Detecting inter-sectional accuracy differences indriver drowsiness detection algorithms.

Citation for published version:
Ngxande, M, Tapamo, J-R & Burke, M 2020, Detecting inter-sectional accuracy differences indriver drowsiness detection algorithms. in *Proceedings of the Southern African Universities Engineering Conference/RobMech/Pattern Recognition Association of South Africa 2020.* Institute of Electrical and Electronics Engineers (IEEE), pp. 1-6, Southern African Universities Engineering Conference/RobMech/Pattern Recognition Association of South Africa 2020, Cape Town, South Africa, 29/01/20. https://doi.org/10.1109/SAUPEC/RobMech/PRASA48453.2020.9041105

Digital Object Identifier (DOI):
10.1109/SAUPEC/RobMech/PRASA48453.2020.9041105

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the Southern African Universities Engineering Conference/RobMech/Pattern Recognition Association of South Africa 2020

Detecting inter-sectional accuracy differences in driver drowsiness detection algorithms.

Mkhuseli Ngxande
CSIR Defence, Peace, Safety and Security
Optronic Sensor Systems,
Pretoria, South Africa
Email: mngxande@csir.co.za

Jules-Raymond Tapamo
School of Electrical, Electronic and Computer Engineering
University of Kwa-Zulu Natal,
Durban, South Africa
Email: tapamoj@ukzn.ac.za

Michael Burke
Mobile Intelligent Autonomous Systems
Modelling and Digital Sciences
Council for Scientific and Industrial Research
Pretoria, South Africa
Email: michaelburke@ieee.org

Abstract*—Convolutional Neural Networks (CNNs) have been used successfully across a broad range of areas including data mining, object detection, and in business. The dominance of CNNs follows a breakthrough by Alex Krizhevsky which showed improvements by dramatically reducing the error rate obtained in a general image classification task from 26.2% to 15.4%. In road safety, CNNs have been applied widely to the detection of traffic signs, obstacle detection, and lane departure checking. In addition, CNNs have been used in data mining systems that monitor driving patterns and recommend rest breaks when appropriate. This paper presents a driver drowsiness detection system and shows that there are potential social challenges regarding the application of these techniques, by highlighting problems in detecting dark-skinned drivers faces. This is a particularly important challenge in African contexts, where there are more dark-skinned drivers. Unfortunately, publicly available datasets are often captured in different cultural contexts, and therefore do not cover all ethnicities, which can lead to false detections or racially biased models. This work evaluates the performance obtained when training convolutional neural network models on commonly used driver drowsiness detection datasets and testing on datasets specifically chosen for broader representation. Results show that models trained using publicly available datasets suffer extensively from over-fitting, and can exhibit racial bias, as shown by testing on a more representative dataset. We propose a novel visualisation technique that can assist in identifying groups of people where there might be the potential of discrimination, using Principal Component Analysis (PCA) to produce a grid of faces sorted by similarity, and combining these with a model accuracy overlay.

***Index Terms*—CNNs, Road Safety, Drowsiness Detection, Biased models.**

I. Introduction

The convolutional neural network (CNN) has rapidly gained popularity in many social aspects and has been applied across a range of areas, including self-driving cars, collision detection, identification of criminal activities, and to aid the granting of bank loans. Historically, these tasks were generally performed by humans, but the advancement of machine learning is leading to the automation of these processes [1]. CNNs are a multistage mechanism that learns data representations in order to fulfil a specific goal. However, these can suffer from challenges with regard to generalisation. For example, a system that is trained to detect road lanes in an urban environment and then deployed in rural areas could lead to false detections. Furthermore, driver drowsiness detection systems predominately trained on a certain race or ethnicity may not perform well when tested across multiple races. This can potentially result in algorithmic discrimination if the trained model is unable to handle differing skin complexions and facial features [2]. This raises concerns in African contexts, where many cars with driver drowsiness detection systems are imported [3].

For example, the majority (80.8%) of South African citizens identify as black nationals [4]. Deploying a system that is trained in different contexts, for example, using a dataset captured in Asia, could result in failure if trained models learn to use skin complexion for decision-making. The taxi industry dominates public transportation used in South Africa. Statistics South Africa [5] report that about 76% of citizens in the country use public transportation to get to their destinations, with private minibus taxis a primary mode of transport (51.0%), followed by busses at 18.1 % and trains at 7.6%.

The alarming statistics of road accidents in South Africa has led to the investigation of technologies to reduce these high numbers of accidents. A Statistics South Africa report showed that in 2015, there was a 2% increase in mortality over the 2014 financial year, with about 12944 deaths caused by accidents [6]. Furthermore, in 2016 there were about 14 071 deaths, which was a 9% increase over 2015 [6].

Drowsiness detection systems that are currently implemented are typically available only in high-end vehicles, which disadvantages citizens using public transport. As a result, a number of researchers have aimed to develop similar systems on mobile phones, which are more easily accessible [7]. In addition to easily accessible systems, researchers also focused on augmenting vehicle control units with machine learning techniques to help reduce road accidents [8, 9]. However, if a large benchmark dataset that is representative of all race ethnicities is not used for training, systems like these can easily fail. Sikander and Anwar conducted a review of existing technologies for detection of fatigue in drivers, where various techniques and features were examined [10]. Of the 23 fatigue detection systems reviewed here, 12 relied on machine learning techniques. Moreover, in [11] it is shown that CNNs tended to outperform other technologies for driver drowsiness detection. However, the authors note that there is no large benchmarking dataset covering a wide range of ethnicities with which to conclusively test the efficiency of CNNs against other technologies.

This paper aims to highlight the challenges of using unrepresentative images to train vision-based driver drowsiness detection systems. In this article, we use a range of pre-trained convolutional neural networks, modifying the last layers by retraining these on a number of popular drowsiness detection datasets including the ULg Multimodality Drowsiness Dataset (DROZY), the National Tsinghua University Drowsy Driver Detection database (NTHU-DDD), and the Closed Eyes in the Wild dataset (CEW). We test models trained using these datasets on a test set more suited to South African contexts. Results show that the three evaluated datasets produce high drowsiness detection accuracies when tested on held out portions of the original datasets, but that the accuracies obtained decreased substantially when these models were evaluated using the more representative South African test set. This decrease in performance is due to models overfitting. Overfitting models can be a particular challenge, as it can be difficult to establish where models are failing to generalise. This work introduces a new visualisation technique to identify potential population groups for whom additional training data may be required, so as to rectify the problem of unrepresentative models in driver drowsiness detection systems.

This paper is structured as follows. Section II provides an overview of related work, which is followed by a discussion on algorithmic bias and convolutional neural networks. This is followed by an introduction of the proposed visualisation technique in Section III, and a description of the experimental methods, including the architectures and datasets investigated in Section IV. Finally, results and the conclusions are provided in Section V and VI respectively.

II. BACKGROUND AND RELATED WORK

This section briefly summarises previous approaches to driver drowsiness detection and existing benchmarking datasets used for testing these algorithms. Related work and advances in convolutional neural network architectures are also discussed.

A. DRIVER DROWSINESS DETECTION SYSTEMS

A number of drowsiness detection systems rely on convolutional neural networks. Sanghyuk et al. [12] proposed a deep architecture called deep drowsiness detections (DDD). This architecture consists of three deep convolutional neural networks including AlexNet [13], VGGNet-FaceNet [14], and FlowImageNet [15]. The output of these networks is concatenated and fed into a softmax classification layer for drowsiness detection. The DDD system was tested on the NTHU-drowsy driver detection dataset, but the authors noted that the NTHU-drowsy lacked reliable ground truth labeling, which led them to use a substitute evaluation dataset for testing. The authors also noted that there was a lack of previously benchmarked datasets to compare with the publicly available NTHU-drowsy dataset.

Reddy et al. [16] proposed a compressed deep neural network model that can be deployed on an embedded board. The authors note that for their focus, the NTHU-drowsy dataset had an unsuitable capture angle and inappropriate class labels. In addition, the authors noted that the DROZY dataset was also unsuitable because the images contain sensor patches attached to a subjects face, which could interfere with the results obtained. Their solution was to use a custom dataset and compare the efficacy of their approach to a number of convolutional neural network architectures, including faster RCNN, VGG-16, and AlexNet.

Lyu et al. [17] proposed a sequential multi-granularity deep framework for detection of driver drowsiness. This framework consists of two components, a multi-granularity CNN and a deep long-short-term memory network (deep LSTM). A contribution of this work was to utilise a group of parallel CNN extractors. The deep LSTM was applied on facial representations to identify long-term features of drowsiness over a sequence of frames. The model was evaluated on the NTHU-drowsy dataset in addition to a new dataset named Forward Instant Driver Drowsiness Detection (FI-DDD). The FI-DDD is a re-labeled NTHU-drowsy dataset, as the authors note that it is difficult to locate drowsy states temporally with high precision using the NTHU-drowsy labels.

Following a different approach, Dwivedi et al. [18] introduced a more diverse dataset that includes persons with different skin tones, eye shapes and eye sizes. This dataset was used to test a CNN with a final softmax classification layer, but unfortunately, the dataset is not publicly available for comparison.

A recent study by Kim et al. proposed a deep CNN based on the classification of opened and closed eyes using a visible light camera sensor [19]. They used the ZJU eye blink dataset in addition to their own dataset collected for performance analysis. Here, the ResNet-50 [20] architecture was adopted, with a modified fully connected layer. The system outperformed AlexNet [13], GoogleNet [21], VGGFace fine-turning [14], and HOG-SVM [22].

B. ALGORITHMIC BIAS

It is clear that a number of modern drowsiness detection systems rely on convolutional neural networks, and many of these models are trained and tested on only a few datasets. These datasets do not always cover a wide range of different races and ethnicities with varying facial features. As a result, these systems are vulnerable to problems regarding algorithmic bias. This paper evaluates the efficacy of the NTHU-drowsy, DROZY, and CEW datasets in a South African context, where racial bias is likely to have a significant impact, using a more representative dataset.

Unfortunately, the road safety community is not the only field that is affected by algorithmic bias caused by using unrepresentative datasets for training. Buolamwini [23] have documented extensive algorithmic bias in face detection systems, which fail on faces with darker skin-tones, while Renda et. al have highlighted bias in predictive policing [24], by showing that a system called PredoPol used to send police to crime hotspots tends to send police to areas where there are large numbers of dark-skinned people or Muslims. Wen et al. [25] analysed a face spoof detection algorithm, designed to recognise fake faces using image distortion analysis and reported that most current systems mis-classify individuals with dark-skinned faces as spoof attacks.

Furthermore, Brauneis and Goodman added to the discussion of how to deploy AI-based systems, evaluating a number of scenarios where dark-skinned people could be misclassified [26]. Zou and Schiebinger [27] shows that there is often bias in machine learning algorithms, which can be caused by a variety of factors including imbalanced training datasets, representation of the datasets and also algorithms themselves. They suggest that datasets should include information on how they were collected and the demographics of participants therein using meta-data. However, this process is also problematic, as it requires the classification of people into different ethnic groups or categories, which is itself a subjective and questionable task. In an attempt to address this, we propose a visualisation technique that can identify groups or individuals on whom algorithms fail, without the need for pre-classification or meta-data.

C. CONVOLUTIONAL NEURAL NETWORKS

CNNs have dominated many computer vision tasks since the breakthrough shown by Alex Krizhevsky in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2012 [13]. Convolutional neural networks are a collection of sequentially stacked layers, which typically consist of convolution, pooling, and fully connected layers, with model parameters trained using gradient descent. The convolutional layer takes in a three-dimensional image tensor, comprising d channels of feature maps, sized $h \times w$ pixels. Here, h denotes the height and w the width of the input image tensor. Convolutional layers extract low level features from an input tensor I by means of a convolution with a two-dimensional kernel K,

$$\begin{aligned} c(i,j) &= (K * I)(i,j) \\ &= \sum_m \sum_n I(i-m, i-n)K(m,n) + b_{m,n} \end{aligned} \qquad (1)$$

where $b_{m,n}$ is a bias parameter, and i, j denote the coordinates of a feature map pixel. After convolution, an activation function is applied to introduce non-linearity and produce an output feature map $\mathbf{a}$, comprising elements, $a(i,j) = f(c(i,j))$. A number of activation functions can be used, but the ReLU function,

$$f(c) = \begin{cases} c, & \text{if } c > 0, \\ 0, & \text{if } c \leq 0, \end{cases} \qquad (2)$$

is typically used for convolutional neural networks in order to avoid vanishing gradients in deeper networks.

Pooling layers are often applied after activations. Here, downsampling is applied to reduce the dimensionality of the image by a grouping operation over activations in small spatial regions of the input image. For example, max pooling returns the maximum value of the input region. Max pooling is also used to control over-fitting. Downsampling can also be achieved by using a convolutional layer with larger filtering strides.

A fully connected output layer is typically the final layer in a convolutional neural network, producing a network output:

$$\mathbf{Z} = \mathbf{W}\mathbf{a} + \mathbf{b} \qquad (3)$$

where $\mathbf{W}$ is a weight matrix and $\mathbf{a}$ the input from the previous layer. This is typically followed by a final activation layer. The activation function that is used for this paper is the sigmoid function, which is commonly used for binary classification tasks.

Convolutional neural networks are trained using gradient descent to find model parameters that minimise some loss. For binary classification tasks like drowsiness detection, the binary cross entropy loss given by

$$J = -\sum_{i=1}^{r} y_i log(o_i), \qquad (4)$$

is typically applied, where y is a vector of one-hot encoded labels, and o is the output probability produced by the final sigmoid layer. This loss function is typically minimised using stochastic gradient descent schemes to adjust model weight and bias parameters, with the Adam optimiser [28] often favoured.

A number of convolutional neural network architectures have been developed. The following section briefly highlights some of the improvements made thus far, by highlighting results in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC):

ZFNet (2013) In 2013 the winners of the ILSVRC were Matthew Zeiler and Rob Fergus from New York University [29]. Their model achieved an 11.2% error rate, improving upon 2012 error rate of 15.4% obtained by AlexNet. Although ZFNet is similar to AlexNet, ZFNet was modified by introducing a deconvNet and decreasing filter sizes from 11 x 11 pixels to 7 x 7 pixels.

GoogleNet (2014) GoogleNet is a 22 layer convolutional neural network that won the 2014 ILSVRC challenge. The novelty of this work was the introduction of the Inception module [21]. This module aims to reduce computational costs while increasing the width and depth of the network. The inception module has been extended several times with recent iterations including Inception-V3 [30] and Inception-V4 [31] models.

VGGNet (2014) The VGGNet developed by Karen Simonyan and Andrew Zisserman consists of two versions (VGG16 and VGG19 models), and took second place in the

ILSVRC 2014 challenge [32]. The two network architectures have depths of 16 and 19 layers respectively. VGGNet decreased the filter sizes of ZFNet to 3x3 with the motivation that these smaller filter sizes are capable of gathering more information from input images.

ResNet (2015) This network, developed by Microsoft Research Asia, won the 2015 ILSVRC with an error rate of 3.6% [20]. This model uses a residual learning framework that aims to simplify the training of deeper networks and yield higher accuracy. This network consists of 152 layers and was extended to 1001 layers on CIFAR-10, achieving an error rate of 4.62% [33].

It is clear that there is a trend of increasing the depth of the network, producing increasing performance, while reducing computational costs. However, these trends can make convolutional neural networks more vulnerable to over-fitting. Strategies for avoiding over-fitting include Batch Normalization [34] and Dropout [35].

III. VISUALISING CONVOLUTIONAL NEURAL NETWORKS

Visualisation is a commonly used technique to interpret trained convolutional neural networks, so as to improve the architecture or identify model failures. For example, saliency visualisation helps to identify which image areas contribute strongly to the network output. This technique was introduced by Simonyan et al. who presented two approaches to visualise what a neural network learns [36]. Here, the gradient of the class score with respect to image pixels is computed to determine the contribution of each pixel to the final output prediction [37]. Many previous approaches [29, 38, 39, 40, 41] tried to come up with visualisation solutions to better understand what each layer learns, but none of these directly address the challenge of potential racial bias in machine learning.

In this work, we introduce a visualisation technique building on Principal Component Analysis (PCA) to help identify population groups where models fail to perform well. Here, we use PCA to project images onto a 2-dimensional grid such that images are located near other images of similar appearance. PCA is a dimension reduction technique that can be used to compress a large number of features to a smaller number while retaining dominant information [42]. This is done by transforming data into an orthogonal subspace where axes (Principal components) align with the directions of maximum variance in the data. In this work we use singular value decomposition (SVD) to perform PCA. Let $\mathbf{X}$ be the matrix of images, formed by reshaping images $\mathbf{x}_i$ into row vectors (where $i = 1...N$, and N is the number of images in the dataset) and stacking these vertically to form an N x P matrix. Here, P denotes the number of pixels in each image. PCA starts by mean centering the matrix of images, which is accomplished by subtracting the mean image $\mu_i = \frac{1}{N}\sum_i^N \boldsymbol{X}_i$ from each 1 x P dimensional row vector, $\mathbf{X}_i$ in the matrix of images,

$$\hat{\mathbf{X}}_i = \mathbf{X}_i - \mu_i$$

The mean centred matrix N x P dimensional matrix of images $\hat{\mathbf{X}}$ is then decomposed using singular value decomposition (SVD),

$$\hat{\mathbf{X}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$$

Here, $\mathbf{U}$ is a $P \times N$ unitary matrix, $\mathbf{V}$ is a $P \times P$ unitary matrix and $\boldsymbol{\Sigma}$ is a diagonal matrix comprising the singular values of $\hat{\mathbf{X}}$ in decreasing order [43]. A reduced dimensional representation of $\hat{\mathbf{X}}$ can be obtained by discarding columns of $\mathbf{U}$ and $\mathbf{V}$,

$$\hat{\mathbf{X}} \approx \mathbf{U}_{0:j}\boldsymbol{\Sigma}_{0:j,0:j}\mathbf{V}_{0:j}^{\mathrm{T}}$$

Here, j denotes the number of columns retained. As shown above, PCA can project data into a low dimensional coordinate system, with axes provided by the columns of $\mathbf{U}_{0:j}$, and data coordinates given by $\mathbf{V}_{0:j}$.

In this work, we retain only two columns ($j = 2$), and project images into a two dimensional coordinate system. Figure 1 shows the 2D projection (coordinates obtained from $\mathbf{V}_{0:2}$) of facial images in our test dataset. We use this projection to construct a grid of images, grouped by similarity. Algorithm 1 describes this process. We create a uniform coordinate grid and search for the closest image (in the reduced dimensional coordinate system) to each point in the grid. We assign each image a corresponding point, and ensure that no image is duplicated, by removing it from the list of available images once allocated a grid coordinate in order to produce a grid of images that groups individuals by facial similarity, as shown in Figure 2. It is clear that this process successfully groups faces of similar skin tone together, with darker skinned individuals located towards the top of the image, and lighter skinned individuals towards the bottom. For each image selected, we calculate the error in prediction, to produce a saliency map indicating model quality for the constructed grid of images.

IV. EXPERIMENTAL METHOD

We examine the potential of CNN-based drowsiness detection algorithms to exhibit algorithmic bias by training a variety of popular classification models on a number of publically available drowsiness detection datasets. A number of strategies were applied to prevent overfitting, so as to ensure a fair analysis.

A. *MODEL ARCHITECTURE*

The architectures of the networks used for testing were based on a variety of pre-trained network models (VGG-Face, VGG, and ResNet), with the final layers modified as shown in Figure 3, with Batch Normalisation (BN) applied and three fully connected layers added to the network.

Pre-trained models (trained for general image classification) were used as feature extractors to lower the number of parameters to be trained and reduce training time. In addition, lower level features are already learned for the pre-trained models, which can prevent overfitting to smaller datasets. We made use of the Adam optimizer and a binary cross entropy

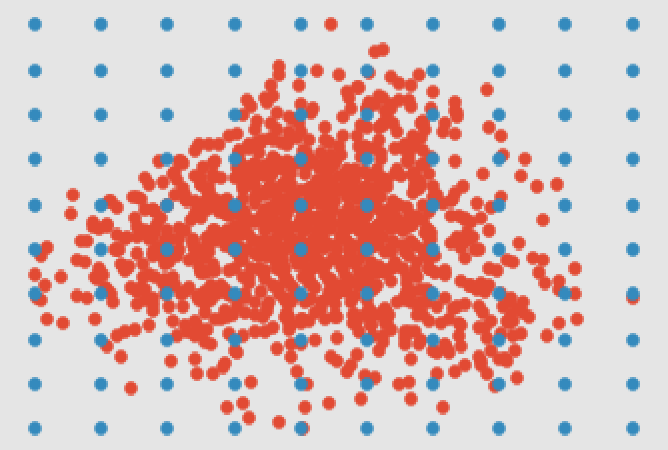


Figure 1. The figure shows 2-dimensional grid projection coordinates obtained after applying the linear PCA transformation into 2-dimensional subspace. A uniformly spaced grid is placed over the projected image coordinates, and images are assigned a grid position by finding the closest image coordinate to each grid position, ensuring each image can only be used once. The blue dots represent grid position and the red dots represent PCA projections.

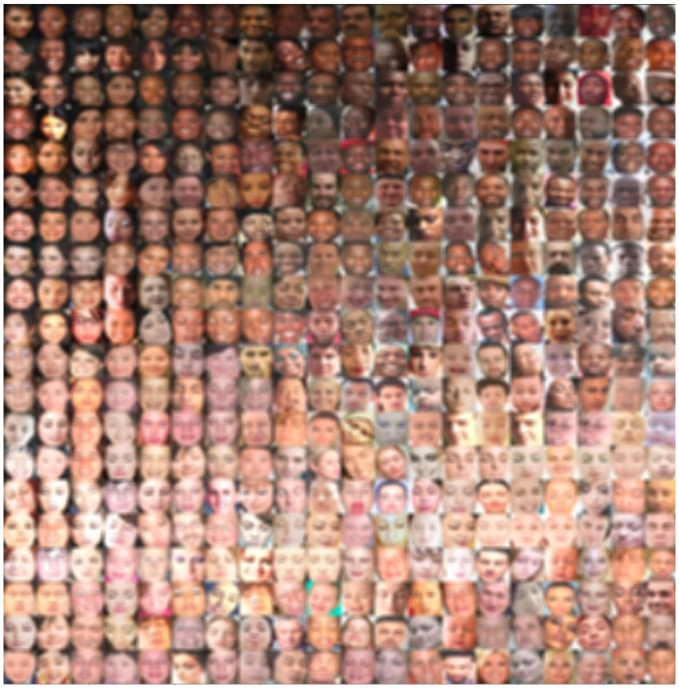


Figure 2. The proposed visualisation strategy uses PCA to sort faces by similarity without requiring meta-data.

loss in the training process. Data augmentation was also used in an attempt to prevent over-fitting. Here, re-scaling, shearing, zooming, and horizontal flipping was applied to extend the size of datasets used for training.

Furthermore, zooming was applied to images because the face is of greater interest in drowsiness detection. Horizontal flipping was also applied to generate different angles of the drivers' faces. Dropout ($\alpha = 0.5$) was applied between fully connected layers to reduce the chances of over-fitting even further.

Algorithm 1 Image overlay generation

```
Let $\mathbf{p}$ be the $N \times 2$ matrix $\mathbf{V}_{0:2}$
**Input:** $\mathbf{p}$, list of images $\mathbf{x}_i$, where $i = 1 \ldots N$, labels$_i$
x-min = min($\mathbf{p_0}$)
x-max = max($\mathbf{p_0}$)
y-min = min($\mathbf{p_1}$)
y-max = max($\mathbf{p_1}$)
image-grid = [ ][ ]
overlay-grid = [ ][ ]
j = 0
**for** pos-x in x-min : d1 : x-max **do**
    k = 0
    **for** pos-y in y-min : d2 : y-max **do**
        min-dist = 10000
        best-idx = 0
        **for** i in range(0,N) **do**
            dist = $\sqrt{(p[i,0]-j)^2 + (p[i,1]-k)^2}$
            **if** dist $<$ min-dist **then**
                min-dist = dist
                best-idx = i
            **end if**
            k = k +1
            image-grid[j,k] = $\mathbf{x}_i$
            overlay-grid$[i,k] = |$labels$_i -$ cnn-model$(\mathbf{x}_i)|$
            remove image $\mathbf{x}_i$ from image-list
        **end for**
        j = j + 1
    **end for**
    **Output:** Returns an overlayed saliency image
**end for**
```

B. DATASETS

This section describes the datasets that were used for training and testing. For this work, the NTHU-drowsy, DROZY, and CEW datasets are used.

NTHU-drowsy was introduced at the 13th Asian Conference on Computer Vision (ACCV2016) [44]. The dataset is split into test and training sets. For training, there are 18 participants (10 men and 8 women) pretending to drive, with 5 scene scenarios for each participant including no-glasses, glasses, glasses at night, no glasses at night, and sunglasses. For evaluation, there are images of 2 men and 2 women. Videos combining drowsy, normal and sleepy states are provided.

DROZY consists of 14 participants (3 males and 11 females) [45]. Each video is approximately 10 minutes long and is accompanied by the results of psychomotor vigilance tests (PVTs) regarding the drowsiness state. For each participant, the dataset contains a time-synchronized Karolinska Sleepiness Scale (KSS) score [45].

CEW is a collection of online images of different races (for example Asians and non-Asians with light-skinned faces) and contains about 2423 participants [46]. Among the

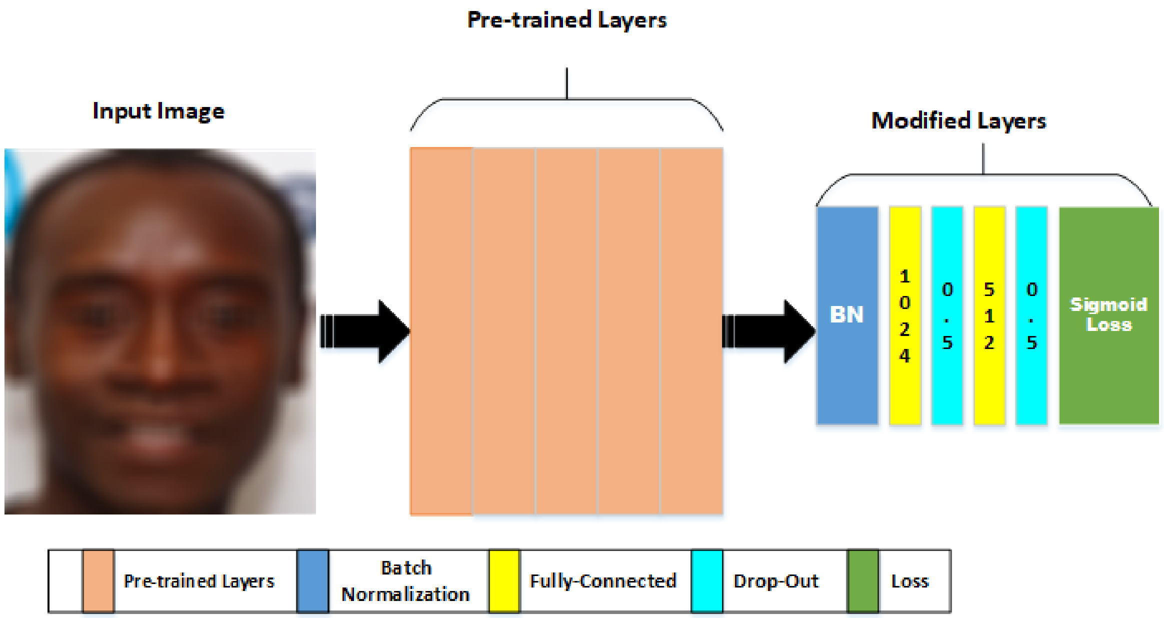


Figure 3. The figure shows the neural network architecture used for detection of drowsiness. Pre-trained layers follow the VGG, VGG-face and RestNet architectures, which are used as feature extraction layers. The modified layers are trained to perform drowsiness detection. Dropout layers are used to prevent over-fitting.

participants, 1192 have both eyes closed and 1231 have their eyes open. These images were selected from the labeled faces in the wild database.

Our Test Dataset was prepared from a collection of online videos of South African faces. There are 30348 images comprising a variety of different races and ethnicities represented in South Africa. Images range from dark to light-skinned faces of multiple genders to provide a diverse testing dataset.

The drowsiness detection models were trained and evaluated on these three datasets, which all consist of two classes (awake and drowsy). Three models were trained on each dataset (on over 300k images) and evaluated on 50k of images that were held out from each of the training datasets used. Finally, the South African test dataset was used to test the three trained models.

We prepared all the images from the three datasets in the same manner for training. All images were resized to 150x150 pixels, before applying augmentation and feeding the data in batches into the model. The Adam optimizer was used to train the model and training was performed for 30 epochs. The batch size was kept constant at 32 as it was observed that using a larger batch size degraded the model's quality.

V. Results

The accuracy obtained when testing on data held out during training for each dataset is shown in Figure 4. All the training datasets include images of light-skinned individuals, but only the CEW contains images of a small number of dark-skinned individuals. All facial images are blurred for confidentiality.

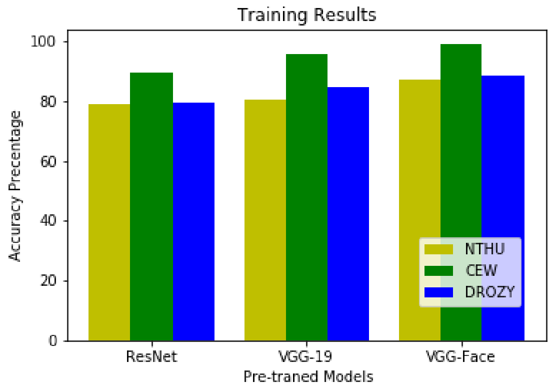


Figure 4. The figure shows validation accuracies obtained for each model when testing on data held out from the training datasets. These accuracies seem to indicate that all the models will generalize well.

The testing dataset was prepared in the same way as the training dataset and using the same parameters for data augmentation. The loss and accuracy were recorded for both the training and testing phases. Pre-trained models performed well when tested on data held out from training sets. However, all models showed decreased performance when tested on our representative dataset, as shown in Figure 5, although the decrease in performance was marginal for the CEW dataset. It is clear that the models trained using NHTU-drowsy and DROZY completely overfit to these datasets, and failed entirely when tested on our dataset. As a result, the NHTU-drowsy and DROZY are excluded from further analysis.

Figure 6 shows a selection of saliency maps obtained when

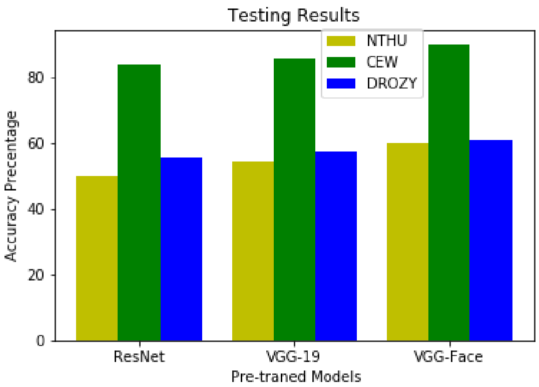
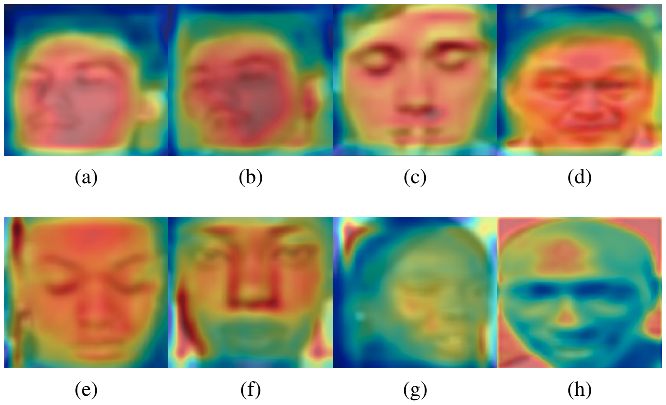


Figure 5. When testing on the representative South African dataset, both the NTHU-drowsy and DROZY models failed to generalise, but the CEW model seems to perform well.

the VGG-Face models trained using the CEW dataset were tested using both light-skinned individuals (dominant in the training sets) and dark-skinned individuals (dominant in our test set). Here, red areas denote image pixels that contributed significantly to the algorithm output. Interestingly, the model seems to focus on facial regions for the lighter skinned individuals shown here, but fails to do so for darker skinned individuals, indicating a potential failure case.


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

(e) (f) (g) (h)

Figure 6. The saliency map overlays highlight pixels in the input image that contribute to the network's final output. Areas marked in red contribute significantly, while blue regions contribute little to the final classification decision. Images (a) to (d) are from the validation set, and the saliency map highlight facial features, as would be expected for a drowsiness detector. Images (e) to (h) were sampled from our test dataset. The saliency visualisation failed to highlight facial features in the images (g) and (h).

We also applied the proposed PCA-saliency visualisation strategy (Figure. 7). Although the accuracy measures highlighted previously showed that the CEW models performed well, the proposed visualisation shows that the CEW models seem to struggle to predict drowsiness for darker-skinned individuals at the top of the image, potentially indicating a population group for which additional data is required to train a better model.

The key findings of these experiments are as follows:

- All training experiments performed well when tested on data held out from training datasets (85.3% - 98.7%).
- All three models showed a decrease in performance when tested on our more representative test dataset, indicating some overfitting.
- Models trained using NTHU and DROZY datasets completely failed to generalise.
- Models trained using the CEW dataset fail to perform well for certain dark-skinned individuals indicating a need for additional training data covering these population groups.

VI. Conclusion

Results presented in this paper have showed that models trained using publicly available datasets for drowsiness detection do not generalise well when tested on dark-skinned races. The 50% accuracy obtained when testing the NTHU and DROZY models on a more representative dataset shows that the network is simply guessing the drowsiness state of the driver, which could lead to system failure and endanger drivers if these models were deployed. In contrast, the CEW models appear to perform well, but further examination shows that they are failing systematically on certain population groups.

This paper has highlighted the potential for racial discrimination by machine learning models when the datasets used for training do not cover the demographics present where the system might be deployed. Going forward, it is crucial that balanced training datasets, covering all races and ethnicities, are used to train systems for driver aid. This work has introduced a visualisation strategy that can be used to identify population groups on which an algorithm is failing, without the need for meta-data regarding race or ethnicity. Furthermore, this work has shown that there is a strong need to evaluate vision-based driver drowsiness detection systems in the countries where they will be deployed, in order to prevent unintentional discrimination.

References

[1] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

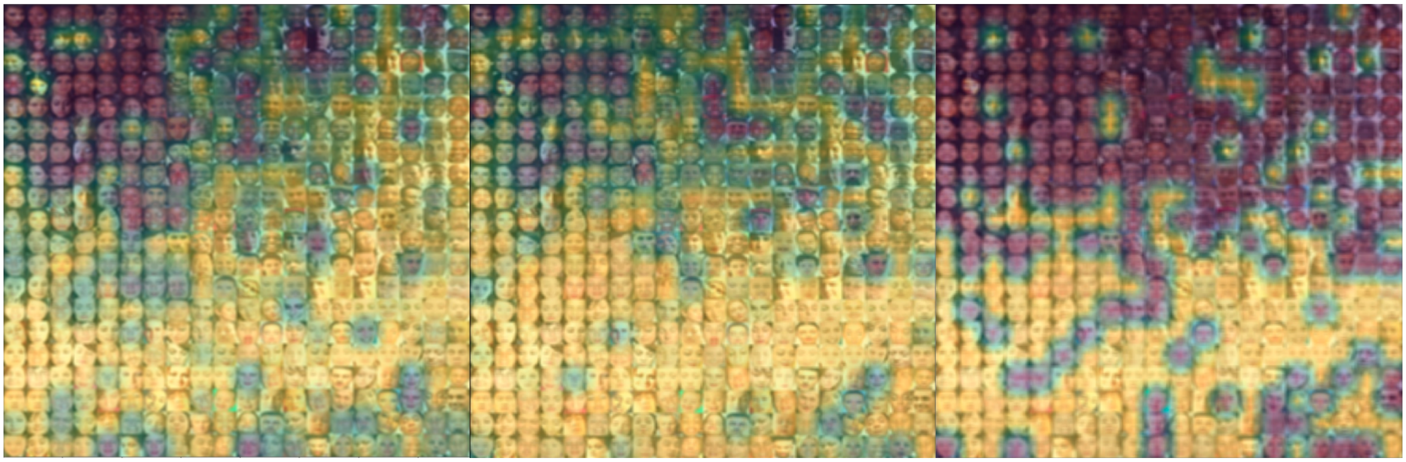
[2] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in Neural Information Processing Systems*, 2016, pp. 4349–4357.

[3] National Association of Automobile Manufacturers of South Africa, "Quarterly Review of Business Conditions," *Quarterly Review of Business Conditions: Motor Vehicle Manufacturing Industry*, no. 012, pp. 1–7, 2017.

[4] Statistics South Africa, "Mid-year population estimates," no. July, pp. 1–22, 2017. [Online]. Available: http://www.statssa.gov.za/publications/P0302/P03022017.pdf

[5] Stats South Africa, "Mbalo brief," no. 01, 2016.

[6] The Road Traffic Management Corporation (RTMC), "Road safety annual report 2017," Tech. Rep., 2017.



(a) ResNet-CEW (b) VGG-CEW (c) VGGFace-CEW

Figure 7. The figure shows images produced using the proposed visualisation strategy. All the trained models appear to be failing on the population groups on the upper part of the image. The yellow shaded parts indicate where the model performs well, while failures are indicated by the purple shaded parts, which appear mostly on the upper part of the images. The green shaded parts show that the model is also performing well, but with lower probability (0.65 to 0.80).

[7] H. J, S. Roberson, B. Fields, J. Peng, S. Cielocha, and J. Coltea, "Fatigue Detection using Smartphones," *Journal of Ergonomics*, vol. 03, no. 03, 2013. [Online]. Available: http://www.omicsgroup.org/journals/fatigue-detection-using-smartphones-2165-7556.1000120.php?aid=21520

[8] X. Li, W. Wang, and M. Roetting, "Estimating driver's lane-change intent considering driving style and contextual traffic," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2018.

[9] W. Wang, J. Xi, and D. Zhao, "Driving style analysis using primitive driving patterns with bayesian nonparametric approaches," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2018.

[10] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2018.

[11] M. Ngxande, J.-R. Tapamo, and M. Burke, "Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech), 2017*. IEEE, 2017, pp. 156–161.

[12] S. Park, F. Pan, S. Kang, and C. D. Yoo, "Driver drowsiness detection system based on feature representation learning using various deep networks," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 154–164.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[14] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[16] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang, "Real-time driver drowsiness detection for embedded system using model compression of deep neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 438–445.

[17] J. Lyu, Z. Yuan, and D. Chen, "Long-term multi-granularity deep framework for driver drowsiness detection," *arXiv preprint arXiv:1801.02325*, 2018.

[18] K. Dwivedi, K. Biswaranjan, and A. Sethi, "Drowsy driver detection using representation learning," in *Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014, pp. 995–999.

[19] K. W. Kim, H. G. Hong, G. P. Nam, and K. R. Park, "A study of deep cnn-based classification of open and closed eyes using a visible light camera sensor," *Sensors*, vol. 17, no. 7, p. 1534, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Multimedia Tools and Applications*, pp. 1–17, 2017.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[22] L. Pauly and D. Sankar, "Detection of drowsiness based on hog features and svm classifiers," in *Research in Computational Intelligence and Communication Networks*

(ICRCICN), 2015 IEEE International Conference on. IEEE, 2015, pp. 181–186.

[23] J. A. Buolamwini, "Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[24] A. Renda, "Ethics, algorithms and self-driving cars–a csi of the trolley problem. ceps policy insights no 2018/02, january 2018," 2018.

[25] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[26] R. Brauneis and E. P. Goodman, "Algorithmic Transparence for the Smart City," vol. 20, pp. 103–175, 2018.

[27] J. Zou and L. Schiebinger, "Design AI so that it's fair," *Nature*, vol. July 18, p. Comment, 2018. [Online]. Available: https://www.nature.com/articles/d41586-018-05707-8

[28] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," pp. 1–15, 2017.

[29] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

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

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization." in *ICCV*, 2017, pp. 618–626.

[38] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[39] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *arXiv preprint arXiv:1704.03296*, 2017.

[40] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," *arXiv preprint arXiv:1711.00867*, 2017.

[41] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau, "A cti v is: Visual exploration of industry-scale deep neural network models," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 88–97, 2018.

[42] L. I. Smith, "A tutorial on Principal Components Analysis," *Statistics*, vol. 51, p. 52, 2002.

[43] G. Stewart, "On the early history of the singular value decomposition," *SIAM Review*, vol. 35, no. 4, pp. 551–566, 1993. [Online]. Available: https://doi.org/10.1137/1035134

[44] "NTHU CVlab - Driver Drowsiness Detection Dataset," 2016. [Online]. Available: http://cv.cs.nthu.edu.tw/php/callforpaper/datasets/DDD/

[45] Q. Massoz, T. Langohr, C. François, and J. G. Verly, "The ulg multimodality drowsiness database (called drozy) and examples of use," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–7.

[46] "The Closed Eyes in the Wild (CEW) dataset." [Online]. Available: http://parnec.nuaa.edu.cn/xtan/data/ClosedEyeDatabases.html