Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2025) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

DRISTI: A Spatio-Temporal Deep Learning Framework for Real-Time Driver Drowsiness Detection

Abstract

Driver drowsiness is one of the leading causes of road accidents worldwide, particularly in resource-constrained and high-variability environments where real-time monitoring is critical. To address this challenge, we present **DRISTI** (*Driver Real-time Intelligent Spatio-Temporal Inference*), a novel deep learning framework designed for efficient and accurate drowsiness detection. Unlike traditional convolution-based systems, DRISTI leverages a Vision Transformer (ViT) encoder to extract global spatial representations from facial frames, enabling the capture of fine-grained cues such as eye closure, yawning, and gaze shifts with reduced computational cost. These embeddings are modelled through a Long Short-Term Memory (LSTM) network to learn temporal behavioural dynamics, while an additive attention mechanism highlights the most informative segments, ensuring reliable recognition of both transient and sustained drowsiness. A dynamic frame selection strategy further reduces redundancy by retaining only behaviourally distinct frames, improving efficiency without compromising accuracy. Optimised for edge deployment, DRISTI achieves a strong balance between speed, robustness, and interpretability, consistently delivering state-of-the-art accuracy and F1-scores across diverse driving conditions. By unifying transformer-based spatial encoding with attention-driven temporal reasoning, DRISTI represents a deployable, scalable, and proactive solution for next-generation driver assistance systems, with clear potential to reduce fatigue-related accidents and enhance road safety.

Keywords: Driver Drowsiness Detection; Spatio-Temporal Deep Learning; Vision Transformer (ViT); Long Short-Term Memory (LSTM); Real-Time Monitoring Systems;

1. Introduction

Driver drowsiness is a major global road safety concern, causing thousands of preventable accidents each year. In the U.S., the National Highway Traffic Safety Administration (NHTSA) reports over 90,000 crashes, nearly 800 deaths, and 50,000 injuries annually due to drowsy driving [1]. Fatigue is responsible for about 15% of fatal heavy vehicle crashes in Australia [2], nearly 40% of road collisions in Iran [3], and an estimated 40% of highway accidents in India [4]. Long-haul drivers are especially vulnerable due to irregular schedules and extended hours. Recognising these risks, the European Union has mandated drowsiness detection systems in all new vehicles from 2024.

Prior work spans three broad categories. *Physiological* methods (EEG/EOG/ECG) are accurate but intrusive, costly, and difficult to scale. *Behavioural* approaches analyse facial landmarks, blink dynamics (e.g., eye aspect ratio, EAR), and head pose, offering non-intrusive, real-time monitoring at low cost but degrading under occlusion and poor illumination. *Vehicle-based* signals (lane position, steering, speed) provide complementary cues yet are highly sensitive to road quality, traffic density, and driver style, which limits generalisability in heterogeneous environments. Critically,

many solutions treat frames independently or rely on short temporal windows, underserving the gradual dynamics that precede full drowsiness. Traditional monitoring methods often fail to capture subtle behavioural cues, making them unreliable in real-world scenarios. With the increasing availability of advanced deep learning techniques, there is a strong need to develop intelligent systems capable of detecting drowsiness accurately and in real time, ultimately enhancing road safety and saving lives.

We address these limitations with **DRISTI** (Driver Real-time Intelligent Spatio-Temporal Inference), a holistic deep framework for real-time drowsiness detection in resource-constrained settings. DRISTI couples a **Vision Transformer (ViT)** for global, fine-grained spatial encoding of facial cues with a **Long Short-Term Memory (LSTM)** network to model temporal evolution, and applies **additive attention** to highlight behaviourally informative segments such as sustained eye closure or yawning. Our contributions are fourfold:

- **Integrated spatio-temporal pipeline:** ViT-based embeddings fused with LSTM sequence modelling capture both static and evolving indicators of fatigue.
- **Attention-augmented reasoning:** An additive attention head emphasises salient temporal segments, improving sensitivity to subtle, pre-drowsy transitions.
- **Efficiency and robustness:** Dynamic frame selection via cosine similarity retains behaviourally distinct frames, cutting redundancy and computation while preserving semantic richness; the lightweight design targets low-power hardware.
- **Practical applicability:** By balancing accuracy, interpretability, and speed, DRISTI is well-suited to next-generation driver monitoring, including the constraints of India's transportation landscape.

The remainder of this paper is structured as follows. Section 2 reviews related work in drowsiness detection, temporal analysis, and transformer-based architectures. Section 3 describes the dataset, annotation pipeline, and preprocessing strategies employed. The section 4 presents the proposed methodology, including ViT-based spatial encoding, LSTM sequence modelling, and the attention mechanism. Section 5 reports experimental configurations, evaluation metrics, and performance analysis. Finally, Section 6 concludes the paper by summarising contributions and outlining directions for future work.

2. Related Work

Driver drowsiness detection has become a critical research domain within computer vision and intelligent transportation systems because of its direct implications for road safety. Existing approaches can be broadly divided into four categories: physiological signal-based techniques, behavioural feature-based methods, vehicle dynamics-based measures, and, more recently, deep spatio-temporal learning frameworks utilising advanced neural architectures.

Physiological monitoring techniques such as EEG, ECG, and EOG offer reliable indicators of driver fatigue. For example, Awais et al. [5] combined EEG spectral features with Support Vector Machines, achieving **91.3%** accuracy, while Jap et al. [6] reported sensitivities above **90%**. Despite their effectiveness, these methods remain intrusive, expensive, and impractical for large-scale or real-world applications. Non-intrusive behavioural approaches analyse facial cues such as blinks, yawns, and expressions. Soukupová and Čech [7] introduced an eye aspect ratio (EAR)-based blink detection system with **95%** accuracy, but performance declined under poor lighting or occlusion. Similarly, Abtahi et al. [8] developed a yawning detection method that achieved **92%** accuracy, yet struggled under unconstrained driving conditions. Vehicle-based indicators such as steering angle, lane deviation, and speed variation provide complementary signals. Dong et al. [9] reviewed such methods, reporting up to **85%** accuracy in highway settings. However, their dependence on road quality, traffic density, and driver variability restricts generalizability in heterogeneous environments like India.

With deep learning advances, spatio-temporal models have gained prominence. Zhang et al. [10] achieved **93.8%** accuracy using CNN-LSTMs, while Zhu et al. [11] introduced temporal attention for improved robustness, reaching **96.2%**. Attention-enhanced methods [12] further demonstrated the value of capturing subtle micro-sleep transitions. The emergence of Vision Transformers (ViTs) [13] has transformed this field, offering global feature representation through self-attention. Recent works [14, 15] achieved above **95%** accuracy with ViT-based hybrids. Building on these advances, we propose **DRISTI**, a unified framework that integrates ViT-based spatial encoding, LSTM-driven

temporal modelling and additive attention. This design enables robust, efficient, and scalable drowsiness detection suitable for deployment in resource-constrained real-world environments.

3. Dataset and Preprocessing

We employ a heterogeneous, large-scale driver drowsiness video corpus originally collected and curated by **Dr. Reza Ghoddoosian** [16] (Honda Research Institute USA, Inc.). The dataset comprises recordings from **52 subjects** spanning diverse ethnicities and geographies, captured under wide variation in illumination, camera pose, facial morphology, eyewear, and video quality. This diversity is crucial for training models that generalise to real-world conditions rather than over-fitting to narrow laboratory settings.

3.1. Raw Dataset Acquisition

Videos are sourced from a public repository and, for each subject, captured in three canonical alertness states: a neutral state (Video 0), where the subject is awake and alert with eyelids open; a transient drowsy state (Video 5), which corresponds to the intermediate pre-sleepiness phase; and a full drowsy state (Video 10), characterized by sustained eye closure and micro-sleep episodes. To standardise downstream processing, all .mov assets were batch-converted to .mp4 using FFmpeg, ensuring consistent codecs and container formats across the corpus.

3.2. Annotation Pipeline

Because frame-level labels were not provided, we developed an automated annotation pipeline grounded in the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR) as in Soukupová and Čech [7]. Each video was decoded at 30 FPS. Using **DLib** (68-point landmark model), we localised periocular and oral landmarks per frame and computed EAR and MAR to quantify eyelid aperture and yawn extent, respectively (Equations 1 and 2):

$$\mathrm{EAR} = \frac{|P_2 - P_6| + |P_3 - P_5|}{2 \cdot |P_1 - P_4|} \tag{1}$$

$$\mathrm{MAR} = \frac{|P_{63} - P_{67}| + |P_{62} - P_{66}| + |P_{61} - P_{65}|}{3 \cdot |P_{60} - P_{64}|} \tag{2}$$

The Eye Aspect Ratio (EAR) is defined with the following landmark indices:

$$\textbf{Right Eye}: \quad P_1 = 36, P_2 = 37, P_3 = 38, P_4 = 39, P_5 = 40, P_6 = 41$$
$$\textbf{Left Eye}: \quad P_1 = 42, P_2 = 43, P_3 = 44, P_4 = 45, P_5 = 46, P_6 = 47$$

The resulting rules produce consistent, scalable frame-level pseudo-labels for the three states and enabled rapid corpus-wide annotation (workflow in Fig. 1).

3.3. Frame Selection and Redundancy Reduction

Decoding videos at 30 FPS across several minutes produces massive redundancy and high computational cost. To address this, we adopt a **dynamic frame sampling** strategy driven by **cosine similarity** of lightweight frame embeddings (Eq. 3):

$$\mathrm{CosSim}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}. \tag{3}$$

Fig. 1: Annotation workflow using facial landmarks and EAR/MAR thresholds to derive frame-wise labels.

Frames are retained only when they are sufficiently dissimilar from their predecessors, capturing salient changes such as head pose shifts, sustained eye closure, or yawning. A dissimilarity counter ensures selection after a persistence threshold (e.g., ≥ 5 frames), filtering out trivial variations like single blinks. This reduces temporal density, eliminates near-duplicates, and preserves behaviourally meaningful transitions. The overall pipeline is depicted in Fig. 2.

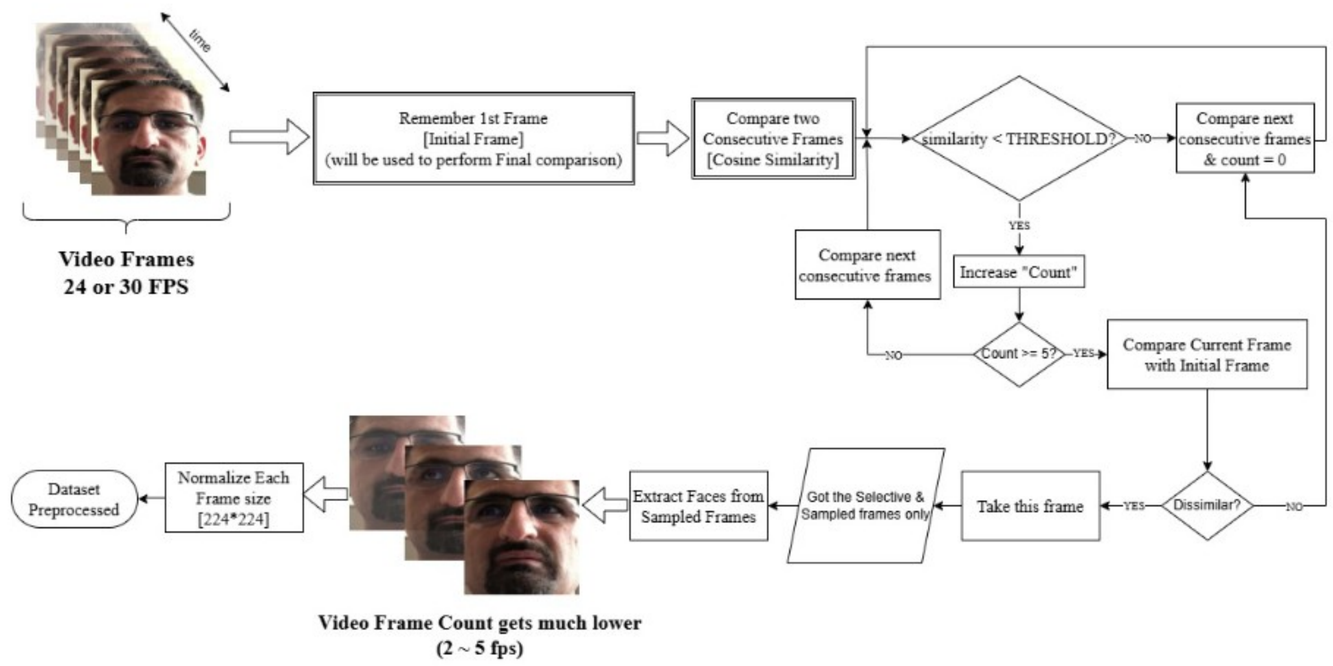


Fig. 2: Preprocessing pipeline: decoding, EAR/MAR-based annotation, cosine-similarity sampling, and face normalisation.

3.4. Face Region Extraction and Normalisation

To focus learning on driver-centric cues, we extract face crops using DLib landmarks to form tight bounding boxes around the visage. Crops are resized and normalised to 224×224 pixels, matching the Vision Transformer input specification. This step standardises scale, improves batch uniformity, and reduces background-induced variance, benefiting both training stability and inference efficiency.

Remarks. The combination of EAR/MAR-derived supervision and cosine similarity sampling strikes a pragmatic balance: EAR/MAR serves as a robust, interpretable proxy for drowsiness indicators, while similarity-based filtering curbs redundancy without complex motion modelling. Together, these steps produce sequences that are compact, semantically dense, and well aligned with the downstream spatio-temporal architecture.

4. Methodology

We introduce **DRISTI**, a unified deep architecture for driver drowsiness detection that jointly creates models on spatial facial cues and temporal dynamics. DRISTI couples a fine-tuned Vision Transformer (ViT) for frame-wise representation learning with a Long Short-Term Memory (LSTM) network and an additive attention head for sequence reasoning. This section summarises the workflow, core modules, and training objective.

4.1. Overview of DRISTI Workflow

The pipeline (Fig. 3) processes a short clip of facial frames to produce a three-way label (*Not Drowsy*, *Transient*, *Drowsy*). First, each frame is embedded by a **fine-tuned ViT** that captures global facial structure (eyes, mouth, head pose) robust to illumination and occlusion. These embeddings form a temporal sequence consumed by an **LSTM** to encode short/mid-range dynamics (e.g., sustained eyelid closure). An **additive attention** module highlights the most informative time-steps before a lightweight classifier outputs the final prediction.

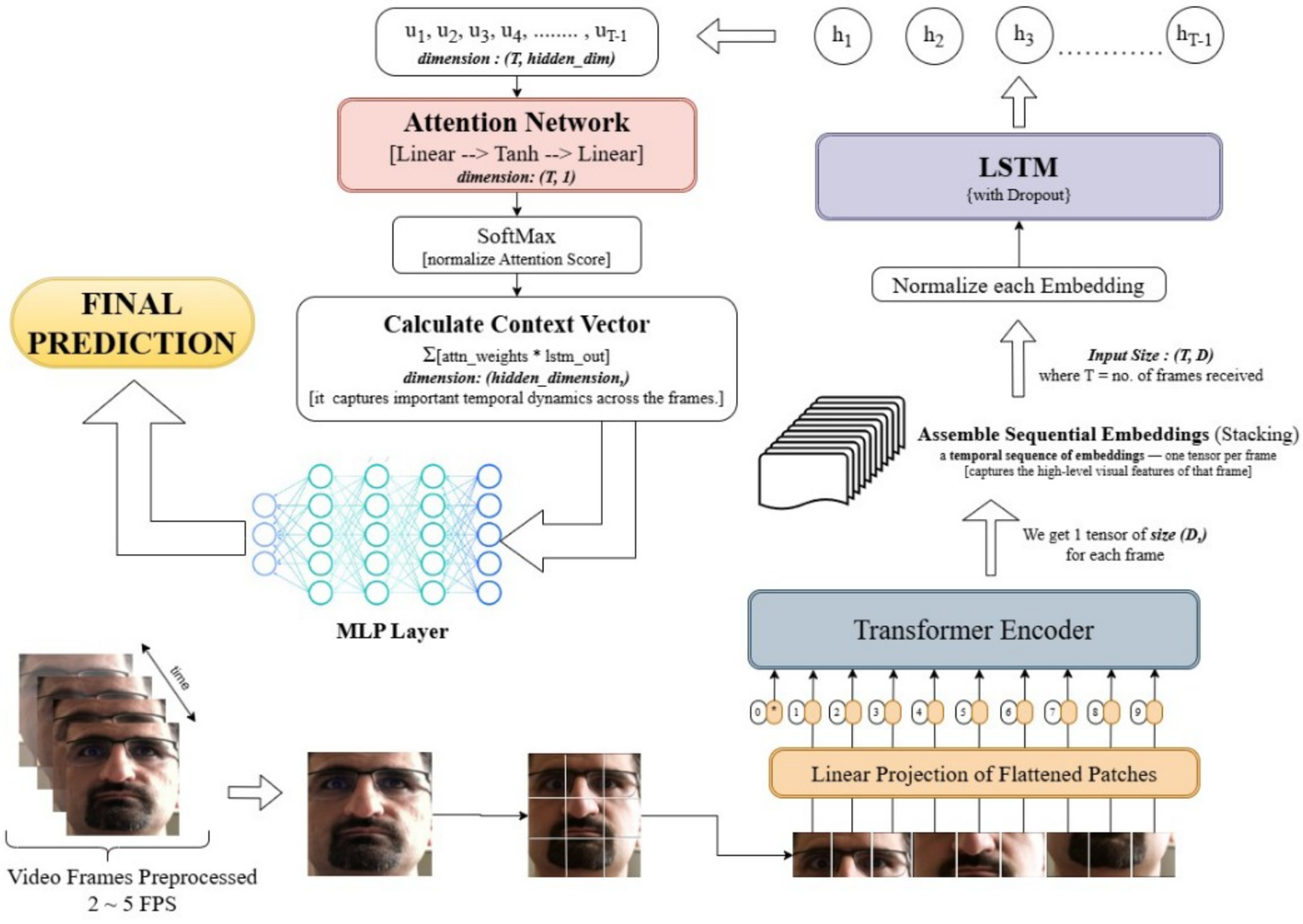


Fig. 3: Overall architecture of the DRISTI framework.

4.2. Spatial Feature Extraction: Vision Transformer Encoder

Why ViT. Unlike CNNs that rely on local receptive fields, ViT partitions each frame into fixed-size patches and applies self-attention, enabling: (i) *global* spatial reasoning; (ii) sensitivity to subtle, distributed cues (partial eye closure, micro-expressions, slight head tilts); (iii) improved robustness to pose and lighting; and (iv) efficient transfer via pre-training. We initialise from ImageNet-pretrained weights and fine-tune on our curated drowsiness imagery. Given a frame $I_t \in \mathbb{R}^{H\times W\times 3}$, ViT produces an embedding $x_t \in \mathbb{R}^D$ (typically D=768). Over a clip of T frames we obtain a sequence

$$\mathcal{X} = \{x_1, x_2, \ldots, x_T\}, \quad x_t \in \mathbb{R}^D.$$

4.3. Temporal Sequence Modelling: LSTM Network

Drowsiness is a *temporal* phenomenon. To model evolution within the clip, we pass $\mathcal{X}$ through an LSTM with hidden size H (we use H=512). At each step t,

$$(h_t, c_t) = \mathrm{LSTM}(x_t, h_{t-1}, c_{t-1}), \qquad h_t, c_t \in \mathbb{R}^H,$$

where h_t summarises the history up to t and c_t is the cell state. The LSTM captures both brief micro-sleep events and longer fatigue build-up.

4.4. Temporal Attention Mechanism

To emphasise informative moments (e.g., sustained eye closure, yawns) and down-weight redundant frames, we apply additive (Bahdanau-style) attention over $\{h_t\}_{t=1}^{T}$:

$$e_t = v^\top \tanh(W_h h_t + b_h), \quad \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \quad c = \sum_{t=1}^{T} \alpha_t h_t, \tag{4}$$

with learnable $W_h \in \mathbb{R}^{A\times H}$, $v \in \mathbb{R}^A$, $b_h \in \mathbb{R}^A$ and attention size A (we use A=128). The context vector $c \in \mathbb{R}^H$ is a temporally pooled representation focused on the most diagnostic frames.

4.5. Output Layer and Objective

A shallow MLP maps c to class logits, followed by softmax:

$$\hat{y} = \text{Softmax}(W_c c + b_c), \quad W_c \in \mathbb{R}^{3\times H}, \; b_c \in \mathbb{R}^3. \tag{5}$$

We optimise the categorical cross-entropy loss for C=3 classes:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i). \tag{6}$$

4.6. Optimisation (AdamW)

We train end-to-end with AdamW [17], combining adaptive moments and decoupled weight decay. Let Θ denote all trainable parameters and $\nabla_\Theta \mathcal{L}$ the gradient. With learning rate η, decay rates β_1, β_2, numerical constant ϵ, and weight decay λ,

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\nabla_\Theta \mathcal{L}, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2)\left(\nabla_\Theta \mathcal{L}\right)^2, \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \quad \Theta \leftarrow \Theta - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} + \lambda\Theta\right). \end{aligned} \tag{7}$$

4.7. Mathematical Summary and Input Representation

For completeness, we summarise the mapping for a sequence $\mathcal{V}=\{I_t\}_{t=1}^{T}$:

$$x_t = \text{ViT}(I_t) \in \mathbb{R}^D, \quad (h_t, c_t) = \text{LSTM}(x_t, h_{t-1}, c_{t-1}), \quad c = \sum_{t=1}^{T} \alpha_t h_t, \quad \hat{y} = \text{Softmax}(W_c c + b_c),$$

with attention weights α_t as defined above. In practice, $H = W = 224$, $D = 768$, $H = 512$, and $A = 128$.

Algorithm 1: DRISTI Model Training Pipeline

Input: Preprocessed video sequence $\mathcal{F}_{selected} = \{I_1, I_2, ..., I_T\}$
Output: Predicted label and Trained model parameters Θ

for *each frame* $I_t \in \mathcal{F}_{selected}$ **do**
    Extract spatial embedding via Vision Transformer:
    $x_t \leftarrow \text{ViT}(I_t)$ where $x_t \in \mathbb{R}^D$;
Assemble sequence of embeddings: $\mathcal{X} = \{x_1, x_2, ..., x_T\}$;
Obtain temporal encoding via LSTM:
$h_t \leftarrow \text{LSTM}(x_t, h_{t-1})$ where $h_t \in \mathbb{R}^H$;
Compute attention scores for each timestep:
$e_t \leftarrow \mathbf{v}^\top \tanh(\mathbf{W}_h h_t + \mathbf{W}_s s)$ (Bahdanau attention);
Normalise attention weights:
$\alpha_t \leftarrow \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$;
Compute context vector:
$c \leftarrow \sum_{t=1}^{T} \alpha_t h_t$;
Feed context vector to dense classifier:
$\hat{y} \leftarrow \text{Softmax}(\mathbf{W}_c c + b_c)$;
Compute cross-entropy loss: $\mathcal{L}_{CE}(y, \hat{y})$;
Update parameters $\Theta \leftarrow \text{AdamW}(\Theta, \nabla_\Theta \mathcal{L}_{CE})$;

5. Experimental Analysis

This section presents an extensive evaluation of the proposed **DRISTI** framework. The analysis was conducted in three phases: (i) empirical determination of preprocessing thresholds, (ii) end-to-end model training, and (iii) final performance assessment through quantitative metrics and qualitative visualisations.

5.1. Experimental Setup

Preprocessing involved tuning EAR/MAR thresholds (0.21–0.25, 0.68) and cosine similarity threshold ($\theta = 0.85$) through controlled iterations, with a persistence count ($C_{th} = 5$) to ensure robust frame selection. Videos were downsampled to 4 FPS, resized to 224×224, and segmented into 20-frame sequences. A subject-wise split (70/15/15) maintained person-independent evaluation.

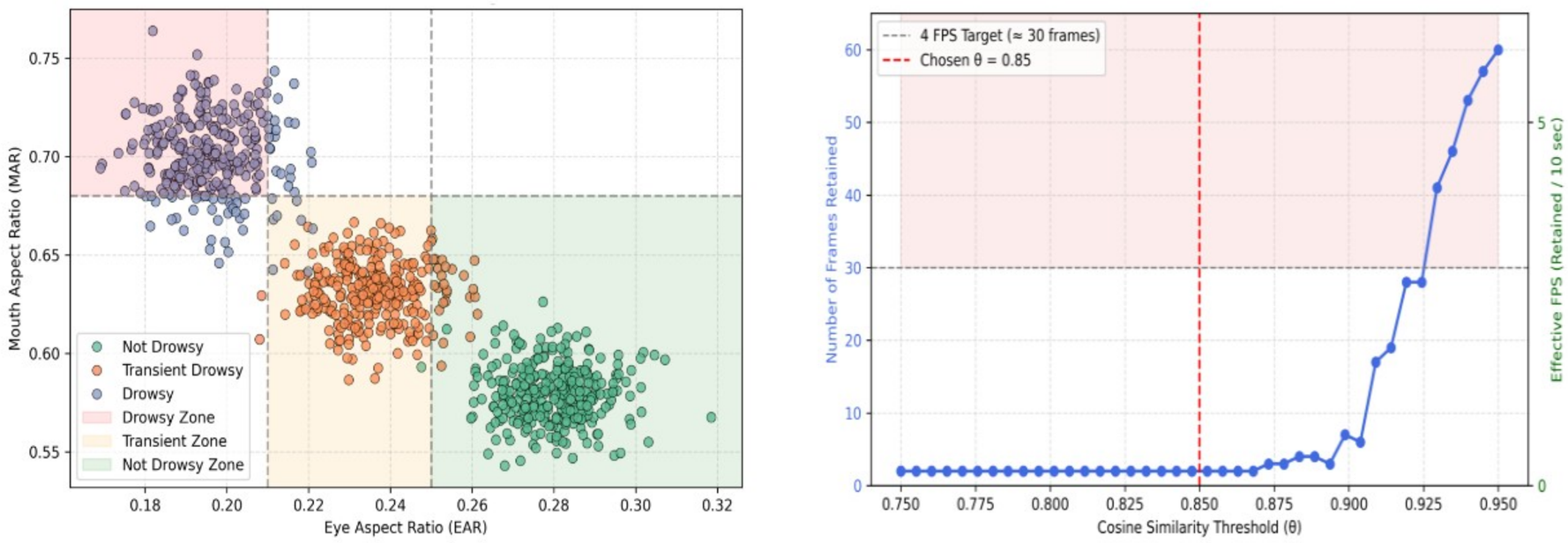
DRISTI was trained using AdamW (1×10^{-4}), batch size 32, categorical cross-entropy, and early stopping (patience=10), with dropout (0.3) in LSTM and attention layers. Experiments ran on an Nvidia V100 GPU (64 GB RAM) with PyTorch 2.1, Python 3.10, and supporting libraries. Thresholds were validated both visually and quantitatively.

5.2. Results Analysis

The final model was trained on balanced class distributions. Label smoothing was applied to reduce prediction overconfidence. Validation monitored macro F1-score with early stopping. We have used Accuracy Macro/Weighted F1, and ROC-AUC as evaluation metrics.

5.3. Performance Analysis of Proposed Architecture

On the held-out test set, **DRISTI** delivers robust overall performance with balanced behaviour across classes (Table 1). As shown in Table 1(a), the model attains an **accuracy of 89.9%**, indicating strong aggregate correctness. The **macro F1-score of 91.2%** evidences class-balanced effectiveness by averaging performance equally across the three labels, while the **weighted F1-score of 93.5%** reflects solid performance when accounting for class frequencies. In addition, **macro precision (90.4%)** and **macro recall (92.0%)** suggest a good trade-off between false positives and false



(a) EAR-MAR Distribution

(b) Threshold Sensitivity

Fig. 4: (a) EAR/MAR-based labelling heuristics. (b) Effect of cosine similarity threshold on frame retention; $\theta = 0.85$ balances diversity and redundancy.

negatives, with slightly higher recall indicating sensitivity to drowsiness-related states—desirable for safety-critical applications. Table 1(b) details **class-wise F1-scores**. The model performs best on the **Not Drowsy** class (94.5%), followed by **Drowsy** (92.2%). The **Transient Drowsy** class (89.2%) is comparatively harder, which is expected given its intermediate nature and partial overlap with the two extremes. Nonetheless, the attention-enhanced temporal modelling helps disambiguate transient episodes from isolated blinks or brief posture changes, improving stability over short temporal windows. The **confusion matrix** in Fig. 5 corroborates these findings. Misclassifications are concentrated between *Transient Drowsy* and its neighbouring classes, while confusions between *Not Drowsy* and *Drowsy* are minimal. This pattern indicates that DRISTI captures the salient distinctions between stable states effectively and that residual errors primarily arise during gradual state transitions, a known challenge in fatigue detection. Receiver Operating Characteristic analysis (Fig. 6) further demonstrates separability, with **AUCs of 0.98** for *Not Drowsy*, **0.96** for *Drowsy*, and **0.92** for *Transient Drowsy*. High AUC values indicate strong ranking quality independent of the operating threshold, and the gap for the transient class reflects its intrinsic overlap. In deployment, threshold tuning could prioritise higher recall for *Transient/Drowsy* to favour early intervention.

Finally, training across multiple random seeds exhibits low variance (±0.7%), indicating **stability and reproducibility**. We attribute this to the combination of *cosine-similarity frame filtering*, which reduces redundancy and noise, and *attention-based temporal pooling*, which focuses the model on behaviourally informative segments.

Table 1: Performance of DRISTI on the Test Set

Metric	**Value (%)**
Accuracy	89.9
Macro F1-Score	91.2
Weighted F1-Score	93.5
Precision (Macro)	90.4
Recall (Macro)	92.0

(a) Overall Metrics

Class	**F1 Score (%)**
Not Drowsy	94.5
Transient Drowsy	89.2
Drowsy	92.2

(b) Class-wise F1 Scores

5.4. Ablation Studies

To further validate the design choices in DRISTI, we conducted ablation studies focusing on the spatial encoder and temporal modelling components. These experiments help disentangle the contributions of each module toward the overall system performance.

5.4.1. Spatial Encoder Comparison

Table 2 compares different spatial feature extractors. Traditional CNN-based models such as ResNet-18, ResNet-50, and VGG-19 performed competitively, with accuracies ranging from 82.1% to 85.6%. However, these models

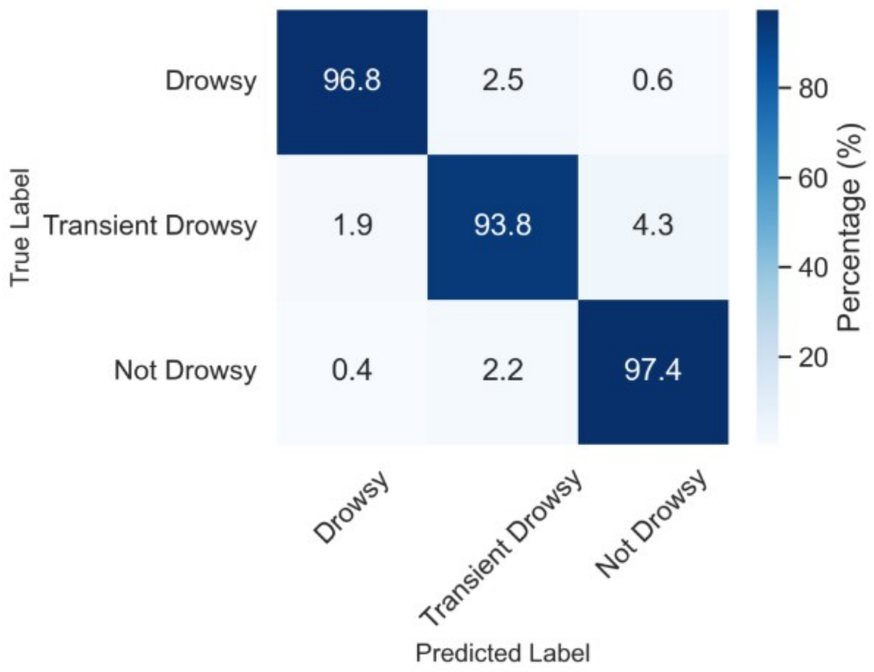


Fig. 5: Confusion matrix of DRISTI predictions.

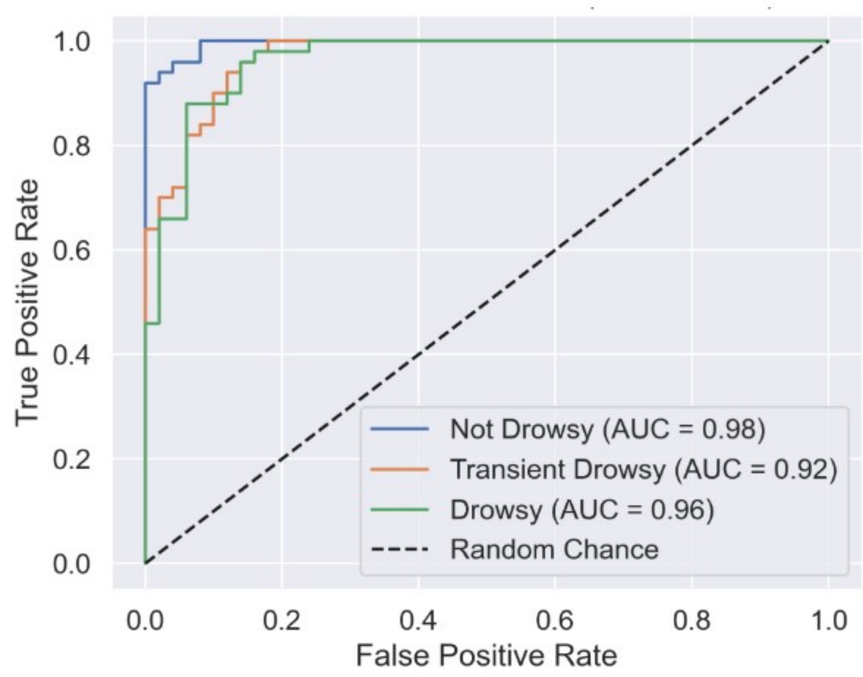


Fig. 6: ROC Curves and AUC values per class.

were more sensitive to illumination changes, head pose variations, and occlusions, which reduced their reliability in unconstrained settings. Lightweight architectures like MobileNetV2 and EfficientNet-B0 achieved greater parameter efficiency (3.4M–5.3M parameters) but sacrificed accuracy, falling below 85%. The Vision Transformer (ViT) demonstrated clear advantages by capturing global dependencies across facial regions, achieving 89.2% accuracy and 90.9% macro F1 even in its baseline form. Importantly, when fine-tuned on the curated drowsiness dataset, ViT improved further to 90.4% accuracy and 91.1% macro F1, while remaining more parameter-efficient than heavier CNNs like VGG-19. These findings highlight the value of attention-based spatial representation in handling diverse visual conditions.

Table 2: Spatial Encoder Ablation

Model	**Accuracy (%)**	**Macro F1 (%)**	**Params (M)**
ResNet-18	82.1	83.4	11.7
ResNet-50	85.6	86.8	25.6
VGG-19	84.3	85.1	39.1
EfficientNet-B0	84.3	87.2	5.3
MobileNetV2	80.6	82.7	3.4
ViT (baseline)	87.2	88.9	14.8
ViT (fine-tuned)	**88.4**	**90.1**	12.5

5.4.2. Temporal Modelling Impact

We next evaluated the effect of temporal modelling by comparing ViT-only with the full DRISTI model (ViT+LSTM). As shown in Table 3, the ViT-only model achieved slightly higher raw accuracy (90.4%) due to its strong ability to classify static frames. However, the macro F1-score of the full DRISTI framework was superior (91.2% vs. 91.1%), reflecting improved balance across all three drowsiness classes. This improvement underscores the importance of temporal reasoning: while ViT alone captures spatial features, it cannot represent behavioural dynamics such as prolonged eye closure or gradual yawning. By incorporating LSTM, DRISTI enhances robustness in real-world driving conditions where temporal context is essential. Thus, even with a marginal dip in raw accuracy, the combined ViT+LSTM design offers more consistent and reliable classification across sequences.

Table 3: Temporal Module Ablation

Model	**Accuracy (%)**	**Macro F1 (%)**
ViT-only	88.4	90.1
ViT + LSTM (DRISTI)	90.9	**91.2**

6. Conclusion

This work presented **DRISTI**, a spatio-temporal deep learning framework for real-time driver drowsiness detection in resource-constrained environments. By combining **Vision Transformer (ViT)**-based spatial encoding with **Long Short-Term Memory (LSTM)** temporal modelling, and enhancing it with a lightweight **additive attention mechanism**, DRISTI effectively captures both global facial patterns and fine-grained temporal cues such as prolonged eye closure, yawning, and gaze deviations. A notable contribution of this study is the introduction of a **three-class detection scheme**: **Not Drowsy, Transient Drowsy, and Drowsy**. This fine-grained formulation enables **early intervention** by detecting transitional states before full drowsiness, thus addressing a critical limitation of conventional binary models. The proposed annotation pipeline, leveraging **EAR-MAR thresholds** and cosine similarity-based frame selection, ensures semantically rich and computationally efficient training data. Experimental results demonstrate nearly **90% accuracy** and a macro F1-score above **91%**, validating robustness across diverse conditions. Future work will extend DRISTI to real-world in-vehicle deployments, addressing challenges such as **occlusions, accessories, driver-specific adaptation, and alert calibration** for long-term reliability.

References

[1] National Highway Traffic Safety Administration. Drowsy driving. https://www.nhtsa.gov/risky-driving/drowsy-driving, 2024. Accessed: 2025-08-20.

[2] National Transport Commission, Australia. Fatigue management: Road safety report. https://www.ntc.gov.au/regulatory-compliance/fatigue-management, 2021. Accessed: 2025-08-20.

[3] K. Jahanbakhsh, M. Ahadi, and H. Soltani. Drowsy driving and road traffic accidents in iran: A review study. *Journal of Injury and Violence Research*, 8(2):65–72, 2016.

[4] Central Road Research Institute (CRRI). Highway safety and driver fatigue study. https://www.crridom.gov.in/, 2023. Accessed: 2025-08-20.

[5] Muhammad Awais, Nasreen Badruddin, and Micheal Drieberg. Driver drowsiness detection using eeg power spectrum analysis. In *2014 IEEE Region 10 symposium*, pages 244–247. IEEE, 2014.

[6] Budi Thomas Jap, Sara Lal, Peter Fischer, and Evangelos Bekiaris. Using eeg spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2):2352–2359, 2009.

[7] Tereza Soukupova and Jan Cech. Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia*, volume 2, page 4, 2016.

[8] Mona Omidyeganeh, Shervin Shirmohammadi, Shabnam Abtahi, Aasim Khurshid, Muhammad Farhan, Jacob Scharcanski, Behnoosh Hariri, Daniel Laroche, and Luc Martel. Yawning detection using embedded smart cameras. *IEEE Transactions on Instrumentation and Measurement*, 65(3):570–582, 2016.

[9] Yanchao Dong, Zhencheng Hu, Keiichi Uchimura, and Nobuki Murayama. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE transactions on intelligent transportation systems*, 12(2):596–614, 2010.

[10] Wei Zhang, Yan Wang, and Jie Chen. Driver drowsiness detection using facial expression recognition and temporal analysis. *IEEE Access*, 8:124586–124595, 2020.

[11] Xiaoming Zhu, Wei Li, Yuxuan Zhao, and Yan Chen. Temporal attention-based drowsiness detection using facial dynamics. *Neural Computing and Applications*, 34(16):13421–13432, 2022.

[12] Jaemin Choi, Hyung Il Kim, and Sang Woo Lee. Attention-based recurrent neural network for driver drowsiness detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Jie Sun, Lei Wang, Min Chen, and Yifan Zhao. Vision transformer-gru hybrid network for driver drowsiness detection. *IEEE Transactions on Intelligent Vehicles*, 2023.

[15] Hao Li, Yiming Zhang, Wei Liu, and Xinyu Chen. Transferable vision transformer for driver fatigue detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2330–2339. IEEE, 2023.

[16] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. A realistic dataset and baseline temporal model for early drowsiness detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.