ENHANCING DISTRACTED DRIVER DETECTION WITH HUMAN BODY ACTIVITY RECOGNITION USING DEEP LEARNING

F. Zandamela[1,2*], F. Nicolls[2], D. Kunene[3] & G. Stoltz[3]

ARTICLE INFO

Article details
Presented at the 24th Annual International Conference of the Rapid Product Development Association of South Africa (RAPDASA) Institute for Industrial Engineering, held from 30 October to 2 November 2023 in Pretoria, South Africa

Available online 14 Dec 2023

Contact details
* Corresponding author
fzandamela@csir.co.za

Author affiliations
1 Smart Places, CSIR, Council for Scientific and Industrial Research, Pretoria, South Africa

2 Department of Electrical Engineering, University of Cape Town, Cape Town, South Africa

3 Defence and Security, CSIR, Council for Scientific and Industrial Research, Pretoria, South Africa

ORCID® identifiers
F. Zandamela
https://orcid.org/0000-0003-2201-1985

F. Nicolls
https://orcid.org/0000-0002-8483-412X

D. Kunene
https://orcid.org/0000-0002-9531-850X

G. Stoltz
https://orcid.org/0000-0001-8500-1798

DOI
http://dx.doi.org//10.7166/34-4-2983

ABSTRACT

Deep learning has become popular owing to its high accuracy and ability to learn features automatically from input data. Various approaches are proposed in the literature to detect distracted drivers. However, the performance of these algorithms is typically limited to image datasets that have a similar distribution to the training dataset, which makes it difficult to apply them in real-world scenarios. To address this issue, this paper proposes a robust approach to detecting distracted drivers, based on recognising the unique body movements involved when a driver operates a vehicle. Experimental results indicate that this method outperforms current deep learning algorithms for detecting distracted drivers, resulting in a 6% improvement in classification accuracy and a two-fold improvement in overall performance (F1 score).

OPSOMMING

Diepleer het gewild geword as gevolg van diepleer se hoë akkuraatheid en vermoë om kenmerke outomaties van invoerdata te leer. Verskeie benaderings word in die literatuur voorgestel om bestuurders wat nie konsentreer op die taak, te identifiseer. Die werkverrigting van hierdie algoritmes is egter tipies beperk tot beelddatastelle wat 'n soortgelyke verspreiding as die opleidingdatastel het, wat dit moeilik maak om dit in werklike scenario's toe te pas. Om hierdie probleem aan te spreek, stel hierdie artikel 'n robuuste benadering voor om bestuurders wat afgelei is op te spoor, gebaseer op die herkenning van die unieke liggaamsbewegings wat betrokke is wanneer 'n bestuurder 'n voertuig bestuur. Eksperimentele resultate dui daarop dat hierdie metode beter presteer as huidige diepleeralgoritmes vir die opsporing van afgeleide bestuurders, wat lei tot 'n 6% verbetering in klassifikasie akkuraatheid en 'n tweevoudige verbetering in algehele prestasie (F1-telling).

1. INTRODUCTION

Distracted driving is a major cause of traffic accidents, injuries, and fatalities worldwide, necessitating effective techniques to detect distracted drivers and enhance road safety. Deep learning has shown remarkable success in various real-world applications, such as number plate recognition for vehicle access control. As a result, significant attention has been devoted to using deep learning, specifically convolutional neural networks (CNNs), to address the issue of distracted driver detection. CNNs are favoured for their ability to extract image features automatically and to perform classification [1]. Promising results have been reported in the literature, demonstrating high accuracy within individual datasets. For instance, Leekhaa et al. [2] proposed a CNN-based approach that achieved an average accuracy of 98.48% on the State Farm Distracted Driver Detection dataset [3]. However, despite such great success, the performance of current algorithms remains limited in cross-dataset testing scenarios. In our recent work [4] we found

that current deep learning algorithms for distracted driver detection do not generalise well on unknown datasets, particularly CNN models that use the entire image for prediction. This poses a problem, since deploying such models in the real world may result in catastrophic events.

To tackle the problem of poor cross-dataset performance in CNN-based distracted driver detection, this work aims to address the following primary research question: Could the performance of CNN-based distracted driver detection be improved across different datasets by detecting driver body parts and classifying their activities? To answer this question, we propose a straightforward yet effective approach that focuses on detecting and recognising specific activities of critical human body parts that are involved in driving. Furthermore, the proposed approach is evaluated on three distinct datasets to assess its cross-dataset performance. The contributions of this paper can be summarised as follows:

1. We introduce a new approach that exhibits superior cross-dataset performance compared with existing CNN-based distracted driver detection algorithms.
2. We evaluate the cross-dataset performance of the proposed approach on two public datasets. In addition, we provide experimental results that compare the performance of our approach with existing deep learning algorithms for distracted driver detection.
3. We also provide the performance results of the proposed algorithm on a custom dataset created by the CSIR. This should indicate the readiness of the proposed approach for deployment and integration in real-world applications.

2. RELATED WORK

Several distracted driver detection algorithms rely on convolutional neural networks (CNNs), owing to their success in real-world vision tasks. Initially, CNN-based approaches for distracted driver detection involved extracting driver features using CNN architectures and classifying them with a multilayer perceptron (MLP) [5, 6]. For instance, Yan et al. [5] introduced the first CNN-based approach, using a CNN model to extract features from the driver's face region, localised using the Face++ Research Toolkit. The extracted features were then forwarded to a multilayer perceptron (MLP) classifier for classification. Since then, various approaches have been proposed; they can be broadly categorised as: (1) transfer learning [7, 8]; (2) modifying CNN architecture to improve accuracy and speed [9, 10] or using an ensemble of CNNs [11, 12]; (3) combining different features with CNN features to improve detection accuracy [13-15]; (4) using advanced data augmentation techniques such as generative adversarial networks (GANs) to improve the diversity of the training data [16]; (5) considering both spatial and spectral, using hybrid CNN-RNN models [17] and 3D CNNs [18, 19]; (6) using background noise removal techniques to force CNNs to focus only on the driver [2, 9, 20]; and (7) using object detection to detect specific regions and features [9], [21-24]. Furthermore, the introduction of transformer architectures in computer vision has led to the adoption of vision transformer (ViT) methods for distracted driver detection [25].

Transfer learning approaches typically use CNN architectures that are pre-trained large compute-vision image datasets such as ImageNet [26]. In these approaches, the fully connected (FC) layers of the pre-trained network are replaced with new FC layers and fine-tuned to recognise classes in distracted driver detection. Commonly used pre-trained models are AlexNet, VGG16, VGG19, ResNet50, InceptionV3, Xception, and DenseNet. While transfer learning approaches can easily achieve highly accurate results, they often suffer from overfitting because of the limited diversity of current distracted driver detection datasets [4]. These datasets are predominantly created through simulation experiments in simulators or real car environments [16]. The captured images exhibit similar backgrounds and are concentrated within a narrow range of distracted driving scenarios. To address this problem, Ou et al. [16] proposed an advanced data augmentation-based approach that employs GANs to generate synthetic data, thereby increasing the diversity of the training dataset. The authors claim that they collected a diverse dataset of drivers in different driving conditions and activity patterns from the internet and trained generative models for multiple driving scenarios. By sampling from these generative models, they augmented the collected dataset with new training samples, and trained a CNN model for distraction recognition. However, it is worth noting that some researchers have reported difficulties in training GANs [27].

To address the overfitting issue of CNNs without augmenting the training data, researchers have explored alternative approaches. One is to train CNNs on histogram of orientation (HOG) images instead of RGB images [10]. The HOG image is generated by applying a HOG feature descriptor to an RGB image, and then a CNN architecture is used to extract features. Researchers have also proposed combining CNN features with HOG features [13, 15], pose-estimation features [14, 28], and vehicle telemetry data [29]. Various

methods leveraging multiple CNN architectures have also been proposed. Huang et al. [12] argue that relying on a single pre-trained CNN model for distracted driver detection is prone to overfitting, and could result in detection failures. As a solution, some researchers have adopted an ensemble approach using multiple CNNs. For instance, Abouelnaga et al. [11] introduced a genetically weighted ensemble of CNNs trained on five different image sources: raw images, skin-segmented images, face images, hands images, and 'face+hands' images. The study employed four pre-trained deep learning models (AlexNet, Inception-V, ResNet50, and VGG-network) and fine-tuned them for driver distraction identification. While ensembles have been successful in mitigating overfitting, other researchers have raised concerns about the computational complexity and cost of training multiple CNN models such as VGG16, AlexNet, ResNet50, and Inception-V3 [17], particularly in real-time inference, which is crucial for autonomous driving applications.

In addition to such approaches, several methods have been introduced that focus on using specific regions of the image rather than the entire image for prediction. One prominent technique involves employing instance segmentation algorithms to eliminate background noise and to restrict CNN models to learning features solely from the driver [2, 9, 20]. For instance, Ezzouhri et al. [20] presented a CNN-based approach that employs a human segmentation algorithm called cross-domain complementary learning (CDCL) to pre-process RGB images. The resulting pre-processed dataset is then used to train a CNN architecture. By leveraging instance segmentation, these methods aim to enhance the model's focus on driver-related features while reducing the influence of irrelevant background information.

Furthermore, to enhance the robustness of distracted driver detection, researchers have explored the use of object detection techniques to identify specific regions and objects [21-24], [30], which aligns with our proposed approach. Yan et al. [30] proposed a vision-based approach that used a modified R*CNN framework [31] with two input regions: the primary region encompassing the entire driver image, and the secondary region consisting of skin-like regions extracted using a Gaussian mixture model (GMM). The two regions were then forwarded to a deep convolutional neural network called R*CNN to generate driver action labels. However, the approach is not an end-to-end deep learning model, as it requires region proposals generated by the GMM model and subsequent classification. Such complex pipelines are slow and difficult to optimise, since each model needs separate training [32]. In addition, the cross-dataset performance of the algorithm was not evaluated.

Another study by Le et al. [21] proposed a multiple scale faster-RCNN approach that employed a standard region proposal network (RPN) to generate region proposals. The approach incorporated feature maps from shallower convolutional layers for ROI pooling, and aimed to detect individual objects such as hands, cell phones, and steering wheels. However, this study focused solely on cell phone distraction, and neglected the driver's attention to the road. Similarly, another related study [22] introduced a distracted driver detection method that was primarily focused on cell phone detection, overlooking other aspects.

A very similar distracted driver detection method to our proposed approach was presented by Sajid et al. [24]. The proposed method uses the EfficientDet model to detect distraction objects and the ROI of the driver body parts and use an EfficientNet model for classification. In summary, the approach involves the following steps: extract CNN features on an input image, classify the image, detect distraction objects and the ROI of the driver body parts, and finally combine the classification label with the detection label to obtain the final prediction. The approach combines both image classification and object detection. There is one major difference between this method and the approach proposed in this study. Instead of combining the full image and the driver ROI for driver behaviour recognition, the approach proposed in this study uses the Yolov7 model to detect driver body parts and classify their state into different activities in one forward pass. The final prediction is made by evaluating two conditions: "eyes on the road" and "both hands on the steering wheel." In addition, the cross-dataset performance of the proposed approach is evaluated on three distinct distracted driver detection image datasets. Further, the performance of the proposed approach is evaluated on a custom dataset to assess its potential for deployment in the real world.

3. METHODOLOGY

3.1. Proposed approach

The proposed approach consists of two key steps. In the first step, the driver's body parts are detected, and the state of each body part is classified into an activity. The second step uses the detected activities to make a final prediction, using a simple decision tree-based approach. Inspired by the findings of Yang et

al. [9], the distracted driver classification problem in the second step is treated as a binary classification problem, focusing on determining whether the driver is distracted. The specific tasks involved in each step are detailed below. Figure 1 provides an overview of the proposed approach.

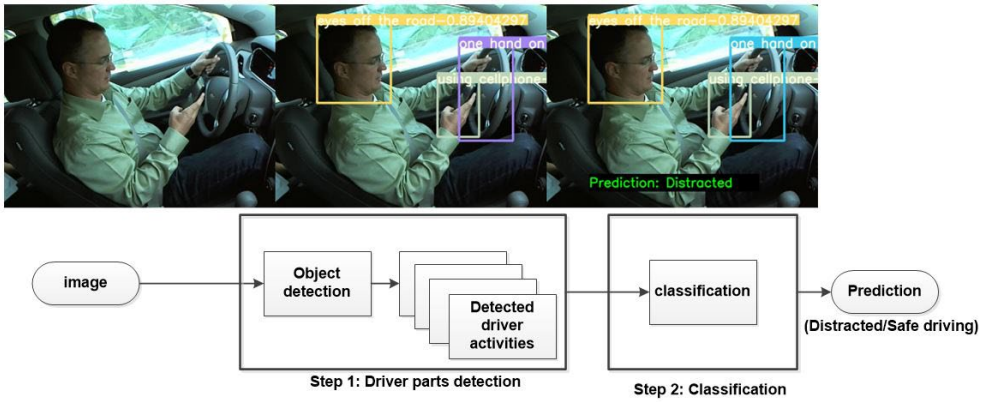

Figure 1: Proposed approach to distracted driver detection

In step 1, the driver's head and hands are detected. The state of the driver's head is classified as either 'eyes on the road' or 'eyes off the road'. The positions of the left and right hands in relation to the steering wheel are used to classify whether both hands are on the wheel or if only one hand is on the wheel. Furthermore, common distracting activities that a driver may engage in while driving are recognised. These activities include cellphone usage, drinking, hands on the face, and talking on the phone. The activities are identified by detecting the presence of a cellphone or any object that could be used for drinking and classifying the activity based on the detected object. In the second step of the approach, the final prediction is made by evaluating two conditions: 'eyes on the road' and 'both hands on the steering wheel'. If both conditions are satisfied, the driver is considered to be in a safe driving position. Conversely, if one of the conditions is not met, the driver is considered to be distracted.

3.2. Implementation

The proposed approach was implemented following the three-stage methodology outlined in Figure 2: pre-processing, training the object detection model on a custom dataset, and predicting the driver behaviour using the model. The Yolov7 object detection model was selected for its superior speed and accuracy compared with other state-of-the-art models during our experiments [33]. We adapted the Yolov7 model for the task of distracted driver detection, using the source code provided by the authors in their official GitHub repository. The details of the Yolov7 model are left to the reader to look up, as they go beyond the scope of this paper.

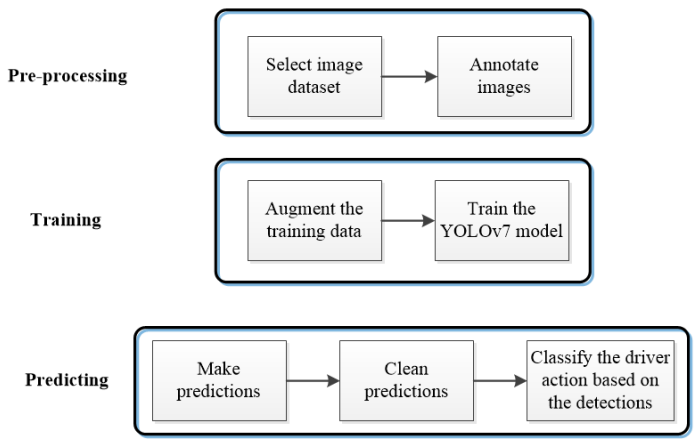

Figure 2: Three-stage methodology: pre-processing the images, training the Yolov7 model, predicting the driver action

3.2.1. Pre-processing

The Yolov7 model was trained on the State Farm (STF) distracted driver detection image dataset [3], following the three-stage methodology illustrated in Figure 2. The dataset consists of 22,457 images, including nine distracted driver classes and a safe driving class [3] (see Table 1 for all of the class names that are in the STF image dataset). As our approach focused on specific regions for prediction, we randomly selected 3,346 images for training, reserved 1,000 images for validation, and retained the original test split of 2,257 images for testing. The selected training and validation images were annotated using nine labels: eyes on the road, eyes off the road, hands on the wheel, hands on the face, no hands on the wheel, one hand on the wheel, using cell phone, talking on the phone, and drinking. The images were annotated using the labelling tool. Figure 3 shows the class-representative images from the STF image dataset.

Table 1: Classes in the STF, EZZ2021, AUC2, and CSIR datasets.

Class	Driver action
C0	Safe driving
C1	Text right
C2	Talk right
C3	Text left
C4	Talk left
C5	Adjust radio
C6	Drinking
C7	Reach behind
C8	Make-up
C9	Talking to passenger

In addition to the STF image dataset, we evaluated the cross-dataset performance of the proposed approach, using three additional test datasets for distracted driver detection. The test datasets include original test splits from the AUC2 dataset [11], the driver distraction dataset introduced by Ezzouhri *et al*. (EZZ2021) [20], and a custom distracted driver detection dataset created by the CSIR. All three datasets consisted of the same ten classes as the STF image dataset.

Table 2: Properties of the test image datasets used.

Image Dataset	Environment	Types of distraction	Participants	Image samples
AUC2	Real	1 save driving, 9 distracted activities	44	1,074
EZZ2021	Real	1 save driving, 9 distracted activities	9	3,719
State Farm (STF)	Real	1 save driving, 9 distracted activities	26	2,247
CSIR*	Real	1 save driving, 9 distracted activities	4	512

***Dataset not yet released to the public**

Figure 3: Class-representative images randomly sampled from the STF dataset

3.2.2. Training the YOLO model

To adapt the Yolov7 model for distracted driver detection, we performed fine-tuning using the annotated dataset. Initially, a standard Yolov7 model pre-trained on the Microsoft COCO dataset was used. The specific hyper-parameters employed during training varied across the different experiments; comprehensive information about the experiment-specific hyper-parameters is presented in Section 4.

3.2.3. Predicting

During the prediction stage, the trained Yolov7 model was loaded and used to make predictions. However, it was observed that the model often produced multiple detections on a single driver feature. To address this issue, a custom Python module was developed specifically for cleaning the predictions. This module operates by considering the confidence scores associated with each prediction. Predictions with low confidence scores are discarded, while predictions with high confidence scores are retained, resulting in a refined set of predictions.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Multi-class object detection vs multiple class-specific object detectors

4.1.1. Training and evaluation

The main goal of this experiment was to assess whether the performance of distracted driver detection in the proposed approach could be enhanced by employing multiple class-specific object detectors instead of a single multi-class object detector. To achieve this objective, two approaches were implemented: the first used a multi-class object detector, while the second relied on class-specific object detectors. By comparing the results of these two approaches, the effectiveness of using class-specific object detectors could be determined.

In the multi-class object detector-based approach, the Yolov7 model was trained using the annotated STF dataset, which consisted of nine classes, which encompassed various driver activities, such as eyes on the

road, eyes off the road, hands on the wheel, one hand on the wheel, no hands on the wheel, hands on the face, using a cell phone, talking on the phone, and drinking. On the other hand, the class-specific approach involved training three separate Yolov7 models. The first model focused specifically on the driver's head to detect whether their eyes were on the road. The second model concentrated on the driver's hands and the steering wheel to identify two driver activities: both hands on the wheel, and one hand on the wheel. Last, the third model was trained to detect four distracting activities that a driver might engage in while driving, namely hands on the face, using a cell phone, talking on the phone, and drinking. The hyper-parameters used for training all the models, as well as their performance measured in terms of mean average precision (mAP) on the validation dataset, are presented in Table 3. The hyper-parameters were obtained through a series of hyper-parameter-tuning experiments. The specific details of these experiments were not included in this paper because of space limitations.

Table 3: Training details of the multi-class object detection vs multiple class-specific object detectors experiment

Experiment		epochs	img_size	iou_t	anchor_t	mAP@0.5	mAP@0.95
Yolov7_multi-class		150	680	0.50	9.0	0.964	0.704
Yolov7_class-specific	head	100	1280	0.50	4.0	0.902	0.7438
	hands	100	1280	0.20	4.0	0.988	0.7929
	distractions	150	1280	0.20	4.0	0.925	0.7038

4.1.2. Results

Table 4 presents the accuracy performance of two approaches: the Yolov7 multi-class approach and the Yolov7 class-specific approach. Based on the results, it is evident that the Yolov7 multi-class approach performs well on the AUC2 and CSIR test datasets but performs poorly on the EZZ2021 and STF test datasets when compared with the Yolov7 class-specific approach. In summary, the Yolov7-Multi-class approach maintains a slightly higher average accuracy rate of 85.44% than the Yolov7-Class-specific approach, which achieves an average accuracy of 84.96%.

Table 4: Performance of the multi-class object detector and the class-specific object detectors

	Accuracy [%]				
Algorithm	**AUC2**	**CSIR**	**EZZ2021**	**STF**	**Average**
Yolov7 multi-class	91.62	62.63	90.91	96.62	85.44
Yolov7 class-specific	88.45	60.57	94.16	96.66	84.96
Improvement	**-3.46%**	**-3.29%**	**3.57%**	**0.04%**	**-0.56%**

In this study, distracted driver detection was treated as a binary problem. Consequently, the test datasets used in the evaluation were imbalanced, with only one class representing safe driving and nine classes representing distracted driving. This class imbalance introduced a significant difference in the number of instances between the safe driving and distracted classes. Thus relying solely on accuracy as a performance measure can be misleading. To address this issue, the F1 score was employed as a balanced measure, combining precision and recall into a single metric. Precision quantifies the accuracy of positive predictions, while recall assesses the model's ability to identify all positive instances correctly. The F1 score ranges from 0 to 1, with a score of 1 indicating perfect precision and recall and a score of 0 indicating poor performance. The F1 score was computed using the safe driving class as the positive class.

Table 5 shows the F1 scores of the proposed approach, using both the Yolov7 multi-class approach and the Yolov7 class-specific approach. Based on the F1-score performance, it could be observed that the Yolov7 class-specific approach outperformed the Yolov7 multi-class approach only on the EZZ2021 test dataset, exhibiting a 91.67% improvement. However, the Yolov7 class-specific approach showed a superior overall F1 score across all four test datasets.

Table 5: F1 scores of the proposed approach: multi-class object detector vs class-specific object detector

	F1-score				
Algorithm	**AUC2**	**CSIR**	**EZZ2021**	**STF**	**Average**
Yolov7 multi-class	0.75	0.42	0.36	0.86	0.60
Yolov7 class-specific	0.61	0.40	0.69	0.85	0.64
Improvement	-18.67%	-4.76%	91.67%	-1.18%	6.67%

Based on the analysis above, the Yolov7 class-specific approach did not enhance distracted driver detection accuracy, as it showed a slight decrease of 0.56% in overall accuracy. However, the F1-score results suggest that the Yolov7 class-specific approach significantly improved the overall balanced performance by 6.67%. In conclusion, the findings indicate that, while using multiple class-specific object detectors leads to a minor decrease in overall distracted driver detection accuracy (0.56%), there is a notable enhancement of 6.67% in the overall balanced performance, as indicated by the F1 score. For further insight into the performance of the algorithms, readers are encouraged to refer to the confusion matrices presented in Appendix A of this paper. These matrices provide a detailed breakdown of the models' performance in respect of correctly or incorrectly predicting each class.

4.2. Cross-dataset performance evaluation

The primary goal of the current work was to improve the cross-dataset performance of CNN-based distracted driver detection. In this experiment, the cross-dataset performance of the proposed approach and three other CNN-based approaches were evaluated and compared. The three evaluated approaches were a ResNet50 classification model [34], an EffecientNetB0 classification model [35], and a CNN-based model that used a background removal algorithm (Leekha_GrabCut) [2]. The details of the three approaches are left to the reader, as they go beyond the scope of this paper.

The specifics of the training and evaluation of the approaches can be found in Section 4.2.1. The results and analysis of the cross-dataset performance of the algorithms are provided in Section 4.2.2.

4.2.1. Training and evaluation

The ResNet50 and EfficientNetB0 architectures, initially pre-trained on the ImageNet dataset, were fine-tuned using the STF training dataset through the transfer learning framework. The top layers of these architectures were replaced with a GlobalAveragePooling2D layer, followed by a Dropout layer and a fully connected layer with two neurons. The hyper-parameters used for training are presented in Table 6.

Table 6: Hyper-parameters used for training the three comparison algorithms

Algorithm	**Learning rate**	**Epochs**	**Optimiser**	**Dropout**
ResNet50	Head: 0.001	15	Adam	0.2
	Fine-tuning: 1e-5	70	Adam	-
EfficientNetB0	Head: 0.001,	15	Adam	0.2
	Fine-tuning: 1e-5	70	Adam	-
Leekha_GrabCut	Head: 0.001	15	Adam	0.2
	Fine-tuning: 1e-5	70	Adam	-

As for the Leekha_GrabCut algorithm, it consisted of an EfficientNetB0 model that was pre-trained on ImageNet and subsequently fine-tuned using the STF training dataset. However, the GrabCut background removal algorithm was incorporated as a pre-processing stage within the data pipeline used for training the Leekha_GrabCut algorithm. To assess and compare the cross-dataset performances, each approach was tested on all four distracted driver detection test datasets.

4.2.2. Results

Figure 4 and Table 7 show the performance of the proposed approach, both Yolov7 multi-class and Yolov7-Class-specific, and of the other three CNN-based approaches. From the results, the following observations can be made:

- All of the approaches performed well on the STF test dataset, with an average accuracy of 95.68%.
- All of the approaches did not perform well on the CSIR test dataset, with an average accuracy of 56.88%.
- Comparing the proposed approach, encompassing both the Yolov7 multi-class and the Yolov7 class-specific approach, with the three CNN-based approaches, it outperformed them. Specifically, when comparing our approach with the Yolov7 multi-class approach, it showed a superior overall cross-dataset accuracy performance of 6% compared with the Leekha_GrabCut algorithm, which attained the highest accuracy among the three CNN-based approaches.
- Among the evaluated approaches, the ResNet50 classification model demonstrated the poorest overall cross-dataset performance, with an average accuracy of 79.93%.

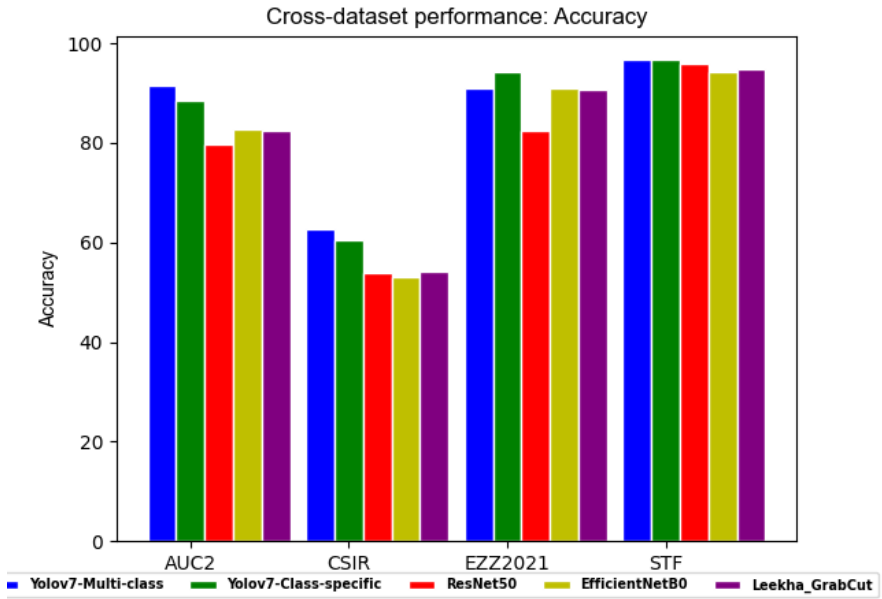

Figure 4: Cross-dataset performance of the algorithms on the four datasets measured using accuracy

Table 7: Performance of the algorithms on all four test datasets

	Accuracy [%]				
Algorithm	**AUC2**	**CSIR**	**EZZ2021**	**STF**	**Average**
Yolov7 multi-class	91.62	62.63	90.91	96.62	85.44
Yolov7 class-specific	88.45	60.57	94.16	96.66	84.96
ResNet50	79.70	53.80	82.33	95.91	79.93
EfficientNetB0	82.68	53.18	90.94	94.35	80.29
Leekha_GrabCut	82.50	54.21	90.73	94.84	80.57
Average	84.99	56.88	89.81	95.68	82.24

For further analysis, the F1 score was used to evaluate and compare the balanced performance of the algorithms. Figure 5 and Table 8 show the F1 score results of all of the approaches. Based on the results, the following observations can be made:

- All of the approaches demonstrated commendable performance on the STF test dataset, achieving an average F1 score of 0.78.
- Conversely, all of the approaches exhibited poor performance on the CSIR test dataset, with an average F1 score of 0.19.
- The proposed Yolov7 class-specific approach had the best overall balanced distracted driver detection performance.
- The ResNet50 classification model obtained an F1 score of 0 on the AUC2 test dataset. Similarly, the EfficientNetB0 model attained an F1 score of 0 on the CSIR test dataset. These scores indicated that both models did not make any correct predictions for the positive class, which is the safe driving class. In addition, the ResNet50 model demonstrated extremely low F1 scores on the CSIR and EZZ2021 test datasets, measuring 0.02 and 0.05 respectively. Consequently, the ResNet50 model showed the poorest balanced performance for distracted driver detection, followed by the EfficientNetB0 model.

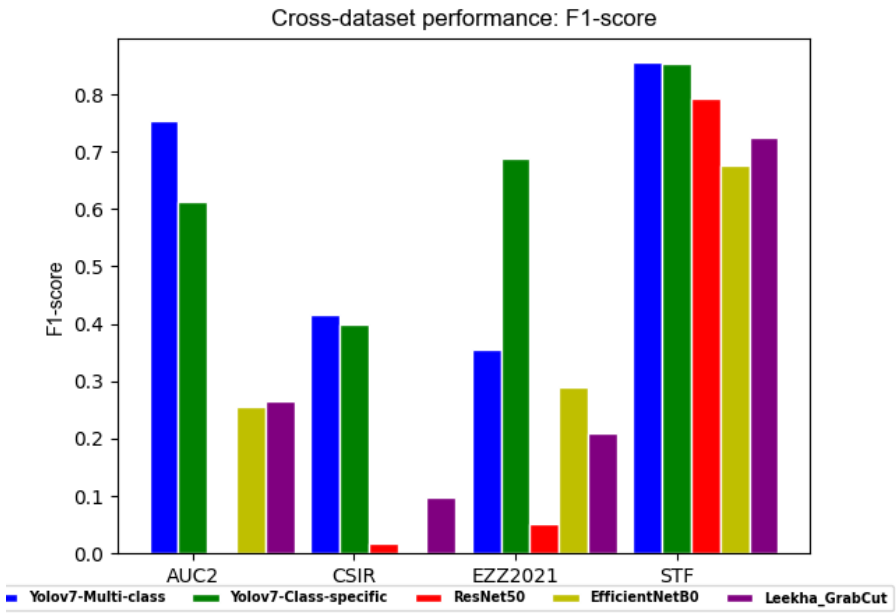

Figure 5: F1 scores of the approaches on the four datasets

Table 8: F1 scores of the approaches on the four datasets

	F1 score				
Algorithm	**AUC2**	**CSIR**	**EZZ2021**	**STF**	**Average**
Yolov7 multi-class	0.75	0.42	0.36	0.86	0.60
Yolov7 class-specific	0.61	0.40	0.69	0.85	0.64
ResNet50	0.00	0.02	0.05	0.79	0.22
EfficientNetB0	0.26	0.00	0.29	0.68	0.31
Leekha_GrabCut	0.27	0.10	0.21	0.73	0.32
Average	0.38	0.19	0.32	0.78	0.42

The results and analysis presented above reveal the following insights:

- The proposed approach demonstrated a significant improvement in the cross-dataset performance of CNN-based distracted driver detection. In comparison with the best-performing approach among the three CNN-based methods (Leeka_GrabCut), our approach achieved an overall increase of 6% in classification accuracy performance. Moreover, the overall balanced performance, as measured by the F1 score, was substantially improved by a factor of two. We attribute these improvements to the fact that our approach focused on crucial driver body parts and their associated activities.
- All approaches exhibited satisfactory performance on the STF test dataset, which was expected, since all of the algorithms were trained on the STF training dataset. The similarity in characteristics between the STF training and STF test datasets, such as camera viewpoint, drivers used, and cars used, contributed to this favourable performance.
- However, all of the algorithms encountered difficulties when dealing with the custom CSIR test dataset. This can be attributed to the low image quality of the images in the CSIR test dataset, particularly because of problematic lighting conditions (see Figure 6).
- Based on the poor performance of all of the algorithms on the CSIR dataset, there is still much work to be done in order to make the algorithms more suitable for real-world deployment and integration.

Overall, these findings highlight the effectiveness of the proposed approach in enhancing the cross-dataset performance of distracted driver detection, while also emphasising the impact of dataset characteristics on algorithm performance.

Figure 6: Sample qualitative results from the CSIR dataset

5. CONCLUSIONS AND FUTURE WORK

The purpose of this study was to investigate whether the performance of convolutional neural network-based distracted driver detection could be improved by detecting driver body parts and classifying their state into activities across different datasets. To achieve this, we proposed an object detection-based approach that used the Yolov7 model to detect and classify driver body parts' activities. The experimental results demonstrated that our proposed approach significantly improved cross-dataset performance. We observed an accuracy improvement of 6% in classification. Most importantly, we observed a significant overall balanced (F1 score) performance improvement of a factor of two.

The results indicated that the proposed approach outperformed the other CNN-based approaches. The approach's emphasis on important driver body parts and activities contributed to its superior performance. In addition, all of the algorithms showed satisfactory results on the STF test dataset, which shared similar characteristics with the training dataset. However, problems were encountered when dealing with the custom CSIR test dataset, mainly as a result of lower image quality and difficult lighting conditions.

Future work could focus on several aspects. First, the poor performance on the CSIR dataset could be investigated further by using a qualitative evaluation approach. Second, conducting user studies and real-world deployments of the developed models would provide valuable insights into their practical effectiveness and their potential for integration into real-time driver monitoring systems. In addition, creating larger and more diverse datasets, encompassing different driving scenarios and environmental

factors, could help to improve the generalisability and robustness of the proposed approach. Finally, specialised domain generalisation approaches could be explored for deep learning distracted driver detection.

REFERENCES

[1] **Y. Lecun, K. Kavukcuoglu, and C. Farabet**, *Convolutional Networks and Applications in Vision*, in Proceedings of 2010 IEEE International Symposium on Circuits and Systems, ISCAS, 30 May-02 June 2010, Paris, France, 2010.

[2] **M. Leekha, M. Goswami, R. R. Shah, Y. Yin, and R. Zimmermann**, *Are you paying attention? Detecting distracted driving in real-time*, in Proceedings of 2019 IEEE Fifth International Conference on Multimedia Big Data, BigMM, 11-13 September 2019, Singapore, 2019.

[3] **A. Montoya, D. Holman, T. Smith, and W. Kan**, *State farm distracted driver detection | Kaggle*, available from: https://www.kaggle.com/c/state-farm-distracted-driver-detection (accessed Mar. 28, 2022).

[4] **F. Zandamela, T. Ratshidaho, F. Nicolls, and G. Stoltz**, *Cross-dataset performance evaluation of deep learning distracted driver detection algorithms*, in Proceedings of 2022 Rapid Product Development Association of South Africa – Robotics and Mechatronics – Pattern Recognition Association of South Africa – South African Advanced Materials Initiative, RAPDASA-RobMech-PRASA-CoSAAMI, 9-11 November 2022, Somerset West, South Africa, 2022.

[5] **C. Yan, H. Jiang, B. Zhang, and F. Coenen**, *Recognizing driver inattention by convolutional neural networks*, in Proceedings of 2015 8th International Congress on Image and Signal Processing, CISP, 14-16 October 2015, Shenyang, China, 2015.

[6] **C. Yan, F. Coenen, and B. Zhang**, *Driving posture recognition by convolutional neural networks*, in Proceedings of 2015 11th International Conference on Natural Computation, ICNC, 15-17 August 2015, Zhangjiajie, China, 2015.

[7] **Z. A. Varaich and S. Khalid**, *Recognizing actions of distracted drivers using Inception v3 and Xception convolutional neural networks*, in Proceedings of 2019 2nd International Conference on Advancements in Computational Sciences, ICACS, 18-20 February 2019, Lahore, Pakistan, 2019.

[8] **F. R. da Silva Oliveira and F. C. Farias**, *Comparing transfer learning approaches applied to distracted driver detection*, in 2018 IEEE Latin American Conference on Computational Intelligence, LA-CCI, 07-09 November 2018, Guadalajara, Mexico, 2018.

[9] **Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, and F. Y. Wang**, Driver activity recognition for intelligent vehicles: A deep learning approach, *IEEE Transactions on Vehicular Technology*, 68, pp. 5379–5390, 2019.

[10] **B. Qin, J. Qian, Y. Xin, B. Liu, and Y. Dong**, Distracted driver detection based on a CNN with decreasing filter size, *IEEE Transactions on Intelligent Transportation Systems*, 23, pp. 1–12, 2021.

[11] **H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa**, Driver distraction identification with an ensemble of convolutional neural networks, *Journal of Advanced Transportation*, 2019, 4125865, 2019.

[12] **C. Huang, X. Wang, J. Cao, S. Wang, Y. Zhang**, A hybrid CNN framework for behavior detection of distracted drivers, *IEEE Access*, 8, pp. 109335–109349, 2020.

[13] **M. R. Arefin, F. Makhmudkhujaev, O. Chae, and J. Kim**, *Aggregating CNN and HOG features for real-time distracted driver detection*, in Proceedings of 2019 IEEE International Conference on Consumer Electronics, ICCE, 11-13 January 2019, Las Vegas, NV, USA, 2019.

[14] **M. Wu, X. Zhang, L. Shen, and H. Yu**, *Pose-aware multi-feature fusion network for driver distraction recognition*, in Proceedings of 2020 25th International Conference on Pattern Recognition, ICPR, 10-15 January 2021, Milan, Italy, 2021.

[15] **M. H. Alkinani, W. Z. Khan, Q. Arshad, and M. Raza**, A hybrid scheme for the detection of distracted driving through fusion of deep learning and handcrafted features, *Sensors*, 22, 1864, 2022.

[16] **C. Ou and F. Karray**, Enhancing driver distraction recognition using generative adversarial networks, *IEEE Transactions on Intelligent Vehicles*, 5, pp. 385–396, 2020.

[17] **J. Mafeni Mase, P. Chapman, G. P. Figueredo, and M. Torres Torres**, *A hybrid deep learning approach for driver distraction detection*, in Proceedings of 2020 International Conference on Information and Communication Technology Convergence, ICTC, 21-23 October 2020, Jeju, South Korea, 2020.

[18] **F. Nel and M. Ngxande**, *Driver activity recognition through deep learning*, in Proceedings of 2021 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa, SAUPEC/RobMech/PRASA, 27-29 January 2021, Potchefstroom, South Africa, 2021.

[19] **N. Moslemi, R. Azmi, and M. Soryani**, *Driver distraction recognition using 3D convolutional neural networks*, in Proceedings of 2019 4th International Conference on Pattern Recognition and Image Analysis, 06-07 March 2019, Tehran, Iran, 2019.

[20] **A. Ezzouhri, Z. Charouh, M. Ghogho, and Z. Guennoun**, Robust deep learning-based driver distraction detection and classification, *IEEE Access*, 9, pp. 168080–168092, 2021.

[21] **T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides**, *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, CVPRW, 11-12 June 2016, Las Vegas, Nevada, 2016.

[22] **Q. Xiong, J. Lin, W. Yue, S. Liu, Y. Liu, and C. Ding**, *A deep learning approach to driver distraction detection of using mobile phone*, in Proceedings of 2019 IEEE Vehicle Power and Propulsion Conference, VPPC,14-17 October 2019, Hanoi, Vietnam, 2019.

[23] **J. Wang, Z. C. Wu, F. Li, and J. Zhang**, A data augmentation approach to distracted driving detection, *Future Internet*, 13(1), pp. 1-11, 2021.

[24] **F. Sajid, A. R. Javed, A. Basharat, N. Kryvinska, A. Afzal, and M. Rizwan**, An efficient deep learning framework for distracted driver detection, *IEEE Access*, 9, pp. 169270–169280, 2021.

[25] **Y. Li, L. Wang, W. Mi, H. Xu, J. Hu, and H. Li**, *Distracted driving detection by combining ViT and CNN*, in Proceedings of 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design, CSCWD, 04-06 May 2022, Hangzhou, China, 2022.

[26] **O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla**, M. Bernstein, and A.C. Berg, ImageNet large scale visual recognition challenge, *International Journal of Computer Vision*, 115, pp. 211-225, 2015.

[27] **E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell**, *Adversarial discriminative domain adaptation*, in Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 21-26 July 2017, Honolulu, HI, USA, 2017.

[28] **M. Cetinkaya and T. Acarman**, *Driver activity recognition using deep learning and human pose estimation*, in Proceedings of 2021 International Conference on INnovations in Intelligent SysTems and Applications, INISTA, 25-27 August 2021, Kocaeli, Turkey, 2021.

[29] **C. Streiffer, R. Raghavendra, T. Benson, and M. Srivatsa**, *DarNet: A deep learning solution for distracted driving detection*, in Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial track, 11-15 December 2017, Las Vegas, Nevada, 2017.

[30] **S. Yan, Y. Teng, J. S. Smith, and B. Zhang**, *Driver behavior recognition based on deep convolutional neural networks*, in Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD, 13-15 August 2016, Changsha, China, 2016.

[31] **G. Gkioxari, R. Girshick, and J. Malik**, *Contextual action recognition with R∗CNN*, in Proceedings of the IEEE International Conference on Computer Vision, ICCV, 13-16 December 2015, Santiago, Chile, 2015.

[32] **J. Redmon, S. Divvala, R. Girshick, and A. Farhadi**, *You only look once: Unified, real-time object detection*, in Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 27-30 June, Las Vegas, Nevada, 2016.

[33] **C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao**, *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 18-22 June 2023, Vancouver, Canada, 2023.

[34] **K. He, X. Zhang, S. Ren and J. Sun**, *Deep residual learning for image recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 27-30 June, Las Vegas, Nevada, 2016.

[35] **M. Tan and Q. V. Le**, *EfficientNet: Rethinking model scaling for convolutional neural networks*, in Proceedings of the 36th International Conference on Machine Learning, ICML, 11-13 June 2019, Long Beach, California, 2019.

APPENDIX A: CONFUSION MATRICES OF THE ALGORITHMS

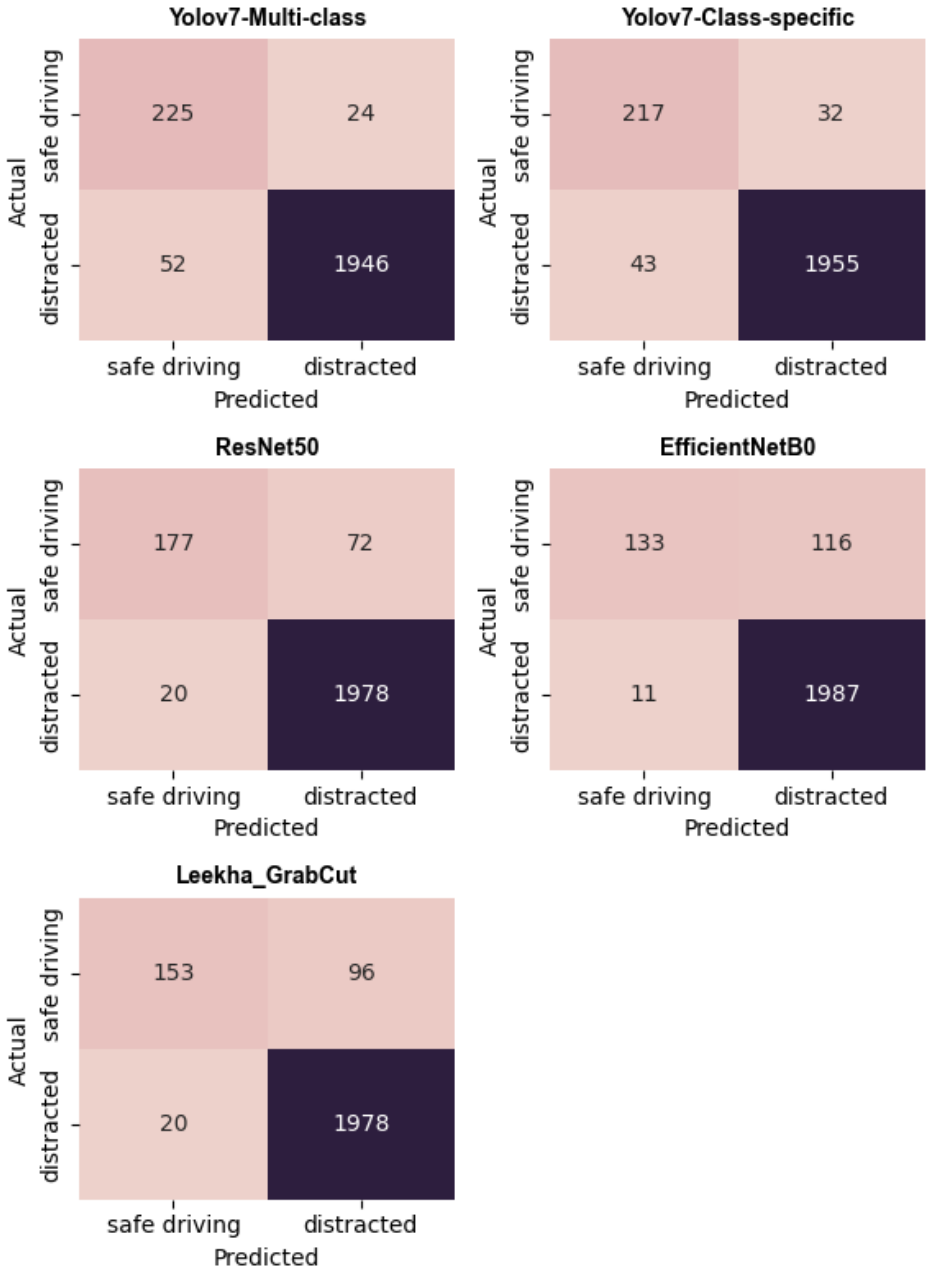

Figure 7: Confusion matrices of the algorithms on the STF-test dataset.

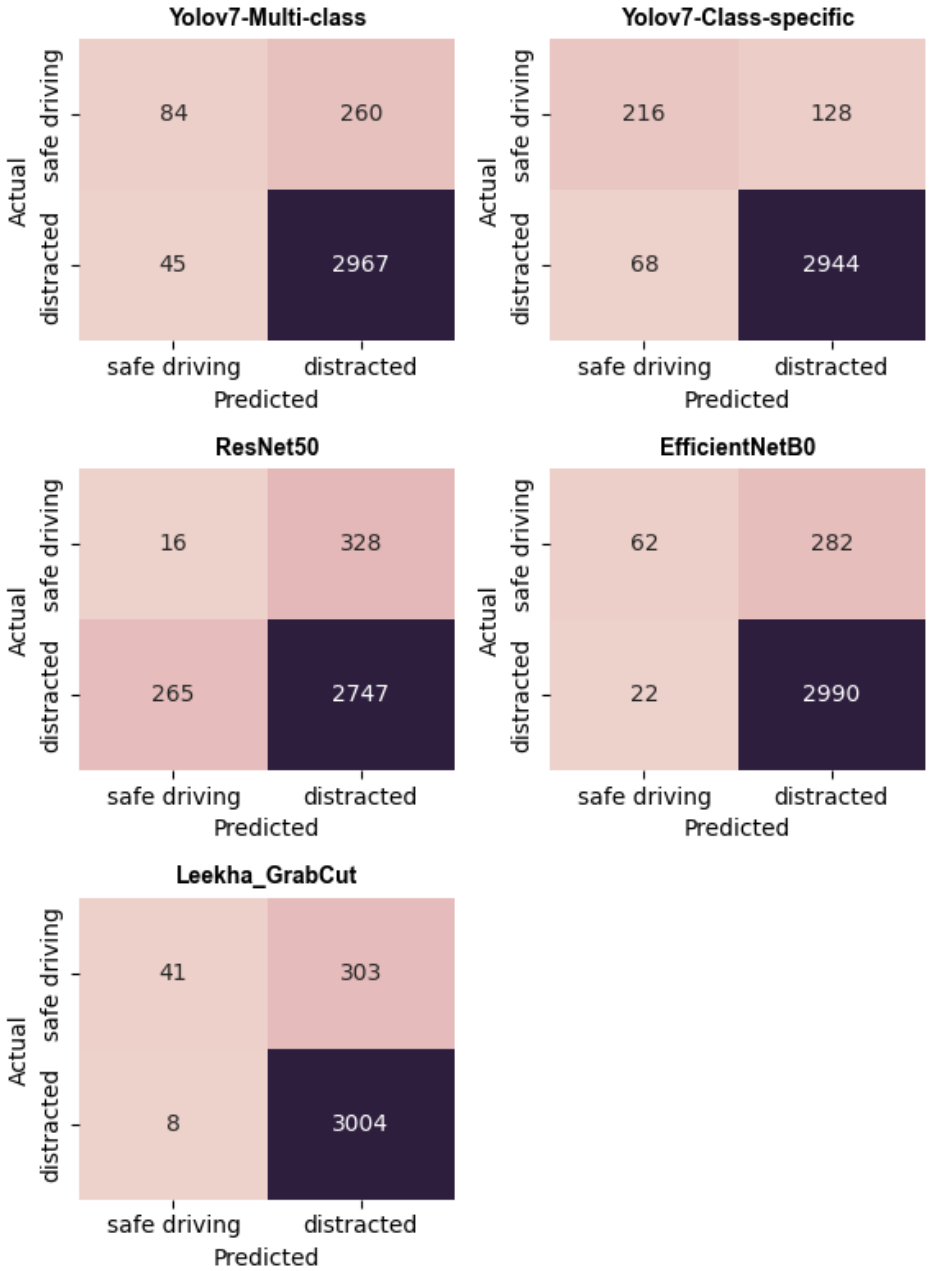

Figure 8: Confusion matrices of the algorithms on the EZZ2021-test dataset.

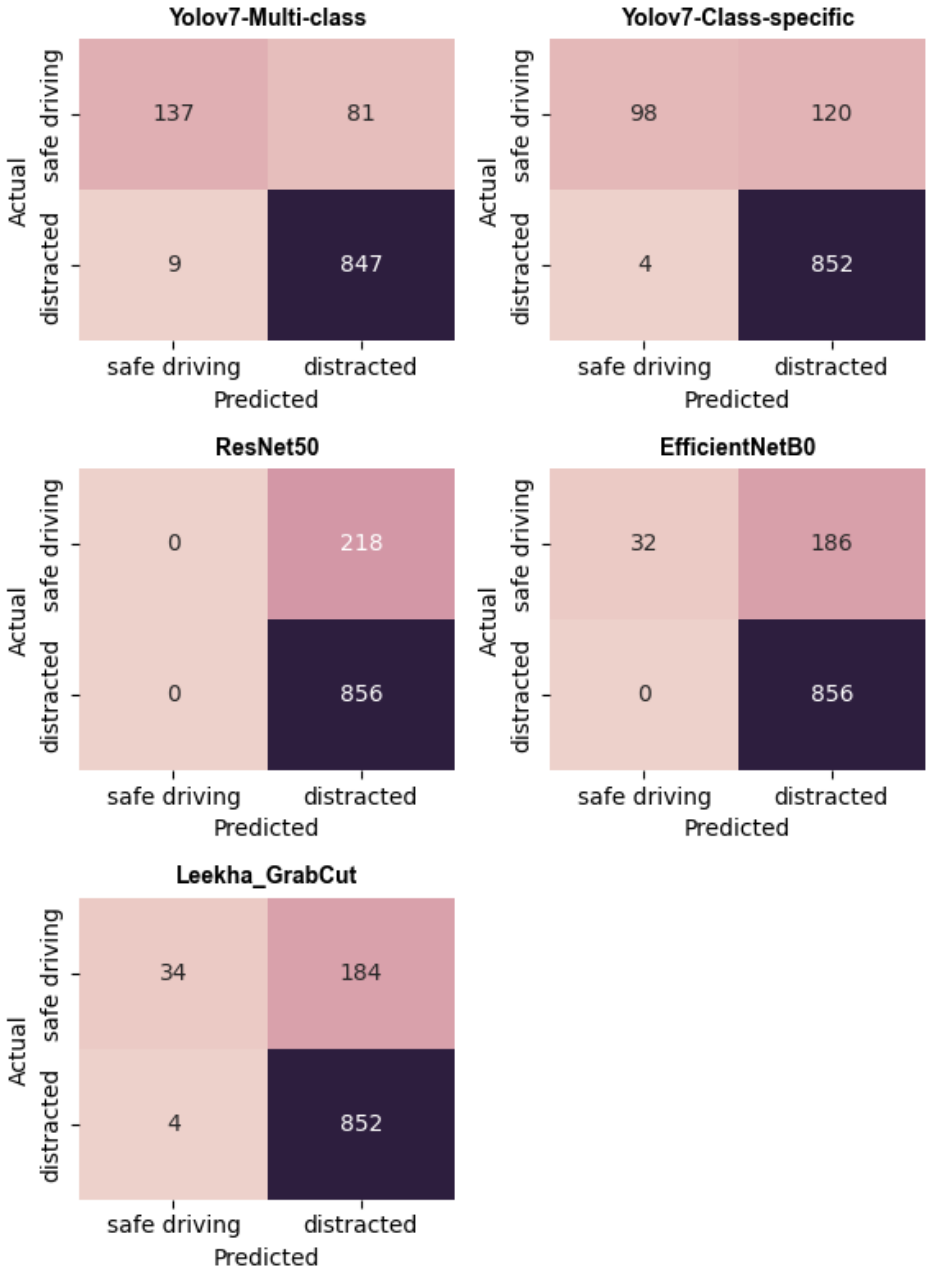

Figure 9: Confusion matrices of the algorithms on the AUC2-test dataset.

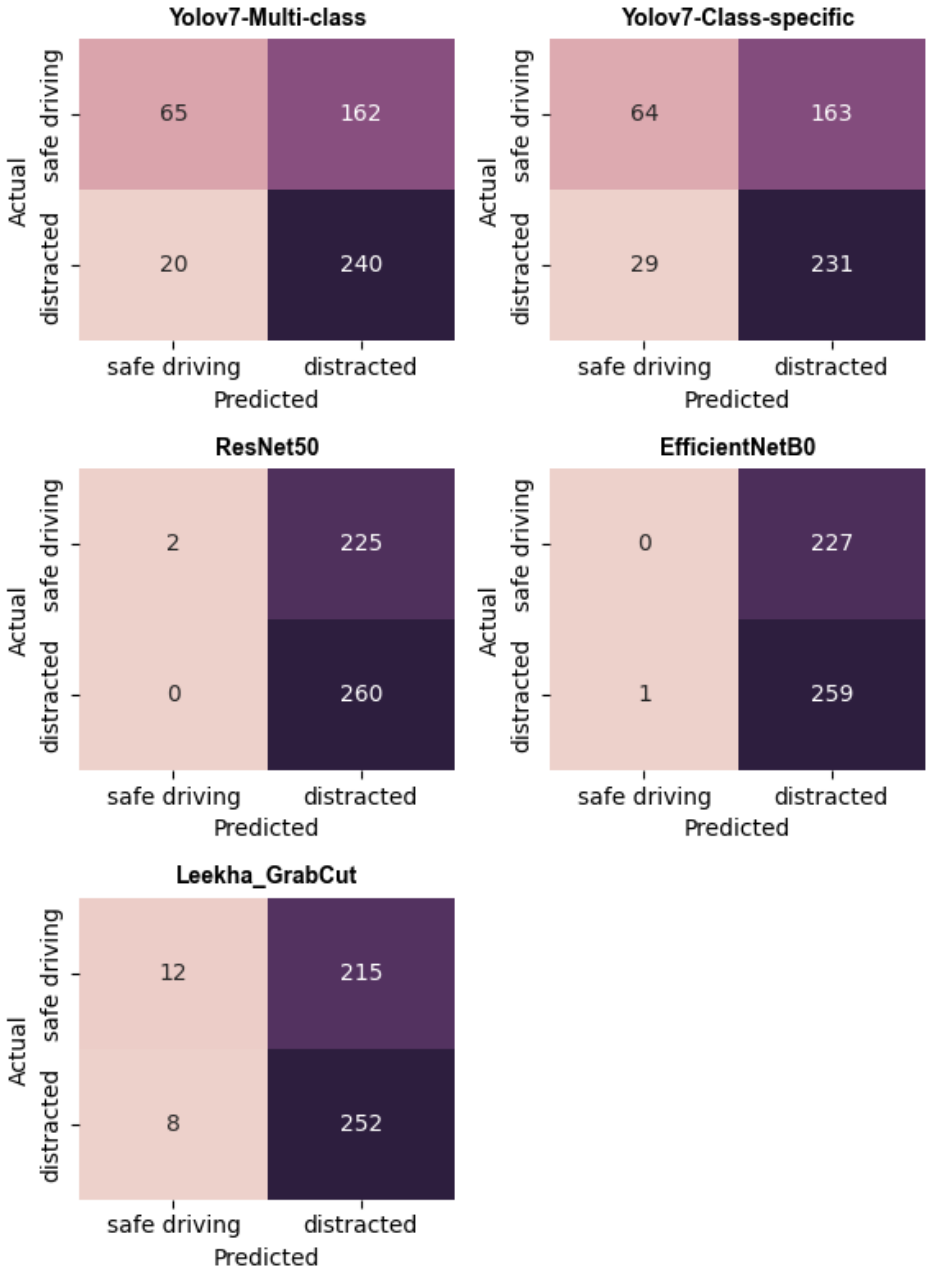

Figure 10: Confusion matrices of the algorithms on the CSIR-test dataset.