, Brake master cylinder pressure, Relative distance to front vehicle, Distance from a wheel to a lane, Steering wheel torque, angle, speed	12	–	GMM, ANN	–
Artan et al. (2014)	2014	External camera	High Occupancy Vehicle (HOV) and High Occupancy Tolling (HOT) NIR images – image descriptors extracted from ROI around the face	–	Visual, Manual, Cognitive	SVM	Acc: 0.8619
Xu and Loce (2015)	2015	External camera	High Occupancy Vehicle (HOV) and High Occupancy Tolling (HOT) NIR images	–	Visual, Manual, Cognitive	DPM+FV	Acc: 0.95
Bo et al. (2016)	2016	Smartphone	Inertial sensors	–	Cognitive	HMM	Acc: 0.8718
Bingham et al. (2016)	2016	CAN	Acceleration pedal use, driving distance, driving speed and lane offset in both time and frequency domain	108 (54/54)	Visual, Cognitive, Manual	HMM	Acc: 0.59–0.88
Son and Park (2016)	2016	Vehicle Simulator	Standard deviation of lane position, steering wheel reversal rate	15 (15/0)	Visual, Cognitive	RBPNN	Acc: 0.9310
Ye et al. (2017)	2017	IMU	Speed, longitudinal acceleration, lateral acceleration, yaw rate, and throttle position (from SHRP-2 dataset)	–	Manual, Cognitive, Visual	ANN	Acc: 0.9810–0.9980
Aksjonov et al. (2017)	2017	Vehicle Simulator	Car Location, Speed	18 (13/5)	Visual, Manual, Cognitive	ANFIS, FL	–
Saleh et al. (2017)	2017	External camera, IMU, GPS	IMU – Acceleration along xyz axis, Roll, Pitch, Yaw angle; GPS – Vehicle speed; Camera – Distance to ahead vehicle, Number of detected vehicles (UAH-DriveSet)	6 (5/1)	–	MLP DT Stacked LSTM	F1: 0.48 F1: 0.80 F1: 0.91
Li et al. (2017)	2017	CAN	Steering entropy, absolute mean of speed prediction error	16	Visual, Manual	SVM	Acc: 0.9330–0.9833
Goel et al. (2018)	2018	Wearable	Accelerometer, gyroscope data	16 (10/6)	Visual, Cognitive, Manual	NB SVM DT RF	F1: 0.618–0.817 F1: 0.752–0.805 F1: 0.842–0.934 F1: 0.863–0.946
Aksjonov et al. (2018)	2018	Vehicle Simulator	Car Location, Speed	18 (13/5)	Visual, Manual, Cognitive	ANFIS, FL	–
Iranmanesh et al. (2018)	2018	CAN	Gas pedal position, range with neighbouring vehicles and variation rate, turn signal, yaw rate, brake flags, longitudinal acceleration, velocity, heading	–	Cognitive	SVM NN	Acc: 0.756 Acc: 0.78
Xie et al. (2018)	2018	Smartphone	Velocity, Gyro, Acceleration	24 (16/8)	Cognitive	KNN+RF+ LR+NB	F1: 0.87
Torres et al. (2019)	2019	Smartphone	Touches, speed, acceleration, gyroscope	13 (8/5)	Visual, Manual, Cognitive	DT SVM RF GB	F1: 0.882 F1: 0.930 F1: 0.932 F1: 0.939
Xie et al. (2021)	2021	Wearable IMU	Acceleration data	20 (12/8)	Visual, Cognitive, Manual	CNN LSTM ConvLSTM	F1: 0.76 F1: 0.85 F1: 0.87
Sun et al. (2021)	2021	Wearable IMU	Accelerations, Angular velocities, Euler angles, Quaternion algebra	20 (14/6)	Manual	HMM	Acc: 0.9663
Tavakoli et al. (2021a)	2021	Wearable	PPG, Gyroscope, Heart Rate, Light, Acceleration. (Data collected from HARMONY)	15	Manual, Visual, Cognitive	RF	F1: 0.9099–0.9455
Echanobe et al. (2021)	2021	CAN, IMU, Pedal Sensor	Steering wheel angle, Steering wheel relative speed, vehicle speed, percent gas pedal, engine RPM, break pedal pressure, gas pedal pressure, rool, pitch and yaw rate, xyz-axis accelerometer (Data from UYANIK)	101	–	Multi-Obj GA Fisher Score SVM-RFE Lasso	Acc: 0.7821–0.8465 Acc: 0.6282–0.7436 Acc: 0.7064–0.8302 Acc: 0.6931–0.7983

CE: Classification Error; Acc: Accuracy
* – represent data not stated in the article.
SD: standard deviation; Q1: First Quartile; Q2: Second quartile; Q3: Third quartile; IQR: Interquartile Range.
[a]Due to different metrics used across different study, we named this column as "effectiveness".

model was created by fine-tuning methods. In the second stage, a LSTM or RNN model is created to fuse the sensor and image data together.

Besides combining driving performance and driver face, Costa et al. (2019) utilized IR camera to track eye and head and Ancho Radar (contactless biometric sensor) to collect heart and respiration rate. The vehicle's telematics, such as the presence of the driver's hands on the steering wheel and data related to time-of-day and automation levels of the vehicle, is collected to enhance the detection process further.

Streiffer et al. (2017) proposed DarNet framework, which can detect distracted driving behaviour. DarNet is made up of two main components, which are a data collection system and an analysis engine. They collect images from an inward-facing camera and IMU from a smartphone. The images and IMU signals are applied to a CNN and RNN, respectively. A bayesian network then combines the outputs of the two networks to classify the driver behaviours. DarNet achieves an accuracy of 87.02%.

5.4.3. Visual and physiological data

Riani et al. (2020) explored both attention and alertness together based on a multiclass classification scheme using multiple physiological modalities such as BVP, skin conductance, skin temperature, and respiration data. They found that multimodal feature learning showed a significant improvement in overall individual modalities for both multiclass classification schemes. Early modality fusion led to improved performance compared to individual modalities for multiclass classification with an overall accuracy of 79.73% for the three-class scheme and 55.12% for the four-class approach.

Papakostas et al. (2021) evaluated visual and physiological information and explore the potential of multimodal modelling for distraction recognition. They used RGB, thermal, and NIR camera to collect visual data, a physiological sensor to collect BVP, skin temperature, skin conductance, and respiration rate, as well as the audio data. However, the audio data is not used. The participants carried out four actions, and the data were recorded and analysed individually before fusing for distraction recognition. They used RF and GB with late fusion and achieved a maximum F1-score of 94%. They indicated that the visual data did not characterize cognitive inattention well as physiological signals since it is proved to be more effective and robust.

5.4.4. Visual, physiological and external data

Chen et al. (2021) proposed a fine-grained detection method for driver distraction based on neural architecture search (NAS). They used a palm EDA sensor, adrenergic sensor, thermal facial camera, visual facial camera, FaceLAB eye-tracking system, and driving parameter extractor to collect visual, physiological, and external data. They first designed an automatic construction algorithm for deep CNN based on NAS, which automatically searches for the optimal deep CNN architecture without human interference. Then, they fused driver-related multisource perception information and used an automatically constructed deep CNN to extract high-dimensional mapping features, then implemented fine-grained detection of various types of driver distraction states. The results on the dataset demonstrated that their method could efficiently search an optimal deep CNN, which can

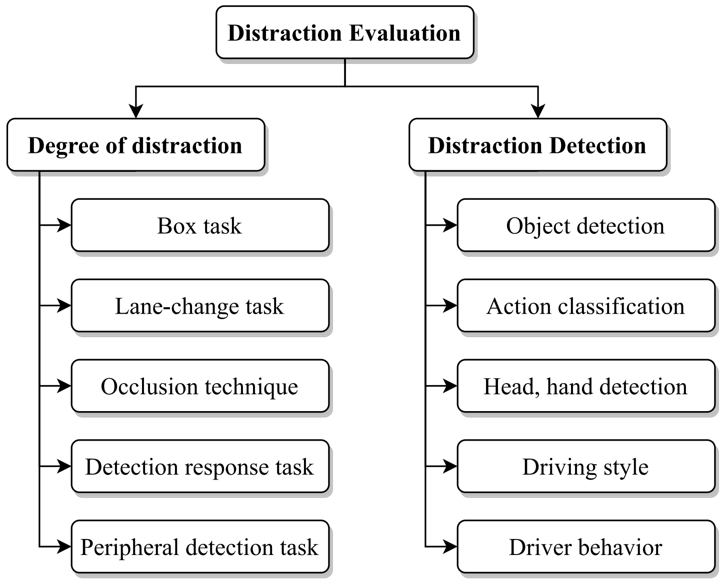

Fig. 7. Classification of driver distraction evaluation.

quickly converge and accurately detect the considered types of driver distraction states. The average detection accuracy reaches 99.7796%.

Discussion. The usage of multiple sensors in recognizing distraction is highly recommended to further reduce false alarms (see Table 19). However, the involvement of multiple sensors not only increases the cost, but also the design of ML models to adapt to various modalities. Among the fusion group, combining visual and external sensors are more realistic and optimum. This is because the external data, such as vehicle telemetry data and external images could better understand the state of the driver, which can serve as a piece of extra information for images. For example, the sudden change of speed of the vehicle could signal the model that a possible distraction is going on, and can then be verified with the visual images for final prediction.

In terms of the proposed algorithms, since they are dealing with different data modalities, many works tend to conform them into a single modal in order to adapt to the algorithm. This has therefore introduces error when transforming data into another representation. A simple solution to overcome this is to perform optimization and hybridization, which will be discussed later, in order to use the current modalities without much modification, while maintaining its accuracy.

5.5. Patents involving distraction detection

There are several patents filed involving innovative methods in detecting distraction while driving. We summarized the patents in Table 20.

6. Driver distraction evaluation

In this section, the various evaluation techniques are discussed. We classify the evaluation of driver distraction into two classes, which are the degree of distraction and distraction detection, as shown in Fig. 7.

6.1. Degree of distraction evaluation

There are several methods to evaluate the degree of driver distraction. Among them are the box task, lane-change task, occlusion technique, detection response task, and peripheral detection task. They are used to determine the degree of distraction caused by engagement in non-driving related secondary tasks (Morgenstern et al., 2020).

Box task method is one of the evaluations used to measure the degree of driver distraction. It aims to measure the amount of cognitive load imposed by various distracting tasks. Participants are asked to use pedals and a steering wheel to control the size and position of the screen. Around the box presented on the screen, some boundaries change their size and shape. Participants must keep the controlled box inside the boundaries. The system will then record several box intersections with the boundaries and the standard deviation of box size and position from the ideal size and position. While carrying on the experiments, the participants may and may not be distracted. The drop in measured performance represents the cognitive load that various additional tasks impose on the driver (Morgenstern et al., 2020).

A more driving-based method is called the lane-change test (Mattes, 2003). The participants are asked to drive on a simulated three-lane road with no traffic present while reacting to signs indicating lane changes at certain times. When a lane change sign appears, the participant must change the lane as fast as possible. The standard deviation between the ideal and actual paths is measured to evaluate the participant's performance.

The occlusion technique (Senders et al., 1967) is employed to determine the visual demand associated with a specific task. The task's interruption level is determined by periodically obstructing subjects' view while engaged in a secondary task. Subjects are asked to wear special glasses that block the field of view at certain times. When the glasses are blocked, the driver cannot see anything, which occludes the driver's vision. The subject is directed to drive as fast as possible while maintaining a safe speed. The maximal safe speed and the period of not blocking the driver's vision provide information about the necessary visual information for safe driving.

Detection response task (Stojmenova and Sodnik, 2018) is used to access the attentional effects of cognitive load in driving conditions. Drivers are stimulated with a sensory stimulus every 3–5 s and are asked to respond to it by pressing a button attached to their finger. The indicators in this evaluation technique are the response time and hit rate. It is simple to carry out and relatively cheap. It is deemed a more reliable way to detect the effects of cognitive load compared to questionnaires. Peripheral detection task (Martens and Van Winsum,

Table 19
Summary of studies utilizing multimodal data as input.

Ref	Year	Input sensor	Features	Task	ML algorithm	Effectiveness[a]
Solovey et al. (2014)	2014	Physiological, External	Physiological – ECG (heart rate, skin conductance level); External – CAN (vehicle speed, steering wheel angle)	Distraction classification	KNN NB DT LR MLP	Acc: 0.694 Acc: 0.750 Acc: 0.750 Acc: 0.755 Acc: 0.757
Li and Busso (2014)	2014	Visual, External	Visual – road camera, driver camera; External – CAN (Vehicle speed, steering wheel angle and jitter, brake pedal pressure), microphone	Binary classification of visual and cognitive distraction	LDC KNN SVM QDC RUSB	F1: 0.772–0.794 F1: 0.681–0.723 F1: 0.670–0.790 F1: 0.747–0.764 F1: 0.731–0.791
				Distraction classification	LDC KNN SVM QDC RUSB	F1: 0.513 F1: 0.681–0.723 F1: 0.670–0.790 F1: 0.747–0.764 F1: 0.549
Liang and Lee (2014)	2014	Visual, External	Visual – Blink frequency, Mean and SD of fixation duration, pursuit duration, pursuit distance, pursuit direction, and pursuit speed, percentage of the time spent on performing pursuit movements in each time window, Mean and SD of horizontal and vertical fixation location coordinates; External – SD of steering wheel position, mean steering error, SD of lane position	Distraction detection	DBN Layered algorithm SVM	Acc: 0.88 Acc: 0.88 Acc: 0.90
Li and Busso (2015)	2015	Visual, External	Visual – road camera, driver camera; External – CAN (Vehicle speed, steering wheel angle and jitter, brake pedal pressure), microphone	Mirror checking action detection	SVM KNN RUSBoost	F1: 0.652–0.677 F1: 0.797–0.873 F1: 0.877–0.914
				Distraction classification	LDC	F1: 0.655–0.906
Yang et al. (2015)	2015	Visual, External	Visual – 3D angles of head rotation, blink frequency, PERCLOS, pitch/yaw angles for left/right eye gaze; External – CAN (steering wheel angles, speed, left/right lane offsets)	Distraction recognition	SVM ELM	Acc: 0.829 Acc: 0.870
Liao et al. (2016)	2016	Visual, External	External – Steering entropy, SD of lane position, Min speed, Speed across the stop line, Max acceleration and deceleration, Crossing time, Brake on timing, Compliance time; Visual – Head rotation, Gaze position, Saccade, Pupil diameter	Distraction detection	SVM-RFE	F1: 0.9350–0.9600
Streiffer et al. (2017)	2017	Visual, External	Visual – Image, External – IMU	Distraction classification	CNN CNN+SVM CNN+RNN	Acc@1: 0.7388 Acc@1: 0.8623 Acc@1: 0.8702
Du et al. (2018)	2018	Visual, External	Visual – Facial landmarks, Gaze vectors, 18 Action Units, Head pose; External – Prosody, Voice-Quality, Frame Energy, Voice Activity Detection, F0 fundamental frequency, Syllables per second, Speed of the vehicle, Steering wheel position, Gas pedal position, Break pedal position	Distraction detection	SVM NN MPF (Single Feature) MPF	Acc: 0.7542 Acc: 0.8015 – 0.8046 Acc: 0.7491–0.7749 Acc: 0.8139
McDonald et al. (2018)	2018	Physiological, External	Physiological – breathing rate, heart rate, perinasal perspiration; External – brake force, lane offset, speed, steering angle	Distraction recognition	RF, DT, NB, KNN, SVM, NN	Acc: >0.39
de Salis et al. (2019)	2019	Physiological, External	Physiological – Perinasal EDA, Heart Rate, Breathing Rate, Palm EDA; External – Speed, Acceleration, Brake Force, Steering, Lane Offset	Distraction detection	KNN RF CNN	F1: 0.85 F1: 0.92 F1: 0.9848
Costa et al. (2019)	2019	Visual, External	Visual – IR image for eye and head activity; External – Heart Rate, Respiration Rate, Force Sensitive Resistors (FSR), Vehicle's Telematics	Fatigue Detection	SVM DT	F1: 0.91 F1: 0.93
				Cognitive Distraction Detection	DT SVM	F1: 0.88 F1: 0.90
				Distraction recognition	DT SVM	F1: 0.76 F1: 0.91
Riani et al. (2020)	2020	Visual, Physiological	Visual – RGB, Thermal; Physiological – Blood Volume Pulse (BVP), Skin Conductance (SC), Skin Temperature (ST), a Respiration Rate (RR)	Drowsiness detection	DT Early Fusion DT Late Fusion	Acc: 0.6409 Acc: 0.8204
				Distraction recognition	DT Early Fusion DT Late Fusion	Acc: 0.5512 Acc: 0.3512
Omerustaoglu et al. (2020)	2020	Visual, External	Visual – SD3 image, mobile phone video; External – Mobile phone (gyroscope, accelerometer, GPS), OBD (engine load & RPM, throttle position, fuel level, coolant temperature, speed)	Distraction recognition	LSTM-RNN	0.85
Chen et al. (2021)	2021	Visual, Physiological, External	Visual – thermal and visual facial image, eye gaze direction; Physiological – EDA signal, heart rate, breathing rate; External – speed, acceleration, braking, steering, and lane position	Distraction detection	ResNeXt ResNet-20 AlexNet LeNet NIN VGG19 LSTM Deep SAE CNN-NAS	Acc: 0.257685 Acc: 0.257687 Acc: 0.257823 Acc: 0.260272 Acc: 0.261633 Acc: 0.264626 Acc: 0.345664 Acc: 0.867055 Acc: 0.997796
Papakostas et al. (2021)	2021	Visual, Physiological	Visual – Top-view RGB, Face closeup RGB, Face closeup NIR, Face thermal image; Physiological – Blood Volume Pulse (BVP), skin temperature, skin Conductance, respiration	Distraction detection (Visual)	RF GB	F1: 0.68–0.73 F1: 0.65–0.75
				Distraction detection (Physiological)	RF GB	F1: 0.53–0.87 F1: 0.53–0.84
				Distraction recognition (Visual)	RF GB	F1: 0.47–0.85 F1: 0.47–0.88
				Distraction recognition (Psychological)	RF GB	F1: 0.31–0.90 F1: 0.31–0.88
				Distraction recognition	RF + GB	F1: 0.46–0.94

CE: Classification Error; Acc: Accuracy
* – represent data not stated in the article. SD: standard deviation; Q1: First Quartile; Q2: Second quartile; Q3: Third quartile; IQR: Interquartile Range.
[a]Due to different metrics used across different study, this column is named as "effectiveness".

2000) is also a variant in detection response task, used to measure driver's mental workload. Subjects must react to visual stimuli periodically presented in the drivers' peripheral view. Usually, drivers can detect fewer targets (visual stimuli presented by the device) and react slower when performing secondary tasks while driving.

The advantages and disadvantages of each evaluation discussed above are summarized in Table 21. Note that all of these methods are usually carried out in a simulated/lab environment. Therefore, such evaluation is not used in real driving conditions to evaluate if a driver is experiencing distraction.

Table 20
Patent related to driver distraction detection techniques using machine learning.

Ref	Year	Filed company	Title	Data source	Brief summary
Deruyck and McLaughlin (2017)	2017	Trimble Inc	Detection of driver behaviours using in-vehicle systems and methods	Microphone, Camera, Motion Sensor	Driver motion data, audio data and optionally driver context, driver history, driver behaviour, and/or vehicle information data are used to analyse the driver behaviours.
Shenoy et al. (2018)	2018	Lightmetrics Technologies Pvt Ltd	Method and system for driver monitoring by fusing contextual data with event data to determine context as cause of event	Camera, Vehicle IMU, OBD, location data sensors	Event data from vehicle sensors and audio or video feed received from camera are fused to monitor the driver.
Sicconi and Stys (2019)	2019	Telelingo LLC	Method to analyse attention margin and to prevent inattentive and unsafe driving	Microphone, Speaker, Camera, Biosensor, Car dynamics data	Extracted features from a driver-facing and road-facing camera; driver's behaviour including head and eyes movement, speech and gestures; and telemetry features from vehicle; driver's biometrics are fed to a decision engine to determine the driver's attention and emotional state and the associated risks.
Levkova et al. (2019)	2019	Nauto Inc	System and method for driver distraction determination	Camera, user device (phone), vehicle parameter	Usage of interior camera, exterior camera, vehicle information (vehicle sensors such as velocity and pedal position), user's device parameter (interaction with phone) and external data (e.g weather and traffic) to determine driver distraction through ML techniques.
Madkor et al. (2020)	2020	Com-Iot Technologies	Distracted driver detection	Camera	Video frame is collected to train a classifier and alert is sent if driver is distracted.
Yu et al. (2021)	2021	FutureWei Technologies Inc	Primary preview region and gaze based driver distraction detection	Camera	Detect if the driver's gaze is inside the primary preview region (PPR)
Renbo and Daqian (2021)	2021	Shanghai Sensetime Intelligent Technology Co Ltd	Driving state analysis method and apparatus, driver monitoring system and vehicle	Camera	Performing fatigue and distraction state detection on a given image. Alarm will be sounded if either one or both state achieve certain threshold.

Table 21
Pro and cons of various degree of distraction measurement.

Technique	Pro	Cons
Box task	Applied to all approaches of distraction detection	Unrelated to driver
Lane-change task	Applied to all approaches of distraction detection	Only considers lane-changing task
Occlusion technique	Duration of eye-off-road is measured	Cognitive distraction is not considered
Detection response task	Driver reaction time is calculated	Not applicable in driving scenario

6.2. Distraction detection evaluation

As discussed in the previous section, we can briefly group all the studies discussed into five categories: object detection, action classification, head and hand detection, driving style, and driver behaviour.

In object detection, usually, we define driver distraction through the interaction with other objects presented in the cabin. For example, when we would like to detect if a driver is using a cell phone, we can detect the presence of a cell phone in the driver's hand. This type of detection is usually done through object detection algorithms, such as Faster R-CNN or YOLO.

The second category, action classification, is one of the most active fields of study in the driver distraction detection area. This is one of the most straightforward tasks, where the research just pass a set of labelled dataset to an ML algorithm for training. The trained model is then tested with the unseen dataset to obtain the final test classification accuracy. Some famous model includes ResNet, VGGNet, AlexNet, and EfficientNet.

In head and hand detection, similar to object detection, the goal is to detect the location of the head and hand of the driver to infer if they are being distracted. Generally, while driving, a driver should have his eyes on the road and hands on the steering wheel. By using this notion, the algorithm can then detect if the driver is being distracted if any of the criteria is not met.

In driving style and driver behaviour, this is usually done by specifying certain requirements to be met to make sure that the driver is not distracted. For example, the model recognizes a driver's specific behaviour and driving style when driving safely, and if there is a sudden change in the driving style, the driver is probably distracted. However, such a model is feasible, since the model only recognizes the pattern learnt from the training dataset.

All of the distraction detection algorithms that is done with the usage of ML used certain bounded evaluation criteria to train the model. The evaluations that are commonly used are explained in the next subsection.

Table 22
Confusion matrix.

	Actual positive	Actual negative
Predicted positive	True positive (TP)	False negative (FN)
Predicted negative	False positive (FP)	True negative (TN)

6.3. Machine learning evaluation

6.3.1. Evaluation metrics

Claims of the effectiveness of specific ML algorithms often detail the algorithm's quality using a core set of performance measurements.

Most of the studies in the field deal with classification problems to classify if a driver is experiencing distraction. Specifically, classification problems can be divided into binary (distracted vs. non-distracted driver), multiclass (smoking, eating, reaching back, etc.), and multi-labelled (multiple actions in one scene) classification. In a typical data classification problem, the evaluation metric is employed in two stages, namely the training stage and the testing stage. In the training stage, the evaluation metric optimizes the classification algorithm, while in the testing stage, the evaluation metric evaluates the effectiveness of produced classifier when tested with the unseen data.

For binary classification problems (distracted vs. non-distracted), the discrimination evaluation of the optimal solution during the classification training can be defined based on the confusion matrix as shown in Table 22. The row of the table represents the predicted class, while the column represents the ground truth.

From confusion matrix, several commonly used metrics are inferred, and are tabulated in Table 23 (Powers, 2020; Fawcett, 2006; Hossin and Sulaiman, 2015). They are used to evaluate the performance of the classifier with a different goal of evaluations. All of them are used in a binary classification problem, and some of the metrics are extended for multiclass problems, as shown in the last five rows of the table.

There are other metrics, such as Matthews correlation coefficient (MCC) and Fowlkes–Mallows index (FM), that are not included in Table 23 because they are not used in any articles discussed in the previous section.

We found that accuracy is the most used evaluation metric for both binary and multiclass classification problems from all the articles reviewed. The quality of the produced algorithm is evaluated based on the percentage of correct predictions over total instances. However, accuracy is a good measure when the class distribution is similar. In almost all cases, the distribution of classes is not similar. Therefore the usage of F1-Score is more favourable.

6.3.2. ROC curve

The receiver operating characteristic (ROC) curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: the true positive rate (TPR) and false positive rate (FPR). A ROC curve plots TPR against FPR at different classification thresholds. Reducing the classification threshold results in more items regarded as positive, therefore increasing both False Positives and True Positives. A sample of the curve is shown in Fig. 8.

A more common metric derived from the ROC curve is its area under the curve, namely the area under the ROC curve (AUC). AUC represents the degree of separability, telling how much the model is capable of distinguishing between classes. An optimum model will have an AUC near 1, while a poor model has an AUC near 0. When AUC is 0.5, the model has no class separation capacity whatsoever.

ROC curves are typically used in binary classification to study the output of a classifier. To extend the usage in multiclass problems, ROC curves can be plotted with the methodology of using one class versus the rest, whereby one ROC curve is drawn per label.

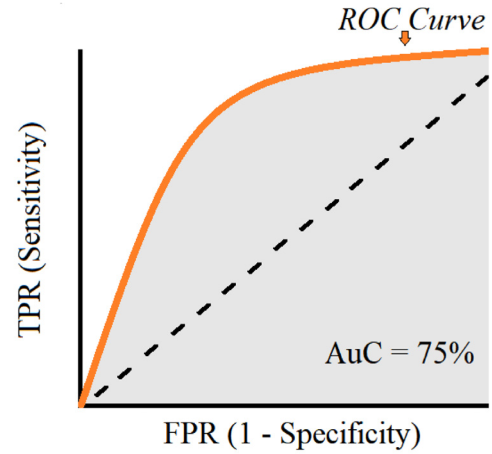

Fig. 8. Example of ROC curve.

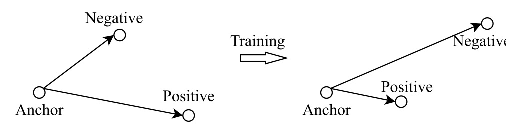

Fig. 9. Triplet loss illustration.

6.3.3. Loss function

Loss functions are used in supervised ML to minimize the differences between the predicted output of the model and the ground truth labels. It is used to measure how well a model can correctly predict a sample from the dataset. We identified two commonly used loss functions from all the literature – Cross Entropy loss and triplet loss. We also include other loss functions for completeness.

Cross entropy loss. Cross Entropy loss, or log loss, is a measure used to define the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverge from the actual label.

Suppose we have a binary classification problem with the label of 0,1. Then, the loss function for a single sample in the dataset is expressed as:

$$-y\log(p) - (1-y)\log(1-p)\ , \tag{1}$$

where y is the sample's label, and p is the predicted probability of the sample belonging to class 1.

Suppose we have a multiclass classification problem with K classes. The loss function is now:

$$-\log\frac{\exp(x_k)}{\sum_{i=0}^{K-1}\exp(x_i)} \tag{2}$$

where x is the original prediction for the sample.

Triplet loss. There are three main components in each triplet, a positive, an anchor, and a negative sample, as shown in Fig. 9. Triplet loss aims to minimize the distance between the anchor and the positive during the learning process and simultaneously increases the distance between the anchor and the negative during the learning process to improve the classification accuracy of deep networks (Okon and Meng, 2017).

The equation for triplet loss is:

$$\sum_{i}^{N}\max\left(0, f(x_i^a, x_i^p) - f(x_i^a, x_i^n) + \alpha\right) \tag{3}$$

where x_i^a represents the anchor feature vector, x_i^p is the positive feature vector and x_i^n is the negative feature vector, and α is the forced margin

Table 23
Evaluation metrics.

Metrics	Equation	Description
Precision/Positive predicted value, P	$\frac{TP}{TP+FP}$	The positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Negative predictive value	$\frac{TN}{TN+FN}$	The negative patterns that are correctly predicted from the total predicted patterns in a negative class.
Recall/Sensitivity/ True-positive rate, R	$\frac{TP}{TP+FN}$	The fraction of positive patterns that are correctly classified.
Specificity/True negative rate	$\frac{TN}{TN+FP}$	The fraction of negative patterns that are correctly classified.
Miss rate/false negative rate	$\frac{FN}{FN+TP}$	The ratio between the number of positive class wrongly categorized as negative and the total number of actual positive class.
False positive rate	$\frac{FP}{FP+TN}$	The ratio between the number of negative class wrongly categorized as positive and the total number of actual negative class.
Error rate	$\frac{FP+FN}{TP+FP+TN+FN}$	The ratio of incorrect predictions over the total number of instances evaluated.
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	The ratio of correct predictions over the total number of instances evaluated.
F1-score	$\frac{2\times P\times R}{P+R}$	The harmonic mean between recall and precision values.
Averaged accuracy	$\frac{\sum_{i=1}^{l}\frac{TP_i+TN_i}{TP_i+FP_i+TN_i+FN_i}}{l}$	Averaged accuracy of all classes.
Averaged error rate	$\frac{\sum_{i=1}^{l}\frac{FP_i+FN_i}{TP_i+FP_i+TN_i+FN_i}}{l}$	Averaged error rate of all classes.
Averaged precision	$\frac{\sum_{i=1}^{l}\frac{TP_i}{TP_i+FP_i}}{l}$	Averaged precision of all classes.
Averaged recall	$\frac{\sum_{i=1}^{l}\frac{TP_i}{TP_i+FN_i}}{l}$	Averaged recall of all classes.
Averaged F1-score	$\sum_{i=1}^{l}\frac{2*P_i*R_i}{P_i+R_i}$	Averaged F1-score of all classes.

Table 24
Loss functions.

Loss function	Equation	Description
0-1 loss	$L(y,f(x)) = \begin{cases} 0 & ,\text{if } yf(x) < 0 \\ 1 & ,\text{if } yf(x) \geq 0 \end{cases}$	If the predicted value of the sample has the same sign with the ground truth, the loss value is 0; otherwise, the loss value is 1.
Perceptron loss	$L(y,f(x)) = \max\{0 - yf(x)\}$	If the predicted value of the sample has the same sign with the ground truth, the loss value is 0; otherwise, the loss value is the absolute value of the predicted value.
Logarithmic loss	$L(y,\tilde{y}) = -\log\tilde{p}$, where $\tilde{p} = \begin{cases} p & ,\text{if } y = 1 \\ 1-p & ,\text{if } y \neq 1 \end{cases}$	The greater the probability of the sample being predicted as its label, the smaller the corresponding loss value.
Hinge loss	$H_s(z) = \max\{0, s-z\}$, where s is constant	It defines the margin (represented by the parameter s) near the decision boundary, where the correctly classified samples are in the middle of the two margin boundaries, while all misclassified samples are penalizes.
Ramp loss	$R_s(z) = H_1(z) - H_s(z)$, where $s < 1$ is constant	It limits the maximum loss value, which limits the influence of outliers to some extent, and the model is more robust to outliers.
Pinball loss	$L_\tau(u)\max\{u, -\tau u\}$, where $\tau \in [0,1]$ is a constant	It penalizes all correctly classified samples and makes the model less sensitive to noise near the decision boundary, thus having stronger resampling stability and more robust to outliers.

between the anchor-to-positive distance and the anchor-to-negative distance.

Other loss functions that can be used in the field of distraction detection are summarized in Table 24. We refer the reader to a review article by Wang et al. (2020) for a detailed analysis of various loss functions used in ML.

7. Discussion

From all the articles reviewed above, we could summarize them into a framework, as shown in Fig. 10.

From Fig. 10, we map the route of distraction detection from sensors used to the inferred driver distractions. We can describe the distraction

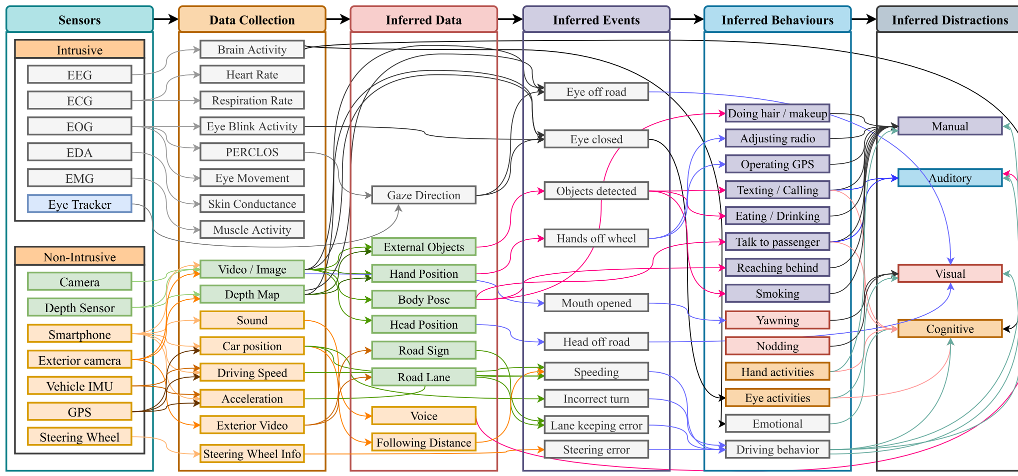

Fig. 10. Classification of driver distraction evaluation.

detection mechanism by the sensors used, the collected data/features, the inferred data from pre-processed data, the inferred events from the features, the inferred behaviours from events and features, and finally, the type of distractions. All of these are discussed in detail in previous sections.

Our study finds that it is impossible to use a single feature (visual, external, or physiological) to detect driver distraction in all environmental conditions. Some research suggests the usage of hybrid systems, which combines multiple features to get higher accuracy. Usually, visual and external features are preferred, since it is easy to set up. Combining driver image and vehicle performance data could easily infer and detect if a driver is being distracted.

For physiological data, it is observed that the sensors are usually attached to the driver's body, which is intrusive and may cause an extra cognitive burden to the driver. Such experiments are usually carried out in the simulated environment and are not verified in naturalistic driving settings. However, we observed that using a smartwatch to collect physiological data started as early as 2016. Using a smartwatch to collect these data is non-intrusive yet easy to set up at a lower cost. No microcontroller and extensive sensor pads are required, and such experiments could be carried out in real-life settings safely.

In this study, most of the distraction detection systems are proposed based on visual data. Thanks to the publicly available dataset, many works perform their benchmark on this dataset. Many of the works are just fine-tuning the dataset using one of the state-of-the-art models. This technique guarantees high accuracy. However, many of the studies did not disclose their training hyperparameters. Therefore the studies could not be duplicated to validate their claims. Besides, we believe that most of the datasets are not diverse, which only involve a small number of participants and distraction actions and have the same background.

In terms of the ML algorithms used in all studies discussed above, we found that most of the authors used CNN, followed by traditional ML and RNN, as illustrated in Fig. 11. Specifically, traditional ML techniques are primarily used in non-visual data since it is easily implemented and many comprehensive libraries are available. As for visual data, most researchers prefer CNN since CNN has been proven its ability to obtain high accuracy in image data. RNN techniques are less preferred since it has high complexity and computational requirements.

We further categorized CNN techniques used in all studies reviewed above in Fig. 12. We found that many authors prefer ResNet, VGG, and InceptionNet when training visual data. Among them, most of the studies perform TL on these models. With the usage of TL, less epochs are required, as observed in Tables 14 and 15.

Zooming into visual data, we observed that the usage of DL models topped over ML methods, as shown in Fig. 13. Given the two famous datasets, which are StateFarm and AUCDDD, we observed that the accuracy of all models averaged more than 80%, signalling that the usage

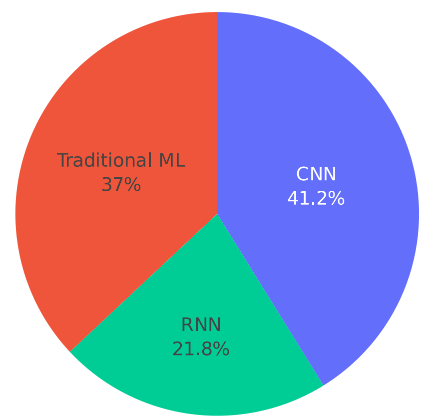

Fig. 11. Summary of all studies in terms of algorithm used.

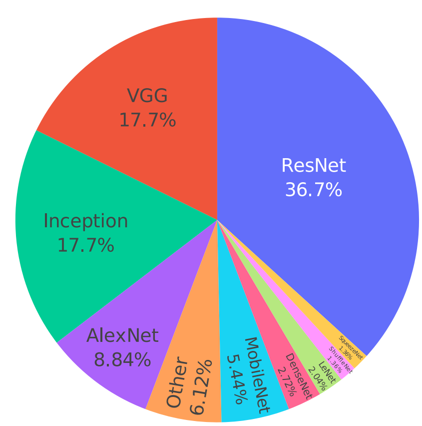

Fig. 12. Summary of all studies in terms of DL models.

of DL in image classification is the best choice. The same performance is also observed in the private dataset. As for the traditional ML method, which is only seen in private datasets, they usually came in lower average accuracy.

7.1. Implementation in commercial vehicles

Ideally, introducing a distraction detection system is to help create a safe environment for all road users when more autonomous vehicles

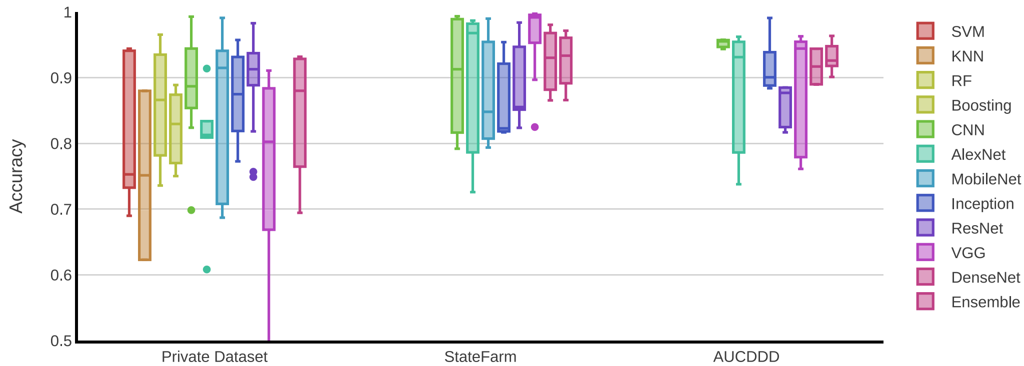

Fig. 13. Box plot for all models used in driver distraction detection using visual sensors.

are introduced on the road. Before that, there is a need to understand vehicle automation and how it is currently employed in commercial vehicles. In this subsection, the definition of the level of automation and its relevance to distraction detection on the different levels of automation is discussed.

7.1.1. Currently available system

While many companies produce eye- and face-tracking systems, few car manufacturers apply such systems for driver monitoring purposes. To the best of our knowledge, three car manufacturers use in-vehicle driver-facing cameras to assess distraction: Lexus, DS Automobiles, and General Motor.

DS Automobiles introduced Driver Attention Monitoring (Automobiles, 2021), which performs tracking on the driver's face to detect signs of fatigue or inattention. Lexus's Driver Attention Monitor (Lexus, 2021) uses cameras placed between the steering wheel and the dashboard to monitor the driver's gaze. Recently, GM introduced their detection system within their partially autonomous highway driving system. The system is named Super Cruise (General Motors Corporate Newsroom, 2021), which operates at Level-2 and requires the driver to be engaged in monitoring the driving environment. The driver attention detection system, which uses an IR front-facing camera to monitor the driver's gaze direction, serves as an advisory device to ensure the driver performs the driving role at all times.

7.1.2. Vehicle automation definition

SAE International describes six levels of vehicle automation, ranging from "no automation" to "full automation" (On-Road Automated Driving (ORAD) committee, 2016).

- **Level 0: No automation**. A human driver is required to perform all aspects of the dynamic driving task.
- **Level 1: Driver assistance**. The automated system controls either the steering wheel or the acceleration/deceleration (but not both) of the vehicle in specific driving modes. The human driver monitors the driving environment and performs everything else involved in the dynamic driving task.
- **Level 2: Partial automation**. A driver is a necessity but is not required to monitor the environment at all times.
- **Level 3: Conditional automation**. The automated system performs everything involved in the dynamic driving task during a specific driving mode. However, the automated system can still fall back on the human to take control if needed, but it should allow some time between the request for intervention and the critical situation that requires the intervention.
- **Level 4: High automation**. The ability of the automated system to carry out the dynamic driving task even the human does not respond to a request to take control of the vehicle.
- **Level 5: Full automation**: The highest level of automation, where the automated system is expected to carry out the entire dynamic driving task during all driving modes in which a human driver could.

Currently, Level 2 cars are available at the mass on the road. With the mass evolution to the next automation level (Level 3), the system shall take over the driving task from the human driver in some scenarios and drive autonomously. In Level 3, a human driver is still required to take back control when the automation fails. Automation allows the changes in vehicle design, such as integrated interior, technology, and service design (Detjen et al., 2021).

7.1.3. On the relevance of distraction detection on different level of vehicle automation

In Level 0 vehicles, none of the driving tasks is automated, and the driver is expected to be fully controlling the vehicle at all times. Therefore, the distraction detection methods reviewed above, which are usually investigated in Level 0 vehicles, can be directly applied to vehicles at this level of automation.

However, the implementation of distraction detection systems becomes nuanced as soon as some parts of the driving tasks are automated (for Level 1 and 2 vehicles). Thus, indicators used to detect distraction that relies on specific driving performance measurements (vehicle IMU, CAN data, etc.) become irrelevant since some of the driving tasks are automated. Some of the works discussed before use the measurement of lateral control over the vehicle to assess the degree of driver distraction. Besides, some commercially available vehicle has included systems that can assess distraction or fatigue based on signs such as drifting outside the current lane. These measurements become uninformative when the automated system takes over lateral control of the vehicle. Therefore, the usage of other sensors should be deployed to collect the driver's state. Generally, distraction detection at Levels 1 and 2 has little to no difference from Level 0 in terms of the consequences of distraction. The main challenge for Level 1 and 2 vehicle distraction detection is the added complexity for the system since automation can now obscure sources of information critical to its detection. The usage of visual and psychological sensors are much more suitable for these level of a vehicle.

Starting at Level 3 vehicles, the driver is not expected to be actively attending to the driving environment at all times. While the system may request human intervention, the vehicle system should be given the driver adequate time to attend to the driving task. This differs from the immediate intervention required for lower-level vehicles besides the need for attention at all times. Therefore, Level 3 vehicles can tolerate distraction. Hence, the main challenge arises if a driver could react to the request in the given time to intervene in the driving task. This means the monitoring system in the vehicle may need to look for different signs to detect distractions. For example, the driver should still be in the driver's seat even when the automated system requests no intervention.

Level 4 and 5 vehicles are designed with capabilities to control the entire driving task without any intervention from the human. Thus, distraction detection should find no application in these levels of vehicles.

The distraction detection system is only suitable for vehicles in Level 0 to Level 2. As vehicles move to a higher level of automation, distraction detection becomes obsolete. A whole new research direction — transfer of control is more suitable for these vehicles. Transfer of control studies looks into transferring control of the vehicle from an automated system to the human. Ultimately, it has been shown that automation has negatively correlated with drivers' attention to the road ahead (Carsten et al., 2012). Besides, the automated system can reduce the demand for the driver's cognitive resources, allowing the driver to be in a mental underload condition (Johns et al., 2015).

Currently, Level 0, Level 1 and Level 2 vehicles are still largely on the road. Therefore, a distraction detection system is still relevant as the number of fatal crashes is expected to increase if such an early warning preventive system is not being adopted. Even if the vehicles are almost fully autonomous, some sort of human attention should still be needed. Ultimately, a human driver has vast knowledge and response to unknown cases, while models are trained to act according to what is being "taught". Studies reveal that drivers who are involved in other tasks (being distracted) are less aware of road hazard (Zangi et al., 2022). If the vehicle wants to transfer the task back to the driver, the driver might not have enough reaction time to react accordingly, which would eventually cause fatal crashes. Therefore, we believe that such a system is still needed under partial autonomous vehicles (Level 1 and Level 2).

7.2. Simulated and naturalistic driving environment

Many works have illustrated its relatively high accuracy yet efficiency in terms of detecting driver distraction. However, such results might only be valid for a specific scenario. For example, models trained and tested on the publicly available visual dataset, such as StateFarm and AUC-DDD, might not apply to real-world applications. Note that almost all datasets are collected in a simulated driving condition to safeguard the participant when collecting distracting actions. This, therefore, produces one of the most challenging tasks in ML — validate the proposed algorithm with the wide variety of driving scenarios (Velez and Otaegui, 2015). To integrate such an algorithm into ADAS, some of the key points need to be addressed.

"Naturalistic driving" refers to driving that is not constrained by strict experimental design. Several hardware and software algorithms are being developed mainly in simulated environments rather than naturalistic driving to monitor the driver and the driving behaviour. This is because of the possible danger of collecting distraction data in naturalistic driving settings (Sahayadhas et al., 2012). The usage of the simulator presented several advantages, such as the experimental control, efficiency, safety, and ease of data collection (Konstantopoulos et al., 2010). It is shown that driving simulators can create a driving environment relatively similar to real-world scenarios (Mayhew et al., 2011; Johnson et al., 2011; Auberlet et al., 2012). However, some considerations need to be taken since the simulator could produce contradictory results. One common issue is that real-world danger and its consequences do not reflect in a driving simulator, giving rise to a false sense of safety, responsibility, or competence (De Winter et al., 2012).

A study on distraction in both simulated and real environments found that the driver's physiological activity showed a significant difference in different settings (Engström et al., 2005). The authors found that physiological workload and steering activity was much higher under real driving conditions than in simulated environments. Bach et al. (2008) found that controlled driving yielded more frequent and more prolonged eye glances than the simulated driving setting, and driving errors were more common in simulated driving. Östlund et al. (2006) found that the driver's heart rate changed significantly while performing the visual task in real-world driving compared to the baseline condition, suggesting that visual task performance in actual driving was more stressful.

Since most of the studies are carried out in a simulator, the study should be validated and tested on actual driving conditions since there are many external factors that the simulator did not consider. Illuminations, surrounding noise, and changing backgrounds are challenges when applying to an actual driving situation. Besides, a helpful system shall adapt to the variability of every end-user and detect in real-time. Therefore, it is vital to ensure that the driving simulation settings are close to the actual driving situation. A public dataset covering simulated and natural driving environments should be made available soon.

7.3. Limited dataset availability

There is an increasing need for multiple large datasets to undertake a successful training process. We observed that the go-to dataset is either StateFarm or AUC-DDD for distraction recognition, while other specific distraction detection methods usually collected their dataset. Many of the datasets are not large-scale, in terms of classes and instances that could impact the training accuracy of ML algorithms. The available dataset is usually small with limited variability in environments. However, we have observed that the most recent datasets started to collect many datasets with many more actions considered as distractions, besides having multiple modalities. As discussed in the previous subsection, the collection of driver distraction datasets might not help accelerate the research since the paradigm shift of vehicle automation. A better approach is to collect a large-scale naturalistic driving dataset, compromising the road condition, driver behaviour, and vehicle metrics, allowing scholars to propose a better system for autonomous vehicles.

7.4. Privacy issue involved in driving data

One key challenge for vehicle and driving data exploitation is how to safeguard the privacy of the driver (Kaiser et al., 2018). Placing a camera inside a car imposes significant privacy concerns that might deter individuals from using the application (Streiffer et al., 2017). Despite general agreement that intelligent vehicles would increase safety, studying driver behaviour to assess intelligent vehicles required a massive amount of naturalistic driving data. However, there is a scarcity of publicly available naturalistic driving data in the current literature due to individual privacy. It should also be emphasized that a real-time visual-based distraction detection system does not necessarily require the video stream to be saved. As a result, privacy concerns are particularly prevalent in research projects in which video feed is recorded and kept for subsequent analysis.

"De-identification" is used to describe the process of protecting an individual's privacy in a video sequence (Newton et al., 2005). Although this will help safeguard drivers' identities, it will obstruct the goal of sensorizing cars to regulate both drivers and their behaviour. In an ideal world, a de-identification system would safeguard drivers' identities while keeping enough information to deduce their actions (e.g., eye gaze, head pose, or hand activity) (Martin et al., 2014c).

Martin et al. (2014b) used de-identification filters to protect the privacy of drivers while preserving sufficient details to infer their behaviour. One de-identification filter is used to preserve the mouth region of the driver to monitor yawning or talking actions, and the other de-identification filter is used to preserve the eye regions of the driver to detect fatigue or gaze direction. Specifically, they implemented and compared de-identification filters made up of a combination of preserved eye regions for fine gaze estimation, superimposing head pose encoded face masks for spatial context, and replacing the background with black pixels to ensure privacy protection. The study revealed that gaze zone estimation accuracies are 65%, 71%, and 85% for One-Eye, Two-Eyes, and Mask with Two-Eyes, respectively.

Baragchizadeh et al. (2017) evaluated the effectiveness of the personalized supervised bilinear regression method for Facial Action Transfer (FAT) (Huang and De La Torre, 2012) de-identification algorithm. In their work, de-identification was examined for driver videos taken from Head Pose Validation (HPV) dataset modelled on the SHRP2-NDS data. Specifically, this work aimed to determine if a driver could be recognized after the face was masked. The experiment showed that the algorithms substantially reduced the accuracy of human identification of drivers. A follow-up study was performed in Orsten-Hooge et al. (2019).

Most of the privacy issues evolved around visual data, where the driver's face is recorded and detected. As for other data, such as vehicle and physiological data, although they pose some concern over the exposure of personal data or behaviour, these data could be easily masked. Compared to the visual dataset, masking a human's eye would ultimately cause a reduction in accuracy where eye gaze location and eyelid is one of the dominant features to determine if a driver is looking away from the road. While masking other than the eye would be possible, but usually not applied since the eye is the most prominent feature to identify a person. Blurring the face is possible, but that would confuse the ML model even more.

7.5. Challenges in distraction detection using visual data

Besides privacy issues, the main challenge faced by visual data is the variation of lighting conditions. Changes in illumination occur frequently depending on the weather, day of time, and driving location of the vehicle (e.g., entering tunnels, high-rise buildings, etc.). Ideally, we want the dataset to encapsulate all possible driving scenarios, including daytime and nighttime. If a system is built to detect facial features and hand location to determine distraction, various facial occlusions can occur due to hand activities. Considering the highly deformable nature of hands, they change in appearance frequently. Thus, detection and tracking of hands for recognition of driver distraction actions is a challenging problem.

Different viewpoints should be considered when collecting the dataset. Many datasets only collect images from a single viewpoint, which is not ideal for real-life implementation. The model will not be able to regularize well when deploying on a different viewpoint. This poses a challenge when implementation, where the exact viewpoint should be set up to ensure correct classification.

7.6. Challenges of DL architecture

The rapid development of DL architectures, including hybrid learning and transfer learning, is a recent trend in ML for driver distraction detection. Many of the studies have shown promising results with the DL algorithm. However, to detect distractions in real-time, the DL architectures face the challenges of computational cost, the complexity of the network, and system performance. Therefore, a robust DL-based architecture is necessary to develop for feature extraction and classification tasks.

As illustrated in Fig. 12, we observed that many works leverage ResNet, VGG, and Inception. Thanks to the widely available pretrained networks and adoption in almost every framework, these models are the go-to for classification tasks. However, the reported accuracy is not always reliable, as "cheating" might occur if the same image is present in training and testing datasets. Besides, when hyperparameters and augmentations are not clearly stated in the articles, we could not validate if the claims are correct. Moreover, many studies did not include the computational cost in terms of floating-point operations (FLOPs) and inference time.

In terms of a multimodal dataset, the most commonly used technique for incorporating knowledge is feature-level fusion. The main advantage of using early fusion is that it exploits the similarity between the modalities at an early stage. However, time synchronization between feature sets is a significant constraint of the function level fusion, as the data is collected at various rates. Most works either utilized feature-level fusion or decision fusion, hence losing the chance to fuse rich representations of mid-level features available in a CNN-based architecture. Thus, a new deep architecture with fusion frameworks needs to be explored to resolve the shortcomings mentioned above and exploit various fusion strategies.

Generally, DL is extremely data-hungry, considering it also involves representation learning (Karimi et al., 2019). DL requires a large amount of data to achieve a sub-par performance model, which as the data increases, a more optimum performance model can be achieved (Alzubaidi et al., 2021). One of the solutions is to perform transfer learning, as many of the articles discussed above utilized. Note that while the transferred data will not directly augment the actual data, it will help enhance the original input representation of data and its mapping function.

A balanced dataset is required to allow the DL model to perform accurately. Usually, negative samples (normal driving) are much more than positive samples (distracted actions) due to the challenges in collecting distracting images. Therefore, it is necessary to employ the correct criteria for evaluating the loss and the prediction result. The model should perform well in small classes as well as larger ones. The model should employ AUC as the resultant loss as well as the criteria (Wang et al., 2015b) to ensure a fair evaluation of the proposed model.

7.7. Model performance

While almost all papers promised good performance using various ML or DL models. However, the underlying support for this performance is little been known. This is because most of the work only validates their model on a very small dataset, which might not be enough to picture the real-world scenario. Besides, when a random shuffle is used to split the dataset into train, validation and test set, the metrics might no longer be a good measure of its generalization ability. An easy test would be to give unseen data to the model and ask for its prediction. The model which does not generalize well from observed data to unseen data is known as overfitting (Ying, 2019b). Underfitting, on the other hand, tends to not capture the detail of the dataset, leading to lower accuracy. Because of the existence of overfitting, the model performs perfectly on the training set or sometimes the whole dataset if the split was not done correctly while fitting poorly on unseen data. This is due to the over-fitted model having difficulty coping with pieces of the information in the unseen data, which may be different from those in the training set. Typically, over-fitted models tend to memorize all the data, including unavoidable noise on the training set, instead of learning the discipline hidden behind the data. To avoid overfitting, several techniques could be applied. Among them, early stopping, network reduction, expansion of dataset, and regulation are the most common technique to mitigate overfitting. We refer the reader to the work by Ying (2019b) for further explanation of the techniques.

In training a supervised ML model, the objective is to achieve good generalization performance on unseen data (Cunningham and Delany, 2020). This is important for a model to be trustworthy.

Besides overfitting and underfitting, gradient vanishing and exploding is one of the threats to ML model due to the nonlinear character of the activation function. The two commonly used activation functions — sigmoid function and ReLU are highly vulnerable to gradient vanishing and exploding. The sigmoid function is vulnerable to the vanishing gradient problem, while ReLU has a special vanishing gradient problem that is called dying ReLU problem (Hu et al., 2021). Besides, the shattered gradient problem is one of the obstacles in obtaining high-performance model (Balduzzi et al., 2017).

While using a state-of-the-art DL model in classifying images is the trend now, beware that there is a high tendency of being over parameterized. Many DL models are trained on a large-scale dataset, and therefore requires more parameters to represent them. When porting them to a smaller dataset, they might be redundant, which brings lower accuracy. Therefore, the usage of model compression techniques, knowledge distillation, as well as network architecture search could be the solution to this issue.

7.8. Future works

As the automotive industry is looking to shift its mass production from Level-2 vehicles to Level-3 vehicles, many issues are not adequately addressed. We split this subsection into several branches, ranging from data collection strategy, and inferencing strategy to technological advancement. Then, we provide some research gap outlook in the field.

7.8.1. Naturalistic driving study

Naturalistic driving study (NDS) is generally a type of study that systematically collects video, audio, vehicle telemetry, and other sensor data that captures various aspects of driving for long periods (Fridman et al., 2019). The collected data are usually acquired as close to real-world conditions as possible, which drivers typically drive "in the wild". Often, a driver's vehicle is instrumented, and the driver is asked to continue using their vehicle as they ordinarily would. This, therefore, solves the simulated environment's first issue — unpredicted danger. The collection of data in real-world conditions would allow the ML system to model and learn as close to natural conditions as possible.

Although there is a limited number of NDS to date, they have been conducted across different countries (e.g., United States, Europe, Canada, China, Japan, and Australia). Currently only three big-scale studies, HARMONY (Tavakoli et al., 2021b), MIT-AVT (Fridman et al., 2019) and ANDS (Williamson et al., 2015) are still ongoing, while others have been halted or finished (100-car naturalistic study (Neale et al., 2005), SHRP2 (Campbell, 2012; Dingus et al., 2015), UDRIVE (Barnard et al., 2016), Shanghai NDS (Zhu et al., 2018), Canadian NDS (Hankey et al., 2014), Japanese NDS (Uchida et al., 2010), Candrive (Marshall et al., 2013b,a)). Since such studies usually involve a lot of drivers, and each driver is usually compensated with stipends, therefore not much research groups are interested in investing in the projects. Besides, the equipment required to record those data are expensive.

Most of the NDS driver sensing is solely conducted through features and behaviours extracted through in-cabin videos. Therefore, to better understand driver behaviour, which could later be used to infer distractions, more modalities of data should be collected. We summarize some of the future improvements when collecting the NDS dataset as follows:

1. The dataset should consist of data inside and outside the cabin and the driver's physiological state, vehicle state, and location data. These can be collected through smartwatch usage to collect physiological data, camera, vehicle IMU, CAN, and smartphone to collect location data.
2. Asynchronous Sensor Recording. Recording all sensors data streams so that each data sample is timestamped using a centralized, reliable time-keeper.
3. Expand the data collection throughout the country. This will aid in modelling a more accurate model for the country.
4. Diversification of participants (age, gender, race, socio-economic level, etc.) to better capture individual differences.
5. Inclusion of semi-automated vehicles such as commercially available TESLA vehicles in the data collection process.

7.8.2. Edge intelligence

We observed that DL is preferred over ML for a specific detection algorithm, especially when dealing with image data. However, the computational requirements for DL are huge, and therefore, DL models are typically trained on GPUs. Even after finishing the model training process, computational requirements for inference on unseen data remain high. Many applications are currently looking toward cloud computing, where the computation is done on remote computing infrastructure with the necessary processing power. However, using cloud infrastructure as the centralized processing server increases the frequency of communication between user devices and the geographically distant data centres (Plastiras et al., 2018). This is the main limiting factor for real-time distraction detection; higher bandwidth is required to send full-length data to the server, thus increasing the latency.

The combination of AI and edge computing is expected since there is an intersection between them. Specifically, edge computing targets to coordinate multiple collaborative edge devices and servers to process the generated data in proximity, while AI strives for simulating intelligent human behaviour in devices/machines by learning from data (Zhou et al., 2019). Note that there are differences between edge computing and edge intelligence. When data is processed physically close to where the data is produced, we refer to it as edge computing (Garcia Lopez et al., 2015). When data is acquired, stored, and processed with ML algorithms at the network edge, we refer to it as edge intelligence (International Electrotechnical Commission, 2017).

The main difference between cloud- and edge-based systems is the processing of raw data. In the cloud paradigm, the data needs to be transferred to the remote infrastructure for processing to produce the prediction and stored in a database. For the edge paradigm, all of those are done offline, except for storing data in the database. Therefore, we can immediately see the benefit in terms of performance and latency.

Putting in driver distraction context, sensors from the camera, CAN, IMU, etc., can be gathered in-vehicle and processed on edge devices. Thus, for practical deployment of NN on mobile devices, it is necessary to have low complexity models that can run on embedded processors. The proposed algorithms need to leverage the fixed-point operations to speed up the prediction process. Moreover, edge intelligence devices usually have limited resources in terms of computation and memory for storage and data access. ML algorithms often require storing and accessing several parameters that describe the model architecture and weight values that form the classification model. With a deep model, more memory is required to retrieve the weight and parameters. Therefore, one of the challenges for deploying an ML algorithm on a resource constraint device is to reduce the memory access and keep the data locally to avoid costly reads and writes.

We summarize some of the future improvements when proposing ML techniques for edge devices:

1. Model compression. Balancing resource-hungry ML models and resource-poor end devices is vital to reducing the model complexity and resource requirement, enabling local and fast inference. Weight pruning (Han et al., 2015a,b) is one of the widely used techniques to compress models.
2. Model partitioning and model early exit. Model partition is used to distribute the computational-intensive part to the edge server or nearby devices, while the model early exit is used to leverage output data of the early layer to get the predictions.
3. Multitasking support. ADAS usually have multiple model operating simultaneously for vehicle detection, pedestrian detection, traffic sign recognition, driver distraction detection, lane line detection, etc. Thus, multiple models would compete for the limited resource. Therefore, careful design should be made for multitasking.
4. Selection of features. While more features warrant higher accuracy, selecting the most prominent feature for detection recognition is more realistic for resource-constraint edge devices.

7.8.3. Model hybridizing and optimization

While many time-series data can achieve high accuracy with either traditional ML or CNN algorithm, they might not be efficient nor robust. It was shown that the performance of ML could be improved through hybridization with other ML methods and a more robust and efficient model can be obtained (Mosavi et al., 2018). Specifically, hybridization aims to combine and optimize different knowledge schemes and learning strategies to solve a computational task. This is especially useful for multimodal data, whereby the knowledge from different modalities could be hybridized in order to produce a more accurate predictions. As opposed to transform different modalities data into a single common modal, the process of conforming the data has induced extra losses. Therefore, we believe that hybridization could be explored in the future works, especially those using data from combination of sensors.

As for optimization techniques, the goal is to find a more efficient and cost-effective procedures to produce an optimal solution (Abbaszadeh Shahri et al., 2021). This applies to all ML models. Putting aside the common optimization techniques in training DL, we focus on how to further optimize the model in terms of efficiency and robustness. Specifically, there are many models that are large in size, and therefore model compression techniques could be used to compress the model. The usage of knowledge distillation and neural architecture search are some of the commonly used techniques to compress a given models.

7.8.4. Uncertainty quantification and sensitivity analysis

Uncertainty arises when the test and training data are mismatched, while data uncertainty occurs when class overlap or due to the presence of noise in the data (Phan, 2019). This is especially true when dealing with physiological data and vehicle telemetric data. Therefore, one way to mitigate this is through uncertainty quantification (UQ). Predictions made without UQ are usually not trustworthy (Abdar et al., 2021).

UQ methods play an important role in reducing the impact of uncertainties during both optimization and decision-making processes. It covers different dimensions of uncertainty and aims to enhance the model's reliability by producing the output in a probabilistic framework, where a confidence interval can be placed to estimate the model's robustness. Currently, two commonly used UQ methods are Bayesian approximation and ensemble learning techniques (Abbaszadeh Shahri et al., 2022; Volodina and Challenor, 2021).

Here, we identified several uncertainties that should be quantified:

1. The selection of sensors and collection of the raw data. The selected sensors should be operated in the specified optimum working condition as stated in the datasheet of the sensor when collecting raw data.
2. The completeness of the data. Collected data should be of the same condition for all scenarios or subjects involved. For example, the placing of physiological sensors are the same places for all involved subjects and recorded with the same condition. Data that are not tallied, such as different lengths of timestamps, should be discarded.
3. The understanding of DL or ML model with the performance bounds and limitations.
4. The uncertainties corresponding to the performance of the model based on operational data.

However, the usage of UQ in driver distraction detection has yet to be identified. Therefore, the usage of UQ should be applied in future works to build a trustworthy model.

7.8.5. Vision transformer

Transformer (Vaswani et al., 2017) offers an alternative approach to solving vision tasks. A Transformer is primarily made up of self-attention blocks and allows us to leverage specific information relevance. Interestingly, multi-head self-attention (MSA) layers work like a convolution layer (Cordonnier et al., 2019). Thanks to the flexibility of the Transformer, it can maintain a long-range relationship. However, this incurs much higher computational costs.

Despite the excellent performance of Transformer models and their interesting salient features, several challenges are associated with their applicability to practical settings. The main bottlenecks include the requirement of a large training dataset and high computational costs. There have also been some challenges to visualize and interpret Transformer models.

The usage of vision Transformers in driver distraction detection is not widely explored yet. We only identified one article related to the field (Koay et al., 2021a). Therefore, we hope to see more articles exploring the usage of Vision Transformers in detecting driver distraction.

7.8.6. Research gap

After having reviewed as many relevant articles as been published, we conclude that there is still a significant research gap to implement ML architecture for driver distraction detection. We have summarized the future improvement for implementing ML in driver distraction detection. We use "detection" and "recognition" to represent binary class (distracted vs. non-distracted) classification and multi-class (different actions deemed to be distracted actions) classification, respectively.

1. Description of hyperparameters should be well-defined. Some of the articles did not specify the training procedure. Therefore other scholars may not be able to replicate the exact studies and validate the claim. At the bare minimum, the training epoch, learning rate, optimizer, batch size, and framework used to train the model should be described in the articles.
2. Usage of a fair split between training and testing datasets. While randomly splitting the dataset to train, test, and validation datasets is the common practice, this practice should not be implemented in mission-critical tasks, which in this case, distraction recognition. A better approach is to split the dataset by driver/subject to ensure the model did not learn unwanted features from the dataset.
3. Data augmentation should be deployed with care. Although deploying augmentation techniques such as flipping helps the model generalize well, these images would never appear in real life. Augmentation techniques should be used with care because some of the augmented images might not be realistic, even though it helps to improve the accuracy. However, there is no research to validate if some augmentation would hurt the classification accuracy in real life.
4. Different kinds of DL architecture should be investigated, such as few-shot learning and semantic-based learning (SBL). This kind of model can significantly improve identification accuracy with fewer training samples and help classify unseen image classes well. From our observation, there is little research on this area.
5. Selection of multimodal features and fusion approach is critical in classifying distractions. When there are too many features to select from, a deep analysis of every feature should be done to select only the most helpful feature for model training. While more features do not usually bring higher accuracy, optimizing and selecting the most prominent features simplifies the model training and its parameters and paves the way for edge device prediction.
6. Visual features, such as PERCLOS and mouth opening, and physiological features that look into the anomaly in data, should be investigated with care. This is because the individual response to distractions differs for everyone, especially when the involved participant is in a small amount.
7. Usage of non-intrusive devices to collect physiological data. With the emergence of smart health devices, such as a smartwatch, which can collect heart data, physiological data can be carried out non-intrusively. However, we see only a few articles leverage this technology.

8. Real-time distraction detection in real driving conditions is not being studied yet. After developing the algorithm, the next step should have been testing the algorithm in real-world settings. This would help to validate if the algorithm could generalize well in non-seen data.
9. Combining both spatial and temporal data in determining distraction. Such techniques are not widely adopted, and we believe there are insights into spatial and temporal data.
10. Usage of model compression to produce a smaller and more accurate detection model.
11. Quantifying and performing sensitivity and uncertainty analysis with the collected data. This analysis helps determine the effectiveness of input parameters on produced outputs. These methods help find a simplified yet robust calibrated model from a large number of parameters and identifying important connections between observations and model output.

8. Conclusion

Driver distraction detection and driver monitoring systems could help mitigate some of the dangerous consequences to the driver and other road users. Most road accidents could be prevented if a driver is not being distracted. In this paper, a comprehensive review of the scientific literature on distracted driving detection from 2014–2021 was presented. We started first to understand the causes of distraction and how sensors could be deployed to detect such distraction. We evaluated various sensors used in the literature and grouped them into three categories: physiological, visual, and external sensors. We presented a summary of the publicly available datasets for each group of sensors and multimodal datasets, compromising two or more different categories of the sensor. Then, we provide a general review of ML techniques and the concept of transfer learning. We generalized the ML techniques into traditional ML, CNN, and RNN, tailored to the distraction detection field. It was observed that traditional ML was extensively used in modelling non-visual data, while CNN and RNN were commonly used for visual data.

In this article, we divided the works based on the sensors used since they provided more systematic ways of understanding the advancement in the field. We observed that most of the studies we reviewed are from visual sensors. This is due to the nature of visual data since it provides images of drivers, which could be easily inferred if a driver is distracted. Besides, we also discussed various patents related to distraction detection by various institutions. For comprehensiveness purposes, we include the evaluation of distraction and evaluation metrics used for ML models. We provide insights on implementing the distraction detection system in commercially available vehicles and the relevance of such a system for a different level of vehicle automation. Furthermore, we look into the issues related to visual images, the differences between simulated and naturalistic environments, challenges of DL architectures and provide a general outlook of future works in this field.

We hope future work could better emphasize large-scale naturalistic driving data to model driver behaviours and other road users. Besides, as we are moving towards the autonomous vehicle era, more factors could harm the driver. Therefore, distraction detection alone is not enough for future proof. We believe that the future work would be more varied, with much work shifted to focus more on enhancing autonomous vehicle safety.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbas, Q., Alsheddy, A., 2021. A methodological review on prediction of multi-stage hypovigilance detection systems using multimodal features. IEEE Access 9, 47530–47564.

Abbaszadeh Shahri, A., Pashamohammadi, F., Asheghi, R., Abbaszadeh Shahri, H., 2021. Automated intelligent hybrid computing schemes to predict blasting induced ground vibration. Eng. Comput. 1–15.

Abbaszadeh Shahri, A., Shan, C., Larsson, S., 2022. A novel approach to uncertainty quantification in groundwater table modeling by automated predictive deep learning. Nat. Resour. Res. 1–23.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf. Fusion 76, 243–297.

Abou Elassad, Z.E., Mousannif, H., Al Moatassime, H., Karkouch, A., 2020. The application of machine learning techniques for driving behavior analysis: A conceptual framework and a systematic literature review. Eng. Appl. Artif. Intell. 87, 103312.

Abouelnaga, Y., Eraqi, H.M., Moustafa, M.N., 2017. Real-time distracted driver posture classification. arXiv preprint arXiv:1706.09498.

Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., Hariri, B., 2014. YawDD: A yawning detection dataset. In: Proceedings of the 5th ACM Multimedia Systems Conference. pp. 24–28.

Abut, H., Erdoğan, H., Erçil, A., Çürüklü, B., Koman, H.C., Taş, F., Argunşah, A.O., Coşar, S., Akan, B., Karabalkan, H., et al., 2009. Real-world data collection with "UYANIK". In: In-Vehicle Corpus and Signal Processing for Driver Behavior. Springer, pp. 23–43.

Aksjonov, A., Nedoma, P., Vodovozov, V., Petlenkov, E., Herrmann, M., 2017. A method of driver distraction evaluation using fuzzy logic: Phone usage as a driver's secondary activity: Case study. In: 2017 XXVI International Conference on Information, Communication and Automation Technologies (ICAT). IEEE, pp. 1–6.

Aksjonov, A., Nedoma, P., Vodovozov, V., Petlenkov, E., Herrmann, M., 2018. Detection and evaluation of driver distraction using machine learning and fuzzy logic. IEEE Trans. Intell. Transp. Syst. 20 (6), 2048–2059.

Ali, S.F., Hassan, M.T., 2018. Feature based techniques for a driver's distraction detection using supervised learning algorithms based on fixed monocular video camera. KSII Trans. Internet Inf. Syst. (TIIS) 12 (8), 3820–3841.

Ali, J., Khan, R., Ahmad, N., Maqsood, I., 2012. Random forests and decision trees. Int. J. Comput. Sci. Issues (IJCSI) 9 (5), 272.

Alizadeh, V., Dehzangi, O., 2016. The impact of secondary tasks on drivers during naturalistic driving: Analysis of EEG dynamics. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 2493–2499.

Alkinani, M.H., Khan, W.Z., Arshad, Q., 2020. Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges. IEEE Access 8, 105008–105030.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A., Asari, V.K., 2019. A state-of-the-art survey on deep learning theory and architectures. Electronics 8 (3), 292.

Alotaibi, M., Alotaibi, B., 2019. Distracted driver classification using deep learning. Signal Image Video Process. 1–8.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8 (1), 1–74.

Artan, Y., Bulan, O., Loce, R.P., Paul, P., 2014. Driver cell phone usage detection from HOV/HOT NIR images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 225–230.

Auberlet, J.-M., Rosey, F., Anceaux, F., Aubin, S., Briand, P., Pacaux, M.-P., Plainchault, P., 2012. The impact of perceptual treatments on driver's behavior: From driving simulator studies to field tests—First results. Accid. Anal. Prev. 45, 91–98.

Automobiles, D., 2021. DS driver attention monitoring. Accessed: Oct. 5, 2021. URL https://www.dsautomobiles.co.uk/inside-ds/ds-news/ds-automobiles-anti-fatigue-technology.

Azim, T., Jaffar, M.A., Mirza, A.M., 2014. Fully automated real time fatigue detection of drivers through fuzzy expert systems. Appl. Soft Comput. 18, 25–38.

Azman, A., Ibrahim, S.Z., Meng, Q., Edirisinghe, E.A., 2014. Physiological measurement used in real time experiment to detect driver cognitive distraction. In: 2014 International Conference on Electronics, Information and Communications (ICEIC). IEEE, pp. 1–5.

Bach, K.M., Jæger, M.G., Skov, M.B., Thomassen, N.G., 2008. Evaluating driver attention and driving behaviour: comparing controlled driving and simulated driving. In: People and Computers XXII Culture, Creativity, Interaction 22. pp. 193–201.

Baheti, B., Gajre, S., Talbar, S., 2018. Detection of distracted driver using convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1032–1038.

Baheti, B., Talbar, S., Gajre, S., 2020. Towards computationally efficient and realtime distracted driver detection with mobilevgg network. IEEE Trans. Intell. Veh. 5 (4), 565–574.

Balakrishnama, S., Ganapathiraju, A., 1998. Linear discriminant analysis-a brief tutorial. Inst. Signal Inf. Process. 18 (1998), 1–8.

Balamurugan, M., Kalaiarasi, R., 2021. Dimensionally improved residual neural network to detect driver distraction in real time. In: Journal of Physics: Conference Series, Vol. 1964. IOP Publishing, 042037.

Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K.W.-D., McWilliams, B., 2017. The shattered gradients problem: If resnets are the answer, then what is the question? In: International Conference on Machine Learning. PMLR, pp. 342–350.

Baragchizadeh, A., Karnowski, T.P., Bolme, D.S., O'Toole, A.J., 2017. Evaluation of automated identity masking method (AIM) in naturalistic driving study (NDS). In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, pp. 378–385.

Barnard, Y., Utesch, F., van Nes, N., Eenink, R., Baumann, M., 2016. The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. Eur. Transp. Res. Rev. 8 (2), 14.

Behera, A., Keidel, A., Debnath, B., 2018. Context-driven multi-stream LSTM (M-LSTM) for recognizing fine-grained activity of drivers. In: German Conference on Pattern Recognition. Springer, pp. 298–314.

Behera, A., Wharton, Z., Keidel, A., Debnath, B., 2020. Deep CNN, body pose and body-object interaction features for drivers' activity monitoring. IEEE Trans. Intell. Transp. Syst..

Belkin, M., Niyogi, P., Sindhwani, V., 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. 7 (11).

Benedek, M., Kaernbach, C., 2010. A continuous measure of phasic electrodermal activity. J. Neurosci. Methods 190 (1), 80–91.

Berri, R.A., Silva, A.G., Parpinelli, R.S., Girardi, E., Arthur, R., 2014. A pattern recognition system for detecting use of mobile phones while driving. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Vol. 2. IEEE, pp. 411–418.

Billah, T., Rahman, S.M., 2016. Tracking-based detection of driving distraction from vehicular interior video. In: 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 423–428.

Billah, T., Rahman, S.M., Ahmad, M.O., Swamy, M., 2018. Recognizing distractions for assistive driving by tracking body parts. IEEE Trans. Circuits Syst. Video Technol. 29 (4), 1048–1062.

Bingham, C.R., Bao, S., Flannagan, C., Pradhan, A.K., et al., 2016. Using Naturalistic Driving Performance Data to Develop an Empirically Defined Model of Distracted Driving. Tech. rep., Nextrans.

Bo, C., Jian, X., Jung, T., Han, J., Li, X.-Y., Mao, X., Wang, Y., 2016. Detecting driver's smartphone usage via nonintrusively sensing driving dynamics. IEEE Internet Things J. 4 (2), 340–350.

Borghi, G., Frigieri, E., Vezzani, R., Cucchiara, R., 2018. Hands on the wheel: a dataset for driver hand detection and tracking. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, pp. 564–570.

Borghi, G., Venturelli, M., Vezzani, R., Cucchiara, R., 2017. Poseidon: Face-from-depth for driver pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4661–4670.

Braithwaite, J.J., Watson, D.G., Jones, R., Rowe, M., 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology 49 (1), 1017–1034.

Braunagel, C., Kasneci, E., Stolzmann, W., Rosenstiel, W., 2015. Driver-activity recognition in the context of conditionally autonomous driving. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE, pp. 1652–1657.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Caird, J.K., Willness, C.R., Steel, P., Scialfa, C., 2008. A meta-analysis of the effects of cell phones on driver performance. Accid. Anal. Prev. 40 (4), 1282–1293.

Campbell, K.L., 2012. The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. Tr News (282).

Cañas, P., Ortega, J.D., Nieto, M., Otaegui, O., 2021. Detection of distraction-related actions on DMD: An image and a video-based approach comparison.. In: VISIGRAPP (5: VISAPP). pp. 458–465.

Cao, Z., Chuang, C.-H., King, J.-K., Lin, C.-T., 2019. Multi-channel EEG recordings during a sustained-attention driving task. Sci. Data 6 (1), 1–8.

Carney, C., McGehee, D., Harland, K., Weiss, M., Raby, M., 2015. Using naturalistic driving data to assess the prevalence of environmental factors and driver behaviors in teen driver crashes.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308.

Carsten, O., Lai, F.C., Barnard, Y., Jamson, A.H., Merat, N., 2012. Control task substitution in semiautomated driving: Does it matter what aspects are automated? Hum. Factors 54 (5), 747–761.

Cattan, G., Rodrigues, P.L.C., Congedo, M., 2018. EEG alpha waves dataset. http://dx.doi.org/10.5281/zenodo.2348892.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chen, J., Jiang, Y., Huang, Z., Guo, X., Wu, B., Sun, L., Wu, T., 2021. Fine-grained detection of driver distraction based on neural architecture search. IEEE Trans. Intell. Transp. Syst..

Chen, S., Kuhn, M., Prettner, K., Bloom, D.E., 2019. The global macroeconomic burden of road injuries: estimates and projections for 166 countries. Lancet Planet. Health 3 (9), e390–e398.

Chhabra, R., Verma, S., Krishna, C.R., 2017. A survey on driver behavior detection techniques for intelligent transportation systems. In: 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence. IEEE, pp. 36–41.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Choi, I.-H., Hong, S.K., Kim, Y.-G., 2016. Real-time categorization of driver's gaze zone using the deep learning techniques. In: 2016 International Conference on Big Data and Smart Computing (BigComp). IEEE, pp. 143–148.

Chuang, M.-C., Bala, R., Bernal, E.A., Paul, P., Burry, A., 2014. Estimating gaze direction of vehicle drivers using a smartphone camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 165–170.

Chui, K.T., Alhalabi, W., Liu, R.W., 2019. Head motion coefficient-based algorithm for distracted driving detection. Data Technol. Appl..

Cordonnier, J.-B., Loukas, A., Jaggi, M., 2019. On the relationship between self-attention and convolutional layers. arXiv preprint arXiv:1911.03584.

Costa, M., Oliveira, D., Pinto, S., Tavares, A., 2019. Detecting driver's fatigue, distraction and activity using a non-intrusive ai-based monitoring system. J. Artif. Intell. Soft Comput. Res. 9 (4), 247–266.

Craye, C., Karray, F., 2015. Driver distraction detection and recognition using RGB-D sensor. arXiv preprint arXiv:1502.00250.

Cunningham, P., Delany, S.J., 2020. Underestimation bias and underfitting in machine learning. In: International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning. Springer, pp. 20–31.

Cunningham, P., Delany, S.J., 2020a. K-Nearest neighbour classifiers: (with Python examples). arXiv preprint arXiv:2004.04523.

Das, N., Ohn-Bar, E., Trivedi, M.M., 2015. On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. IEEE, pp. 2953–2958.

de Salis, E., Baumgartner, D.Y., Carrino, S., 2019. Can we predict driver distraction without driver psychophysiological state? a feasibility study on noninvasive distraction detection in manual driving. In: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings. pp. 194–198.

De Winter, J., van Leeuwen, P.M., Happee, R., et al., 2012. Advantages and disadvantages of driving simulators: A discussion. In: Proceedings of Measuring Behavior, Vol. 2012. Citeseer, p. 8th.

Dehzangi, O., Rajendra, V., 2019. Wearable galvanic skin response for characterization and identification of distraction during naturalistic driving. In: Advances in Body Area Networks I. Springer, pp. 15–27.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Deo, N., Trivedi, M.M., 2019. Looking at the driver/rider in autonomous vehicles to predict take-over readiness. IEEE Trans. Intell. Veh. 5 (1), 41–52.

Deruyck, J., McLaughlin, B.R., 2017. Detection of driver behaviors using in-vehicle systems and methods. US Patent 9, 714, 037.

Deshmukh, S.V., Dehzangi, O., 2019. Characterization and identification of driver distraction during naturalistic driving: an analysis of ECG dynamics. In: Advances in Body Area Networks I. Springer, pp. 1–13.

Detjen, H., Faltaous, S., Pfleging, B., Geisler, S., Schneegass, S., 2021. How to increase automated vehicles' acceptance through in-vehicle interaction design: A review. Int. J. Human–Comput. Interact. 37 (4), 308–330.

Diaz-Chito, K., Hernández-Sabaté, A., López, A.M., 2016. A reduced feature set for driver head pose estimation. Appl. Soft Comput. 45, 98–107.

Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., McGraw, D., Perez, M., Stowe, L., 2015. Naturalistic Driving Study: Technical Coordination and Quality Control, no. SHRP 2 Report S2-S06-RW-1.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

DriveRisk, 2020. Reducing the risks of distracted drivers. URL https://driverisk.com.au/reducing-risks-distracted-drivers.

Du, Y., Raman, C., Black, A.W., Morency, L.-P., Eskenazi, M., 2018. Multimodal polynomial fusion for detecting driver distraction. arXiv preprint arXiv:1810.10565.

Echanobe, J., Basterretxea, K., del Campo, I., Martínez, V., Vidal, N., 2021. Multi-objective genetic algorithm for optimizing an ELM-based driver distraction detection system. IEEE Trans. Intell. Transp. Syst..

El Khatib, A., Ou, C., Karray, F., 2019. Driver inattention detection in the context of next-generation autonomous vehicles design: A survey. IEEE Trans. Intell. Transp. Syst. 21 (11), 4483–4496.

Engström, J., Johansson, E., Östlund, J., 2005. Effects of visual and cognitive load in real and simulated motorway driving. Transp. Res. F 8 (2), 97–120.

Eraqi, H.M., Abouelnaga, Y., Saad, M.H., Moustafa, M.N., 2019. Driver distraction identification with an ensemble of convolutional neural networks. J. Adv. Transp. 2019.
European Commission, 2015. Driver Distraction. Tech. rep., European Commission, Directorate General for Transporn.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.
Passengers and Drivers reading while driving. 2019. http://dx.doi.org/10.6084/m9.figshare.8313620.v1. URL https://figshare.com/articles/dataset/Passengers_and_Drivers_reading_while_driving/8313620/1.
Fridman, L., Brown, D.E., Glazer, M., Angell, W., Dodd, S., Jenik, B., Terwilliger, J., Patsekin, A., Kindelsberger, J., Ding, L., et al., 2019. MIT advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation. IEEE Access 7, 102021–102038.
Fridman, L., Langhans, P., Lee, J., Reimer, B., 2016. Driver gaze region estimation without use of eye movement. IEEE Intell. Syst. 31 (3), 49–56.
Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. Devise: A deep visual-semantic embedding model.
Fukushima, K., 1988. Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Netw. 1 (2), 119–130.
Gao, Y., Mosalam, K.M., 2018. Deep transfer learning for image-based structural damage recognition. Comput.-Aided Civ. Infrastruct. Eng. 33 (9), 748–768.
Garcia Lopez, P., Montresor, A., Epema, D., Datta, A., Higashino, T., Iamnitchi, A., Barcellos, M., Felber, P., Riviere, E., 2015. Edge-centric computing: Vision and challenges.
General Assembly, 2015. Sustainable development goals. In: SDGs Transform Our World 2030.
General Motors Corporate Newsroom, General Motors Corporate Newsroom, 2021. GM introduces new super cruise features to 6 model year 2022 vehicles. Accessed: Oct. 5, 2021. URL https://media.gm.com/media/us/en/gm/news.detail.html/content/Pages/news/us/en/2021/jul/0723-gm-supercruise.html.
Goel, B., Dey, A.K., Bharti, P., Ahmed, K.B., Chellappan, S., 2018. Detecting distracted driving using a wrist-worn wearable. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). IEEE, pp. 233–238.
Goodman, M., Benel, D.C., Lerner, N., Wierwille, W.W., Tijerina, L., Bents, F.D., 1997. An Investigation of the Safety Implications of Wireless Communications in Vehicles.
Gumaei, A., Al-Rakhami, M., Hassan, M.M., Alamri, A., Alhussein, M., Razzaque, M.A., Fortino, G., 2020. A deep learning-based driver distraction identification framework over edge cloud. Neural Comput. Appl. 1–16.
Han, S., Mao, H., Dally, W.J., 2015a. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.
Han, S., Pool, J., Tran, J., Dally, W.J., 2015b. Learning both weights and connections for efficient neural networks. arXiv preprint arXiv:1506.02626.
Hankey, J.M., et al., 2014. Canadian naturalistic driving study.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Hedlund, J., Simpson, H.M., Mayhew, D.R., 2006. International Conference on Distracted Driving: Summary of Proceedings and Recommendations: October 2-5, 2005. CAA.
Hoang Ngan Le, T., Zheng, Y., Zhu, C., Luu, K., Savvides, M., 2016. Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 46–53.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Horberry, T., Anderson, J., Regan, M.A., Triggs, T.J., Brown, J., 2006. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. Accid. Anal. Prev. 38 (1), 185–191.
Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manage. Process 5 (2), 1.
Hssayeni, M.D., Saxena, S., Ptucha, R., Savakis, A., 2017. Distracted driver detection: Deep learning vs handcrafted features. Electron. Imaging 2017 (10), 20–26.
Hu, Y., Lu, M., Lu, X., 2019. Driving behaviour recognition from still images by using multi-stream fusion CNN. Mach. Vis. Appl. 30 (5), 851–865.
Hu, Y., Lu, M., Lu, X., 2020. Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network. Signal Process., Image Commun. 81, 115697.
Hu, Z., Zhang, J., Ge, Y., 2021. Handling vanishing gradient problem using artificial derivative. IEEE Access 9, 22371–22377.
Huang, D., De La Torre, F., 2012. Facial action transfer with personalized bilinear regression. In: European Conference on Computer Vision. Springer, pp. 144–158.
Huang, J., Liu, Y., Peng, X., 2022. Recognition of driver's mental workload based on physiological signals, a comparative study. Biomed. Signal Process. Control 71, 103094.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.
Huang, G., Song, S., Gupta, J.N., Wu, C., 2014. Semi-supervised and unsupervised extreme learning machines. IEEE Trans. Cybern. 44 (12), 2405–2417.
Huang, X., Zhang, B., 2018. Research on method of driver distraction state based on mouth state. In: 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, pp. 301–304.
Hyundai Motor Group Tech, 2021. ADAS – at the epicenter of safety technology development. URL https://tech.hyundaimotorgroup.com/mobility-device/adas/.
Im, S., Lee, C., Yang, S., Kim, J., You, B., 2014. Driver distraction detection by in-vehicle signal processing. In: 2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS). IEEE, pp. 64–68.
International Electrotechnical Commission, 2017. Edge Intelligence. Tech. rep., IEC Market Strategy Board.
Iranmanesh, S.M., Mahjoub, H.N., Kazemi, H., Fallah, Y.P., 2018. An adaptive forward collision warning framework design based on driver distraction. IEEE Trans. Intell. Transp. Syst. 19 (12), 3925–3934.
Jain, A., Koppula, H.S., Raghavan, B., Soh, S., Saxena, A., 2015. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3182–3190.
Jakkula, V., 2006. Tutorial on Support Vector Machine (svm), Vol. 37. School of EECS, Washington State University.
Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A., 2018. Safe driving: Driver action recognition using SURF keypoints. In: 2018 30th International Conference on Microelectronics (ICM). IEEE, pp. 60–63.
Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A., 2019. MDAD: A multimodal and multiview in-vehicle driver action dataset. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 518–529.
Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A., 2020a. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD. Signal Process., Image Commun. 88, 115960.
Jegham, I., Khalifa, A.B., Alouani, I., Mahjoub, M.A., 2020b. Soft spatial attention-based multimodal driver action recognition using deep learning. IEEE Sens. J. 21 (2), 1918–1925.
Joachims, T., et al., 1999. Transductive inference for text classification using support vector machines. In: Icml, Vol. 99. pp. 200–209.
Johns, M., Miller, J.B., Sun, A.C., Baughman, S., Zhang, T., Ju, W., 2015. The driver has control: Exploring driving performance with varying automation capabilities.
Johnson, M.J., Chahal, T., Stinchcombe, A., Mullen, N., Weaver, B., Bédard, M., 2011. Physiological responses to simulated and on-road driving. Int. J. Psychophysiol. 81 (3), 203–208.
Kaiser, C., Steger, M., Dorri, A., Festl, A., Stocker, A., Fellmann, M., Kanhere, S., 2018. Towards a privacy-preserving way of vehicle data sharing–a case for blockchain technology? In: International Forum on Advanced Microsystems for Automotive Applications. Springer, pp. 111–122.
Kaplan, S., Guvensan, M.A., Yavuz, A.G., Karalurt, Y., 2015. Driver behavior analysis for safe driving: A survey. IEEE Trans. Intell. Transp. Syst. 16 (6), 3017–3032.
Kapoor, K., Pamula, R., Murthy, S.V., 2020. Real-time driver distraction detection system using convolutional neural networks. In: Proceedings of ICETIT 2019. Springer, pp. 280–291.
Karimi, H., Derr, T., Tang, J., 2019. Characterizing the decision boundary of deep neural networks. arXiv preprint arXiv:1912.11460.
Kashevnik, A., Shchedrin, R., Kaiser, C., Stocker, A., 2021. Driver distraction detection methods: A literature review and framework. IEEE Access 9, 60063–60076.
Kavi, R., Kulathumani, V., Rohit, F., Kecojevic, V., 2016. Multiview fusion for activity recognition using deep neural networks. J. Electron. Imaging 25 (4), 043010.
Khan, M.Q., Lee, S., 2019. A comprehensive survey of driving monitoring and assistance systems. Sensors 19 (11), 2574.
Kim, W., Choi, H.-K., Jang, B.-T., Lim, J., 2017. Driver distraction detection using single convolutional neural network. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, pp. 1203–1205.
Koay, H.V., Chuah, J.H., Chow, C.-O., 2021a. Shifted-window hierarchical vision transformer for distracted driver detection. In: 2021 IEEE Region 10 Symposium (TENSYMP). pp. 1–7. http://dx.doi.org/10.1109/TENSYMP52854.2021.9550995.
Koay, H.V., Chuah, J.H., Chow, C.-O., Chang, Y.-L., Rudrusamy, B., 2021b. Optimally-weighted image-pose approach (OWIPA) for distracted driver detection and classification. Sensors 21 (14), 4837.
Koesdwiady, A., Bedawi, S.M., Ou, C., Karray, F., 2017. End-to-end deep learning for driver distraction recognition. In: International Conference Image Analysis and Recognition. Springer, pp. 11–18.
Konstantopoulos, P., Chapman, P., Crundall, D., 2010. Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers' eye movements in day, night and rain driving. Accid. Anal. Prev. 42 (3), 827–834.
Köpüklü, O., Wei, X., Rigoll, G., 2019. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. arXiv preprint arXiv:1911.06644.
Kornblith, S., Shlens, J., Le, Q.V., 2019. Do better imagenet models transfer better? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2661–2671.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Kumar, A., Sangwan, K.S., et al., 2021. A computer vision-based approach for driver distraction recognition using deep learning and genetic algorithm based ensemble. arXiv preprint arXiv:2107.13355.
Lamble, D., Kauranen, T., Laakso, M., Summala, H., 1999. Cognitive load and detection thresholds in car following situations: safety implications for using mobile (cellular) telephones while driving. Accid. Anal. Prev. 31 (6), 617–623.
Lampert, C.H., Nickisch, H., Harmeling, S., 2013. Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. 36 (3), 453–465.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.
Lee, J.D., Moeckli, J., Brown, T.L., Roberts, S.C., Schwarz, C., Yekhshatyan, L., Nadler, E., Liang, Y., Victor, T., Marshall, D., et al., 2013. Distraction Detection and Mitigation Through Driver Feedback: Appendices (Report No. DOT HS 811 547B). National Highway Traffic Safety Administration.
Lee, B.G., Park, J.-H., Pu, C.C., Chung, W.-Y., 2015. Smartwatch-based driver vigilance indicator with kernel-fuzzy-c-means-wavelet method. IEEE Sens. J. 16 (1), 242–253.
Lee, J.D., Young, K.L., Regan, M.A., 2008. Defining driver distraction. In: Driver Distraction: Theory, Effects, and Mitigation, Vol. 13. pp. 31–40, (4).
Leekha, M., Goswami, M., Shah, R.R., Yin, Y., Zimmermann, R., 2019. Are you paying attention? detecting distracted driving in real-time. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, pp. 171–180.
Lemley, J., Bazrafkan, S., Corcoran, P., 2017. Transfer learning of temporal information for driver action classification. In: MAICS. pp. 123–128.
Levkova, L., Heck, S., Alpert, B.O., Satzoda, R.K., Sathyanarayana, S., Sekar, V., 2019. System and method for driver distraction determination. US Patent 10, 246, 014.
Lexus, 2021. LEXUS safety system+ 2.0. Accessed: Oct. 5, 2021. URL https://www.lexus.com/models/LS/safety.
Li, Z., Bao, S., Kolmanovsky, I.V., Yin, X., 2017. Visual-manual distraction detection using driving performance indicators with naturalistic driving data. IEEE Trans. Intell. Transp. Syst. 19 (8), 2528–2535.
Li, N., Busso, C., 2014. Predicting perceived visual and cognitive distractions of drivers with multimodal features. IEEE Trans. Intell. Transp. Syst. 16 (1), 51–65.
Li, N., Busso, C., 2015. Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition. IEEE Trans. Intell. Transp. Syst. 17 (4), 980–992.
Li, G., Yan, W., Li, S., Qu, X., Chu, W., Cao, D., 2021a. A temporal-spatial deep learning approach for driver distraction detection based on EEG signals. IEEE Trans. Autom. Sci. Eng..
Li, P., Yang, Y., Grosu, R., Wang, G., Li, R., Wu, Y., Huang, Z., 2021b. Driver distraction detection using octave-like convolutional neural network. IEEE Trans. Intell. Transp. Syst..
Li, L., Zhong, B., Hutmacher Jr., C., Liang, Y., Horrey, W.J., Xu, X., 2020. Detection of driver manual distraction via image-based hand and ear recognition. Accid. Anal. Prev. 137, 105432.
Liang, Y., Lee, J.D., 2014. A hybrid Bayesian network approach to detect driver cognitive distraction. Transp. Res. C 38, 146–155.
Liang, Y., Lee, J.D., Reyes, M.L., 2007. Nonintrusive detection of driver cognitive distraction in real time using Bayesian networks. Transp. Res. Rec. 2018 (1), 1–8.
Liao, Y., Li, S.E., Wang, W., Wang, Y., Li, G., Cheng, B., 2016. Detection of driver cognitive distraction: A comparison study of stop-controlled intersection and speed-limited highway. IEEE Trans. Intell. Transp. Syst. 17 (6), 1628–1637.
Lipton, Z.C., Berkowitz, J., Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.
Liu, D., Yamasaki, T., Wang, Y., Mase, K., Kato, J., 2021. TML: A triple-wise multi-task learning framework for distracted driver recognition. IEEE Access.
Liu, T., Yang, Y., Huang, G.-B., Yeo, Y.K., Lin, Z., 2015. Driver distraction detection using semi-supervised machine learning. IEEE Trans. Intell. Transp. Syst. 17 (4), 1108–1120.
Lu, M., Hu, Y., Lu, X., 2020. Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Appl. Intell. 50 (4), 1100–1111.
Madkor, A., Elqattan, Y., S. AbdElHamid, A., 2020. Distracted driver detection. US Patent 10, 769, 461.
Mafeni Mase, J., Chapman, P., Figueredo, G.P., Torres Torres, M., 2020. A hybrid deep learning approach for driver distraction detection. In: 2020 International Conference on Information and Communication Technology Convergence (ICTC). pp. 1–6. http://dx.doi.org/10.1109/ICTC49870.2020.9289588.
Majdi, M.S., Ram, S., Gill, J.T., Rodríguez, J.J., 2018. Drive-net: Convolutional network for driver distraction detection. In: 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). IEEE, pp. 1–4.
Mandal, D., Narayan, S., Dwivedi, S.K., Gupta, V., Ahmed, S., Khan, F.S., Shao, L., 2019. Out-of-distribution detection for generalized zero-shot action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9985–9993.
Marija, P., 2022. 13 Crucial texting and driving statistics for Canada in 2021. URL https://reviewlution.ca/resources/texting-and-driving-statistics-canada/.
Marshall, S.C., Man-Son-Hing, M., Bedard, M., Charlton, J., Gagnon, S., Gelinas, I., Koppel, S., Korner-Bitensky, N., Langford, J., Mazer, B., et al., 2013a. Protocol for Candrive II/Ozcandrive, a multicentre prospective older driver cohort study. Accid. Anal. Prev. 61, 245–252.
Marshall, S.C., Wilson, K.G., Man-Son-Hing, M., Stiell, I., Smith, A., Weegar, K., Kadulina, Y., Molnar, F.J., 2013b. The Canadian Safe Driving Study—Phase I pilot: Examining potential logistical barriers to the full cohort study. Accid. Anal. Prev. 61, 236–244.
Martens, M., Van Winsum, W., 2000. Measuring distraction: the peripheral detection task. In: TNO Human Factors, Soesterberg, Netherlands.
Martin, S., Ohn-Bar, E., Tawari, A., Trivedi, M.M., 2014a. Understanding head and hand activities and coordination in naturalistic driving videos. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, pp. 884–889.
Martin, M., Popp, J., Anneken, M., Voit, M., Stiefelhagen, R., 2018. Body pose and context information for driver secondary task detection. In: 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 2015–2021.
Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., Stiefelhagen, R., 2019. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2801–2810.
Martin, S., Tawari, A., Trivedi, M.M., 2014b. Balancing privacy and safety: Protecting driver identity in naturalistic driving video data. In: Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. pp. 1–7.
Martin, S., Tawari, A., Trivedi, M.M., 2014c. Toward privacy-protecting safety systems for naturalistic driving videos. IEEE Trans. Intell. Transp. Syst. 15 (4), 1811–1822.
Martin, S., Yuen, K., Trivedi, M.M., 2016. Vision for intelligent vehicles & applications (viva): Face detection and head pose challenge. In: 2016 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1010–1014.
Mase, J.M., Chapman, P., Figueredo, G.P., Torres, M.T., 2020. Benchmarking deep learning models for driver distraction detection. In: International Conference on Machine Learning, Optimization, and Data Science. Springer, pp. 103–117.
Masood, S., Rai, A., Aggarwal, A., Doja, M.N., Ahmad, M., 2020. Detecting distraction of drivers using convolutional neural network. Pattern Recognit. Lett. 139, 79–85.
Massoz, Q., Langohr, T., François, C., Verly, J.G., 2016. The ULg multimodality drowsiness database (called DROZY) and examples of use. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1–7.
Mattes, S., 2003. The lane-change-task as a tool for driver distraction evaluation. In: Quality of Work and Products in Enterprises of the Future, Vol. 57. p. 60.
Mayhew, D.R., Simpson, H.M., Wood, K.M., Lonero, L., Clinton, K.M., Johnson, A.G., 2011. On-road and simulated driving: Concurrent and discriminant validation. J. Saf. Res. 42 (4), 267–275.
McDonald, A., Carney, C., McGehee, D.V., 2018. Vehicle Owners' Experiences with and Reactions to Advanced Driver Assistance Systems. Tech. rep., URL https://aaafoundation.org/vehicle-owners-experiences-reactions-advanced-driver-assistance-systems/.
McEvoy, S.P., Stevenson, M.R., Woodward, M., 2006. The impact of driver distraction on road safety: results from a representative survey in two Australian states. Injury Prev. 12 (4), 242–247.
Min, J., Wang, P., Hu, J., 2017. The original EEG data for driver fatigue detection. http://dx.doi.org/10.6084/m9.figshare.5202739.v1, [Online]. Available: https://figshare.com/articles/dataset/The_original_EEG_data_for_driver_fatigue_detection/5202739/1.
Morgenstern, T., Wögerbauer, E.M., Naujoks, F., Krems, J.F., Keinath, A., 2020. Measuring driver distraction–Evaluation of the box task method as a tool for assessing in-vehicle system demand. Applied Ergon. 88, 103181.
Mosavi, A., Ozturk, P., Chau, K.-w., 2018. Flood prediction using machine learning models: Literature review. Water 10 (11), 1536.
Moslemi, N., Azmi, R., Soryani, M., 2019. Driver distraction recognition using 3d convolutional neural networks. In: 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA). IEEE, pp. 145–151.
Moslemi, N., Soryani, M., Azmi, R., 2021. Computer vision-based recognition of driver distraction: A review. Concurr. Comput.: Pract. Exper. e6475.
Murugan, S., Selvaraj, J., Sahayadhas, A., 2020. Detection and analysis: driver state with electrocardiogram (ECG). Phys. Eng. Sci. Med. 43 (2), 525–537.
National Highway Traffic Safety Administration, 2019. Distracted Driving in Fatal Crashes, 2017. (Traffic Safety Facts Research Note. Report No. DOT HS 812 700). Tech. rep., National Center for Statistics and Analysis.
National Highway Traffic Safety Administration, 2020. Distracted driving 2018 (Research Note. Report No. DOT HS 812 926). Tech. rep., National Center for Statistics and Analysis.
National Highway Traffic Safety Administration, 2021. Distracted driving. URL https://www.nhtsa.gov/risky-driving/distracted-driving.
National Safety Council, 2010. Understanding the Distracted Brain: Why Driving while using Hands-Free Cell Phones is Risky Behavior. Tech. rep..
Neale, V.L., Dingus, T.A., Klauer, S.G., Sudweeks, J., Goodman, M., 2005. An overview of the 100-car naturalistic study and findings. In: National Highway Traffic Safety Administration, Paper 5. Citeseer, p. 0400.
Nel, F., Ngxande, M., 2021. Driver activity recognition through deep learning. In: 2021 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA). IEEE, pp. 1–6.
Newton, E.M., Sweeney, L., Malin, B., 2005. Preserving privacy by de-identifying face images. IEEE Trans. Knowl. Data Eng. 17 (2), 232–243.

Noor, N.M.M., Mustafa, M.A.M., 2016. Eye movement activity that affected the eye signals using electrooculography (EOG) technique. In: 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE). IEEE, pp. 91–95.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J., 2013. Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650.

Ohn-Bar, E., Martin, S., Tawari, A., Trivedi, M.M., 2014. Head, eye, and hand patterns for driver activity recognition. In: 2014 22nd International Conference on Pattern Recognition. IEEE, pp. 660–665.

Ohn-Bar, E., Trivedi, M.M., 2014. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. IEEE Trans. Intell. Transp. Syst. 15 (6), 2368–2377.

Ojsteršek, T.C., 2019. Eye tracking use in researching driver distraction: A scientometric and qualitative literature review approach. J. Eye Mov. Res. 12 (3).

Okon, O.D., Meng, L., 2017. Detecting distracted driving with deep learning. In: International Conference on Interactive Collaborative Robotics. Springer, pp. 170–179.

Olson, R.L., Hanowski, R.J., Hickman, J.S., Bocanegra, J., et al., 2009. Driver Distraction in Commercial Vehicle Operations. Tech. rep., United States Department of Transportation, Federal Motor Carrier Safety Administration.

Omerustaoglu, F., Sakar, C.O., Kar, G., 2020. Distracted driver detection by combining in-vehicle and image data using deep learning. Appl. Soft Comput. 96, 106657.

On-Road Automated Driving (ORAD) committee, 2016. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. URL https://doi.org/10.4271/J3016_201609.

Orsten-Hooge, K.D., Baragchizadeh, A., Karnowski, T.P., Bolme, D.S., Ferrell, R., Jesudasen, P.R., Castillo, C.D., O'Toole, A.J., 2019. Evaluating the effectiveness of automated identity masking (AIM) methods with human perception and a deep convolutional neural network (CNN). arXiv preprint arXiv:1902.06967.

Ortega, J.D., Kose, N., Cañas, P., Chao, M.-A., Unnervik, A., Nieto, M., Otaegui, O., Salgado, L., 2020. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. arXiv preprint arXiv:2008.12085.

Östlund, J., Nilsson, L., Törnros, J., Forsman, Å., 2006. Effects of Cognitive and Visual Load in Real and Simulated Driving. Statens väg-och transportforskningsinstitut.

Ou, C., Ouali, C., Karray, F., 2018. Transfer learning based strategy for improving driver distraction recognition. In: International Conference Image Analysis and Recognition. Springer, pp. 443–452.

Ou, C., Zhao, Q., Karray, F., El Khatib, A., 2019. Design of an end-to-end dual mode driver distraction detection system. In: International Conference on Image Analysis and Recognition. Springer, pp. 199–207.

Oviedo-Trespalacios, O., Haque, M.M., King, M., Washington, S., 2016. Understanding the impacts of mobile phone distraction on driving performance: A systematic review. Transp. Res. C 72, 360–380.

Papakostas, M., Riani, K., Gasiorowski, A.B., Sun, Y., Abouelenien, M., Mihalcea, R., Burzo, M., 2021. Understanding driving distractions: A multimodal analysis on distraction characterization. In: 26th International Conference on Intelligent User Interfaces. pp. 377–386.

Phan, B.T., 2019. Bayesian Deep Learning and Uncertainty in Computer Vision (Master's thesis). University of Waterloo.

Pickrell, T.M., et al., 2015. Driver Electronic Device Use in 2013. Tech. rep., United States. National Highway Traffic Safety Administration.

Plastiras, G., Terzi, M., Kyrkou, C., Theocharidcs, T., 2018. Edge intelligence: Challenges and opportunities of near-sensor machine learning applications. In: 2018 Ieee 29th International Conference on Application-Specific Systems, Architectures and Processors (Asap). IEEE, pp. 1–7.

Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Qiu, Z., Yao, T., Mei, T., 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541.

Ragab, A., Craye, C., Kamel, M.S., Karray, F., 2014. A visual-based driver distraction recognition and detection using random forest. In: International Conference Image Analysis and Recognition. Springer, pp. 256–265.

Rajendra, V., Dehzangi, O., 2017. Detection of distraction under naturalistic driving using galvanic skin responses. In: 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, pp. 157–160.

Regan, M.A., Hallett, C., Gordon, C.P., 2011. Driver distraction and driver inattention: Definition, relationship and taxonomy. Accid. Anal. Prev. 43 (5), 1771–1781.

Reiß, S., Roitberg, A., Haurilet, M., Stiefelhagen, R., 2020. Activity-aware attributes for zero-shot driver behavior recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 902–903.

Ren, H., Guo, Y., Bai, Z., Cheng, X., 2021. A multi-semantic driver behavior recognition model of autonomous vehicles using confidence fusion mechanism. In: Actuators, Vol. 10. Multidisciplinary Digital Publishing Institute, p. 218.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28, 91–99.

Renbo, Q., Daqian, Y., 2021. Driving state analysis method and apparatus, driver monitoring system and vehicle. US Patent App. 17/031, 030.

Riani, K., Papakostas, M., Kokash, H., Abouelenien, M., Burzo, M., Mihalcea, R., 2020. Towards detecting levels of alertness in drivers using multiple modalities. In: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments. pp. 1–9.

Roitberg, A., Al-Halah, Z., Stiefelhagen, R., 2018. Informed democracy: voting-based novelty detection for action recognition. arXiv preprint arXiv:1810.12819.

Roitberg, A., Ma, C., Haurilet, M., Stiefelhagen, R., 2020. Open set driver activity recognition. In: 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1048–1053.

Romera, E., Bergasa, L.M., Arroyo, R., 2016. Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 387–392.

Roth, M., Gavrila, D.M., 2019. DD-Pose-A large-scale driver head pose benchmark. In: 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 927–934.

Rother, C., Kolmogorov, V., Blake, A., 2004. "GrabCut" interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

Sahayadhas, A., Sundaraj, K., Murugappan, M., 2012. Detecting driver drowsiness based on sensors: a review. Sensors 12 (12), 16937–16953.

Sahayadhas, A., Sundaraj, K., Murugappan, M., Palaniappan, R., 2015. A physiological measures-based method for detecting inattention in drivers using machine learning approach. Biocybern. Biomed. Eng. 35 (3), 198–205.

Saleh, K., Hossny, M., Nahavandi, S., 2017. Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1–6.

Sayer, J., LeBlanc, D., Bogard, S., Funkhouser, D., Bao, S., Buonarosa, M.L., Blankespoor, A., et al., 2011. Integrated Vehicle-Based Safety Systems Field Operational Test: Final Program Report. Tech. rep., United States. Joint Program Office for Intelligent Transportation Systems.

Schneiders, E., Kristensen, M.B., Svangren, M.K., Skov, M.B., 2020. Temporal impact on cognitive distraction detection for car drivers using EEG. In: 32nd Australian Conference on Human-Computer Interaction. pp. 594–601.

Schwarz, A., Haurilet, M., Martinez, M., Stiefelhagen, R., 2017. Driveahead-a large-scale driver head pose dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–10.

Senders, J.W., Kristofferson, A., Levison, W., Dietrich, C., Ward, J., et al., 1967. The attentional demand of automobile driving.

Seshadri, K., Juefei-Xu, F., Pal, D.K., Savvides, M., Thor, C.P., 2015. Driver cell phone usage detection on strategic highway research program (SHRP2) face view videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 35–43.

Shenoy, R., Krishna, A., Putraya, G., Ukil, S., Uliyar, M., Patwardhan, P., 2018. Method and system for driver monitoring by fusing contextual data with event data to determine context as cause of event. US Patent 10, 065, 652.

Sicconi, R., Stys, M.E., 2019. Method to analyze attention margin and to prevent inattentive and unsafe driving. US Patent 10, 467, 488.

Sigari, M.-H., Pourshahabi, M.-R., Soryani, M., Fathy, M., 2014. A review on driver face monitoring systems for fatigue and distraction detection.

Sikander, G., Anwar, S., 2018. Driver fatigue detection systems: A review. IEEE Trans. Intell. Transp. Syst. 20 (6), 2339–2352.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Solovey, E.T., Zec, M., Garcia Perez, E.A., Reimer, B., Mehler, B., 2014. Classifying driver workload using physiological and driving performance data: two field studies. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 4057–4066.

Son, J., Park, M., 2016. Real-Time Detection and Classification of Driver Distraction using Lateral Control Performance. Humanlab. Kr., [Online]. Available: http://www.humanlab.kr/attachment/cfile2.uf@220AED48575F588411225B.pdf.

Sonnleitner, A., Treder, M.S., Simon, M., Willmann, S., Ewald, A., Buchner, A., Schrauf, M., 2014. EEG alpha spindles and prolonged brake reaction times during auditory distraction in an on-road driving study. Accid. Anal. Prev. 62, 110–118.

2016. StateFarm distracted driver detection dataset. URL https://www.kaggle.com/c/state-farm-distracted-driver-detection.

Stojmenova, K., Sodnik, J., 2018. Detection-response task—uses and limitations. Sensors 18 (2), 594.

Stratou, G., Morency, L.-P., 2017. MultiSense—Context-aware nonverbal behavior analysis framework: A psychological distress use case. IEEE Trans. Affect. Comput. 8 (2), 190–203.

Strayer, D.L., Cooper, J.M., Turrill, J., Coleman, J., Medeiros-Ward, N., Biondi, F., 2013. Measuring cognitive distraction in the automobile.

Streiffer, C., Raghavendra, R., Benson, T., Srivatsa, M., 2017. Darnet: a deep learning solution for distracted driving detection. In: Proceedings of the 18th Acm/Ifip/Usenix Middleware Conference: Industrial Track. pp. 22–28.

Stutts, J.C., Reinfurt, D.W., Staplin, L., Rodgman, E., et al., 2001. The role of driver distraction in traffic crashes.

Sun, W., Si, Y., Guo, M., Li, S., 2021. Driver distraction recognition using wearable IMU sensor data. Sustainability 13 (3), 1342.

Swetha, A., Sharma, M., Sunkara, S.V., Kattampally, V.J., Muralikrishna, V., Sankaran, P., 2019. Ensemble methods on weak classifiers for improved driver distraction detection. In: International Conference on Computer Vision and Image Processing. Springer, pp. 233–242.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Taamneh, S., Tsiamyrtzis, P., Dcosta, M., Buddharaju, P., Khatri, A., Manser, M., Ferris, T., Wunderlich, R., Pavlidis, I., 2017. A multimodal dataset for various forms of distracted driving. Sci. Data 4 (1), 1–21.

Taherisadr, M., Dehzangi, O., Parsaei, H., 2017. Single channel EEG artifact identification using two-dimensional multi-resolution analysis. Sensors 17 (12), 2895.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tan, M., Le, Q.V., 2021. Efficientnetv2: Smaller models and faster training. arXiv preprint arXiv:2104.00298.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Springer, pp. 270–279.

Tavakoli, A., Kumar, S., Boukhechba, M., Heydarian, A., 2021a. Driver state and behavior detection through smart wearables. arXiv preprint arXiv:2104.13889.

Tavakoli, A., Kumar, S., Guo, X., Balali, V., Boukhechba, M., Heydarian, A., 2021b. HARMONY: A human-centered multimodal driving study in the wild. IEEE Access 9, 23956–23978.

The National Academies of Sciences, Engineering, and Medicine, 2021. The second Strategic Highway Research Program (2006–2015). Accessed: Oct. 5, 2021. URL http://www.trb.org/StrategicHighwayResearchProgram2SHRP2/Blank2.aspx.

Thorslund, B., 2004. Electrooculogram Analysis and Development of a System for Defining Stages of Drowsiness. Statens väg-och transportforskningsinstitut.

Tjolleng, A., Jung, K., Hong, W., Lee, W., Lee, B., You, H., Son, J., Park, S., 2017. Classification of a Driver's cognitive workload levels using artificial neural network on ECG signals. Applied Ergon. 59, 326–332.

Tkach, D., Huang, H., Kuiken, T.A., 2010. Study of stability of time-domain features for electromyographic pattern recognition. J. Neuroeng. Rehabil. 7 (1), 1–13.

Torres, R.H., Ohashi, O., Garcia, G., Rocha, F., Azpúrua, H., Pessin, G., 2019. Exploiting machine learning models to avoid texting while driving. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4489–4497.

Tran, D., Do, H.M., Sheng, W., Bai, H., Chowdhary, G., 2018. Real-time detection of distracted driving based on deep learning. IET Intell. Transp. Syst. 12 (10), 1210–1219.

Uchida, N., Kawakoshi, M., Tagawa, T., Mochida, T., 2010. An investigation of factors contributing to major crash types in Japan based on naturalistic driving data. IATSS Res. 34 (1), 22–30.

Ugli, I.K.K., Hussain, A., Kim, B.S., Aich, S., Kim, H.-C., 2021. A transfer learning approach for identification of distracted driving. In: 2021 23rd International Conference on Advanced Communication Technology (ICACT). IEEE, pp. 1–4.

United Nations Economic and Social Council, 2021. Progress towards the Sustainable Development Goals: Report of the Secretary-General. High-Level Political Forum on Sustainable Development, Convened under the Auspices of the Economic and Social Council.

Valeriano, L.C., Napoletano, P., Schettini, R., 2018. Recognition of driver distractions using deep learning. In: 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin). IEEE, pp. 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Velez, G., Otaegui, O., 2015. Embedded platforms for computer vision-based advanced driver assistance systems: a survey. arXiv preprint arXiv:1504.07442.

Vicente, F., Huang, Z., Xiong, X., De la Torre, F., Zhang, W., Levi, D., 2015. Driver gaze tracking and eyes off the road detection system. IEEE Trans. Intell. Transp. Syst. 16 (4), 2014–2027.

Vicente, J., Laguna, P., Bartra, A., Bailón, R., 2016. Drowsiness detection using heart rate variability. Med. Biol. Eng. Comput. 54 (6), 927–937.

Volodina, V., Challenor, P., 2021. The importance of uncertainty quantification in model reproducibility. Phil. Trans. R. Soc. A 379 (2197), 20200071.

Vora, S., Rangesh, A., Trivedi, M.M., 2017. On generalizing driver gaze zone estimation using convolutional neural networks. In: 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 849–854.

Wagner, B., Taffner, F., Karaca, S., Karge, L., 2021. Vision based detection of driver cell phone usage and food consumption. IEEE Trans. Intell. Transp. Syst..

Wang, Y.-K., Jung, T.-P., Lin, C.-T., 2015a. EEG-based attention tracking during distracted driving. IEEE Trans. Neural Syst. Rehabil. Eng. 23 (6), 1085–1094.

Wang, Q., Ma, Y., Zhao, K., Tian, Y., 2020. A comprehensive survey of loss functions in machine learning. Ann. Data Sci. 1–26.

Wang, S., Sun, S., Xu, J., 2015b. Auc-maximized deep convolutional neural fields for sequence labeling. arXiv preprint arXiv:1511.05265.

Wang, H., Wang, L., 2017. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 499–508.

Wang, J., Wu, Z., Li, F., Zhang, J., 2021. A data augmentation approach to distracted driving detection. Future Internet 13 (1), 1.

Weng, C.-H., Lai, Y.-H., Lai, S.-H., 2016. Driver drowsiness detection via a hierarchical temporal deep belief network. In: Asian Conference on Computer Vision. Springer, pp. 117–133.

Wharton, Z., Behera, A., Liu, Y., Bessis, N., 2021. Coarse temporal attention network (CTA-Net) for driver's activity recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1279–1289.

Williamson, A., Grzebieta, R., Eusebio, J.E., Zheng, W.Y., Wall, J., Charlton, J., Lenne, M., Haley, J., Barnes, B., Rakotonirainy, A., et al., 2015. The australian naturalistic driving study: From beginnings to launch. In: Proceedings of the 2015 Australasian Road Safety Conference (ARSC2015). Australasian College of Road Safety (ACRS), pp. 1–7.

World Health Organization, 2018. Global status report on road safety 2018: Summary.

Wu, M., Zhang, X., Shen, L., Yu, H., 2021. Pose-aware multi-feature fusion network for driver distraction recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 1228–1235.

Xiao, D., Feng, C., 2016. Detection of drivers visual attention using smartphone. In: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, pp. 630–635.

Xie, J., Hilal, A.R., Kulic, D., 2018. Driver distraction recognition based on smartphone sensor data. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 801–806.

Xie, Z., Li, L., Xu, X., 2021. Real-time driving distraction recognition through a wrist-mounted accelerometer. Hum. Factors 0018720821995000.

Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., Wang, F.-Y., 2019. Driver activity recognition for intelligent vehicles: A deep learning approach. IEEE Trans. Veh. Technol. 68 (6), 5379–5390.

Xing, Y., Lv, C., Zhang, Z., Wang, H., Na, X., Cao, D., Velenis, E., Wang, F.-Y., 2017. Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition. IEEE Trans. Comput. Soc. Syst. 5 (1), 95–108.

Xu, B., Loce, R.P., 2015. A machine learning approach for detecting cell phone usage. In: Video Surveillance and Transportation Imaging Applications 2015, Vol. 9407. International Society for Optics and Photonics, p. 94070A.

Yan, C., Coenen, F., Zhang, B., 2014. Driving posture recognition by joint application of motion history image and pyramid histogram of oriented gradients. Int. J. Veh. Technol. 2014.

Yan, C., Coenen, F., Zhang, B., 2016. Driving posture recognition by convolutional neural networks. IET Comput. Vis. 10 (2), 103–114.

Yang, Y., Sun, H., Liu, T., Huang, G.-B., Sourina, O., 2015. Driver workload detection in on-road driving environment using machine learning. In: Proceedings of ELM-2014 Volume 2. Springer, pp. 389–398.

Ye, M., Osman, O.A., Ishak, S., Hashemi, B., 2017. Detection of driver engagement in secondary tasks from observed naturalistic driving behavior. Accid. Anal. Prev. 106, 385–391.

Ying, X., 2019b. An overview of overfitting and its solutions. In: Journal of Physics: Conference Series, Vol. 1168. IOP Publishing, 022022.

Yu, H., Porikli, F., Yuzhu, W., 2021. Primary preview region and gaze based driver distraction detection. US Patent 11, 017, 249.

Zangi, N., Srour-Zreik, R., Ridel, D., Chasidim, H., Borowsky, A., 2022. Driver distraction and its effects on partially automated driving performance: A driving simulator study among young-experienced drivers. Accid. Anal. Prev. 166, 106565.

Zhang, C., Eskandarian, A., 2020. A survey and tutorial of EEG-based brain monitoring for driver state analysis. arXiv preprint arXiv:2008.11226.

Zhang, C., Li, R., Kim, W., Yoon, D., Patras, P., 2020. Driver behavior recognition via interwoven deep convolutional neural nets with multi-stream inputs. IEEE Access 8, 191138–191151.

Zheng, W.-L., Lu, B.-L., 2017. A multimodal approach to estimating vigilance using EEG and forehead EOG. J. Neural Eng. 14 (2), 026017.

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., Zhang, J., 2019. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proc. IEEE 107 (8), 1738–1762.

Zhu, M., Wang, X., Tarko, A., et al., 2018. Modeling car-following behavior on urban expressways in Shanghai: A naturalistic driving study. Transp. Res. C 93, 425–445.