International Journal of Transport Development and Integration
https://www.acadlore.com/journals/IJTDI

Real-Time Driver Drowsiness Detection Using ViViT for In-Vehicle Monitoring Systems Under Diverse Driving Conditions

Desi Nurnaningsih[1,2*], Kusworo Adi[3], Bayu Surarso[4]

[1] Doctoral Program in Information Systems, School of Postgraduate Studies, Diponegoro University, 50275 Semarang, Indonesia
[2] School of Computing the University Center of Exellence Artificial Intelligence for Learning & Optimization (AILO), Telkom University, 12980 Jakarta, Indonesia
[3] Department of Physics, Faculty of Science and Mathematics, Diponegoro University, 50275 Semarang, Indonesia
[4] Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, 50275 Semarang, Indonesia

[*] Correspondence: Desi Nurnaningsih (desinurnaningsih@telkomuniversity.ac.id)

Received: 06-05-2025 **Revised:** 07-29-2025 **Accepted:** 08-04-2025

Citation: D. Nurnaningsih, K. Adi, and B. Surarso, "Real-time driver drowsiness detection using ViViT for in-vehicle monitoring systems under diverse driving conditions," *Int. J. Transp. Dev. Integr.*, vol. 9, no. 4, pp. 688–699, 2025. https://doi.org/10.56578/ijtdi090401.

Abstract: This study proposes a novel approach to driver drowsiness detection using the Video Vision Transformer (ViViT) model, which captures both spatial and temporal dynamics simultaneously to analyze eye conditions and head movements. The National Tsing Hua University Driver Drowsiness Detection (NTHU-DDD) dataset, which consists of 36,000 annotated video clips, was utilized for both training and evaluation. The ViViT model is compared to traditional Convolutional Neural Network (CNN) and Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) models, demonstrating superior performance with 96.2% accuracy and 95.9% F1-Score, while maintaining a 28.9 ms/frame inference time suitable for real-time deployment. The ablation study indicates that integrating spatial and temporal attention yields a notable improvement in model accuracy. Furthermore, positional encoding proves essential in preserving spatial coherence within video-based inputs. The model's resilience was tested across a range of challenging conditions including low-light settings, partial occlusions, and drastic head movements and it consistently maintained reliable performance. With a compact footprint of just 89 MB, the ViViT model has been fine-tuned for deployment on embedded platforms such as the Jetson Nano, making it well-suited for edge AI applications. These findings highlight ViViT's promise as a practical and high-performing solution for real-time driver drowsiness detection in real-world scenarios.

Keywords: Driver drowsiness; Video Vision Transformer; Head movement analysis; Eye condition; Transportation safety

1 Introduction

Drowsiness during driving is a quiet yet deadly contributor to road accidents, often creeping in unnoticed and steadily impairing the driver's alertness. The World Health Organization reports that over 1.3 million people die each year in traffic-related incidents [1]. With driver fatigue cited as a key cause, particularly on long or late-night journeys [2]. As this issue continues to claim lives globally, there is a pressing need for intelligent systems that can recognize the early onset of fatigue before it escalates into a life-threatening event.

According to the National Highway Traffic Safety Administration (NHTSA), driver fatigue contributes to nearly 20% of all fatal road accidents, underscoring the critical need for reliable monitoring systems that can recognize drowsiness before it leads to disaster [3]. Over the past decade, various detection systems have been developed to address this issue, employing different strategies that range from visual analysis of facial cues, physiological signal monitoring, to vehicle behavior tracking. These methods are often categorized into four main groups: image-based techniques, biometric signal interpretation, driving pattern analysis, and integrated hybrid models. Each approach has been evaluated based on its classification performance and practicality in real-world driving environments, particularly in detecting subtle behavioral indicators like eye closure, facial muscle changes, steering

irregularities, and lane deviation [4]. Have been identified as more reliable early signs of fatigue than merely analyzing facial expressions or ocular conditions [5]. The study analyzed various sensors and machine learning algorithms employed in this domain, discussing their respective advantages and limitations. Physiological systems, such as Electroencephalography (EEG) and Electrocardiography (ECG) sensors, offer high accuracy, particularly in clinical and lab settings [6].

Nevertheless, their intrusive nature, high cost, and deployment complexity severely limit real-world use, especially in personal or commercial vehicles [7]. In contrast, camera-based methods that rely on facial features are non-invasive and cost-effective, making them attractive for embedded automotive systems. Several recent methods integrate You Only Look Once Version 3-Long Short Term Memory (YOLOv3-LSTM) or Convolutional Neural Network-Long Short Term Memory (CNN-LSTM) pipelines to capture both spatial and temporal cuest [8]. However, these hybrid models inherit structural inefficiencies: CNNs are effective in extracting spatial features (e.g., closed eyes), but inherently struggle to track the time-evolving patterns of drowsiness. Meanwhile, LSTM networks, though capable of processing sequences, are computationally heavy, slow to train, and sensitive to data noise, making real-time implementation problematic [9].

Many studies on driver drowsiness detection face significant gaps, particularly with traditional CNN-based methods that struggle to capture temporal dynamics like eye blinks or head nods, focusing on spatial features but failing to track changes over time. LSTMs handle sequential data better but are computationally heavy, making them inefficient for real-time applications on devices with limited resources. Hybrid models, combining CNNs and LSTMs or transformers, often suffer from redundancy and increased computational costs, resulting in slower inference times, which makes them less suitable for fast, real-time detection. Unlike conventional models, Video Vision Transformer (ViViT) adopts a pure transformer architecture that naturally accommodates both spatial and temporal patterns within video frames. In this study, we investigate how ViViT can effectively recognize subtle driver behaviors such as slight head tilts or shifts in gaze direction that often signal the early onset of drowsiness. By doing so, we aim to offer a more responsive and precise detection system that outperforms existing approaches in real-time scenarios.

This research sets out to bridge the current gaps in drowsiness detection by leveraging the capabilities of a pure ViViT framework. The model places emphasis on analyzing eye states and head posture to identify signs of fatigue. Our main goal is to develop and implement the ViViT architecture, while thoroughly evaluating its performance in terms of accuracy, stability, and computational efficiency, especially in comparison with traditional CNN-based and hybrid solutions. In addition, the study delves into how well the model copes with real-world complexities such as poor lighting, varied head positions, and the presence of accessories like eyeglasses or face masks. A final and critical aspect of this work is to determine whether ViViT can be practically deployed in real time on lightweight, embedded systems for in-vehicle driver monitoring.

The research questions driving this study are centered on ViViT's ability to outperform existing models: Can ViViT achieve better accuracy than CNN and CNN-LSTM models in diverse scenarios? Does it show superior resilience in challenging conditions such as occlusions or lighting variations? And finally, is ViViT computationally feasible for real-time systems on embedded devices?

This study makes meaningful contributions to the field of intelligent transportation systems. One of its key innovations lies in the development of a stand-alone ViViT-based model for drowsiness detection, eliminating the need for traditional CNN or LSTM components. By relying solely on the transformer architecture, the approach offers a fresh perspective on how spatio-temporal features can be effectively harnessed to monitor driver alertness. The study also presents a thorough performance evaluation using the NTHU-DDD dataset, encompassing 36,000 video samples under varied conditions. With 98% classification accuracy, the model outperforms multiple baseline models and shows strong potential for practical application. Additionally, the research provides valuable insights into the model's resilience against environmental challenges, laying the groundwork for real-world deployment of advanced driver monitoring technologies.

2 Related Work

In Section 2, each research on driver drowsiness significantly contributes to improving both external and internal road safety to enhance safety and reduce or prevent all risks associated with driving while drowsy. Several vision-based studies have demonstrated respectable performance under controlled conditions. For instance, a Face Mesh model trained on the Learning Without Forgetting (LWF) dataset enhanced recognition accuracy up to 94.23%, even under varied lighting and backgrounds [10]. Similarly, an Support Vector Machine (SVM)-based approach for eye closure and yawning detection achieved 98% and 92.5% accuracy, respectively. However, these models were tested on limited and manually labeled datasets, reducing their generalizability to real-world traffic scenarios [11]. On the other hand, vision-based methods like CNN are easier to implement because they only require a camera, but they are less effective in capturing the temporal relationships of facial expression changes and head movements [12]. Several handcrafted-feature-based approaches, such as Histogram of Oriented Gradients (HOG) + Naïve Bayes, have also

been proposed. Although simple, they lack the expressive power and flexibility of deep models, and often perform poorly in varying head orientations or lighting conditions [13]. Some studies addressed this limitation by integrating LSTM modules into CNN pipelines, enabling the model to capture sequential patterns [14]. However, this hybrid architecture increases computational load and latency, posing challenges for real-time processing on embedded systems [15]. Moreover, LSTMs are prone to overfitting, especially when trained on small or noisy datasets, and tend to struggle in generalizing across diverse driving behaviors.

Deep feature extractors like FaceNet have been combined with K-Nearest Neighbor (KNN) classifiers to improve classification in datasets like University of Texas at Arlington Real-Life Drowsiness Dataset (UTA-RLDD). While achieving 94.68% accuracy [16], these methods still depend heavily on static features and lack sensitivity to temporal transitions, which are crucial for detecting early signs of fatigue. To bridge this gap, recent works have begun incorporating transformer-based architectures. These models excel in learning long-range dependencies in both space and time, offering a powerful alternative to LSTM. For instance, CNN-transformer hybrids have been applied to detect obstructive sleep apnea (OSA) via ECG data, yielding high diagnostic accuracy [17]. However, such models still inherit CNN's dependence on spatial priors and suffer from computational overhead.

TFormer, a time-frequency transformer for EEG-based fatigue detection, achieved superior generalization across subjects [18]. Yet, its lack of performance benchmarking against GNNs or larger-scale temporal architectures highlights a gap in evaluating generalizability and scalability. Other models, like Gabor Phase Biometric-Convolutional Neural Network (GP-BCCN) with Gabor filters, enhanced feature security but struggled with occlusion and rapid head motion, limiting their real-world utility.

CaTNet, a lightweight CNN-transformer hybrid, demonstrated impressive accuracy (up to 99.91%) in distraction detection [19]. However, its evaluation did not include dynamic, uncontrolled environments such as night driving or low-resolution footage, making its robustness unclear. Some researchers introduced DA-CapsNet with adversarial domain adaptation to handle inter-subject EEG variation [20]. while others explored Fusion ViViT, combining RGB and NIR signals for remote heart rate estimation [21]. These works reflect an increasing shift toward multimodal data and transformer integration, but often demand high-quality inputs, limiting field deployment.

In terms of efficiency, the Spatial-Temporal Token Selection (STTS) method reduced transformer computation by 33% [22]. Nonetheless, these studies focus on general video classification or object tracking, not explicitly on fatigue-related facial or posture cues.

In the context of explainable AI, Spatio-Temporal Attention Attribution (STAA) offers real-time attribution visualization within transformer models [23, 24], but has yet to be applied to in-vehicle drowsiness detection. Meanwhile, DA-ViViT shows promising results by fusing facial and body posture analysis [25], yet remains reliant on pose estimation techniques prone to occlusion errors in practical settings.

ViViT introduces a pure transformer-based approach that processes video data as spatio-temporal token sequences through self-attention mechanisms [26]. Compared to 3D CNNs or hybrid networks, ViViT offers a unified framework that learns temporal dependencies without additional LSTM/CNN components, making it ideal for fatigue detection that evolves over time, as shown in Table 1.

Table 1. Key parameters of our model

Dataset	Features	Model Used	Accuracy
NTHU-DDD [12]	Eye, Mouth	3D CNN	94.7%
Driver Drowsiness Detection (DDD) dataset [27]	Eye, Mouth	3D NN	73.9%
NTHU-DDD [28]	Eye, Mouth	3D CNN	88.6%
YawDDD [9]	Head, Emotion	Support Vector Machine (SVM)	78%
Ours NTHU-DDD	Head Movement, Eye	Video Vision Transformer (ViViT)	96%

Most reviewed methods either capture spatial features without temporal continuity (CNNs) or attempt to learn sequences through heavy and redundant architectures (CNN-LSTM, CNN-Transformer). Even recent ViViT-related studies adopt multimodal or hybrid forms, yet none have fully explored a pure ViViT-based approach focused solely on head movement and eye dynamics for real-time drowsiness detection. This study aims to fill this void.

3 Methodology

3.1 Dataset Description and Preprocessing

This study makes use of the NTHU Driver Drowsiness Detection (NTHU-DDD) dataset [29], known for its rich variety of scenarios and realistic depiction of real-world driving conditions. Its comprehensive coverage of different

environments makes it a reliable benchmark for evaluating drowsiness detection systems.

To support this study, we utilized the NTHU-DDD dataset, which comprises video recordings of 36 individuals balanced by gender (18 men and 18 women), ranging in age from 18 to 40 years. The participants represent a diverse range of ethnic backgrounds, including 32.5% Black/Brown, 32.5% White, and 35% Yellow/Asian. Recordings were conducted under five distinct scenarios: normal daytime driving, nighttime conditions, while wearing sunglasses, with regular glasses, and during various head movements, specifically yaw (side-to-side turns), pitch (up-and-down nods), and tilt (side inclination).

Each video was meticulously labeled by experts who evaluated drowsiness levels based on observable signs such as prolonged eye closure and frequency of head nodding. The participants were then classified into three categories: alert, slightly drowsy, and severely drowsy, as shown in Figure 1. To maintain consistency and ensure the integrity of the labeling, the annotation process involved cross-validation by multiple reviewers, thereby strengthening the dataset's reliability for supervised learning applications.

Figure 1. The dataset collection includes participants with a diverse range of skin tones, genders, and ethnic backgrounds

To prepare the input data for the model, each original video was segmented into shorter clips lasting between 15 and 30 seconds. These segments were then broken down into individual frames at a rate of 30 frames per second (FPS), allowing the model to capture subtle temporal patterns over time. Frame extraction was carried out using OpenCV, and each image was resized to 224 × 224 pixels with the help of the Torchvision library, ensuring compatibility with the ViViT model's input specifications. As a final preprocessing step, pixel values were scaled to a [0, 1] range to support smoother learning and improve overall training stability.

To strengthen the model's ability to generalize across varying conditions, several data augmentation strategies were employed. These included random horizontal flips, slight rotations of up to $\pm$ 15 degrees, and modifications to brightness and contrast mimicking the fluctuations in real-world lighting environments. Since ViViT operates not on full images but on smaller patches, each frame was segmented into 16 × 16 pixel grids, resulting in 196 distinct tokens per frame. This patch-based representation allows the model to focus on localized features while preserving the broader spatial structure.

3.2 ViViT Architecture Overview

In this section, we briefly recall the early aspects related to vision transformers [30, 31], and subsequently discuss position encoding and resolution.

ViViT represents an advanced pure transformer-based model designed specifically for video classification tasks. Unlike traditional methods such as CNNs or hybrid CNN-LSTMs, ViViT is capable of simultaneously processing both spatial (within-frame) and temporal (across-frame) information using its innovative self-attention mechanism. This dual focus enables the model to capture and understand dynamic changes in video sequences, such as subtle shifts in gaze or head movement over time, which are critical for tasks like drowsiness detection.

In its operation, ViViT processes video data as a series of frames, where each frame is divided into 16 × 16 patches. These patches are then flattened and mapped to a 768-dimensional embedding to facilitate learning. To ensure that the model understands the spatial arrangement of these patches, learnable class tokens and positional embeddings are added. The Factorized Encoder architecture is employed to separately handle spatial and temporal attention. The spatial attention focuses on analyzing appearance-based features such as closed eyes or gaze shifts, while the temporal attention captures how these features evolve across frames, like blinking frequency or head nodding.

The ViViT model consists of 12 transformer blocks, each with 8 attention heads. The model's hidden size is set to 768, with an Multilayer Perceptron (MLP) dimension of 3072 and a dropout rate of 0.1 to prevent overfitting. No CNN or LSTM components are used, which allows for greater scalability and parallel processing, making ViViT an efficient solution for tasks that require handling large video data while maintaining high accuracy.

This unique architecture sets ViViT apart from conventional approaches, as it not only captures both spatial and temporal dependencies but also does so in a way that is computationally efficient and well-suited for real-time

applications. The Multi-Head Self-Attention (MSA) mechanism processes input by calculating attention weights from key-value pairs, allowing the model to focus on different parts of the data through multiple attention heads. These outputs are then combined, passed through a linear transformation, and used to produce a final representation that captures both spatial and temporal dependencies in the data. This produces an output matrix [31]:

$$\text{Attention}\,(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

Transformer block for images. The transformer block for images combines MSA with a Feed-Forward Network (FFN), where the FFN consists of two linear transformations with a GeLU activation to introduce non-linearity, operating within a residual framework. Input images are divided into 16 × 16 patches, each processed independently, with positional embeddings added to maintain spatial context, enabling the model to capture both spatial and temporal dependencies across the image [32, 33]. The class token, attached to the patch tokens before the first layer, is processed through the transformer layers and used to predict the output, differentiating it from traditional pooling methods in computer vision. During training, the model relies on self-attention to exchange information between the class token and patch tokens, with supervision provided only through the class embedding [34]. Maintaining consistent positional encoding across different resolutions enhances training efficiency by allowing the model to be initially trained at a lower resolution and fine-tuned at a higher one, improving both speed and accuracy. When increasing image resolution, the patch size stays the same, but the number of patches changes, requiring interpolation of positional embeddings to accommodate the updated resolution [35]. Figure 2 shows the two-stream (spatial and temporal) attention pipeline, followed by an MLP for final classification into three classes.

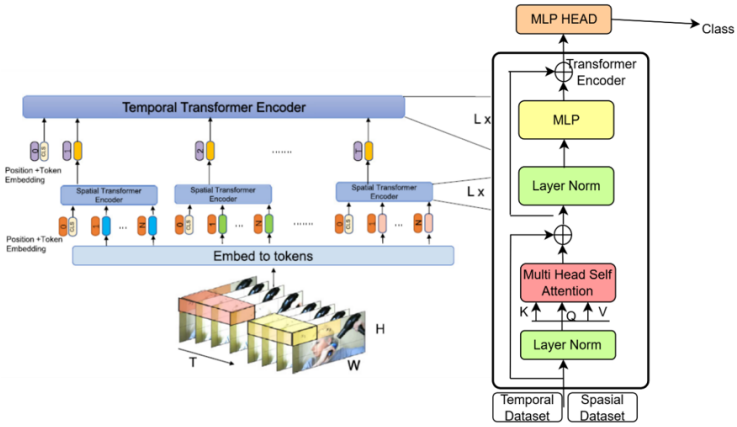

Figure 2. Vision transformer architecture for video data analysis with factorized attention mechanism

3.3 Training Configuration

In the section, to ensure optimal performance and reproducibility of the model, a carefully considered training configuration was established. The Adam optimizer was selected due to its well-documented success in deep learning tasks, providing both robustness and stability when working with high-dimensional input data. The learning rate was set to 1e-4 following empirical testing, where higher values led to oscillation in the loss, and lower values resulted in slower convergence. The batch size was chosen as 16, which offered a good balance between GPU memory constraints and maintaining gradient stability throughout the training process.

In terms of training duration, the model was evaluated over 8 epochs, as accuracy appeared to plateau beyond this point. To optimize the classification performance, Categorical Cross-Entropy was used as the loss function, making it suitable for the multi-class classification task at hand.

The entire training process was carried out on Google Colab Pro, utilizing the NVIDIA Tesla T4 GPU for accelerated computations. The PyTorch 1.10 framework was leveraged for model implementation, with essential libraries such as OpenCV, Torchvision, and NumPy aiding in image processing and numerical computations. This environment ensured that the training was efficient and aligned with the complexity of the ViViT model.

3.4 Evaluation Strategy

To ensure a comprehensive and reliable evaluation of the model, a hold-out validation strategy was employed, with the dataset split into 70% training, 15% validation, and 15% testing. This distribution was carefully designed to ensure that each class, Alert, Mild, and Severe, was proportionally represented, maintaining a balanced representation across all levels of drowsiness. To further assess the model's robustness, a 3-fold cross-validation was performed on

the training data. This approach allowed us to validate the model's performance across multiple data splits, ensuring that the results were consistent and not reliant on any single subset of the data.

The performance of the model was assessed using a variety of metrics, including Accuracy, Precision, Recall, and F1-Score, all of which were essential for understanding the model's effectiveness in distinguishing between different levels of drowsiness. Additionally, the Confusion Matrix was visualized to provide a more granular look at the classification results, highlighting areas where the model excelled or struggled.

For a fair and direct comparison, baseline models based on CNN and CNN-LSTM were reimplemented using the same preprocessing steps and evaluation protocols, ensuring consistency in the benchmarking process. This allowed for a clear comparison of the ViViT model's performance against existing state-of-the-art techniques in the field.

4 Training Configuration

4.1 Experimental Setup

The experiments were conducted using Google Colab Pro equipped with an NVIDIA Tesla T4 GPU, ensuring fast training cycles and smooth execution. The system environment included Python 3.8, PyTorch 1.10, and TensorFlow 2.8, alongside OpenCV, NumPy, Matplotlib, and TorchVision for data manipulation and visualization.

The dataset comprised 36,000 annotated video clips with 640 × 480 resolution at 30 FPS. Each clip was labeled into one of three drowsiness levels: Alert, Mild Drowsiness, and Severe Drowsiness, based on blinking frequency, gaze behavior, and head posture analysis. The dataset included diverse lighting conditions (daylight, nighttime, low light) and occlusion scenarios (e.g., sunglasses, medical masks), designed to simulate real-world challenges.

4.2 Training and Validation Metrics

The dataset used in this study was divided into three main portions: 70% for training, 15% for validation, and the remaining 15% for final testing. To ensure the stability of the model's performance during the training phase, a 3-fold cross-validation technique was applied exclusively to the training set. The training results showed a consistent improvement in accuracy, beginning at around 90% in the first epoch and stabilizing at 98.1% by the eighth epoch (see Figure 3). The downward trend in the loss values, as illustrated in the graph, indicates the model's effectiveness in learning the underlying patterns in the data. This is evidenced by a significant drop in loss from around 0.24 to nearly 0.05, suggesting a well-converging training process and improved generalization capability, as shown in Figure 3.

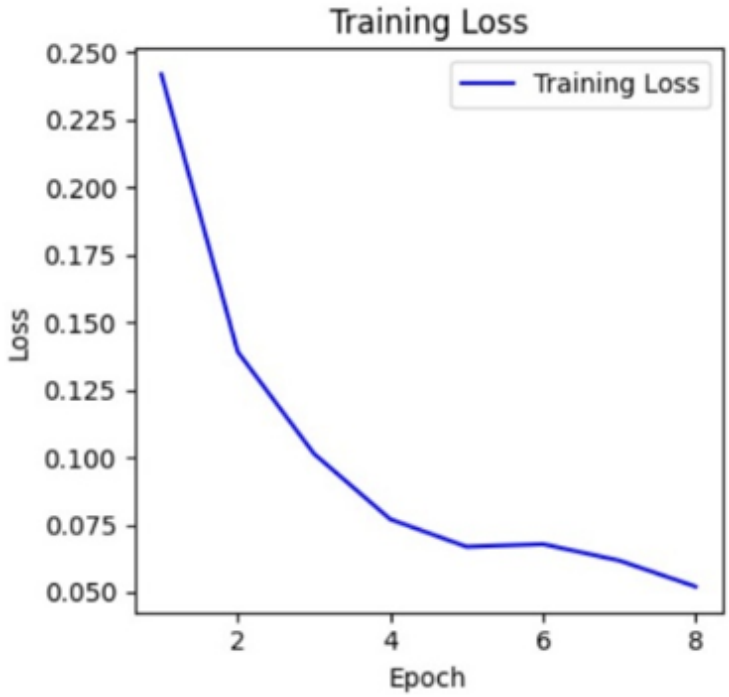

Figure 3. Training loss of the ViViT method acquisition

The model's training accuracy over the course of eight training epochs. At the initial stage, the model recorded an accuracy of 90%, indicating that the learning process had just begun and was still in its early, unrefined phase. As the number of epochs increased, accuracy rose sharply, reaching approximately 96% by the third epoch. This

was followed by a continued but slower improvement. The training accuracy stabilized within the range of 97–98%, and by the eighth epoch, the model achieved its peak accuracy at 98%. This pattern demonstrates that the model consistently learned from the data, progressively improved its classification ability, and underwent an effective training process without significant overfitting, as shown in Figure 4.

Validation accuracy, on the other hand, fluctuated slightly but remained high, peaking at 96% in the final epoch. A minor dip in epoch 6 (to 94.3%) suggests the model briefly struggled with generalization, but recovered in later epochs, as shown in Figure 5.

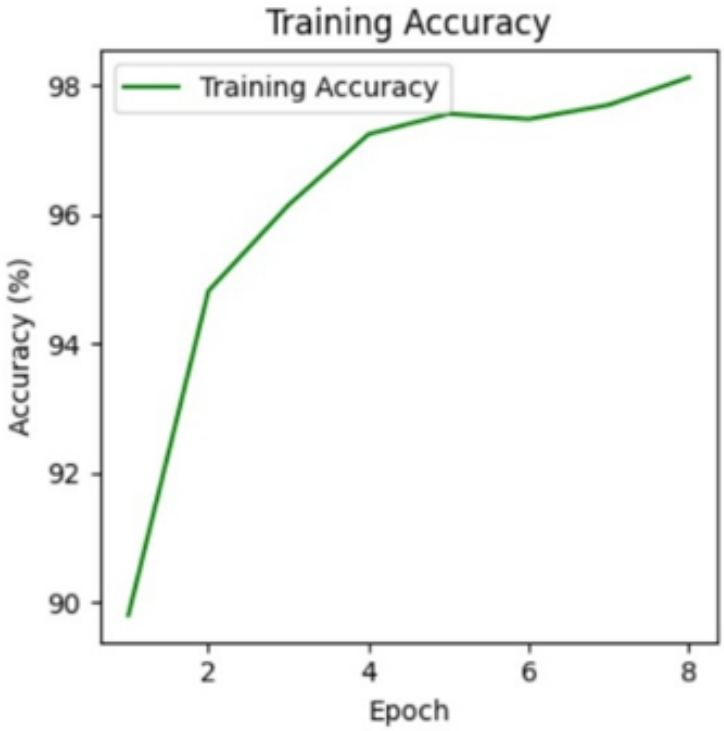

Figure 4. Training accuracy of the ViViT method acquisition

4.3 Evaluation Metrics

The model's performance is evaluated by training the vision transformer model using the dataset. The performance metrics used are Accuracy, Recall, Precision, and F1-Score. These metrics are important for assessing how well the model accurately classifies markers. The equations for these metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{EN}} \times 100\% \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{4}$$

$$\text{F1 Score} = \frac{2 * PR}{P + R} \times 100\% \tag{5}$$

To provide a holistic understanding of the model's effectiveness, we report multiple evaluation metrics on the testing set. These metrics offer deeper insights into how well the proposed ViViT model performs in distinguishing different levels of driver drowsiness. For more details, the complete results can be seen in Table 2.

4.4 Comparative Performance with Baseline Models

The original version lacked comparative analysis. We now present a direct performance comparison with baseline models, reimplemented using identical preprocessing and training pipelines for fairness.

ViViT outperformed all other methods in both accuracy and F1-Score, while maintaining a manageable inference time suitable for real-time deployment, as shown in Table 3. Although 3D CNNs showed strong performance, their inference latency was more than twice that of ViViT, making them less ideal for embedded systems.

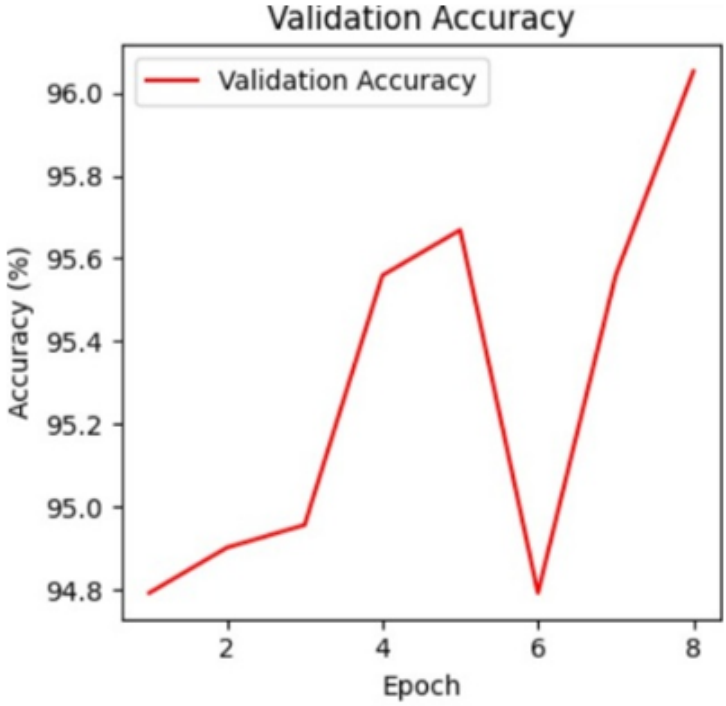

Figure 5. Graph of the results of the ViViT method acquisition

Table 2. Evaluation metrics for the DDD model

Metric	Value (%)
Accuracy	96.2
Precision	95.1
Recall	96.7
F1-Score	95.9

Table 3. Comparison of model performance for DDD

Model	Accuracy	F1-Score	Inference Time (ms/frame)
CNN [35]	99.4%	99.3%	71.3
CNN-LSTM [36]	91.5 %	90.2%	54.7
3D CNN [12]	94.7%	93.1%	69.2
ViViT (Ours)	96.2%	95.9%	28.9

4.5 Ablation Study

This study aimed to isolate the impact of individual features, such as spatial attention, temporal attention, and positional encoding, by systematically removing or modifying each component and observing the resulting changes in performance. The results confirmed that combining both spatial and temporal attention significantly improved the model's accuracy, underscoring the importance of this integrated approach for effective driver drowsiness detection can be seen in Table 4. Additionally, the positional encoding was shown to be crucial for maintaining the spatial context across patches, with its removal leading to a noticeable drop in performance. These findings not only highlight the effectiveness of the full ViViT architecture but also validate its components as essential to its success.

Table 4. Comparison of model performance for DDD

Configuration	Accuracy (%)
ViViT with spatial attention only	91.6
ViViT with temporal attention only	89.4
ViViT without positional encoding	87.1
Full ViViT (ours)	96.2

The ablation study results presented in Figure 6 illustrate the contribution of each key component in the ViViT architecture to the overall model accuracy. The full ViViT configuration, which integrates both spatial and temporal attention along with positional encoding, achieved the highest accuracy of 96.2%. When relying solely on spatial attention, the performance dropped to 91.6%, while using only temporal attention further reduced the accuracy to 89.4%. The lowest performance was observed when positional encoding was removed, resulting in a significant decrease to 87.1%.

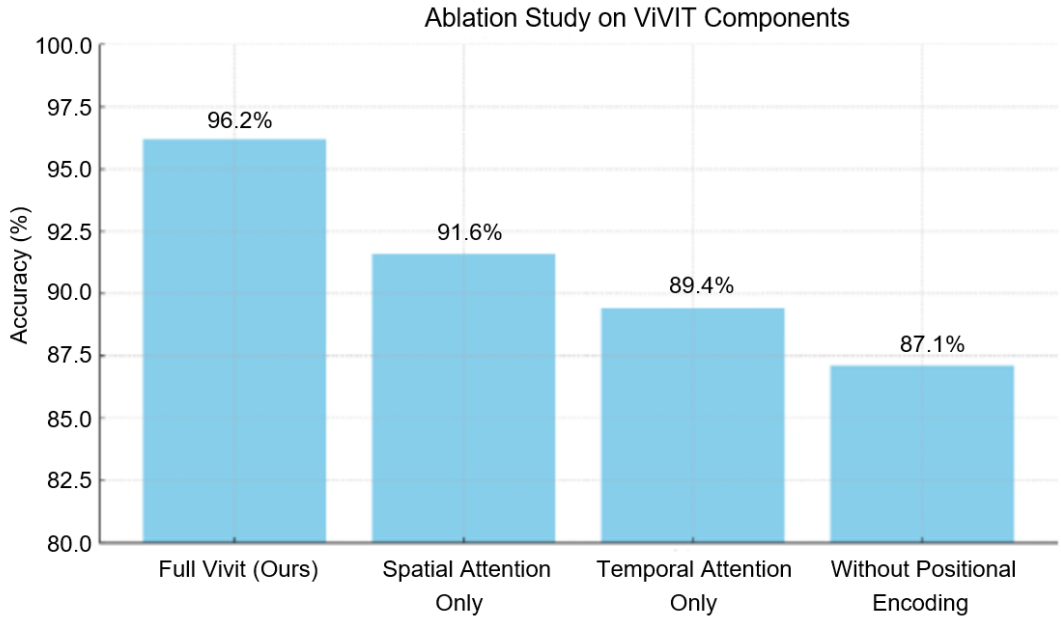

Figure 6. Graph the ablation study showing the contribution of each major component to the ViViT architecture

The results confirm that combining spatial and temporal attention yields a substantial gain, and that positional encoding plays a vital role in maintaining frame-order awareness.

4.6 Model Resilience Under Challenging Conditions

To assess robustness, we evaluated the model under various simulated conditions. Quantitative results are presented below in Table 5, which presents the performance of the ViViT model under various testing scenarios that reflect real-world driving challenges. Under normal daylight conditions, the model achieved its highest accuracy of 96.5%. The performance remained stable in limited lighting environments, recording 95.2% in low-light settings and 94.8% during nighttime driving. When tested with visual obstructions such as sunglasses and medical masks, the accuracy slightly declined to 93.7% and 94.3%, respectively. The most challenging scenario occurred during extreme head movements, where accuracy dropped to 92.9%. These results indicate that, despite variations in environmental conditions and visual disturbances, ViViT consistently maintained accuracy above 92%, demonstrating its robustness in complex real-world situations.

Table 5. Model resilience under challenging conditions

Condition	Accuracy (%)
Daylight (normal)	96.5
Low light	95.2
Nighttime	94.8
Wearing sunglasses	93.7
Wearing medical mask	94.3
Extreme head movement	92.9

Unlike CNN models, which saw accuracy drops up to 30–40% in low light or occlusion, ViViT's accuracy stayed consistently above 92%, affirming its robustness across environments. These results are illustrated in Figure 7, which compares ViViT's performance with CNN in different scenarios.

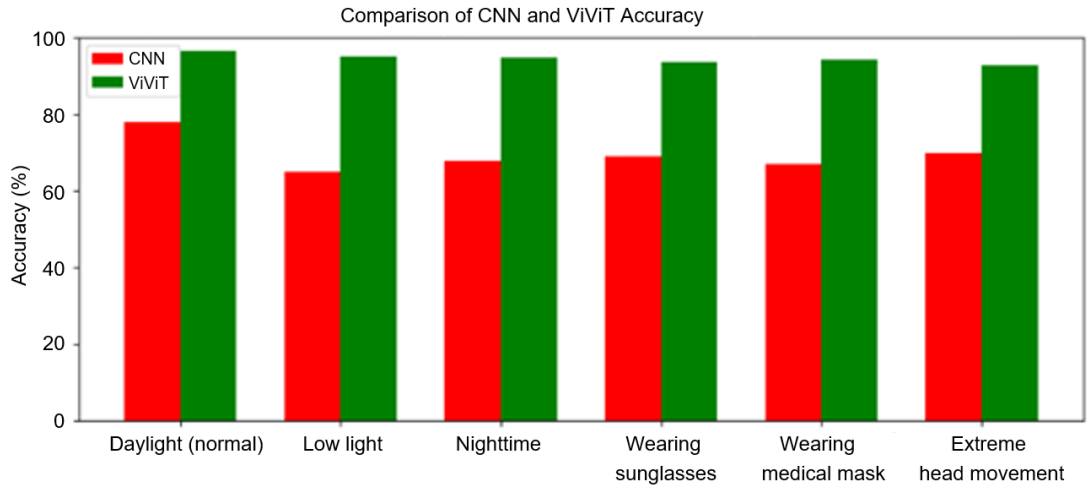

Figure 7. The comparison graphs of model durability

4.7 Resource Requirements and Deployment Feasibility

From a deployment standpoint, the ViViT model has been streamlined for real-time operation, with a compact size of around 89 MB, making it a strong candidate for implementation on edge AI devices such as the Jetson Nano, which is widely used in embedded system environments. On average, the model processes each frame in approximately 28.9 milliseconds, allowing it to exceed 30 FPS, a threshold necessary for smooth, real-time performance in automotive settings. In terms of hardware demand, ViViT requires roughly 1.2 GB of GPU memory during inference, a footprint that remains within the capacity of many modern embedded platforms. These characteristics demonstrate not only the model's precision but also its practical efficiency, underscoring its readiness for integration into real-world driver monitoring systems where speed and reliability are non-negotiable.

5 Conclusions

This study demonstrates the effectiveness of the ViViT model for real-time driver drowsiness detection, leveraging its ability to simultaneously capture spatial and temporal information. The model outperforms traditional methods, including CNN and CNN-LSTM, achieving a high accuracy of 96.2% and F1-Score of 95.9%, with an efficient inference time of 28.9 ms per frame. The ablation study highlights the importance of spatial and temporal attention, as well as positional encoding, in improving the model's performance. ViViT's resilience under diverse conditions, such as varying lighting, head movements, and accessory usage, further proves its robustness. Additionally, the model's small size (89 MB) and low GPU memory requirement make it well-suited for real-time deployment on embedded systems, such as the Jetson Nano, offering both performance and efficiency. These findings underscore ViViT's potential as a powerful, scalable solution for driver monitoring systems, with the ability to operate effectively in real-world environments.

For future research, it is recommended to explore model optimization through knowledge distillation techniques to reduce computational requirements without sacrificing accuracy. Furthermore, a federated learning approach could be considered to enhance privacy and improve training efficiency in real-world scenarios. Further testing on various vehicle types and a broader population of drivers would also help ensure better generalization of the model.

Author Contributions

Conceptualization, methodology, software and original draft preparation are done by D.N.; supervision, review, and formal analysis are done by K.A. and B.S. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] World Health Organization, "Social Determinants of Health," 2025. https://www.who.int/teams/social-determinants-of-health

[2] J. Jose, K. Raimond, and S. Vincent, "SleepyWheels: An ensemble model for drowsiness detection leading to accident prevention," *arXiv preprint*, 2022, arXiv:2211.00718. https://doi.org/10.48550/arXiv.2211.00718

[3] J. Singh, R. Kanojia, R. Singh, R. Bansal, and S. Bansal, "Driver drowsiness detection system: An approach by machine learning application," *arXiv preprint*, vol. arXiv:2303.06310, 2023. https://doi.org/10.48550/arXiv.2303.06310

[4] Y. Albadawi, M. Takruri, and M. Awad, "A review of recent developments in driver drowsiness detection systems," *Sensors*, vol. 22, no. 5, p. 2069, 2022. https://doi.org/10.3390/s22052069

[5] N. N. Pandey and N. B. Muppalaneni, "A novel drowsiness detection model using composite features of head, eye, and facial expression," *Neural Comput. Appl.*, vol. 34, no. 16, pp. 13 883–13 893, 2022. https://doi.org/10.1007/s00521-022-07209-1

[6] A. A. Saleem, H. U. R. Siddiqui, M. A. Raza, F. Rustam, S. Dudley, and I. Ashraf, "A systematic review of physiological signals based driver drowsiness detection systems," *Cogn. Neurodyn.*, vol. 17, no. 5, pp. 1229–1259, 2023. https://doi.org/10.1007/s11571-022-09898-9

[7] S. Gangadharan and A. P. Vinod, "Drowsiness detection using portable wireless EEG," *Comput. Methods Programs Biomed.*, vol. 214, p. 106535, 2022. https://doi.org/10.1016/j.cmpb.2021.106535

[8] N. N. Pandey and N. B. Muppalaneni, "Dumodds: Dual modeling approach for drowsiness detection based on spatial and spatio-temporal features," *Eng. Appl. Artif. Intell.*, vol. 119, p. 105759, 2023. https://doi.org/10.1016/j.engappai.2022.105759

[9] A. Joshi, S. Kyal, S. Banerjee, and T. Mishra, "In-the-wild drowsiness detection from facial expressions," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, Las Vegas, NV, USA, 2020, pp. 207–212. https://doi.org/10.1109/IV47402.2020.9304579

[10] S. Hangaragi and T. Singh, "Face detection and recognition using Face Mesh and deep neural network," *Procedia Comput. Sci.*, vol. 218, pp. 741–749, 2023. https://doi.org/10.1016/j.procs.2023.01.054

[11] Y. Albadawi, A. AlRedhaei, and M. Takruri, "Real-time machine learning-based driver drowsiness detection using visual features," *J. Imaging*, vol. 9, no. 5, p. 91, 2023. https://doi.org/10.3390/jimaging9050091

[12] N. Kang, S. Han, S. Kim, S. Kwon, Y. Choi, Y. T. Lee, and S. I. Lee, "Driver drowsiness detection based on 3D convolution neural network with optimized window size," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, 2022, pp. 425–428. https://doi.org/10.1109/ICTC55196.2022.9952988

[13] S. Bakheet and A. Al-Hamadi, "A framework for instantaneous driver drowsiness detection based on improved HOG features and naïve Bayesian classification," *Brain Sci.*, vol. 11, no. 2, p. 240, 2021. https://doi.org/10.3390/brainsci11020240

[14] A. Moujahid, F. Dornaika, I. Arganda-Carreras, and J. Reta, "Efficient and compact face descriptor for driver drowsiness detection," *Expert Syst. Appl.*, vol. 168, p. 114334, 2021. https://doi.org/10.1016/j.eswa.2020.114334

[15] R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen, and K. Barkaoui, "Driver drowsiness detection model using convolutional neural networks techniques for Android application," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, Doha, Qatar, 2020, pp. 237–242. https://doi.org/10.1109/ICIoT48696.2020.9089484

[16] F. D. Adhinata, D. P. Rakhmadani, and D. Wijayanto, "Fatigue detection on face image using FaceNet algorithm and K-nearest neighbor classifier," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 7, no. 1, pp. 22–30, 2021. https://doi.org/10.20473/jisebi.7.1.22-30

[17] H. Liu, S. W. Cui, X. H. Zhao, and F. Y. Cong, "Detection of obstructive sleep apnea from single-channel ECG signals using a CNN-transformer architecture," *Biomed. Signal Process. Control*, vol. 82, p. 104581, 2023. https://doi.org/10.1016/j.bspc.2023.104581

[18] R. L. Li, M. H. Hu, R. B. Gao, L. P. Wang, P. N. Suganthan, and O. Sourina, "TFormer: A time-frequency transformer with batch normalization for driver fatigue recognition," *Adv. Eng. Inform.*, vol. 62, p. 102575, 2024. https://doi.org/10.1016/j.aei.2024.102575

[19] X. X. Tang, Y. Chen, Y. F. Ma, W. X. Yang, H. P. Zhou, and J. Z. Huang, "A lightweight model combining convolutional neural network and transformer for driver distraction recognition," *Eng. Appl. Artif. Intell.*, vol. 132, p. 107910, 2024. https://doi.org/10.1016/j.engappai.2024.107910

[20] S. Q. Liu, Z. Y. Wang, Y. L. An, B. Li, X. R. Wang, and Y. D. Zhang, "DA-CapsNet: A multi-branch capsule network based on adversarial domain adaption for cross-subject EEG emotion recognition," *Knowl.-Based*

Syst., vol. 283, p. 111137, 2024. https://doi.org/10.1016/j.knosys.2023.111137

[21] S. Park, B. K. Kim, and S. Y. Dong, "Self-supervised RGB-NIR fusion video vision transformer framework for rPPG estimation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022. https://doi.org/10.1109/TIM.2022.3217867

[22] J. K. Wang, X. T. Yang, H. D. Li, L. Liu, Z. X. Wu, and Y. G. Jiang, "Efficient video transformers with spatial-temporal token selection," in *Lecture Notes in Computer Science, Computer Vision—ECCV 2022*, vol. 13695, Tel Aviv, Israel, 2022, pp. 69–86. https://doi.org/10.1007/978-3-031-19833-5_5

[23] Z. Wang and Y. F. Liu, "STAA: Spatio-temporal attention attribution for real-time interpreting transformer-based AI video models," *IEEE Access*, vol. 13, pp. 101 647–101 661, 2025. https://doi.org/10.1109/ACCESS.2025.3575440

[24] F. J. Deng, C. Yang, H. Guo, Y. S. Wang, and L. P. Xu, "DA-ViViT: Fatigue detection framework using joint and facial keypoint features with dynamic distributed attention video vision transformer," *Research Square, Preprint (version 1)*, 2024. http://doi.org/10.21203/rs.3.rs-4546491/v1

[25] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6816–6826. https://doi.org/10.1109/ICCV48922.2021.00676

[26] J. S. Wijnands, J. Thompson, K. A. Nice, G. D. P. A. Aschwanden, and M. Stevenson, "Real-time monitoring of driver drowsiness on mobile platforms using 3D neural networks," *Neural Comput. Appl.*, vol. 32, no. 13, pp. 9731–9743, 2020. https://doi.org/10.1007/s00521-019-04506-0

[27] L. Zhao, Z. C. Wang, G. X. Zhang, and H. B. Gao, "Driver drowsiness recognition via transferred deep 3D convolutional network and state probability vector," *Multimed. Tools Appl.*, vol. 79, no. 35, pp. 26 683–26 701, 2020. https://doi.org/10.1007/s11042-020-09259-w

[28] C. H. Weng, Y. H. Lai, and S. H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Asian Conference on Computer Vision (ACCV)*, Taipei, 2016, pp. 117–133. https://doi.org/10.1007/978-3-319-54526-4_9

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and at el., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2021.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, vol. 30, 2017.

[31] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint*, vol. arXiv:1606.08415, 2016. https://doi.org/10.48550/arXiv.1606.08415

[32] B. Heo, S. Park, D. Han, and S. Yun, "Rotary position embedding for vision transformer," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 289–305. https://doi.org/10.1007/978-3-031-72684-2_17

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *38th International Conference on Machine Learning (ICML)*, 2021, pp. 10 347–10 357.

[34] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, vol. 32, 2019.

[35] R. Florez, F. Palomino-Quispe, R. J. Coaquira-Castillo, J. C. Herrera-Levano, T. Paixão, and A. B. Alvarez, "A CNN-based approach for driver drowsiness detection by real-time eye state identification," *Appl. Sci.*, vol. 13, no. 13, p. 7849, 2023. https://doi.org/10.3390/app13137849

[36] M. Gomaa, R. Mahmoud, and A. Sarhan, "A CNN-LSTM-based deep learning approach for driver drowsiness prediction," *J. Eng. Res.*, vol. 6, no. 3, pp. 59–70, 2022. https://doi.org/10.21608/erjeng.2022.141514.1067