Robust SleepNets

Yigit Alparslan
Department of Computer Science
Drexel University
Philadelphia, PA, US
ya332@drexel.edu

Edward Kim
Department of Computer Science
Drexel University
Philadelphia, PA, US
ek826@drexel.edu

***Abstract*—State-of-the-art convolutional neural networks excel in machine learning tasks such as face recognition, and object classification but suffer significantly when adversarial attacks are present. It is crucial that machine critical systems, where machine learning models are deployed, utilize robust models to handle a wide range of variability in the real world and malicious actors that may use adversarial attacks. In this study, we investigate eye closedness detection to prevent vehicle accidents related to driver disengagements and driver drowsiness. Specifically, we focus on adversarial attacks in this application domain, but emphasize that the methodology can be applied to many other domains. We develop two models to detect eye closedness: first model on eye images and a second model on face images. We adversarially attack the models with Projected Gradient Descent, Fast Gradient Sign and DeepFool methods and report adversarial success rate. We also study the effect of training data augmentation. Finally, we adversarially train the same models on perturbed images and report the success rate for the defense against these attacks. We hope our study sets up the work to prevent potential vehicle accidents by capturing drivers' face images and alerting them in case driver's eyes are closed due to drowsiness.**

***Index Terms*—adversarial attacks, drowsy sleeping, adversarial defense, adversarial training**

I. Introduction

Recent machine learning breakthroughs help solve many tasks including facial recognition [1], surveillance [2], natural language processing tasks [3], materials discovery [4] and bio-authentication systems [5]. One of these tasks where machine learning models can be utilized is to detect whether a driver is drowsy inside the car and help prevent accidents with an alert system. National Highway Traffic Safety Administration (NHTSA) data shows 37,461 people were killed in 34,436 motor vehicle crashes, an average of 102 per day in 2016. Of all these accidents, alcohol-impaired driving fatal crashes totaled 9477 in 2016 [6]. A potential alert system that detects driver disengagements while driving could prevent these accidents where the cause was alcohol or drowsiness. An estimated 1 in 25 adult drivers (aged 18 or older) report having fallen asleep while driving in the previous 30 days [7] [8]. Additionally, drowsy driving was responsible for 72,000 crashes, 44,000 injuries, and 800 deaths according to the National Highway Traffic Safety Administration (NHTSA) in 2013 [9]. NHTSA also reports that up to 6,000 fatal crashes each year may be caused by drowsy drivers [10] [11] [12]. Moreover, driving after going more than 20 hours without sleep is the equivalent of driving with a blood-alcohol concentration of 0.08%, which is the U.S. legal limit. A person is three times more likely to be in a car crash if fatigued [7]. Many car manufacturers incorporate assisted driving or drowsiness alert systems to their cars. Such cars may have an eye tracking camera built-in that records driver's face while driver is on the wheel. Other form of such alerting systems currently implemented to production level cars by car manufacturers may include pressure sensors on the steering wheel. Pressure sensors can be used to measure any long duration where the driver's hands are not on the steering wheel. In the event that this scenario happens, a car's alerting system would alert the driver via sound, or apply emergency brakes in case the driver is sleepy, not conscious or passed out due to a health incident. Currently such implementations exist for high-end luxury cars [13]. Even though such steering interventions exist for high-end cars, the average car doesn't have such sophisticated drowsiness detection systems or intervention mechanisms. A general solution scalable to not only high-end luxury cars but also to general public cars could be a mobile application which could detect when driver closes their eyes for a certain amount of time and alert the driver via sound.

A previous work [14] by the authors resulted in an initial prototype published as a mobile app for Android users. See here for a working prototype of the said alerting system deployed and published as a mobile application. Such mobile application would utilize a state-of-the-art neural network trained on images of people with closed and open eyes. Due to the nature of the task, it is crucial that such a neural network is agnostic to poor light conditions, out of zoom image captures, blurs, shadows, low resolution images, etc. Therefore, it is essential to have a robust model.

The title of this study, "Robust SleepNets", come from the idea of being able to detect fatigued or drowsy drivers (or any driver that might disengaged with driving for long duration) at the steering wheel. In this study, we assume eye closedness is a direct indicator of drowsiness so we focus on implementing neural network architectures that would detect eye closedness. We attack the models with adversarial attacks to study the robustness of the networks under adversarial conditions. The robustness acts as a proxy to real life conditions which might exhibit poor lightning, out-of-focus camera zoom, overexposed/underexposed shutter, height or width shift in the

All code is open-sourced on GitHub.

frame etc. We also investigate augmenting train data with the parameters described in Table I to study the impact of data augmentation on accuracy and adversarial defense.

In this study, we assume the following two problems are functionally equivalent.

1) Detecting drowsy driving
2) Detecting whether driver closes their eyes. (longer than some threshold that could be determined empirically. such threshold is not the focus of this study)

In other words, we focus on detecting eye closedness to detect drowsy driving. Then, we **treat the real-word driving conditions such as poor light, out-of-focus camera etc as black-box adversarial attacks**. With these assumptions and motivations, we investigate the usage of adversarial training as a means to creating robust models that could detect eye closedness against said adversarial attacks. Additionally, we study the effect of training data augmentation.

This research is organized so that section I introduces the concept of drowsy driving and section II explores what has been done in this field. section III explains the dataset, the models, adversarial attacks, adversarial training and limitations that we use/have in this study. We report the results in section IV and conclude the study in section V and section VI with summarizing what we have done in this study and discussing where the research might go in the future.

II. Related Work

There has been previous results where eye detection was used as a gateway to detect drowsiness [15]. There has been little study to investigate the robustness of these models with adversarial attacks.

Adversarial attacks are inputs that look like the original images but with perturbations added to result in misclassifications in the classifier [16] [17] [18] [19]. Adversarial attacks can be created in the image domain [1] as well as audio domain [20].

Adversarial training is one way of defending against these attacks since using adversarial attacks [21]. we can generate adversarial samples and then use these samples in our training to develop high accuracy models. Adversarial training as a defense depends on model and task at hand significantly.

Alparslan et al. [14] investigated the eye and face models on driver detection and explored data augmentation to simulate real-world black-box adversarial image settings. Even though they did not apply adversarial attacks, or adversarial training, they claimed that adding a robust and systematic data augmentation to the training datasets would represent black-box attacks in a real-world scenario where driver face might be in too much light, or shadow, or it might be blurred if the camera angle is not adjusted.

In this study, we include data augmentation that Alparslan et al. included in their study. In addition to the data augmentation, we also investigate adversarial robustness by attacking the models with Projected Gradient Descent, Fast Gradient Sign and DeepFool attacks. Once we report the accuracy on the adversarially generated dataset, we feed the adversarial inputs back into the training dataset to defend against the attacks to study the possibility of creating robust models that would detect eye closedness and drowsiness in the presence of attacks.

III. Methodology

We develop two models: first on eye images and second on face images. We use Eye-blink dataset [22] for the eye model and we use Closed-Eye in the Wild dataset [23] for the face model.

For the eye model, we use 3,108 images belonging to 2 classes (open and closed) to train and we use 776 images to cross-validate during training. We use 962 images that the model has never seen before to just test the data. For the eye model, all classification results in Table III are reported from these 962 images of test dataset.

For face model, we use 1,559 images belonging to 2 classes (open and closed) to train and we use 389 images to cross-validate during training. We use 485 images that the model has never seen before to just test the data. For the face model, all classification results in Table III are reported from these 485 images of test dataset.

Additionally, in this study, we investigate the possible impact of data augmentation against adversarial attacks. For both models, we report results for the model trained on augmented data as well as non-augmented data. The data augmentations that we apply are summarized below:

1) **Rotation:** Image can be rotated randomly depending on the driver's position and the camera angle.
2) **Width Shift:** Image width might depend on the camera angle. The model needs to mitigate this randomness .
3) **Height Shift:** Image height might depend on the camera angle. The model needs to mitigate this randomness.
4) **Shear angle Shift:** Drivers plane intersects with the plane in which camera is mounted on a car. This creates additional randomness and the model needs to mitigate this randomness.
5) **Zoom:** The camera can be close or far to the driver and the model needs to mitigate this randomness.
6) **Horizontal Flip:** This doesn't correlate to real life setting, but the idea is that the driver's window will be always at its left side, which means the lightning conditions from the left side of the camera will be always poor compared to the right side. The model needs to detect fatigue regardless of the lightning hence the flip.
7) **Image Fill:** Image can be scaled down or up. The model needs to detect fatigue regardless.
8) **Scaling:** Image can be scaled down or up. The model needs to detect fatigue regardless.

Additionally, we apply Projected Gradient Descent [21], Fast gradient Sign Method [24] and DeepFool [25] to attack the models and report their accuracies.

A. Adversarial Attacks

1) PGD: Projected Gradient Descent [21] is a strategy for finding an adversarial example x' for an input x that satisfies

a given norm-bound $\|x' - x\|_p \leq \epsilon$.

Let B denote the ℓ_p-ball of radius ϵ centered at x. The attack starts at a random point $x_0 \in B$, and repeatedly sets

$$x_{i+1} = \text{Proj}_B(x_i + \alpha \cdot g)$$
$$\text{for } g = \arg\max_{\|v\|_p \leq 1} v^\top \nabla_{x_i} L(x_i, y) \ .$$

Here, $L(x, y)$ is a suitable loss-function (e.g., cross-entropy), α is a step-size, $\texttt{Proj}_B$ projects an input onto the norm-ball B, and g is the *steepest ascent* direction for a given ℓ_p-norm. E.g., for the ℓ_∞-norm, $\texttt{Proj}(z)$ is a clipping operator and $g = \texttt{sign}(\nabla_{x_i} L(x_i, y))$.

2) Fast Gradient Sign: The fast gradient sign [24] method optimizes for the L_∞ distance metric and its advantage is fast running tim, which comes at the expense of generating images that are very similar to the original image.

Given an image x the fast gradient sign method sets

$$x' = x - \epsilon \cdot \text{sign}(\nabla \text{loss}_{F,t}(x)),$$

where ϵ is chosen to be sufficiently small so as to be undetectable, and t is the target label. Intuitively, for each pixel, the fast gradient sign method uses the gradient of the loss function to determine in which direction the pixel's intensity should be changed (whether it should be increased or decreased) to minimize the loss function; then, it shifts all pixels simultaneously.

3) DeepFool: Deepfool [25] optimizes over L_2 distance metric with the assumption that neural networks are linear. Since neural networks are not linear, once a hyperplane is found (if found any) that seperates two classes, search terminates. DeepFool takes about 10x more than PGD and FGSM takes on average when we apply for our eye and face models. We refer the authors to the work of Moosavi et al. [25] for more in-depth explanation.

B. Adversarial Training

Adversarial training [26] [27] [21] is the process where the adversarially generated samples are included in the training data in the hopes that the model will recognize the attacks next time sees it. In the current literature, adversarial training is one of the stronger defenses against adversarial attacks especially if it is combined with other defenses [21] [28]. We adversarially train our augmented eye model, non-augmented eye model, augmented face model and finally non-augmented face model. During our adversarial training, we attack entirety of the training dataset via Projected Gradient Descent, Fast Gradient Sign Method and DeepFool attacks. Since we use one attack to adversarially train one model and there are three attacks, we report three adversarially trained models for the Eye-blink dataset and three adversarially trained models for the Closed Eyes in the Wild dataset. We repeat the entire process once for the augmented data case and once for non-augmented data case, which doubles our combinations.

C. Data Augmentation

Data augmentation is the process of generating new training data from the existing samples. Some data augmentation methods include adding rotation, adding text, adding zoom, height shift, width shift. In this study, we also investigate the effect of data augmentation in the presence of adversarial attacks. The data augmentation parameters are the same for both eye and face model and can be seen in Table I. These parameters are the same as the parameters described in [14]

TABLE I: Noisified training data parameters are shown below. Adding noise to the training data to augment it simulates black-box settings in real-world scenarios where a driver face might blurred, occluded, underexposed/overexposed etc. In these scenarios, the data augmentation helps simulate an adversary under black box settings.

Change Type	Change Amount #
rotation range	40^o
width shift range	0.2
height shift range	0.2
shear range	0.2
zoom range	0.2
horizontal flip	True
fill mode	'nearest'
rescale	1./255

D. Models

We use the same model architecture for the Eye-blink dataset and the Closed Eyes in the Wild dataset in order to eliminate the differences in accuracy that might arise due to architecture configurations. Since both dataset represents a binary classification problem, we argue that using the same architecture for both datasets doesn't cause problem to detect eye closedness. For eye model, we use 3,108 images belonging to 2 classes as training data and 776 images belonging to 2 classes as validation data to train and cross validate our models (See Figure 4, Figure 3 to see the training accuracy and loss plots). Then, we report the testing accuracy on the testing dataset 962 images belonging to 2 classes (All numbers are from testing accuracy in Table III). The test dataset consist of images that are never seen before by the model.

For face model, we use 1,559 images belonging to 2 classes as training data and 389 images belonging to 2 classes as validation data to train and cross validate our models (See Figure 6, Figure 5 to see the training accuracy and loss plots). Then, we report the testing accuracy on the testing dataset 485 images belonging to 2 classes (All numbers are from testing accuracy in Table III). The test dataset consist of images that are never seen before by the model.

Model architectures can be seen in Table II.

E. Dataset

In this study, we use Eye-blink Dataset [29] data set to train our eye model. Eye Blink dataset has 2,100 closed and open eye images that are black and white. They are 24x24 pixels and only show eye patches. We use Closed Eyes In

TABLE II: Model Architecture for both models. We used the same architecture on the Eye-Blink and CWE dataset. The eye model and the face model both use binary cross entropy with Adam's optimizer on iterative gradient descent.

Layer type	Output Shape	Param #
Conv2D	(98, 98,6)	60
Average Pooling	(49,49,6)	0
Conv2D	(47,47,16)	880
Average Pooling	(23,23,16)	0
Flatten	8464	0
Dense	120	1015800
Dense	84	10164
Dense	1	85

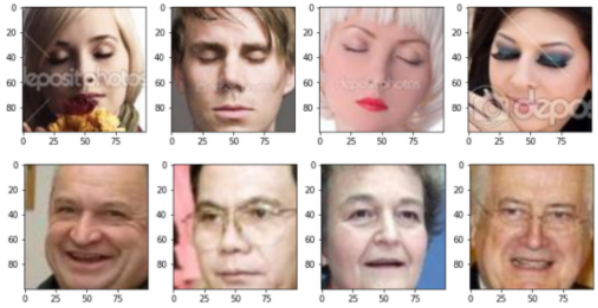


Fig. 1: Samples from Closed Eyes in the Wild dataset. This dataset includes 1,231 opened-eye images of people and 1192 closed-eye images of people. Some difficulties with this dataset include blur, fade, shade and over/underexposing. Due to these innate difficulties, this dataset represents well the environment an actual driver might be in while driving in a real-world scenario.

The Wild (CEW) [29] dataset to train our face model. CEW dataset has 1,192 subjects with both eyes closed and 1,231 subjects with eyes open. Some challenges of this set include amateur photography, occlusions, problematic lighting, pose, and motion blur.

F. Limitations

In this study, we examine many different combinations to see the full effect of attacks and defenses. We have two models and for each model, we repeat the experiments for the case when the training data is augmented and for the case where it is not augmented. We use three attacks on each model and also apply adversarial training for each of the attacks. Because we have about 16 different configurations when all combined in this study, we are generating grayscale adversarial images even though the input can be colored (applying the attack for each channel). This helps us speed up the training, attacking and defending duration. The readers for example will see from the Table III that PGD doesn't defend successfully when attacked. Researchers who might reproduce our combinations in the future might get different results if this nuance is not taken into consideration.

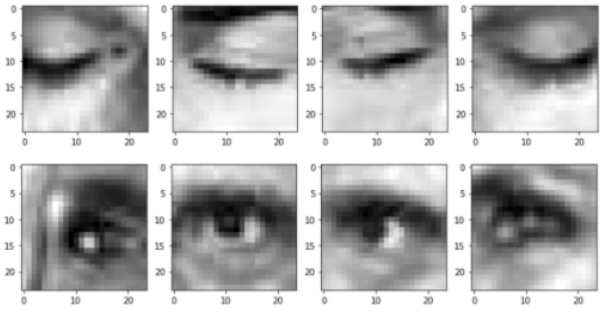


Fig. 2: Samples from Eye-blink dataset. This dataset consists 2100 closed and open 24x24 pixel grayscale eye patch images. Some difficulties with this dataset include very low resolution and shade.

IV. Experiment Results and Evaluation

We report all the results in Table III. In our analysis, we need to consider two aspects of the attacks.

- How successful is the adversarial attack?
- How successful is the defense?

In order to answer the first part, (i.e how successful the attacks are), we attack the entire test data and report how accurate the model is against them. The accuracies for this case are represented in parenthesis in the Table III. For example, for the case of Eye model **without** the data augmentation, PGD attack reports 81.70% accuracy. This means the model is able to accurately assign a class label in the images adversarially altered via PGD attack, 81.70% of the time. Since this accuracy value is fairly high, we conclude that PGD attack was not able to fool the classifier and failed as an *attack*. The same model is then adversarially trained with the images that were altered via PGD attack, and reports 81.70% accuracy, which suggests that the adversarial training doesn't have any success as a defense at all. Readers might be curious why the same accuracy is reported (81.70% vs 81.70%). This is quite possible if adversarial training is an identity function. (i.e adversarial training is not a successful defense since model reports the same accuracy for the attacks before the adversarial training and after the adversarial training).

A high accuracy value in the accuracy column of the Table III does not signify any conclusion regarding the success of the adversarial training. Mainly, the increase from the accuracy value inside the parenthesis tells how successful the adversarial training was in Table III. To answer the second part (i,e how successful the defenses), we calculate and compare that difference.

1) PGD: All PGD rows reports very high accuracy values ($> 81\%$) for all models after the adversarial training, but this does *not* mean the adversarial training was successful since this can be attributed to the fact that PGD attack cannot lower the accuracy of the models to begin with. (All accuracies inside the parenthesis for the PGD rows are ($> 79\%$). We conclude PGD cannot fool the classifier successfully. So,

adversarial training via PGD attack doesn't offer significant improvements. On average PGD attack decreases the classifier accuracy 9.44% for the eye model and 0.1% for the face model.

On average PGD defense increases the classifier accuracy -1.2% for the eye model and 0.2% for the face model (Please note -1.2% increase means 1.2% decrease). This can be empirically checked via the samples generated by PGD in section VII since adversarially generated samples that used PGD method look very similar to the original images. PGD attack/defense failure can be attributed to the fact that eye closedness detection is a very localized task on an image and PGD can't successfully alter such information.

2) FGSM and DeepFool: In the case of FGSM and DeepFool, as it can be seen in Table III, these two attacks can lower the accuracy of the models significantly compared to the baseline. (around 43%)

On average FGSM attack decreases the classifier accuracy 42.24% for the eye model and 32.22% for the face model. On average FGSM defense increases the classifier accuracy 1.1% for the eye model and 0% for the face model.

On average DeepFool attack decreases the classifier accuracy 31.43% for the eye model and 28.77% for the face model. On average DeepFool defense increases the classifier accuracy -9.81% for the eye model and -23.5% for the face model. (Please note a -9.81% increase means 9.81% decrease so adversarially training a model via DeepFool attack actually worked the opposite way of its intent.)

For these attacks, however, we conclude the defense is not successful since adversarially trained models cannot produce accuracy values greater than the accuracy values on the adversarial attacks before the training. We conclude for FGSM and DeepFool that the attacks are extremely strong and defenses are not. We invite the reader to examine the adversarially generated samples in section VII.

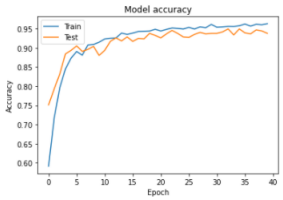
3) Data augmentation: For the face model, accuracy drops 17% when training data is augmented compared to the case when training data is *not* augmented.

However, we also see that data augmentation helps the classifier become more robust. On average for all the attacks, when training data is augmented, classifier reports 1.3% higher accuracy when attacked.

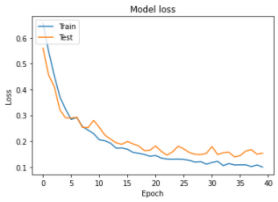
For the eye model, accuracy drops 2% when training data is augmented compared to the case when training data is *not* augmented.

However, we also see that data augmentation helps the classifier become more robust. On average for all the attacks, when training data is augmented, classifier reports 6.14% higher accuracy when attacked.

4) Eye vs Face Model: Accuracy values indicate that eye closedness detection is easier to achieve when done on only eye images. On average, eye models classify 12% more accurately than face model for the base cases. This can be attributed to the fact that face image has redundant information such as hair, mouth, ears etc (i.e all part of the face other than eyes) that the classifier needs to learn whereas eye model does not have such overhead. Because eye detection is such localized task, this could also explain why data augmentation helps the eye model 5x more than the face model when attacked. Defending the area of eyes via augmentation is easier on the eye images than face images.

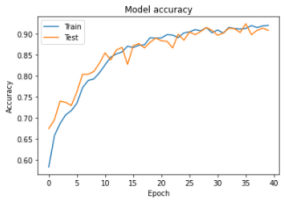


(a) Eye model accuracy for training and testing dataset versus epoch count

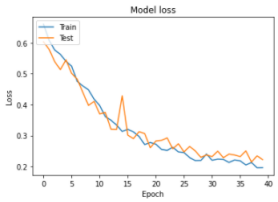


(b) Eye model binary cross entropy loss for training and testing datasets versus epoch count

Fig. 3: Training and testing of eye model with Eye-blink dataset when training data is **non-augmented**.

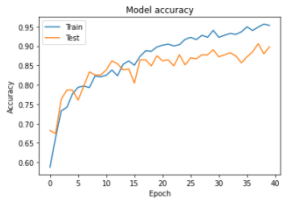


(a) Eye model accuracy for training and testing dataset versus epoch count

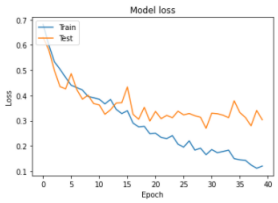


(b) Eye model binary cross entropy loss for training and testing datasets versus epoch count when training data is **augmented**

Fig. 4: Training and testing of eye model with Eye-blink dataset when training data is **augmented**.



(a) Face model accuracy for training and testing dataset versus epoch count

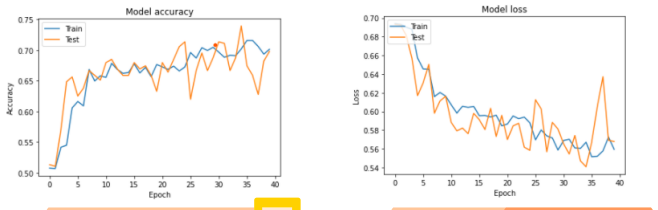


(b) Face model binary cross entropy loss for training and testing datasets versus epoch count

Fig. 5: Training and testing of face model with CEW dataset when training data is **non-augmented**.

V. Conclusion

In this paper, we assumed eye closedness was a gateway to detecting driver fatigue. We trained two deep convolutional neural network models to detect eye closedness: one based on Eye-blink dataset and other based on Closed Eyes in the Wild dataset (CEW). Later, we crafted adversarial attacks via Fast



(a) Face model accuracy for training and testing dataset versus epoch count

(b) Face model binary cross entropy loss for training and testing datasets versus epoch count

Fig. 6: Training and testing of face model with CEW dataset when training data is **augmented**.

TABLE III: Accuracy results for the adversarially trained eye and face model. The third column (Config) reports the configuration of the attack type used during adversarial training. The accuracy column represents accuracy on newly generated adversarial samples after adversarial training. The increase compared to the accuracy value inside the parenthesis tells how successful the adversarial training was. Inside the parenthesis, it reports accuracy on original adversarial samples, which are generated on the model before its adversarial training. A high accuracy value inside the parenthesis means the attack couldn't lower the accuracy successfully. Overall, the eye model has performed better than the face model and data augmentation reduces accuracies more for the face model than eye model.

Model	D.A	Config	Accuracy	Precision	Recall	F-1 Score
Eye	W/O	Base	95	96	96	95.5
		PGD	81.70 (81.70)	84	82	81
		FGSM	54.47 (52.18)	73	54	42
		DeepFool	55.09 (55.69)	64	55	45
	W/	Base	93	92	94	93
		PGD	85 (87.42)	85	85	85
		FGSM	51.14 (51.14)	26	51	35
		DeepFool	50.42 (69.44)	45	50	36
Face	W/O	Base	90	90	90	89
		PGD	81.03 (79.18)	81	81	81
		FGSM	49.28 (49.28)	24	49	33
		DeepFool	52.58(62.47)	53	53	48
	W/	Base	73	74	73	73
		PGD	82.68 (84.12)	83	83	83
		FGSM	50.72 (50.72)	26	51	34
		DeepFool	49.90 (60.0)	54	50	37

Gradient Sign Method, Projected Gradient Descent Method and DeepFool attacks. Highest models for eye detection were baseline models without data augmentation with 95% accuracy for the eye model and 90% for face model. We conclude that PGD attack is not able to decrease classifier accuracy as much as FGSM and DeepFool attacks decrease (9.44%, 42.24%, 31.44% decrease for eye model and 0.1%, 32.22%, -28.77% decrease for the face model, respectively). Additionally, adversarially training the model with PGD attack or FGSM attack donot increase classifier accuracy. PGD reports 1.2% (eye) and 0.2% (face) accuracy decrease, FGSM reports 1.1% (eye) and 0% (face) accuracy increase and DeepFool reports 9.81% (eye) and 23.5% (face) accuracy *decrease* which prove that DeepFool does not succeed at defending when used as adversarial training. Finally, eye model reports 6.14% higher accuracy when it is trained on augmented data when attacked compared to the case where it is trained on non-augmented data. Face model reports 1.3% higher accuracy when it is trained on augmented data when attacked compared to the case where it is trained on non-augmented data. We hope that our robustness study help emphasize the need for robust machine learning models in mission-critical systems in the presence of adversarial attacks. We also hope that this study gives more insight on robust eye closedness detection methods as well as the effect of data augmentation and adversarial attacks as defense tools.

VI. Future Work

In the future, combining the two models in an ensemble learning setting would yield different results for a given image and might be worthwhile to examine. Additionally, another future work might include improving the inference duration to enable eye closedness detection in real-time or on a video.

References

[1] Y. Alparslan, K. Alparslan, J. Keim-Shenk, S. Khade, and R. Greenstadt, "Adversarial Attacks on Convolutional Neural Networks in Facial Recognition Domain," *arXiv e-prints*, arXiv:2001.11137, arXiv:2001.11137, Jan. 2020. arXiv: 2001.11137 [cs.LG].

[2] Y. Alparslan, I. Panagiotou, W. Livengood, R. Kane, and A. Cohen, "Perfecting the Crime Machine," *arXiv e-prints*, arXiv:2001.09764, arXiv:2001.09764, Jan. 2020. arXiv: 2001.09764 [cs.CY].

[3] W. Daelemans and V. Hoste, "Evaluation of machine learning methods for natural language processing tasks," in *3rd International conference on Language Resources and Evaluation (LREC 2002)*, European Language Resources Association (ELRA), 2002.

[4] Y. Kim, E. Kim, E. Antono, B. Meredig, and J. Ling, "Machine-learned metrics for predicting the likelihood of success in materials discovery," *npj, Computational Materials*, 2020. DOI: https://doi.org/10.1038/s41524-020-00401-8.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[6] N. C. for Statistics and A. (M. V. T. C. D. R. Page, *Fatality analysis reporting system*, 2016. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812451.

[7] Wheaton AG, Chapman DP, Presley-Cantrell LR, Croft JB, Roehler DR, *Drowsy driving – 19 states and the district of columbia*, https://www.cdc.gov/mmwr/pdf/wk/mm6151.pdf, Accessed: 2021-02-08, 2013.

[8] Wheaton AG, Shults RA, Chapman DP, Ford ES, Croft JB, *Drowsy driving and risk behaviors—10 states and puerto rico, 2011-2012*, https://www.cdc.gov/mmwr/pdf/wk/mm6326.pdf, Accessed: 2021-02-08, 2014.

[9] National Highway Traffic Safety Administration, *Research on drowsy driving*, https://one.nhtsa.gov/Driving-Safety/Drowsy-Driving/Research-on-Drowsy-Driving, Accessed: 2021-02-08, 2015.

[10] Klauer SG, Dingus TA, Neale VL, Sudweeks JD, Ramsey DJ, *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic study data,* https://www.nhtsa.gov/DOT/NHTSA/NRD/Articles/HF/Reducing%20Unsafe%20behaviors/810594/810594.htm, Accessed: 2021-02-08, 2006.

[11] Tefft BC, AAA Foundation for Traffic Safety, *Prevalence of motor vehicle crashes involving drowsy drivers, united states*, https://aaafoundation.org/prevalence-motor-vehicle-crashes-involving-drowsy-drivers-united-states-2009-2013/, Accessed: 2021-02-08, 2013.

[12] Institute of Medicine., *Sleep disorders and sleep deprivation: An unmet public health problem*, Washington, DC: The National Academies Press, http://www.nationalacademies.org/hmd/Reports/2006/Sleep-Disorders-and-Sleep-Deprivation-An-Unmet-Public-Health-Problem.aspx, Accessed: 2021-02-08, 2006.

[13] Mercedes-Benz, *Assisted driving and collision prevention & detection system*, 2021. [Online]. Available: https://www.mercedes-benz.com/en/innovation/autonomous/by-far-the-best-mercedes-benz-assistance-systems/.

[14] K. Alparslan, Y. Alparslan, and M. Burlick, "Towards Evaluating Driver Fatigue with Robust Deep Learning Models," *arXiv e-prints*, arXiv:2007.08453, arXiv:2007.08453, Jul. 2020. arXiv: 2007.08453 [cs.CV].

[15] P. Chen, "Research on driver fatigue detection strategy based on human eye state," in *2017 Chinese Automation Congress (CAC)*, 2017, pp. 619–623. DOI: 10.1109/CAC.2017.8242842.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199.

[17] E. Kim, J. Rego, Y. Watkins, and G. T. Kenyon, "Modeling biological immunity to adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4666–4675.

[18] D. Schwartz, Y. Alparslan, and E. Kim, "Regularization and sparsity for adversarial robustness and stable attribution," in *Advances in Visual Computing*, G. Bebis, Z. Yin, E. Kim, J. Bender, K. Subr, B. C. Kwon, J. Zhao, D. Kalkofen, and G. Baciu, Eds., Cham: Springer International Publishing, 2020, pp. 3–14, ISBN: 978-3-030-64556-4.

[19] E. Kim, D. Hannan, and G. Kenyon, "Deep sparse coding for invariant multimodal halle berry neurons," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1111–1120.

[20] K. Alparslan, Y. Alparslan, and M. Burlick, "Adversarial Attacks against Neural Networks in Audio Domain: Exploiting Principal Components," *arXiv e-prints*, arXiv:2007.07001, arXiv:2007.07001, Jul. 2020. arXiv: 2007.07001 [cs.LG].

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv e-prints*, arXiv:1706.06083, arXiv:1706.06083, Jun. 2017. arXiv: 1706.06083 [stat.ML].

[22] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, 2007.

[23] F. Song, X. Tan, X. Liu, and S. Chen, "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," *Pattern Recognition*, vol. 47, no. 9, pp. 2825–2838, 2014, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2014.03.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320314001228.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv e-prints*, arXiv:1412.6572, arXiv:1412.6572, Dec. 2014. arXiv: 1412.6572 [stat.ML].

[25] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," *arXiv e-prints*, arXiv:1511.04599, arXiv:1511.04599, Nov. 2015. arXiv: 1511.04599 [cs.LG].

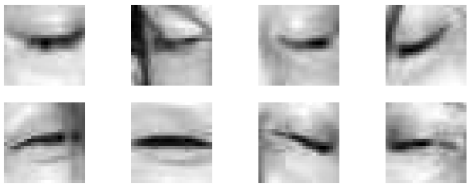
[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, 2014. arXiv: 1412.6572 [stat.ML].

[27] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," *arXiv preprint arXiv:1803.06373*, 2018.

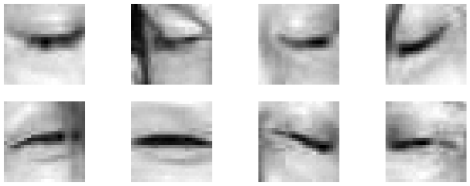
[28] F. Tramer, N. Carlini, W. Brendel, and A. Madry, *On adaptive attacks to adversarial example defenses*, 2020. arXiv: 2002.08347 [cs.LG].

[29] X. F.Song X.Tan and S.Chen, *Eyes Closeness Detection from Still Images with Multi-scale Histograms of Principal Oriented Gradients, Pattern Recognition*, ser. 9. Pattern Recognition, 2014, vol. 47. DOI: https://doi.org/10.1016/j.patcog.2014.03.024..

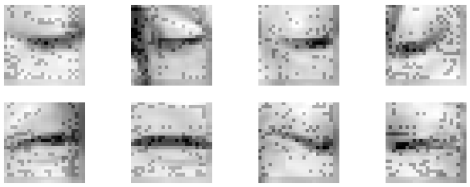
VII. APPENDIX



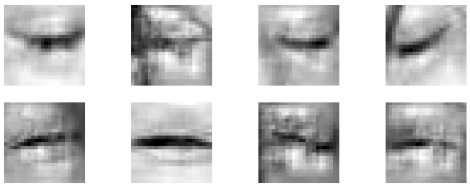
(a) Eye-blink dataset non-augmented samples



(b) Eye-blink dataset non-augmented samples after applying PGD



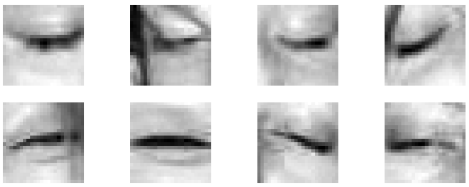
(c) Eye-blink dataset non-augmented samples after applying FGSM



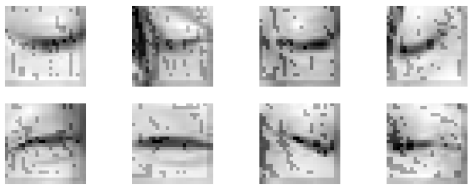
(d) Eye-blink dataset non-augmented samples after applying DeepFool

Fig. 7: Baseline images, FGSM, PGD and DeepFool adversarial attacks on non-augmented Eye-blink dataset.

(a) Eye-blink dataset augmented samples



(b) Eye-blink dataset augmented samples after applying PGD



(c) Eye-blink dataset augmented samples after applying FGSM

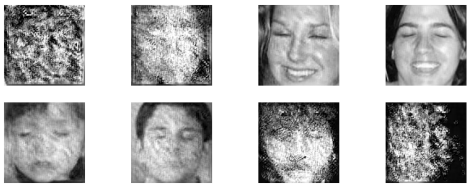
(d) Eye-blink dataset augmented samples after applying DeepFool

Fig. 8: Baseline images, FGSM, PGD and DeepFool adversarial attacks on augmented Eye-blink dataset.

(a) CEW dataset non-augmented samples

(b) CEW dataset non-augmented samples after applying PGD

(c) CEW dataset non-augmented samples after applying FGSM

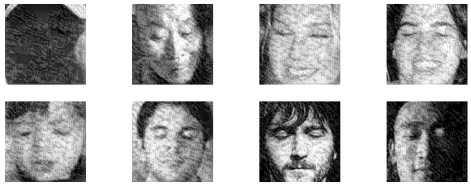


(d) CEW dataset non-augmented samples after applying DeepFool

Fig. 9: Baseline images, FGSM, PGD and DeepFool adversarial attacks on non-augmented Closed Eyes in the Wild (CEW) dataset.

(a) CEW dataset augmented samples

(b) CEW dataset augmented samples after applying PGD



(c) CEW dataset augmented samples after applying FGSM

(d) CEW dataset augmented samples after applying DeepFool

Fig. 10: Baseline images, FGSM, PGD and DeepFool adversarial attacks on augmented Closed Eyes in the Wild dataset.