2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)

A Cross-Dataset Performance Analysis of Vision Transformers for Drowsiness

Bishal Saha
Assistant Professor
Computer Science and Engineering,
SET, MRIIRS,
Faridabad India
bishalsaha.set@mriu.edu.in

Yash Tanwar
Student
Computer Science and Engineering,
Thappar University
Patiala, India
ytanwar_be23@thapar.edu

Tushar Mathur
Student
Computer Science and Engineering,
SET, MRIIRS,
Faridabad India
tusharmathur66@gmaiil.com

Poonam Tanwar
Computer Science and Engineering
SET, MRIIRS,
Faridabad India
poonamtanwar.set@mriu.edu.in

Anshu Sharma
Assistant Professor
Computer Science and Engineering,
SET, MRIIRS,
Faridabad India
anshusharma.set@mriu.edu.in

Advaya Raj Saini
Student
Vasant Valley School,
Vasant Kunj, New Delhi
advayasaini2@gmail.com

***Abstract*— Drowsiness while driving, using machinery, or undertaking other critical activities is a serious hazard, resulting in thousands of accidents each year. Conventional machine learning procedures utilizing handcrafted characteristics are often sensitive to illumination changes, head pose changes, and occlusion. Recently, deep learning, particularly Vision Transformers (ViT), have been shown to better capture global dependencies in visual data. We propose a Vision Transformer-based drowsiness detection system learned using two publicly accessible datasets. The proposed model conducts multi-class classification predicting eye closure, yawning, and normal facial states. The experimental results indicate that our approach achieves 95% on Dataset-A and 96% on Dataset-B; additionally, this is superior to solution using traditional convolutional neural networks. The suggested solution is computationally efficient and appropriate for real-time monitoring of drivers.**

Keywords—Vision Transformer, Drowsiness Detection, Deep Learning, Attention Mechanism, Computer Vision

I. Introduction

Drowsiness is one of the primary causes of accidents in the transportation, industrial, and at-risk work environments. Over time, as society leans on road-based transportation and long-duration drives, the implications of drowsiness on safety have become a serious public safety issue. Studies show that prolonged inattention can significantly delay reaction time, decrease situational awareness, and ultimately relate to poor decision making that threatens the safety of the operator and individuals nearby. Identifying drowsiness early can help reduce the severity of accidents, and can help prevent the loss of life and property.

Historically, drowsiness has been assessed using physiological measurements such as electroencephalogram (EEG), electrocardiogram (ECG), or heart-rate variability (HRV). While it has been shown that these physiological measurements can be accurately used to assess drowsiness, they also require the fixation of specific sensors to the individual to monitor. This can become impractical for a consumer-grade real-time system that can be removed or installed without requiring a specific real-time physiological measurement, such as EEG. Alternatively, behavioral-based evaluations of drowsiness use facial expressions, eyelids closure duration, or yawning frequency as a camera-based, non-contact way to assess. That said, classical computer vision models are limited in performance by lighting conditions, head pose, occlusion, and anatomical variation in faces.

Deep learning has considerably enhanced the reliability of vision-based fatigue detection. Several deep learning models, including Convolutional Neural Networks (CNNs), have been extensively invested in for detecting behavioural indicators such as eye-blink frequency and yawning, i.e., due to their aptitude for spatial feature extraction. CNNs rely on localized receptive field sizes somewhat limiting their ability to model long-range dependencies in the image. This limitation arises as there may be very subtle behavioural indicators scattered within the regions of a face.

Recent advances in transformer-based architectures have transformed the landscape of computer vision tasks. The Vision Transformer (ViT), derived from Natural Language Processing (NLP) models, is capable of learning a purely attention-based mechanism that models global relationships among patches of images. By eliminating convolutional layers, ViTs are capable of learning contextual representations much better than CNNs, and are more robust to the inherent variation in techniques such as illumination, occlusions, and face orientation.

Motivated by these benefits, this study implements a multi-class drowsiness detection framework based on the Vision Transformer framework that can classify normal, eye-closed, or yawning states. The framework is trained on two openly available face datasets and assessed using multiple performance metrics. The aim is to develop a non-contact, real-time, and robust framework for driver assistance

979-8-3315-7767-4/25/$31.00 ©2025 IEEE

systems, surveillance cameras, and industrial monitoring systems.

II. Literature Review

Initial studies on drowsiness detection mainly focused on analyzing physiological signals, including the electroencephalogram (EEG), electrocardiogram (ECG) and electromyography. These studies showed a very high level of accuracy in the assessment of fatigue. However, the intrusive nature of wearable sensors, user discomfort and financial cost of implementation prevented their potential use in practical scenarios. Such continuous monitoring would not be feasible for drivers and industrial workers who require complete freedom of movement.

Afterwards, non-intrusive methods that utilize computer vision combined with facial behavioral characteristics gained momentum. Early methods relied on hand-craft descriptors, such as local binary patterns (LBP), Gabor features or Haar-based eyelid detection to quantify blinking frequency and closure of the eyelid (PERCLOS). These approaches were computationally affordable, however, they were not robust to fluctuations in illumination, facial occlusions, and changes in head pose and performance was significantly degraded under real-world driving scenarios.

The emergence of deep learning has represented a significant step-up in this area of drowsiness detection. Convolutional neural networks (CNN) have been widely introduced due to the power of their spatial feature extraction qualities. Several studies have used CNNs to classify states-of-eye or detect yawning and achieved significantly improved performance when compared to the hand-craft feature-based systems. Some researchers integrated CNNs with Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) units to learn temporal patterns associated with blinking sequences. While effective, these hybrid architectures demanded high computational power and often failed to generalize beyond their training domain.

Although successful, the limiting aspect of CNNs (Convolutional Neural Networks) as they are constructed and used is that they work only through localization associated with a receptive field that bounds their identification of global contextual information over the face. Visual clues that determine state based on fatigue—i.e. drooping eyelids, slight mouth deformation, gradual head lowering, etc—are often spatially distributed on the face. Therefore, It may be difficult for CNN-based architectures to distinguish between a standard blink, prolonged eye closure, and fatigue to produce false predictions.

Recent cohort transformer-based architecture has attempted to improve on this localization constrained by a CNN model with attention mechanisms which resolve long-range relationships found in visual data. The Vision transformer (ViT) adapted from the transformer architecture in Natural Language processing (among other uses) first splits an image into non-overlapping image patches and computes self-attention across those patches. Studies have begun to show that ViT can outperform CNNs on image classification benchmarks when developed on large datasets. ViTs also show superior robustness compared to CNNs when high-level occlusion, pose variation, and noise due to changing environments (more common in the drowsiness literature) are present.

Recent literature related to this systematic review has shown a general move toward employing transformer-based fatigue-detection frameworks. Developers have used ViTs to analyze eye closure, yawn gesture detection, and head pose estimation to achieve state-of-the-art results. A small number of works have employed ViTs with some temporal modules or even used lightweight convolutional stems to improve their performance and prediction accuracy for less than large datasets.

Comparative research demonstrates that multi-head self-attention sufficiently captures subtle spatial dependencies across facial areas allowing for a more accurate classification of drowsiness-related behaviors. In addition, transfer learning with large pre-trained ViT models leads to reduced training time, while addressing potential data sparsity issues.

In conclusion, the literature demonstrates a clear transition from invasive sensor-based methods, to deep learning and now transformer-based methodologies. While CNNs have been the established approach for the past several years, Vision Transformers have a greater contextual understanding, making them a promising new avenue for non-intrusive real-time drowsiness detection systems.

III. Methodology

This system uses vision to detect drowsiness by recognizing facial hints through a pretrained Vision Transformer (ViT). The process contains several components such as the detection of a face, preprocessing the face, patch embedding, self-attention, and a classification. The system will evaluate minor behavioral changes related to fatigue through the visual information such as extended eye closure and yawning without any invasive physiologic sensors.

A. System Overview

The entire pipeline starts with the collection of video frames or still images, then it proceeds with face detection, cropping, and normalization. Once the face has been localized, the Vision Transformer classifies the user state as normal, eye-closed, or yawning. In general, the ultimate goal is to develop a low weight, non-intrusive and reliable system that operates in changing environmental conditions, and is suitable for real-time driver or worker monitoring.

B. Face Detection

The face detection process is located in the first part of the project and uses OpenCV's Haar Cascade classifier, which efficiently finds the facial area from the input frame. Upon detection, the face region is introduced which removes the unneeded background data, which reduces the load on the computational aspect of the model and analytics only to the aspects of data that are relevant, such as the eyes and mouth. This promotes the following stages having a much cleaner, more informative data source than improving its analytics by reducing common environmental clutter of multiple objects and backgrounds that surround the face.

C. Data Preprocessing

The preprocessing of all cropped images will occur prior to feature extraction, and as a means to standardize the image across samples. First, all images will be scaled to the necessary input size of the Vision Transformer model, and the pixel values will also be scaled in order to minimize the effects of lighting inconsistencies. Not only will the image

data be scaled, but the scaling of the standardized RGB channels of the data will also create more uniformity across the datasets. Finally, the data augmentation will include randomly flipped images, random rotation, and brightness adjustment as a way to achieve a greater diversity of samples, and to better generalize the model to see changes that might occur in real life (i.e. head movements, illumination changes, etc.).

D. Patch Embedding

Each image is divided into a set of non-overlapping equal-sized patches after applied preprocessing to the image. The patches are then flattened and projected into a latent vector space with a linear transformation. Because the flattening process destroys any information about spatial arrangements, the embeddings will have positional encodings added to them. The positional encodings will allow the model to understand spatial relationships across regions of the face, which is important to differentiate feature interactions for conditions like simultaneous eyelid drooping and opening of the mouth.

E. Vision Transformer Architecture

The patches placed in the model are processed through multiple layers of the Transformer encoder which consists of multi-head self-attention, feed-forward networks, layer normalization, and residual connections. The self-attention mechanism calculates the interactions of all patches pairwise, which allows the model to perceive fine-grained global relationships in the face. Vision Transformer accounts for long-range relationships to distinguish between momentary blinking and eyelid closure due to fatigue, and to distinguish minute facial deformations caused by yawning, which can be challenging for CNNs with lower receptive fields.

F. Classification Head

During the last stage of the model, output from a specific classification token is obtained. The classification token contains the contextual representation of the entire image. It is input to a fully connected layer with dropout regularization applied to avoid overfitting, and it assigns probability scores for each output class using a softmax, generating an output distribution for the class labels. The predicted behavioral state is converted to the output class with the higher probability.

G. Training Strategy

The model was trained with PyTorch with the Adam optimizer and fairly standard categorical cross-entropy loss. We trained for twenty epochs in batches of thirty-two images. Transfer learning was used because we initialized the Vision Transformer with pre-trained ImageNet weights to promote convergence and better generalization, especially during training on a CPU-only system. We also used augmentations to help with the lack of training diversity.

H. Evaluation Metrics

To evaluate performance, we derive accuracy, precision, recall, and F1-score metrics from the test phase. Confusion matrices were also examined to show what misclassifications were prevalent in the testing phase. All these metrics help provide better insight about the system's performance, particularly about differentiating the normal blinking behavior from similar looking fatigue blinking behavior.

I. Comparative Baseline

A lightweight CNN baseline model was also trained under the same conditions to facilitate comparison. The vision transformer represented a superior option to the CNN model because it demonstrated higher robustness to lighting differences, appropriations, and occlusions. This is anticipated because the transformer is particularly good at modeling global contextual relationships across patches.

J. Real-Time Inference

In real-time operation, alerts are not triggered until multiple frames are analyzed consecutively, which keeps the chances of a false alarm, such as from momentarily blinking or a fluctuation in expression, very low. Continuous monitoring of behavioral patterns provides the methodology needed to create a useful and reliable system for real-world implementation.

IV. Results

The intended drowsiness detection model based on the Vision Transformer was trained and validated on two publicly available facial-state datasets referred to as Dataset-A and Dataset-B. Both datasets contain images of normal, eye-closed, and yawning facial states in a variety of illumination and head poses. After data preprocessing and augmentation, the model was fine-tuned over a period of twenty epochs using a batch size of thirty-two. The training was done within a CPU-only environment where convergence was achieved without overfitting, which could be attributed to the use of dropout regularization as well as the benefits of transfer learning.

The accuracy of overall classification across the two datasets, further demonstrates the advantages of the Vision Transformer in understand subtle facial expressions related to the presence of drowsiness. The proposed model demonstrated an accuracy of 95% on Dataset-A. The model achieved 96% accuracy on Dataset-B indicating good generalization on different visual conditions. Throughout both datasets, the precision and recall scores were also very high indicating that the model accurately detected normal facial display as well as signals of fatigue. The confusion matrices indicate that prolonged eye closure, as well as yawning, were correctly labelled for the majority of test samples, with any mislabeling consisting of brief blinking and partial occlusion.

In a comparative baseline experiment, a standard Convolutional Neural Network (CNN) version of the model produced the lowest accuracy (90%) on both datasets. This lower accuracy indicates that there can be issues with a CNN when learning long-range spatial dependencies, particularly if facial features are distributed or occluded. The accuracy in labeling classification was improved by a strong comprehension of the spacial dynamics of the face through the global attention mechanism.

The training and validation loss curves reiterated the model's stability, showing smooth convergence with no visible fluctuations. Additionally, validation accuracy seemed to mirror training accuracy, suggesting that the proposed architecture was able to generalize to samples it had not seen before. Apart from this, the per-frame inference duration remained low enough to allow monitoring in real

time, even with only a CPU at most, which is a key consideration for applications involving human safety.

Overall, results suggest the Vision Transformer is a viable and dependable approach to multi-class drowsiness detection. Its improved accuracy compared to previous and pre-existing deep learning architectures supports the attested advantages of using attention-based feature extraction. The results generally show it is possible to deploy transformer-based models in safety-critical real-world settings - as illustrated in this proof-of-concept study - especially driver monitoring systems, and emergency-related safety-monitoring frameworks in industry.

V. Conclusion

The current study introduced a method for assessing driver drowsiness in real-time based on vision transformers by analyzing the facial images of drivers. The method effectively captured fine visual clues associated with fatigue, such as eye closure and reduced facial muscle movement, by applying self-attention mechanisms. It was demonstrated through extensive experiments with two publicly available benchmark datasets, to capture high levels of accuracy to be competitive with and surpass several CNN baselines under different protocols, with some experiments conducted using only retrained hardware environments like CPUs.

The comparative experiments support the generalization capacity of ViTs as compared to conventional CNN methods in safety-critical applications, demonstrating robustness to illumination variation, head pose and occlusion. While overall we present a method of good performance, it is reasonable to expect that real world ability will only improve by inclusion of temporal behaviors and sensor fusion to eliminate false positives. Future work based on this paper may also include lightweight versions, on-device inference and the edge-device environments common and integrated in modern cars. Overall, we present a promising, reliable and efficient approach for potential improvements in road safety and decreases in drowsiness-related accidents.

VI. Future Work

While the proposed drowsiness detection system using vision transformers has proven promising results, numerous avenues can be explored to amplify the applicable promise of the technology. For example, improving the temporal modeling by incorporating attention-based time-series modeling to capture gradual change in behavior due to fatigue rather than episodic, frame-level entries would be a valuable improvement. A multi-subject dataset could improve generalizability and bias by capturing greater variation in subjects, environments, and lighting conditions. Furthermore, incorporating multimodal sensor data (yawns, changes in steering pattern, heart rates) would be another meaningful opportunity that would help dramatically reduce false positives in challenging situations.

With respect to deployment, ensuring the model is optimized for the edge can facilitate embedding systems into vehicles and enhance user responsiveness and enhanced privacy. There are important considerations surrounding edge learning here in regard to model compression, light-weight Vision Transformers, and distillation approaches to intend to strike a uniform balance between inference time and accuracy needed from physics limited hardware in fixed location applications. Finally, user studies and real driving field trials will be phenomenally useful in perception wise, as being able to study the practicality of usability regarding system robustness in drive cycles will be important next step to evaluate. Overall, the improvements for the systems as described are intended to enhance fatigue-monitoring for reliable, intelligent, and scalable improved future road safety tech.

VII. References

[1] A. Vaswani, N. Shazeer, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] A. Dosovitskiy, L. Beyer, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[3] Y. H. Zhang, and Q. Li, "Driver Drowsiness Detection Based on Eye State Recognition Using CNN," *IEEE Access*, 2018.

[4] R. Fu, L. Zhang, and J. Xu, "Driver Fatigue Detection through Multiple Cameras and EEG Signals," *IEICE Transactions*, 2016.

[5] S. Abtahi, A. Omidyeganeh, et al., "Driver Drowsiness Monitoring Based on Yawning Detection," *ICIP*, 2014.

[6] Y. Zhang, B. Wu, et al., "Real-Time Driver Fatigue Detection Based on Eye State," *Sensors*, 2019.

[7] G. Patel, and S. Thakkar, "Vision-Based Fatigue Detection Using Deep Learning and Eye Aspect Ratio," *Procedia Computer Science*, 2020.

[8] M. L. Sørensen, et al., "Artificial Intelligence for Vision-Based Drowsiness Detection," *IEEE Intelligent Vehicles Symposium*, 2020.

[9] S. Park, and J. Kim, "Driver Monitoring System for Detecting Drowsiness," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[10] K. He, X. Zhang, et al., "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

[11] A. Howard, M. Zhu, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv*, 2017.

[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *CVPR*, 2005.

[13] P. Viola and M. Jones, "Robust Real-Time Face Detection," *IJCV*, 2004.

[14] T. Nahvi and M. Montazer, "Driver Drowsiness Detection Using Convolutional Neural Networks," *Journal of Computing and Security*, 2017.

[15] B. Sun, et al., "Driver Behavior Analysis Using Computer Vision," *IEEE Transactions on Intelligent Vehicles*, 2020.

[16] A. Muhammad, et al., "Automated Vision-Based Detection of Driver Fatigue," *Transportation Research Record*, 2019.

[17] Z. Li, F. Bao, et al., "Drowsiness Detection Using Facial Landmarks and Deep Learning," *International Journal of Advanced Computer Science*, 2021.

[18] H. Zeng and L. Peng, "Multi-Feature Fusion for Driver Fatigue Recognition," *IEEE Access*, 2020.

[19] C. Wu, et al., "Robust Eye State Detection for Driver Monitoring in Real-Time," *Pattern Recognition Letters*, 2019.

[20] J. Chen, et al., "Transformer-Based Vision Architectures for Driver State Analysis," *Pattern Recognition*, 2022.