

Distracted Driver Detection on Re-trained ResNet Architecture

VinayThandra
Student - AIML
Great Learning
Mumbai, India
vthandra12@gmail.com

ArunSoppimath
Student - AIML
Great Learning
Mumbai, India
asavsmail@gmail.com

Ravi Kumar
Student - AIML
Great Learning
Mumbai, India
ravi.kr.stats@gmail.com

WritankarKundu
Student - AIML
Great Learning
Mumbai, India
writankar.kundu@gmail.com

PiyushTelele
Student - AIML
Great Learning
Mumbai, India
piyushtalele.6491@gmail.com

VaishaliBalaji
Mentor - AIML
Great Learning
Mumbai, India
write2vaishalibalaji@gmail.com

NarayanaDarapaneni
Director - AIML
Great Learning/Northwestern University
Illinois, USA
darapaneni@gmail.com

Anwesh Reddy Paduri
Senior Data Scientist
Great Learning
Hyderabad, India
anwesh@greatlearning.in

Abstract— Road accidents are predominantly caused by the distracted drivers and nearly 1.3 million deaths are automobile accidents, of which, drivers are held responsible for 78% of accidents. There are various reasons for driver distractions which are drinking, operating instruments, mobile usage, interacting with fellow passengers etc. For the scope of this project, we intend to develop a model to successfully identify whether the driver is driving safely or is distracted using a combined dataset from the State Farm Distracted Driver Detection challenge on Kaggle & the AUC (American University in Cairo) Project. Convolutional Neural Network with ResNet architecture was used in developing the model. Grad-CAM technique was used to identify gradients in parts of images which impacted classification of images. Explainable AI can help build better models. This approach was able to provide us with promising accuracy and definite results. Experimental results show that our system achieves an accuracy of 99% on the Kaggle dataset and 82% on the AUC data set.

Keywords—Driver distraction, Deep learning, Convolutional Neural Network, Transfer Learning, Resnet101, GradCAM, Explainable AI.

I. INTRODUCTION

The problem at hand is related to the automotive domain. With approximately 1.3 million deaths every year attributed to motor accidents, India accounts for 11% of global death in road accidents [22]. The motor insurance claims in India amounted to Rs. 58456.9 crores in FY 18-19 [20]. Every year, a significant portion of the country's GDP, about 3-5%, is allocated to road accidents [20]. According to [20], drivers are responsible for 78% of all accidents. Distracted driving has become a significant issue worldwide, and it is expected to worsen before improving. Visual distraction refers to taking one's eyes off the road, while cognitive distraction involves losing focus on driving even though physically present. This can happen due to daydreaming, being lost in thoughts, and other similar reasons. Manual distraction is when the driver takes his/her hands off the wheel while driving to perform different actions like drinking, reaching behind, adjusting the radio, texting, talking on the phone, or conversing with passengers while driving.

We propose to detect such distractions in real time and alert the user to prevent any adversity. This solution can be deployed on an edge device set up on the vehicle to give an instant alarm and it can also interact through IoT devices to process data and give insights in an asynchronous mode. For the scope of this project, the primary focus is to detect the manual distractions when the driver is not primarily focused on driving and accurately classify these activities which lead to driver distraction, through a highly efficient ML model at runtime using computer vision.

II. RELATED WORK

With the advent of Convolutional Neural Networks (CNN) in the early 2000s, deep learning algorithms have registered significant progress in the domain of image recognition. However, zeroing in on the perfect CNN architecture can still be a very difficult task. Among the many architectures proposed in the past, such as VGGNet, AlexNet, GoogLeNet (i.e., Inception), the deep residual network (i.e. - ResNet) was of help to us in this study. Other architectures like the recurrent neural network (RNN) gives impressive results on time series problems and it is a frequently used algorithm. Also problems involving long sequences such as speech recognition and machine translation.

As far as detecting distracted drivers using computer vision based approaches [1] [2], convolutional neural networks (CNNs) have become the most adopted and popular approach. The Distracted Driver Detection through image classification gained traction with the release of a Kaggle competition by State Farm in 2016. In this regard, YehyaAbouelnaga et al [13] [14] had created a new dataset that is often referred to as the AUC Dataset. This proposed a novel system based on posture estimation and achieved a ~96% classification accuracy. Hong Vin Koay et al. [10] focused on exploring the technique that uses both the original image as well as pose estimation images to classify the distraction. The pose estimation images were generated from HRNet and ResNet. The ResNet101 and ResNet50 architectures are used to classify the original and pose estimation images respectively following which a weighted

approach was followed to arrive at the final classification. This resulted in an accuracy of 94.28% and a F1 score of 94.27% on the American University of Cairo (AUC) Distracted Driver Detection dataset.

Another study by NarayanaDarapaneni et al [1] focused on developing the CNN Method of transfer learning on four architectures namely CNN, VGG-16, ResNet50 and MobileNetV2. The model was trained with images from a publicly available dataset containing ten different posture categories. It was observed that the ResNet50 and MobileNetV2 models provided higher accuracies of 94.50% and 98.12% respectively.

Some studies approached the problem by modifying architecture like the one done by Md. UzzolHossaina et al.[18]. In this study, the VGG-16 Architecture was modified using regularization techniques to improve model performance. According to the results, the system achieved an accuracy of 82.5% and processed 240 images per second on a GPU. The pre-trained ImageNet model was used for weight initialisation and the concept of transfer learning was applied.

Two similar studies by H. Varun Chand et.al [4] and Md. TanvirAhammed et al. [5] were done on driver distraction involving drowsiness and fatigue. The study by H.Varun Chand et.al [4] used machine learning with multi-layer perceptrons to detect microsleep and drowsiness using neural network-based methodologies. The accuracy of this paper was improved by using a CNN to classify drowsiness in the facial expressions detected by the camera. The ability to provide a lightweight alternative to heavier classification models with more than 88% accuracy for the category without glasses and more than 85% for the category night without glasses is the accomplishment of this work. In all categories, more than 83% accuracy was achieved on average as well as in usability. Furthermore, the new proposed model had a significant reduction in model size, complexity and storage when compared to the benchmark model (Max Size = 75KB).

Md. TanvirAhammed et al. [5] used MobileNet CNN Architecture with Single Shot Multibox Detector. Based on the output of the SSD MobileNet v1 architecture, a separate algorithm was used. To train the model, a dataset of approximately 4500 images was labeled with the object's face yawn, no-yawn, open eye and closed eye variations. Using the PASCAL VOC metric, 600 randomly selected images were used to test the trained model. The proposed method was intended to improve accuracy and computational efficiency. [23] Bing-Ting Dong et al. used an approach which detected driver fatigue and distracted driving behaviors using a single shot scale-invariant face detector (S3FD). This was first used to detect the face in the image and then the face alignment network (FAN) was utilized to extract facial features. Post that, the facial features were used to determine the driver's yawns, head posture, and eye movements. Finally, to analyze the driving conditions the random forest technique was used. The average accuracies achieved were ~100% for both face and eye detection.

Another experimental study using CNNs was done by RobinsonJime`nez et al.[24] which targeted detection of driver fatigue and emotion analysis of the driver in order to avoid reckless driving. The proposed model had a 93% accuracy rate in detecting the driver state and classified as normal, fatigued, drunken or reckless. A study [24] by Robinson Jime`nez et al. identified driver distraction states by using eye, mouth, and head movement and orientation as parameters for classification. Course segmentation using the Haar classifier techniques were used in conjunction with Adaboost techniques. Rectangular descriptors to detect faces in an image and Fine segmentation using Hough circle detection algorithm and Hough transform were used to determine the position of eye iris. A precision of ~86% was achieved.

III. DATA SET

The Kaggle competition "StateFarm distracted driver detection" published in 2016 [21] is a dataset widely used for different experiments and studies. It comprises ten classes. Creation date and other metadata has been removed from the images. The images are created in a controlled environment where drivers are not actually driving but posing. It is ensured that the driver appearing in train images will not appear in test images. The images are a collection of left-hand drive vehicles only. Each class contains close to 2300 images.

We also used one more recent dataset [13] [14] that was created by students of the American University in Cairo (AUC). Individuals (Males=29; Females=15) were wearing different clothes and videos were produced with different driving conditions. These individuals are from seven different countries, the USA, Palestine, Morocco, Canada, Uganda, Egypt and Germany. The dataset has been divided into train and test datasets distributed over 10 classes with a total of 14,478 images.

TABLE 1 :DATASETS CLASSES

Classes	Driver actions	Kaggle Images	AUC Train	AUC Test
c0	safe driving	2489	2640	346
c1	texting - right	2267	1505	213
c2	talking on the phone - right	2317	1062	194
c3	texting - left	2346	944	180
c4	talking on the phone - left	2326	1150	170
c5	operating the radio	2312	953	170
c6	drinking	2325	933	143
c7	reaching behind	2002	891	143
c8	hair and makeup	1911	898	146
c9	talking to passenger	2129	1579	218
	Total	22424	12555	1923

IV. TECHNICAL APPROACH

From the source datasets mentioned in Table 1 and Table 2, we apportioned the data for the purpose of Training, Validation & Testing with data distributed as shown in Fig 1.

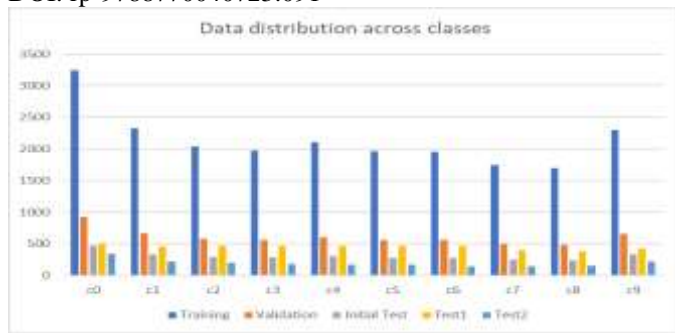


Fig 1. Data distribution across classes

In our approach to build the model, base models were identified namely ResNet101V2, ResNet50 and MobileNet and were trained on an “imagenet” dataset. In our initial findings, the model based on ResNet101V2 had a significantly better accuracy than the other models. Hence ResNet101V2 was selected to further experiment and optimize to achieve better results. Different model variations, Key approaches used for experimenting are described below.

Model 1 and Model 2 have been trained with similar strategy but with a different number of layers that can be retrained. In Model 1, we used transfer learning with 34 last layers trained. The top layers will contain a global average pooling and a dense layer of 10, activated with softmax producing the classification. Categorical cross entropy was used as loss function and adam was used as an optimizer.

In Model2, the number of layers that are retrained are 150. The remaining architecture remains the same as Model 1.

The accuracy on Kaggle data is good with 99% in both the models. However, the model accuracy on AUC dataset has come down to 58% from 66% in Model 1.

To improve the AUC test accuracy, we sought to experiment again the transfer learning approach but by removing some data in AUC test dataset as they were right hand driving images and our model was trained on left hand driving images only. The availability of right hand driving images is very less compared to left hand driving images implying data imbalance problem.

Model 3 & Model 4 were built with similar architecture and strategy as Model 1 & 2. The main difference will be the reduced data by removing the right hand driving images.

Few parameters were changed in our fourth model like transfer learning with top 35 layers with dense 1000 was used keeping all other parameters same as third model

Almost no change was observed in accuracy of both the test data. On kaggle data, validation and test accuracy is 99% in both the models while AUC test accuracy at 58% for third model & 59% for fourth model.

Observations from all four models clearly indicated that our approach of transfer learning was not being useful for AUC data set and experimenting with hyperparameters is not making any impact on the model. To understand why

the model was not performing as expected, a grad cam analysis was then used to understand activation and prediction of the model [15][16][17]. The intent is to understand the model with scientific facts and approach the model building with Explainable AI.

Grad Cam Analysis : The Grad CAM technique uses the feature map of the last convolutional layer and the classification score of the class in interest. It calculates the gradient between them. The larger gradients in the places of the image are the one impacting the final classification score.

In our analysis there were mainly two observations made. First, we could see that most of the activation in our models were wrongly placed implying that our trained model was not correctly identifying parts of the image that could result in correct classification. Sometimes even if the predictions were right the activation on the images were not rightly placed. Examples of the above mentioned observation is shown below in Fig 2 and Fig 3.



Fig 2. Grad-CAM Adjusting Radio

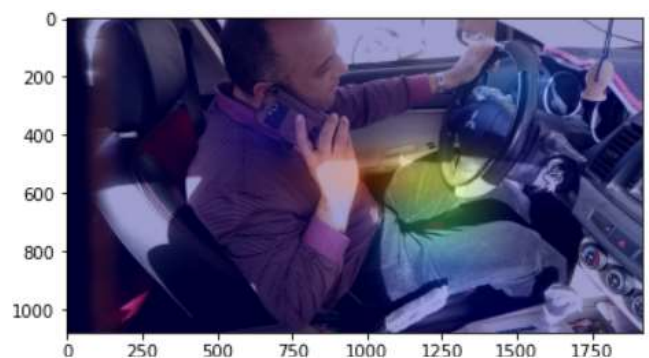


Fig 3. Grad-CAM Talking on phone right

By this study we understood that our trained transfer learning models are not enough to classify our images into the right category. Second, to understand the GradCAM on the dataset with pre-trained ResNet101V2 Model, we did grad cam analysis. Most of the images are classified as minibus, seatbelt, golf cart, etc and the activations are showing away from the person sitting on the driving seat. For example (as shown in fig 4) the activation shows on door windows which will never be useful for classification required for this model. It is inferred that the transfer learning would be inappropriate for the dataset of Distracted Driver Detection.

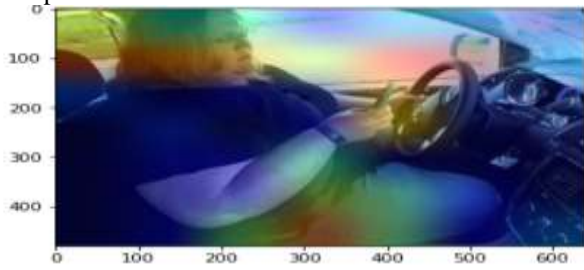


Fig 4. Grad-CAM with ResNet101 pretrained on Imagenet

With all the above stated observations we inferred and acted upon that the usage of transfer learning in our model with few or more layers is not making any significant impact on accuracy of our test datasets. Hence we chose to train the entire network of the ResNet102V2 model.

In Model5, we trained the entire network, without any modifications to the ResNet101V2 architecture. The starting learning rate is chosen as 0.0001 instead of default 0.001.

With this change in approach we could see a significant change in the accuracy of the AUC test data which now improved from 58% in our previous model to 76%.

To validate the model further apart from accuracy of classification, a GradCAM analysis is done on the training data and as well predicted data with the new model. It is observed that the activation on the training data and as well as predicted data is more appropriate such as classification of an image as “operating a Radio” activates the Radio and lower hand of the person. Fig 5 shows the same as below



Fig 5. Grad-CAM Adjusting Radio

On further observing the wrongly classified images by the model, we understood that the model has learnt specific features that alone may not be sufficient to classify an image. For eg., an image with a slight open mouth is considered to be Talking rather than safe driving. Or an image where a cup in the left hand is covered by the right hand position, so the model is unable to understand it as drinking. To give more learning capability to the model, we augmented the data of AUC training set with shear 30 degrees. The idea is to provide more images with a small shift. This augmented approach should improve and generalize the model.

Architecture of our sixth model is exactly the same as the fifth model, except that it is trained with more data (augmented data).

With this change in approach we could see a change in the accuracy of the AUC test data which now improved from 76% to 82%

TABLE 2: KEY PARAMETERS USED IN EACH MODEL

Model	Hyper Parameters
ResNet101V2_Model_1	Activation- leaky_relu Epoch - 20 Batch Size - 50 Dense Layer -10 Optimizer - Adam (beta_1=0.9, beta_2=0.999lr=0.001) Loss - Categorical_crossentropy
ResNet101V2_Model_2	Activation- relu Epoch - 20 Batch Size - 32 Dense Layer -10 Key Change in Approach: Reducing lr for every 5 epochs with 0.1 factor if val_loss is not improving
ResNet101V2_Model_3	Activation- relu Epoch - 20 Batch Size - 32 Dense Layer -10 Key Change in Approach: Global average pooling, flatten and a dense layer of 10
ResNet101V2_Model_4	Activation- relu Epoch - 20 Batch Size - 32 Dense Layer -10 Key Change in Approach: Kaggle and AUC cam1 data. Cam2 data is discarded as it is right wheel driving and has very few images.
ResNet101V2_Model_5	Activation- relu Epoch - 50 Batch Size - 32 Dense Layer -10 Optimizer - Adam (beta_1=0.9, beta_2=0.999lr=0.001) Key Change in Approach: Complete Network retraining.
ResNet101V2_Model_6	Activation- relu Epoch - 50 Batch Size - 32 Dense Layer -10 Optimizer - Adam (beta_1=0.9, beta_2=0.999lr=0.001) Key Change in Approach: Augmented for Cam1 - shear 30 degrees
	Loss - Categorical_crossentropy

IV. RESULTS AND DISCUSSIONS

Based on our experiments, we discovered that our model is able to learn prominent features with a small number of parameters. With sufficient training data the model is able to give accurate results. Moreover, we could see the performance of the proposed ResNet model as shown in Table 4. In the table, it's clearly observed that the Model performed very well with the State Farm Distracted Driver data-set on the pre-trained as well as transfer learning model. However that was not the case when trained on AUC Distracted Driver Set. We found that training the entire network is very optimal for our model. The accuracy increases, prediction and activation is also appropriate.

TABLE 3: ACCURACY ACROSS MODELS

Model	Training Accuracy	Validation Accuracy	Kaggle Test Accuracy	AUC Test Accuracy
ResNet101V2_Model_1	100%	99.23%	99%	66%
ResNet101V2_Model_2	100%	99%	99%	58%

ResNet101V2_Model_3	100%	99.58%	99%	59%
ResNet101V2_Model_4	99.94%	99.23%	99%	59%
ResNet101V2_Model_5	100%	99.53%	100%	76%
ResNet101V2_Model_6	99.99%	99.58%	100%	82%

TABLE 4: CLASS WISE ACCURACY FOR FINAL MODEL

Class	Driver Action	Kaggle Dataset F1-Score	AUC Dataset F1-Score
c-0	safe driving	98.99	66.99
c-1	texting - right	100	85.62
c-2	talking on the phone -right	100	94.95
c-3	texting - left	99.57	84.61
c-4	talking on the phone - left	99.78	98.34
c-5	operating the radio	99.45	98.30
c-6	drinking	99.89	88.88
c-7	reaching behind	100	77.30
c-8	hair and makeup	99.47	77.12
c-9	talking to passenger	99.53	68.76

We would like to mention some pertinent points based on our observations as below:

Data balancing (CAM 2 data had to be removed from the model because of very few images found in right wheel driving)

Data Verification - due to improper data extraction, the AUC data was improperly classified which resulted in significant delays in the initial stages of model training.

The Learning Rate(lr) turned out to be significantly important for convergence due to the lengthy model architectures

Understanding of when to use Transfer Learning & when to use the architectures

Techniques like GRAD CAM were important from a model validation perspective

Availability of more quality data would have helped the model to generalize better

We can conclude that existing state of the art architecture like the ResNet101v2 Model alone can be sufficient if trained and optimized well

Scope for future enhancements

Model training for right seat driving

The solution will be improved by adding more convolutions to the end of the architecture. The ResNet101 for imagenet architecture’s last convolution outputs 7x7x2048 activation maps and it is connected to 1000 dense layers for the classification of 1000 classes. Since there are 10 classes in the problem, connecting 2048 to 10 could have decreased the accuracy. So we will introduce 1 or 2 convolutions before connecting into 10 dense layers.

Based on the observations of the predictions and class activation mappings in the v6 model, we propose to add one more convolutional network trained on imagenet parallel to the existing network. The purpose of this network is to establish face detection with classification of face or not face. Then the classification will be used along with the classification of the original network to give the final

classification. The purpose of this approach is to activate the facial features that contribute to the final accuracy.

At the onset of our study, we had considered a benchmark of 88% on the test set of the KaggleStateFarm dataset with the ResNet50 and 82% on the MobileNet models [11]. Models performed way better by achieving an accuracy of up to 96% on the combined Kaggle and AUC datasets.

ROC Curve for Safe Driving Vs rest of the classes is provided below on ResNet101V2_Model 6. The AUC value of 98% and 100% shows the larger coverage of the model’s on different datasets.

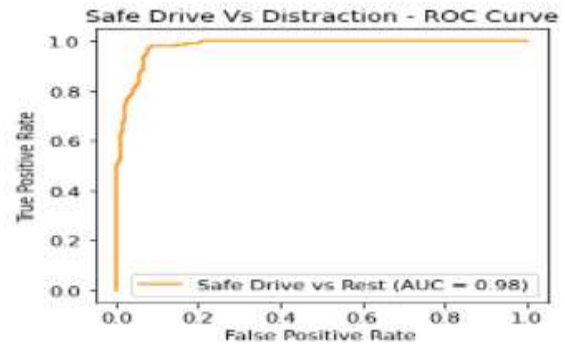


Fig 6: ROC Curve for AUC Test Data

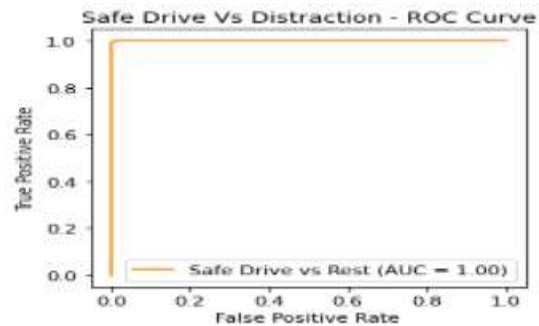


Fig 7: ROC Curve for Kaggle Test Data

REFERENCES

- [1] N.Darapaneni, J. Arora, M.Hazra, N.Vig, S. S. Gandhi,,Gupa, S., and A. R. Paduri, "Detection of distracted driver using convolution neural network," 2022,ArXiv [Cs.CV]. <http://doi.org/10.48550/ARXIV.2204.03371>
- [2] M.Alotaibi, and B. Alotaibi, "Distracted driver classification using deep learning. Signal, Image and Video Processing,"vol. 14, no. 3, pp. 617–624, 2020, <https://doi.org/10.1007/s11760-019-01589-z>.
- [3] D.Ruparel, A.Rajde, S. Shah, and P.Gidwani, "Distracted Driver Detection. Distracted Driver Detection, October 1, 2020, <https://www.jetir.org/view?paper=JETIR2010371>
- [4] H. Chand, J.Karthikeyan, and T. "CNN Based Driver Drowsiness Detection System Using Emotion Analysis," Tech Science Press, January 1, 2021,, <https://doi.org/10.32604/iasc.2022.020008>.
- [5] M. T. A.Dipu, S. S. Hossain, Y. Arafat, and F. B. Rafiq, "Real-time driver drowsiness detection using deep learning," International Journal of Advanced Computer Science and Applications : IJACSA, vol. 12, no. 7, 2021, <https://doi.org/10.14569/ijacsa.2021.0120794>.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, ArXiv [Cs.CV]. <https://doi.org/10.48550/ARXIV.1512.03385>
- [7] I. Mecatrónica, R. J. Moreno, O. Avilés Sánchez, and A. Hurtado, (n.d.). Ingeniería y Competitividad. Redalyc.org. Retrieved March 17, 2023, from <https://www.redalyc.org/pdf/2913/291333276006.pdf>

- [8] M. Gjoreski, M. Z. Gams, M. Lustrek, P. Genc, J. U. Garbas, and T. Hassan, "Machine learning and end-to-end deep learning for monitoring driver distractions from physiological and visual signals", *IEEE Access: Practical Innovations, Open Solutions*, vol. 8, pp. 70590–70603, 2020, <https://doi.org/10.1109/access.2020.2986810>.
- [9] M. Oberoi, and Shah and Anchor Kutchhi Engineering College, "Driver Distraction Detection using Transfer Learning," *International Journal of Engineering Research and Technology (Ahmedabad)*, vol. 9, no. 05, 2020, <https://doi.org/10.17577/ijertv9is050862>
- [10] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," 2018, ArXiv [Cs.CV]. <https://doi.org/10.48550/ARXIV.1804.06208>
- [11] M. Aljasim, and R. Kashef, "E2DR: A deep learning ensemble-based driver distraction detection with recommendations model," *Sensors (Basel, Switzerland)*, vol. 22, no. 5, p. 1858, 2022, <https://doi.org/10.3390/s22051858>
- [12] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., & Manoharan, R. (2021). Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. *IEEE Access*, 9, 74659-74673.
- [13] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, "Driver distraction identification with an ensemble of convolutional neural networks", 2019, ArXiv [Cs.CV]. <https://doi.org/10.48550/ARXIV.1901.09097>
- [14] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification", 2017, ArXiv [Cs.CV]. <https://doi.org/10.48550/ARXIV.1706.09498>.
- [15] B. N. Patro, M. Lunayach, S. Patel, and V. P. Namboodiri, "U-CAM: Visual Explanation using Uncertainty based Class Activation Maps", 2019, ArXiv [Cs.CV]. <https://doi.org/10.48550/ARXIV.1908.06306>
- [16] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - Weakly-supervised learning with convolutional neural networks," 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 685–694, 2015.
- [17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via Gradient-based localization," 2016, ArXiv [Cs.CV]. <https://doi.org/10.48550/ARXIV.1610.02391>.
- [18] Rajesh, M., & Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. *Multimedia Tools and Applications*, 81(1), 873-885.
- [19] B. T. Dong, H. Y. Lin, and C. C. Chang, "Driver fatigue and distracted driving detection using random forest and convolutional neural network," *Applied Sciences (Basel, Switzerland)*, vol. 12, no. 17, p. 8674, 2022, <https://doi.org/10.3390/app12178674>
- [20] L. L. C. Books, "Road accidents in India: Bus accidents in India, level crossing accidents in India, road accident deaths in India", Chhabibiswas, Amir Khan. Books, 2010.
- [21] State farm distracted driver detection. (n.d.). Kaggle.com. Retrieved March 17, 2023, from <https://www.kaggle.com/c/state-farm-distracted-driver-detection>.
- [22] (N.d.). Transportation.gov. Retrieved March 17, 2023, from <https://www.transportation.gov/sites/dot.gov/files/2022-02/USDOT-National-Roadway-Safety-Strategy.pdf>
- [23] B. T. Dong, H. Y. Lin, and C. C. Chang, "Driver fatigue and distracted driving detection using random forest and convolutional neural network," *Applied Sciences (Basel, Switzerland)*, vol. 12, no. 17, p. 8674, 2022, <https://doi.org/10.3390/app12178674>.
- [24] (N.d.). Org.Co. Retrieved March 17, 2023, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-30332014000200006