Grenze International Journal of Engineering and Technology, June Issue

Transformer-CNN Hybrid Model for Real-Time Driver Drowsiness Detection

Prof. Kapil Tajane[1], Prof. Rahul Pitale[2], Utkarsh Gardi[3], Sanskar Ghule[4], Utkarsha Kamble[5] and Aneesh Kanhere[6]
[1-6]Pimpri Chinchwad College of Engineering, Pune, India
Email:{kapil.tajane, rahul.pitale, utkarsh.gardi21, Sanskar.ghule21, utkarsha.kamble21, aneesh.kanhere21}pccoepune.org

***Abstract*—Driver drowsiness is a leading factor in global accident rates, sometimes resulting in significant injuries or even fatalities. Traditional fatigue detection systems depend on behavioural signals, physiological indicators, or vehicle movements. However, these technologies usually yield high false alarm rates and may not always perform in real time. Recently, deep learning models have demonstrated significant accuracy in detecting tiredness just through facial cues. This paper presents a unique transformer-CNN hybrid model that considers both Convolutional Neural Networks (CNNs) for capturing spatial features and Transformers for handling temporal sequences. The model is designed to efficiently process real-time video streams, ensuring it captures both small-scale facial changes and broader eye movement patterns. To evaluate its effectiveness, we compare our model against existing deep learning techniques using well-known datasets like NTHU-DDD and UTA-RLDD. Key performance indicators—accuracy, F1-score, and AUC-ROC—are analysed to see how well the model performs. Our Transformer-CNN hybrid outperforms conventional CNN and RNN-based models, offering better robustness across different lighting conditions and driver behaviours.**

***Index Terms*— Driver Drowsiness Detection, Transformer-CNN Hybrid Model, Deep Learning, Real-Time Monitoring, Feature Extraction, Self-Attention Mechanism, Facial Recognition, Performance Metrics (Accuracy, F1-Score), Eye Movement Analysis, Road Safety Enhancement.**

I. INTRODUCTION

By reason of Road accidents seem to be a major global difficulty, and a key contributing factor to deadly fatalities is driver fatigue. The National Highway Traffic Safety Administration (NHTSA) estimates that sleepy driving causes over 100,000 collisions annually in the US alone. Similarly, research conducted by the German Road Safety Council shows that driver weariness is responsible for about 25% of highway deaths. Real-time fatigue detection serves as vital to preserving traffic safety and mitigating possibly lethal situations. One of the main risk factors for traffic accidents is drowsiness, which affects mental capacity, quickness of response, and alertness to situations. There's a demand for potent and exact detection methods as promptly recognizing and intervening can avert serious outcomes. Numerous techniques are currently put forth to identify tiredness, including behavioral analysis of facial expressions, physiological signals like electroencephalography (EEG) and heart rate monitoring, and vehicle data like steering patterns. A dependable and non-intrusive method for identifying driver tiredness is facial expression analysis.

Grenze ID: 01.GIJET.11.2.542_1
© Grenze Scientific Society, 2025

Facial recognition in computer vision is a complicated area due to the numerous variations in human appearances. But relative to certain other facial features, the eyes provide consistent traits important for identifying fatigue. Important features correlated with drowsiness involve blink rate, gaze vector, and eyelid closure duration. If such traits are correctly identified, a driver's state of awareness can be monitored in real-time, enabling timely alerts and interventions. Deep learning methods have advanced the field of drowsiness detection significantly.

Whereas transformers are also great at detecting sequential dependencies (in time series data) and Convolutional Neural Networks (CNNs) are very good at pulling out spatial relations from images of a face. Transformers explore temporal changes to search for patterns indicative of tiredness, while CNNs generally focus specifically on image data, identifying face landmarks and salient features. To enhance internet of vehicles analysis by detecting driver sleepiness in real time, this study proposes a transformer-CNN hybrid model that integrates the advantages of the two architectures.

The goal of the research project is to examine if this hybrid method increases the discovery rate and reduces false positives and how well it performs in certain environmental conditions like variations in illumination, head motions and face appearances. The research exercise provides details for the practical execution of deep learning-oriented fatigue detection systems by referring on standard datasets and efficiency evaluation metrics.

II. TOOLS AND ALGORITHMS

A. Proposed Approach

Through the capture of both local and global feature dependencies, our hybrid technique guarantees improved efficiency in identifying sleepiness stages. A labelled dataset with four categories—Open, Closed, Yawn, and Non-Yawn—is used for learning the model. Before moving on to fully connected layers for classification, a sequential model was employed for training, in which layers were stacked linearly from convolutional layers. Although this method was somewhat successful, it had trouble picking up complicated feature relationships, particularly when there were minute changes in facial expressions that suggested tiredness. The sequential model was unable to create long-range dependencies over the whole picture and instead depended on local spatial information that the CNN had collected. The CNN-Transformer architecture, on the other hand, produced better outcomes. The Transformer, which employs self-attention to focus on important facial characteristics across many spatial locations, was added to solve the sequential model's inability to adequately represent contextual linkages. This improved the model's ability to discern subtle variations in yawning, eye opening, and other facial indicators. The Transformer's integration gives the model an even more sophisticated feature representation, which enhances its resilience, accuracy, and precision in identifying sleepiness states.

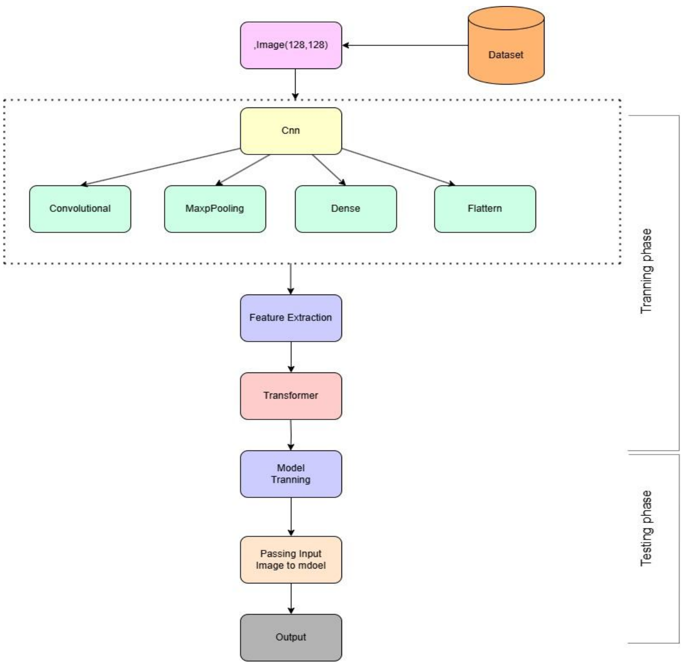

Fig. 1: Proposed approach architecture diagram.

The suggested driver drowsiness detection system's architectural process is depicted in the figure, which combines a Transformer model for classification with Convolutional Neural Networks (CNNs) for feature extraction. Convolutional layers, which extract local spatial characteristics, are followed by max pooling layers, that reduce dimensionality, along with dense layers, that refine features. fter being flattened, the recovered feature maps are fed into the Transformer model, which improves feature representations by capturing long-range relationships through self-attention processes. In order to ensure that drowsiness-related patterns are effectively learned, the model is optimized during the training phase using labeled data. The exact same preprocessing procedures are used to an input image during the testing phase, after which it is fed into the trained model to provide a product prediction that indicates whether the driver is awake or sleepy. By combining the context-aware learning capabilities of Transformers with the local feature extraction strength of CNNs, this hybrid technique improves the accuracy and durability of sleepiness detection.

B. Dataset

There are 2920 photos within the collection, which are separated under four categories: Yawn, Non-Yawn, Closed, and Open. These images are divided into the following:

- Training set: 2487 images
- Testing set: 433

Fig. 2: Dataset sample images.

Preprocessing techniques, including scaling, normalization, and augmentation, are used for every image in order to improve the resilience of the model.

C. Feature Extraction

Our driver drowsiness detection model depends largely on highlight extraction that determines the machines abilities to perceive examples the better in put pictures. In this approach, the processed data directly feeds into a Transformer model, with a Convolutional Neural Network (CNN) being the primary feature extractor. The CNN aims to encode the spatial hierarchy in the image by learning relevant visual patterns from the 2D input space, such as opening eyes, yawning, and face structures.

D. Transformer Model

The Transformer model of our driver drowsiness detection system links and connects to various features being extracted, and hence, it is what we refer to as the most indispensable aspect of the detection. For example, as the transformer has a self-attention mechanism that allows it to process the entire input in parallel (unlike common sequential models that need to move through the input step by step) it appears to be magnificently well-suited for recognizing global patterns. This feature greatly increases classification accuracy by allowing the model to efficiently interpret contextual and spatial information from CNN-extracted data. The scaled dot-product attention, which establishes how much emphasis each element in the input sequence should have on the others, is the central component of the Transformer model. Self-attention is calculated using the query (Q), key (K), and value (V) matrices as inputs:

Attention Formula

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where d_k is the dimensionality of the key vectors. The scaling factor $\sqrt{d_k}$ prevents large dot product values from producing extreme softmax outputs, ensuring stable gradients.

The Multi-Head Self-Attention mechanism, with four attention heads, extends this by computing multiple attention outputs in parallel, concatenating them, and projecting them using a learned weight matrix W^O:

Multi-Head Formula

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^O$$

Each attention head learns different feature dependencies, enhancing the model's ability to distinguish between drowsy and alert states by capturing fine-grained variations in facial features.

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

To further refine the representations, the Transformer incorporates a Feedforward Network (FFN), consisting of two dense layers with ReLU activation:

FFN formula

$$FFN(X) = (0, XW_1 + b_1)\, W_2 + b_2$$

Where W_1, W_2 are weight matrices and b1, b2 are biases are biases. Residual connections are applied to stabilize feature transformations, ensuring gradient flow remains effective:

Xoutput formula

$$X_{output} = LayerNorm\left(X + FFN(X)\right)$$

The final feature representation is obtained through Global Average Pooling, defined as:

Xpooled formula

$$X_{pooled} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

This compresses the extracted features into a compact vector, before classification.

E. Performance Matrix

Once the model has been trained, we evaluated its performance using various metrics to ensure success in detecting drowsy states. Here are the better performance of the CNN Transformer for the CNN based methods.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

III. AUTHORS & AFFILIATION

Jabbar, Rateb, et al. proposed a driver sleepiness detection system based on CNN model using facial features. [1], increasing accuracy but reducing processing needs. NTHU Driver Fatigue Detection Dataset to classify face characteristics with CNN to detect the sleepiness. With an average accuracy of 83.33%. It is the simplest model that runs more efficiently with less cpu and memory footprint than the conventional MLP and Faster-RCNN models.

A Hybrid Driver Drowsiness Detection Using Convolutional Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Rajamohana, S. P., et al.'s presentation was that of Memory (BiLSTM). [2]. Combining CNN for feature extraction and BiLSTM for sequential This method enhances accuracy through eye-tracking analysis. To identify by monitoring a real-time video stream for drowsiness. make face features, track eye features using Euclidean method, and classify states of the eye. First, the classification accuracy without applying BiLSTM is only 85%, while adding BiLSTM achieves 94%.

A driver drowsiness detection system whereby a CNN trained on EEG inputs was utilized was suggested by Chaabene, Siwar, et al. [3]. However, EEG-based recognition which investigates the activity of the brain instantaneously is a better indicator of tiredness compared to visual recognition techniques. Using an Emotiv EPOC+ headset to collect data, the study looks at the effects of alpha and theta brain waves, which are strong indicators of drowsiness. In the proposed CNN model, EEG signals are processed through various convolutional layers. By determining the optimal number of electrodes, which reduces the number of electrodes to seven, this model provides high efficiency while maintaining high accuracy (90.14%). The comparison with existing designs of CNN such as ResNet and WaveNet proves the efficiency of the proposed method, as it achieves higher accuracy.

Kolli Venkata Krishna Kishore, Uyyala Srinivasulu Reddy, Venkata Rami Reddy, and Chirra. [4] proposed a Stacked Deep Convolutional Neural Network (CNN) based on eye state data to detect driver sleepiness. The work employs a psychological method whereby a Deep CNN processes the eye area after the Viola-Jones algorithm recognizes its shape and extracts it to perform categorization. The proposed model consists of four convolutional layers and a fully connected layer, with a SoftMax classifier distinguishing between drowsy and non-drowsy states. The dataset comprises 2,850 images, with the model achieving 96.42% accuracy, outperforming traditional CNN-based methods.

Salem, Dina, and Mohamed Waleed. [5] explores various driver drowsiness detection methods, emphasizing physiological, behavioral, and vehicle-based approaches. Deep learning models like CNNs and Transfer Learning (TL) using InceptionV3 and MobileNetV2 have significantly improved accuracy and robustness. This work builds on existing research by integrating Haar Cascade for feature extraction and Transfer Learning for improved accuracy, achieving up to 99.94% accuracy in real-time applications.

Singh, Dipender, and Avtar Singh. [6] proposed study on driver drowsiness detection have utilized various approaches, including physiological signals (EEG, EOG), behavioral analysis (eye blink rate, yawning, head movements), and vehicle-based measures (steering and braking patterns). CNN-based methods improved accuracy but faced challenges with lighting variations, head angles, and accessories like glasses. This work enhances detection by focusing on eye and mouth regions using CNN with ReLU activation, achieving 94.95% accuracy.

Florez, Ruben, et al. [7] proposed highlights various methods for driver drowsiness detection, including eye-tracking, physiological signals (EEG, EOG), and vehicle-based measures like steering patterns. Convolutional Neural Networks (CNNs) have significantly improved accuracy by focusing on eye region analysis, often incorporating techniques like MediaPipe for feature extraction. Previous works have explored different CNN architectures, including InceptionV3, VGG16, and ResNet50V2, with ResNet50V2 consistently achieving the highest accuracy (99.71%).

IV. RESULTS AND DISCUSSIONS

Experimental results indicate that combining Transformer and CNN significantly improves the accuracy and overall performance of driver sleepiness detection. For the CNN + Transformer method, both accuracy and metrics calculated with predicted and true values in one step show a high improvement over the sequential CNN model (for accuracy, precision, recall and F1). The addition of the Transformers long-distance capture properties with CNN s spatial feature extraction capabilities is increasing the strength of the model.

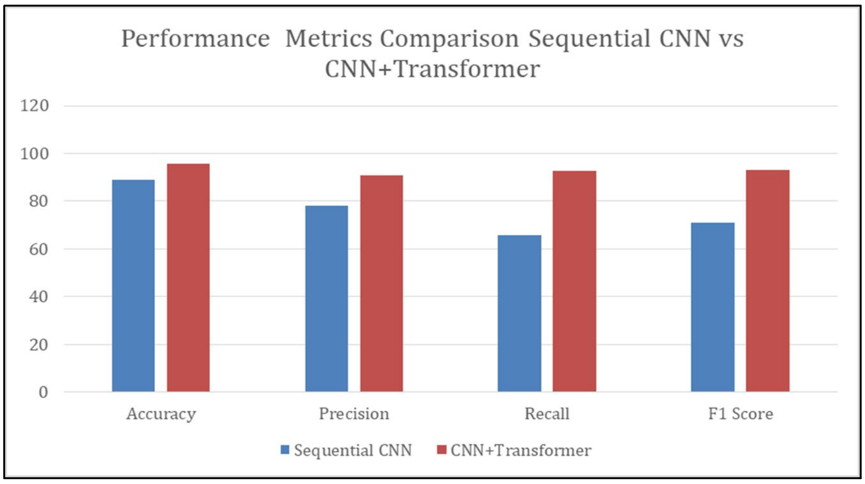

Fig. 3: Performance Metrics of Sequential CNN VS CNN +Transformer

In order to extract spatial features and categorize sleepiness stages, the model was first trained using a Sequential CNN architecture. Recall was 65.9%, accuracy was 88.92%, precision was 78%, and the CNN model's F1-score was 71.2%. While the CNN by itself was successful in identifying patterns of drowsiness, it was not always able to capture complicated feature interactions and long-range dependencies, which resulted in misclassifications.
We combined a Transformer model with the CNN to boost performance, utilizing the self-attention mechanism to improve classification accuracy and feature representation. The suggested CNN-Transformer hybrid model performed exceptionally well, attaining 95.79% accuracy, 91% precision, 92.8% recall, and 93.1% F1-score. The Transformer decreased false positives and false negatives by successfully capturing temporal relationships and minute changes in facial features. According to the data, the CNN-Transformer model performs noticeably better than the CNN alone, giving it a more reliable method of detecting driver drowsiness.

Fig.4: Accuracy and loss of training & testing

The above graphs illustrate the training and testing performance of the CNN + Transformer model for driver drowsiness detection. The left graph represents the accuracy trends over 30 epochs, where both training and testing accuracy steadily improve, indicating that the model is effectively learning and generalizing well to unseen data. Notably, the testing accuracy follows a trajectory similar to the training accuracy, suggesting minimal overfitting. The right graph displays the training and testing loss, where both losses show a downward trend, confirming proper convergence of the model. The testing loss remains slightly higher than the training loss, which is expected in real-world scenarios. The overall results demonstrate that integrating a Transformer with CNN enhances feature extraction, leading to superior accuracy, precision, recall, and F1-score compared to a sequential CNN model.
Experimental results for the CNN + Transformer method, both accuracy and metrics calculated with predicted and true values in one step show a high improvement over the sequential CNN model (for accuracy, precision, recall and F1). The addition of the Transformers long-distance capture properties with CNN s spatial feature extraction capabilities is increasing the strength of the model.
For future study, the model can be further advanced by integrating attention mechanisms, specifically for eye movement monitoring or facial expression analysis. Also, deploying the model on edge devices, such as in-vehicle AI systems, provides emergency real-time drowsiness detection in order to make the roads safer. Exploring larger and more diverse datasets can also improve the model's generalization across different demographic groups and lighting conditions.

V. CONCLUSION

This work presents a new approach for the driver drowsiness detection, by employing a hybrid deep learning architecture, combining the well-known Convolutional Neural Networks (CNN) with a Transformer model. While a CNN alone can be effective at extracting spatial features from image data, the study found that adding a Transformer improves the model’s ability to capture long-range dependencies, which enhances classification performance. The experimental results show that the proposed CNN + Transformer method outperforms conventional sequential CNN models with respect to accuracy, precision, recall, and F1-score, making it a more robust approach for real-time drowsiness detection.
Future work can be done to improve the model architecture for better real-time performance on edge devices and reduced computation complexity. Further investigations can explore the potential of combining additional modalities such as physiological signals or eye-tracking signals to enhance prediction accuracy. In addition, by increasing the variety of the dataset, including features that describe different environmental conditions and

driver behaviours, we will gain a model that generalizes better to different scenarios. Implementing this technology in in-car AI safety systems using the proposed technique has the chance to significantly cut down on fatal traffic collisions caused by driver fatigue.

References

[1] Jabbar, Rateb, et al. "Driver drowsiness detection model using convolutional neural networks techniques for android application." 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT). IEEE, 2020.

[2] Rajamohana, S. P., et al. “Driver drowsiness detection system using hybrid approach of convolutional neural network and bidirectional long short-term memory (CNN_BILSTM).” Materials Today: Proceedings 45 (2021): 2897-2901.

[3] Chaabene, Siwar, et al. "Convolutional neural network for drowsiness detection using EEG Signals” Sensors 21.5 (2021): 1734.

[4] Chirra, Venkata Rami Reddy, Srinivasulu Reddy Uyyala, and Venkata Krishna Kishore Kolli. “Deep CNN: A Machine Learning Approach for Driver Drowsiness Detection Based on Eye State.” Rev.d‘Intelligence Artif. 33.6 (2019): 461-466.

[5] Salem, Dina, and Mohamed Waleed. “Drowsiness detection in real-time via convolutional neural networks and transfer learning.”Journal of Engineering and Applied Science 71.1 (2024): 122.

[6] Singh, Dipender, and Avtar Singh. “Enhanced Driver Drowsiness Detection using Deep Learning.” ITM Web of Conferences. Vol. 54. EDP Sciences, 2023.

[7] Florez, Ruben, et al. “A cnn-based approach for driver drowsiness detection by real-time eye state identification.” Applied Sciences 13.13 (2023): 7849.

[8] Kshatri, Sapna Singh, and Yogesh Kumar Rathore. “Real-Time Driver Distraction Detection Using Fast R-CNN Algorithm.” (2025): 1-6.

[9] Pushkar, Piyush, et al. “Real Time Drowsiness Detection System Using CNN.” International Journal for Research 11 (2023): 1487-1490.

[10] Chand, H. Varun, and J. Karthikeyan. “CNN Based Driver Drowsiness Detection System Using Emotion Analysis.” Intelligent Automation & Soft Computing 31.2 (2022).

[11] Albadawi, Yaman, Maen Takruri, and Mohammed Awad. "A review of recent developments in driver drowsiness detection systems." Sensors 22.5 (2022): 2069.

[12] Singh, Jagendra. "Learning based driver drowsiness detection model." 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2020.

[13] Magán, Elena, et al. "Driver drowsiness detection by applying deep learning techniques to sequences of images." Applied Sciences 12.3 (2022): 1145.

[14] Siddiqui, Hafeez Ur Rehman, et al. "Non-invasive driver drowsiness detection system." Sensors 21.14 (2021): 4833.

[15] Fu, Biying, et al. "A survey on drowsiness detection–modern applications and methods." IEEE Transactions on Intelligent Vehicles (2024).

[16] Satish, K., et al. "Driver drowsiness detection." 2020 international conference on communication and signal processing (ICCSP). IEEE, 2020.

[17] Arakawa, Toshiya. "Trends and future prospects of the drowsiness detection and estimation technology." Sensors 21.23 (2021): 7921.

[18] Ahmed, Mohammed Imran Basheer, et al. "A deep-learning approach to driver drowsiness detection." Safety 9.3 (2023): 65.

[19] Rajkar, Ajinkya, Nilima Kulkarni, and Aniket Raut. "Driver drowsiness detection using deep learning." Applied Information Processing Systems: Proceedings of ICCET 2021. Springer Singapore, 2022.

[20] Saleem, Adil Ali, et al. "A systematic review of physiological signals based driver drowsiness detection systems." Cognitive neurodynamics 17.5 (2023): 1229-1259.

[21] Jasim, Sarah Saadoon, and Alia Karim Abdul Hassan. "Modern drowsiness detection techniques: A review." International Journal of Electrical and Computer Engineering 12.3 (2022): 2986-2995.

[22] Jahan, Israt, et al. "4D: a real-time driver drowsiness detector using deep learning." Electronics 12.1 (2023): 235.