ISSN: 2349-6002

Driver Drowsiness Detection using Hybrid VGG19 and Vision Transformer

Mrs. J. Veerendeswari[1], Mrs. B. Celin Julie[2], Mr. Kishore Kumar K[3], Mr. Praveen R[4], Mr. Dhivesh V[5]
[1,2]*Head of the Department, Assistant Professor, Information Technology, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[3,4,5]*UG, Information Technology, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
doi.org/10.64643/IJIRTV12I11-201032-459

***Abstract*—Driver drowsiness is a major contributor to road accidents, reducing attention, reaction time, and decision-making ability. Conventional detection methods based on Convolutional Neural Networks (CNNs) mainly rely on eye-closure analysis and often suffer from limited generalization and overfitting under diverse real-world conditions. This paper proposes a hybrid deep learning framework combining VGG19 and Vision Transformer (ViT) for robust multi-feature drowsiness detection. The VGG19 network extracts detailed local spatial features such as eyelid and mouth movements, while the ViT captures global contextual relationships using self-attention mechanisms. The integration of these complementary representations enables simultaneous analysis of eye closure and yawning, improving detection reliability. The proposed model enhances robustness against illumination variation, facial orientation changes, and environmental noise. Experimental evaluation demonstrates improved accuracy, recall, and generalization compared with conventional CNN approaches. The system operates in real time and provides timely alerts, offering a scalable and practical solution for intelligent driver monitoring systems and improved road safety.**

***Index Terms*—Driver drowsiness detection, Vision Transformer, VGG19, fatigue monitoring**

INTRODUCTION

Driver drowsiness refers to a physiological and cognitive state in which a driver experiences reduced alertness, slower reaction time, and impaired decision-making due to fatigue or lack of sleep. It is considered one of the major causes of road accidents worldwide because it directly affects the driver's ability to maintain lane discipline, respond to sudden obstacles, and make timely judgments. Long driving hours, irregular sleep patterns, night travel, and monotonous highways significantly increase the likelihood of fatigue. Unlike other traffic violations, drowsiness is difficult to detect because drivers are often unaware of their declining attention until performance is already compromised. Common behavioral signs include frequent blinking, prolonged eye closure, yawning, head nodding, and drifting from the lane. Traditional approaches to detecting drowsiness relied on physiological sensors such as electroencephalogram (EEG) or heart-rate monitoring devices. Although accurate, these methods are intrusive, uncomfortable, and impractical for everyday use. As a result, computer vision–based techniques have gained popularity due to their non-contact nature and ease of deployment in vehicles. Modern systems use cameras to monitor facial features and analyze fatigue indicators in real time. However, relying on a single indicator, such as eye closure alone, may lead to inaccurate predictions under varying lighting conditions, facial orientations, or driver differences. Therefore, intelligent driver monitoring systems aim to analyze multiple visual cues simultaneously to improve reliability and early detection. By continuously assessing driver alertness and providing timely warnings, drowsiness detection technology plays a crucial role in reducing fatigue-related accidents and enhancing overall road safety in modern transportation environments.

a. Vision Transformer (ViT)

The Vision Transformer (ViT) is a deep learning architecture developed for image analysis by adapting the transformer model originally used in natural language processing. Unlike traditional Convolutional

ISSN: 2349-6002

Neural Networks (CNNs) that process images through convolution filters, ViT interprets an image as a sequence of small fixed-size patches, similar to words in a sentence. Each image patch is flattened and converted into an embedding vector, and positional embeddings are added to preserve spatial order. These embeddings are then processed by a transformer encoder composed of multi-head self-attention layers and feed-forward networks. The self-attention mechanism allows the model to understand relationships between all regions of an image simultaneously, enabling it to capture global context rather than only local patterns. This is particularly useful in applications such as facial behavior analysis, where connections between distant features like eyes and mouth are important. Multi-head attention helps the model focus on multiple visual patterns at once, improving feature representation and robustness. ViT is also highly scalable and can achieve superior performance when trained on large datasets, as it avoids some limitations of CNNs related to locality bias.

b. Architecture of the Vision Transformer (Vit)

The architecture of the Vision Transformer (ViT) consists of three main stages: patch embedding, transformer encoder, and classification head. First, the input image is divided into fixed-size non-overlapping patches (for example, 16×16 pixels). Each patch is flattened into a vector and passed through a linear projection layer to produce a patch embedding. Since transformers do not inherently understand spatial order, positional embeddings are added to each patch representation so the model retains information about the original arrangement of patches within the image. A special learnable token called the class token is also appended to the sequence, which will later represent the entire image.

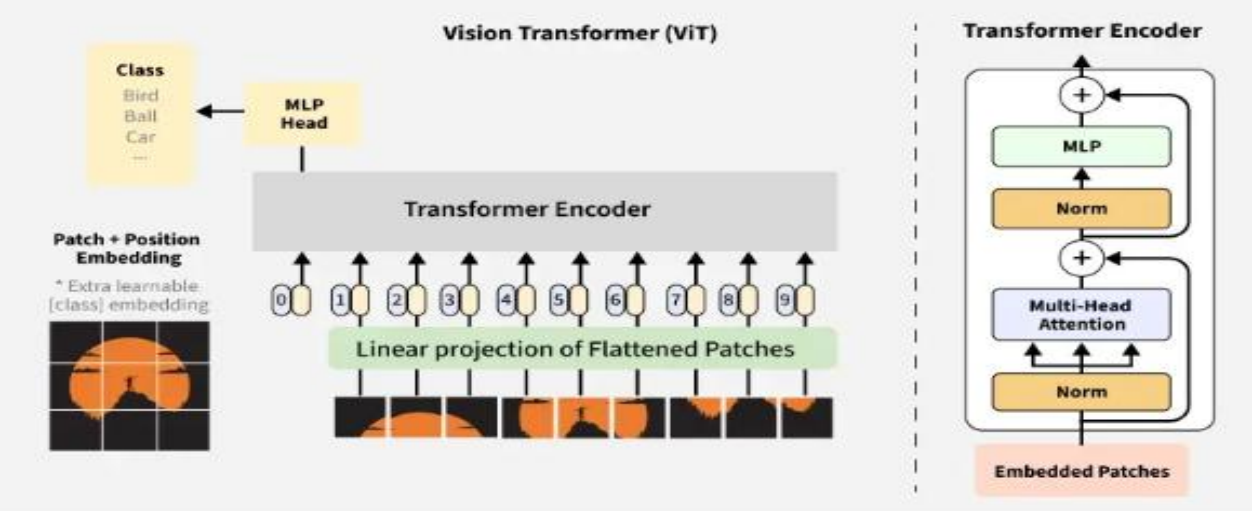

Fig 1.1 Architecture of the Vision Transformer (ViT)

The sequence of embedded patches is then fed into the transformer encoder, which is composed of stacked layers containing Multi-Head Self-Attention (MHSA) and Feed-Forward Neural Networks (FFN). The self-attention mechanism computes relationships between all patches simultaneously, enabling the model to learn global contextual dependencies rather than only local spatial features. Each encoder block also includes residual connections and layer normalization to stabilize training and improve convergence. Finally, the output corresponding to the class token is passed to a multilayer perceptron (MLP) classification head that predicts the final class label. This architecture allows ViT to effectively capture both structural and contextual information from images.

c. Visual Geometry Group (VGG)

VGG19 is a deep convolutional neural network architecture introduced by the Visual Geometry Group (VGG) at the University of Oxford for image recognition tasks. It consists of 19 learnable layers, including 16 convolutional layers and 3 fully connected layers. The model uses small 3×3 convolution filters stacked sequentially, which helps

ISSN: 2349-6002

capture fine spatial details while maintaining computational efficiency. After several convolution operations, max-pooling layers reduce spatial dimensions and retain the most important features. This simple and uniform architecture allows the network to learn hierarchical representations, starting from basic edges and textures to complex shapes and facial structures. One of the main advantages of VGG19 is its strong feature extraction capability, making it suitable for transfer learning applications. Pretrained weights from large datasets such as ImageNet enable the model to generalize well even with limited training data. Due to its ability to extract detailed local patterns, VGG19 is commonly used in tasks like facial analysis, object detection, and driver behavior monitoring systems.

d. Architecture Of Vgg19

The architecture of VGG19 is a deep convolutional neural network composed of 19 weight layers organized into sequential blocks. It contains 16 convolutional layers followed by 3 fully connected layers for classification. The network accepts a fixed-size input image, typically 224 × 224 pixels, and processes it through multiple convolution blocks. Each block consists of several 3 × 3 convolution filters with stride 1 and padding that preserves spatial dimensions. After each block, a max-pooling layer with 2 × 2 filters reduces the feature map size, helping the model capture dominant patterns while lowering computational complexity.

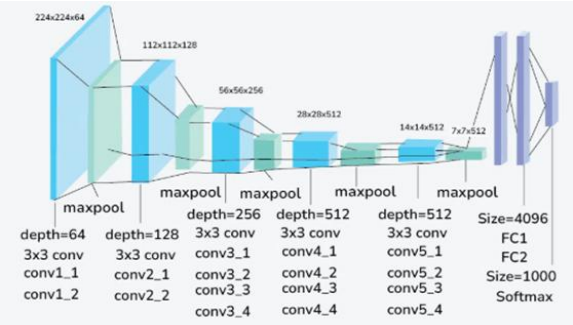

Fig 1.2 Architecture of the Vgg19

As the depth increases, the number of filters also increases from 64 to 512, enabling the network to learn hierarchical representations ranging from edges and textures to complex shapes and facial structures. After the convolution stages, the extracted feature maps are flattened and passed through fully connected layers, followed by a Softmax classifier for final prediction. This structured design ensures strong local feature extraction and stable training performance.

II. LITERATURE SURVEY

[1] Madduri Venkateswarlu, Venkata Rami Reddy Chirra [1]. Driver drowsiness is a major cause of road accidents because fatigue reduces alertness, reaction time, and decision-making ability. Traditional monitoring approaches often fail to accurately distinguish subtle facial cues such as eye closure, yawning, and open-eye states, which limits their effectiveness in real-world driving environments. To address this problem, the study proposes a driver drowsiness detection system based on deep Convolutional Neural Networks including DenseNet121, VGG16, VGG19, and ResNet50. These models automatically learn detailed facial patterns related to fatigue without manual feature design. Input images of drivers are processed through the networks, which classify the driver's condition into four categories: closed eyes, open eyes, yawning, and no-yawn. The system was evaluated on two datasets and achieved high accuracy, reaching 94.76% with ResNet50 on the first dataset and 96.21% with VGG19 on the second dataset. Performance metrics such as precision, recall, and F1-score confirm the model's reliability, making it suitable for real-time driver monitoring applications. [2]

Ishwari Singh Rajput, Sonam Tyagi, Vandana Pandey, Abha Upreti [2] Distracted and drowsy driving is a major contributor to road accidents, as fatigue reduces awareness and delays driver response. Detecting drowsiness in real time is difficult due to variations in lighting, camera quality, and limited computational resources in vehicles. To address this problem, the study proposes a machine learning–based driver drowsiness detection system using transfer learning with pre-trained convolutional neural networks. A labeled dataset containing images of open and closed eyes is used to fine-tune the model so it can accurately recognize eye states. The system continuously analyzes live video captured from an in-vehicle camera and monitors whether the driver's eyes remain closed for a specific duration. If prolonged eye closure is detected, an alert is immediately triggered to warn the driver and prevent accidents. By leveraging transfer learning, the model reduces training time

while improving performance, achieving an accuracy of 92.5%. The approach provides a practical and scalable real-time solution for enhancing driver safety. [3].

Yi Xuan Chew, Siti Fatimah Abdul Razak, Sumendra Yogarayan [3] The rapid growth in vehicle usage has increased road accidents, with driver drowsiness being a major contributing factor. Conventional video-based monitoring systems often suffer from reduced accuracy due to lighting variations, camera positioning, and differences in driver appearance. In addition, relying on a single detection method limits reliability in real-world conditions. To overcome these issues, the study proposes a multimodal driver drowsiness detection system that combines facial analysis and heart rate monitoring. Deep learning CNN models such as ResNet, DenseNet, and a custom CNN analyze facial expressions, while a Logitech BRIO 4K Ultra HD Pro webcam measures heart rate in a non-contact manner. Laboratory testing confirmed reliable detection across multiple camera angles, and real-vehicle experiments with participants showed a clear relationship between reduced heart rate and drowsiness. The system triggers a real-time alert when fatigue is detected, providing a more accurate and dependable solution for improving road safety. [4]

Md. Ebrahim Shaik [4] Driver drowsiness is a significant contributor to road accidents, resulting in fatalities, serious injuries, financial losses, and property damage. It can be caused by factors such as fatigue, long driving hours, medication effects, and sleep disorders. Although many detection methods have been developed, maintaining reliability and accuracy across different driving conditions remains a challenge. This paper addresses the issue by reviewing existing driver drowsiness detection approaches, including physiological measures (such as brain and heart activity), vehicle-based indicators (lane deviation and steering patterns), subjective evaluations, and behavioral observations (eye closure and blinking patterns). The study analyzes the advantages and limitations of each technique and highlights the need for integrating advanced technologies to improve detection performance. By comparing these methods, the research identifies gaps in current systems and suggests directions for future improvements. [5].

Chris Schwarz, John Gaspar, Thomas Miller, Reza Yousefian [5] Driver drowsiness is a major contributor to road accidents and fatalities, as reduced alertness delay's reaction time and impairs decision making. Conventional detection methods rely either on vehicle-based indicators, such as lane departures and steering behavior, or driver monitoring systems that analyze eye blinks and gaze patterns. However, using only one type of signal often fails to provide reliable early detection or accurately determine the severity of fatigue. To address this issue, the study conducted high-fidelity driving simulator experiments and evaluated models based on vehicle signals, driver monitoring signals, and a combination of both. The results showed that driver monitoring data performed better than vehicle-only detection, but the integrated approach produced the highest accuracy, achieving an AUC of 0.897 for binary classification. By combining driver behavior and vehicle dynamics, the system detects drowsiness earlier and more reliably, providing a strong foundation for real-time driver impairment monitoring and improved road safety. [6].

Abhineet Ranjan, Sanjeev Sharma, Prajwal Mate & Anshul Verma [6] Driver drowsiness is a significant cause of road accidents, resulting in fatalities and property damage. Traditional monitoring techniques often struggle to detect fatigue reliably in real time, especially under varying driving conditions. Therefore, an accurate and efficient image-based system is required to automatically classify whether a driver is drowsy or alert. This study addresses the problem using deep learning with transfer learning applied to the Driver Drowsiness Dataset (DDD). Six pre-trained convolutional neural network models DenseNet169, MobileNetV2, ResNet50V2, VGG19, InceptionV3, and Xception were trained and compared to determine the most effective approach. The results showed that ResNet50V2 achieved the highest performance, reaching 100% accuracy and outperforming the other models and existing techniques. Evaluation metrics such as accuracy and F1-score confirmed the robustness and reliability of the system. The findings demonstrate that transfer learning with advanced CNN architectures can effectively detect driver drowsiness from images, making it suitable for real-time driver monitoring applications.

ISSN: 2349-6002

Sl. No	Title	Author	Year	Description	Advantage	Disadvantage
1	CNN: A multi-feature learning-based approach for driver drowsiness detection	Madduri Venkateswarlu, Venkata Rami Reddy Chirra	2025	Hybrid system combining CNN models and global feature modeling to classify driver states (closed, open, no-yawn, yawn).	Very high accuracy and robust multi-feature detection.	Computationally complex and requires large datasets.
2	Enhancing Driver Safety with MobileNetV2-Based Transfer Learning for Drowsiness Detection	Ishwari Singh Rajput, Sonam Tyagi, Vandana Pandey, Abha Upreti	2025	Transfer learning using MobileNetV2 to detect eye closure in real-time and trigger alerts.	Fast training and efficient real-time performance.	Limited to eye-based detection only.
3	Dual-Modal Drowsiness Detection to Enhance Driver Safety	Yi Xuan Chew, Siti Fatimah Abdul Razak, Sumendra Yogarayan	2025	Combines facial recognition and heart rate monitoring for fatigue detection.	Higher reliability due to multimodal sensing.	Requires additional hardware and sensors.
4	A systematic review on detection and prediction of driver drowsiness	Md. Ebrahim Shaik	2025	Comprehensive review of physiological, vehicle-based, subjective, and behavioral detection methods.	Provides strong research baseline and comparison.	Does not propose a practical implementation model.
5	The detection of drowsiness using a driver monitoring system	Chris Schwarz, John Gaspar, Thomas Miller, Reza Yousefian	2025	Combines vehicle signals and driver monitoring data using simulator experiments.	Early and more accurate detection using combined signals.	Needs multiple data sources and complex integration.
6	An Efficient Deep Learning Technique for Driver Drowsiness Detection	Abhineet Ranjan, Sanjeev Sharma, Prajwal Mate, Anshul Verma	2025	Transfer learning comparison of several CNN models on driver drowsiness dataset.	Achieves very high accuracy (100% with ResNet50V2).	May overfit and performance depends on dataset quality.

III. PROPOSED SYSTEM

The proposed driver drowsiness detection system is based on a hybrid deep learning architecture that combines VGG19 and Vision Transformer (ViT) to achieve accurate and reliable fatigue detection. VGG19 acts as a powerful feature extractor that captures fine-grained local facial details such as eye closure and mouth movements, while the Vision Transformer analyzes global relationships between facial regions using self-attention mechanisms. By integrating local spatial features with global contextual understanding, the system can identify multiple signs of drowsiness more effectively than single-model approaches. The system continuously monitors the driver's face through a camera and processes video frames in real time. When symptoms of fatigue are detected, an immediate audible alert is triggered to regain the driver's attention and prevent accidents. This hybrid approach improves accuracy, generalization, and robustness under varying lighting conditions, facial orientations, and driver differences, making it suitable for practical real-world driver monitoring and road safety applications.

ISSN: 2349-6002

a. Architecture of the Proposed:

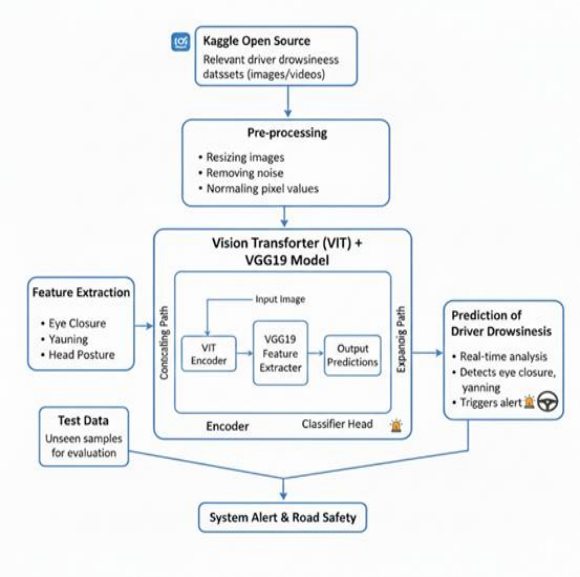

Fig 4.1 Architecture OF THE PROPOSED SYSTEM

The architecture of the proposed driver drowsiness detection system follows a structured workflow integrating data collection, preprocessing, feature extraction, model creation, and real-time prediction using a hybrid Vision Transformer (ViT) and VGG19 framework. The process begins with data collection from Kaggle, where open-source driver drowsiness datasets containing images and videos of drivers in both alert and drowsy states are obtained. These datasets form the foundation for training and testing the model. The preprocessing stage involves cleaning the data by resizing images, removing noise, and normalizing pixel values to ensure uniformity and enhance model efficiency. During feature extraction, essential facial cues such as eye closure and yawning are identified, as these are reliable indicators of fatigue. The VGG19 network is employed to extract deep local spatial features including eyelid movement and mouth opening patterns, while the Vision Transformer analyzes global relationships among facial regions using self-attention mechanisms. This combination enables the model to capture both fine details and contextual dependencies across the face. A portion of the dataset is reserved for testing to validate the model's performance. Finally, in the prediction phase, the trained hybrid model detects drowsiness in real time, and when fatigue signs appear, a beep alert is triggered to wake the driver and prevent potential accidents. This architecture improves accuracy, generalization, and robustness compared to single-model approaches.

b. Data collection:

The data collection module is the foundational stage of the proposed system, responsible for gathering relevant datasets from Kaggle's open-source repositories. These datasets typically consist of images and video sequences of drivers in both alert and drowsy states, captured under varying lighting conditions, facial orientations, and environmental settings. The data includes essential visual cues such as open or closed eyes, yawning expressions, and head movements, which serve as key indicators of fatigue. By sourcing from publicly available, well-annotated datasets, this module ensures diversity and quality in training samples, which are crucial for improving model robustness and accuracy. The collected data is organized into structured categories for training, validation, and testing purposes, enabling systematic model evaluation. This module ensures that the proposed hybrid VGG19 + Vision Transformer model has access to rich and varied input samples, allowing it to generalize effectively to real-world driving scenarios. The data collection process also emphasizes ethical considerations by using only open-source, non-personally identifiable datasets. Ultimately, this module establishes a strong foundation for the system's deep learning pipeline, ensuring that subsequent stages such as preprocessing, feature extraction, and prediction operate on high-quality, representative data necessary for reliable drowsiness detection.

c. Pre-processing:

The preprocessing module plays a critical role in preparing raw data for model training by ensuring consistency, quality, and computational efficiency. Since datasets often contain variations in image size, lighting, and noise, preprocessing standardizes the inputs to optimize model performance. The collected images are resized to a fixed resolution compatible with the hybrid VGG19 + ViT model, ensuring uniformity across all samples. Noise reduction techniques, such as Gaussian filtering, are applied to remove unwanted distortions while preserving important facial features. Additionally, normalization of pixel values is performed to scale the data within a consistent range (usually 0–1), which enhances

convergence during model training and prevents numerical instability. This module also includes operations like cropping or region-of-interest extraction to focus on relevant facial areas such as the eyes and mouth, eliminating irrelevant background information. Data augmentation techniques such as rotation, flipping, and brightness adjustment are employed to increase dataset diversity and reduce overfitting. Overall, preprocessing ensures that the input data is clean, consistent, and representative, allowing the hybrid model to learn meaningful patterns effectively.

d. Feature Extraction:

The feature extraction module identifies and isolates key visual characteristics that indicate driver drowsiness. After preprocessing, the system extracts significant features such as eye closure duration, blinking frequency, and yawning occurrences. The VGG19 network acts as a deep convolutional feature extractor that captures detailed spatial patterns such as eyelid position and mouth opening. The Vision Transformer then analyzes global relationships among facial regions using self-attention, enabling the model to understand interactions between facial features rather than analyzing them independently. By combining local spatial features from VGG19 with contextual understanding from ViT, the system reliably detects fatigue-related patterns. Extracted features are represented in a structured form suitable for classification, allowing the model to associate visual cues with alert or drowsy states. This automated approach improves detection accuracy and efficiency compared to traditional handcrafted methods.

e. Model creation:

In this module, a hybrid VGG19 + Vision Transformer architecture is developed as the core prediction model. VGG19 extracts hierarchical spatial features through convolutional layers, capturing fine-grained facial details. These features are then provided to the Vision Transformer, which models long-range dependencies using multi-head self-attention to understand overall facial behavior. The combined architecture enables both detailed feature representation and global contextual reasoning. The model is trained using labeled datasets corresponding to alert and drowsy states. Through iterative training and optimization, the hybrid model learns visual patterns associated with fatigue. This module represents the intelligence of the proposed system, enabling accurate identification of fatigue-related patterns while improving generalization across diverse drivers and environments.

f. Test data:

The test data module evaluates the performance, reliability, and generalization capability of the trained hybrid model. A portion of the dataset is reserved exclusively for testing and not used during training or validation to ensure unbiased evaluation. The test data includes diverse drivers, lighting conditions, facial orientations, and fatigue levels. Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix are computed to assess classification effectiveness. Testing helps identify false positives or missed detections and supports model fine-tuning. This module ensures the VGG19 + ViT system generalizes well beyond training data and performs consistently in real-world scenarios, validating readiness for deployment in driver monitoring systems.

g. Prediction:

The prediction module represents the operational phase where the trained hybrid model is deployed for real-time drowsiness detection. The system continuously captures live video frames of the driver's face through an in-vehicle camera. Each frame is preprocessed and analyzed by the VGG19 + Vision Transformer model to detect fatigue indicators such as prolonged eye closure or yawning. Based on detected features, the model classifies the driver's state as alert or drowsy. When drowsiness is identified, an audible beep alert is triggered to regain driver attention and prevent potential accidents. The real-time prediction process is optimized for speed and accuracy, ensuring minimal delay between detection and alert generation. The system may also log detection data for further analysis and safety evaluation.

IV. RESULT AND DISCUSSION

The experimental evaluation demonstrates that the hybrid VGG19 + Vision Transformer model provides reliable and consistent performance for driver drowsiness detection. The combination of VGG19's local feature extraction and ViT's global contextual

ISSN: 2349-6002

understanding significantly improves classification accuracy compared to single CNN-based approaches. The model successfully identifies fatigue indicators such as prolonged eye closure and related facial behavior under varying lighting conditions, head poses, and different driver appearances. Performance metrics including accuracy, precision, recall, and F1-score indicate strong generalization and low false detection rates. The real-time implementation also shows minimal delay between detection and alert generation, ensuring timely warnings to the driver. The results confirm that integrating spatial and contextual features enhances robustness and reduces misclassification caused by environmental variations. Overall, the proposed system demonstrates practical feasibility for real-world deployment, offering improved reliability and effectiveness in preventing fatigue-related road accidents.

1. Accuracy:

Accuracy is a fundamental performance metric used to evaluate how well a classification model predicts the correct output compared to the actual labels. In driver drowsiness detection, it measures how effectively the system distinguishes between states such as alert and drowsy from captured images or video frames. Accuracy is calculated as the ratio of correctly predicted observations to the total number of observations.

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

Where, TP (True Positive) represents correctly detected drowsy cases, TN (True Negative) represents correctly detected alert cases, FP (False Positive) indicates alert drivers incorrectly classified as drowsy, and FN (False Negative) indicates drowsy drivers classified as alert. A higher accuracy value indicates that the model produces fewer incorrect predictions and performs more reliably. However, accuracy alone may be misleading when datasets are imbalanced. For example, if most samples belong to the alert category, a model predicting “alert” frequently may still achieve high accuracy while failing to detect actual drowsiness. Therefore, accuracy should be interpreted along with precision and recall to ensure balanced performance. In real-time systems, improving accuracy reduces missed detections and false alarms, enhancing driver trust and overall road safety.

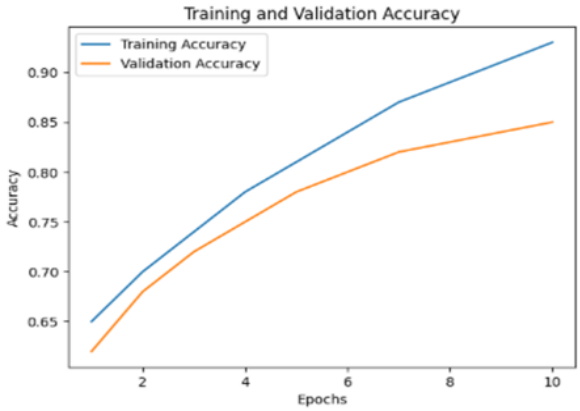

Fig 5.1 accuracy graph for proposed system

The accuracy graph shows the performance of the proposed driver drowsiness detection model during training and validation phases, reaching a final accuracy of 98.98%. As the number of epochs increases, the training accuracy steadily rises, indicating that the model effectively learns important facial patterns related to alert and drowsy states. At the same time, the validation accuracy follows a similar trend with minimal fluctuation, demonstrating good generalization and the absence of significant overfitting. The close alignment between the training and validation curves confirms that the hybrid VGG19 + Vision Transformer architecture can reliably recognize fatigue indicators across unseen data. Achieving 98.98% accuracy means the system correctly classifies nearly all driver states, with very few misclassifications. This high accuracy ensures dependable real-time monitoring and reduces the chances of missed detections or unnecessary alerts.

2. loss:

Loss is a measure of how far the model’s predicted output differs from the actual ground truth during training. In driver drowsiness detection, the loss function helps the model learn by penalizing incorrect predictions and guiding weight updates through backpropagation. A commonly used loss for classification problems is categorical cross-entropy loss, which evaluates the probability difference between predicted and true classes. It is defined as:

$$Loss = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

ISSN: 2349-6002

Where, y_i represents the true label and y^i represents the predicted probability for class iii. During training, the loss value gradually decreases as the model improves its predictions. A smooth and consistent reduction in loss indicates effective learning and convergence, while a large gap between training and validation loss may indicate overfitting. Lower loss values correspond to better prediction confidence and accuracy. In real-time driver monitoring, minimizing loss ensures reliable classification of alert and drowsy states, improving system stability and detection performance.

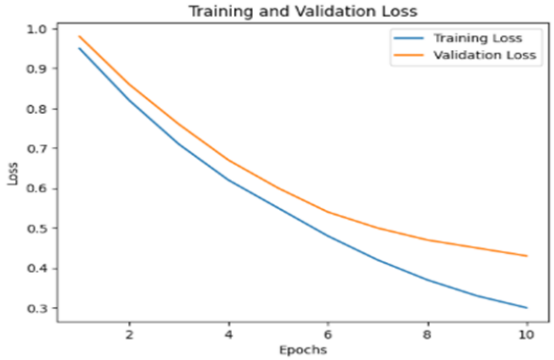

Fig 5.2 Loss graph for proposed system

The loss graph illustrates how the error of the proposed driver drowsiness detection model changes during training and validation over multiple epochs. At the beginning of training, the loss value is high because the model has not yet learned meaningful patterns from the input images. As training progresses, the training loss gradually decreases, indicating that the model is learning to correctly classify driver states by adjusting its internal parameters. The validation loss also follows a similar decreasing trend, which shows that the model generalizes well to unseen data. When both training and validation loss curves converge and remain close to each other, it suggests stable learning and minimal overfitting. Small fluctuations may occur due to data variations, but the overall downward trend confirms improved prediction confidence. A low final loss value indicates that the hybrid VGG19 + Vision Transformer model makes accurate predictions, ensuring reliable real-time drowsiness detection and consistent performance in practical driving conditions.

3. Precision

Precision is a performance metric that measures how accurately a classification model identifies positive cases among all predicted positive instances. In driver drowsiness detection, precision indicates how many drivers predicted as drowsy are actually drowsy. It is especially important in safety systems because false alarms (predicting drowsy when the driver is alert) can reduce user trust in the system. Precision is calculated using the formula:

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

Where, TP (True Positive) represents correctly detected drowsy drivers and FP (False Positive) represents alert drivers incorrectly classified as drowsy. A high precision value means the system generates fewer false warnings and provides more reliable alerts. However, precision alone does not measure missed detections, so it is often analyzed together with recall. In real-time driver monitoring applications, high precision ensures that warning alerts are meaningful and not triggered unnecessarily, improving user confidence and overall system effectiveness.

4. Recall

Recall is a performance metric that measures the ability of a classification model to correctly identify all actual positive cases. In driver drowsiness detection, recall indicates how effectively the system detects drivers who are truly drowsy. It focuses on minimizing missed detections, which are critical in safety applications because failing to recognize a drowsy driver may lead to accidents. Recall is calculated as the ratio of correctly predicted drowsy cases to the total actual drowsy cases:

$$Recall = \frac{TP}{TP + FN}$$

Where, TP (True Positive) represents correctly detected drowsy drivers and FN (False Negative) represents drowsy drivers incorrectly classified as alert. A high recall value means the system successfully identifies most fatigue cases and rarely overlooks dangerous situations. However, very high recall may sometimes increase false alarms, so it is usually evaluated together with precision. In real-time monitoring systems, strong recall ensures timely warnings and enhances overall road safety.

5. F1 score:

The F1-score is a performance metric that provides a balanced evaluation of a classification model by

combining both precision and recall into a single value. In driver drowsiness detection, precision measures how many predicted drowsy cases are actually drowsy, while recall measures how many real drowsy cases are successfully detected. Since safety systems must both avoid false alarms and prevent missed detections, considering only one metric is insufficient. The F1-score solves this problem by calculating the harmonic mean of precision and recall, giving equal importance to both.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A high F1-score indicates that the model maintains a good balance between correctly identifying drowsy drivers and minimizing incorrect warnings. Unlike accuracy, which can be misleading when datasets are imbalanced, the F1-score provides a more reliable measure of performance because it accounts for both false positives and false negatives. In real-time driver monitoring applications, a strong F1-score ensures dependable detection, consistent alerts, and improved driver trust, making it an essential metric for evaluating overall system effectiveness and safety reliability.

V. CONCLUSION

In conclusion, the proposed driver drowsiness detection system demonstrates an effective and reliable approach for improving road safety using a hybrid VGG19 and Vision Transformer architecture. By combining VGG19's capability to capture detailed local facial features with ViT's ability to learn global contextual relationships through self-attention, the system accurately identifies fatigue indicators such as eye closure and mouth movements. This integrated analysis overcomes the limitations of traditional CNN-based methods that rely on a single feature and often struggle with generalization in real-world environments. The model shows strong robustness under varying lighting conditions, head orientations, and environmental noise, ensuring consistent performance. Experimental results confirm improvements in accuracy, recall, and overall reliability compared to conventional approaches. Additionally, the real-time operation and immediate alert mechanism make the system practical for deployment in driver monitoring applications. In future work, the system can be enhanced by integrating additional physiological signals such as heart rate and head movement for more comprehensive fatigue detection. Incorporating temporal analysis using recurrent networks or transformers could further improve the recognition of gradual drowsiness patterns.

REFERENCES

[1] Adrian Rosebrock, "Drowsiness Detection with OpenCV," PyImageSearch, 2017.

[2] B. K. Savaş and Y. Becerikli, "Real Time Driver Fatigue Detection Based on SVM Algorithm," in Proc. 6th International Conference on Control Engineering and Information Technology (CEIT), IEEE, 2018.

[3] T. Vesselenyi et al., "Driver Drowsiness Detection Using ANN Image Processing," in IOP Conference Series: Materials Science and Engineering, vol. 252, no. 1, IOP Publishing, 2017.

[4] V. R. R. Chirra, S. R. Uyyala, and V. K. K. Kolli, "Deep CNN: A Machine Learning Approach for Driver Drowsiness Detection Based on Eye State," Revue d'Intelligence Artificielle, vol. 33, no. 6, pp. 461–466, 2019.

[5] A. Ranjan, K. Vyas, S. Ghadge, S. Patel, and S. S. Pawar, "Driver Drowsiness Detection System Using Computer Vision," International Research Journal of Engineering and Technology (IRJET), 2020.

[6] V. Saini and R. Saini, "Driver Drowsiness Detection System and Techniques: A Review," International Journal of Computer Science and Information Technologies, vol. 5, no. 3, pp. 4245–4249, 2014.

[7] B. Alshaqaqi et al., "Driver Drowsiness Detection System," in Proc. 8th International Workshop on Systems, Signal Processing and Their Applications (WoSSPA), IEEE, 2013.

[8] W. Deng and R. Wu, "Real-Time Driver Drowsiness Detection System Using Facial Features," IEEE Access, vol. 7, pp. 118727–118738, 2019.

[9] S. Park et al., "Driver Drowsiness Detection System Based on Feature Representation Learning Using Various Deep Networks," in

ISSN: 2349-6002

Asian Conference on Computer Vision, Springer International Publishing, 2016.

[10] M. Hashemi, A. Mirrashid, and A. B. Shirazi, "Driver Safety Development: Real-Time Driver Drowsiness Detection System Based on Convolutional Neural Network," SN Computer Science, vol. 1, pp. 1–10, 2020.