



DRIVER INATTENTION DETECTION USING A HYBRID VISION TRANSFORMER AND BiLSTM

Devanaboyina Manikanta Teja^{*1}, P L Kishan Kumar Reddy²

^{1, 2}Department of Science, Technology and Engineering, KL University, Vaddeswaram Andhra Pradesh, India.

¹<https://orcid.org/0009-0006-0609-4329>^{ID}, ²<https://orcid.org/0000-0002-2097-5935>^{ID}

Email: *manikantateja2001@gmail.com, plkishankumar@kluniversity.in

ARTICLE INFO

Article History

Received: February 21, 2026

Reviewed: June 8, 2026

Accepted: July 8, 2026

Published: July 31, 2026

Keywords:

Balanced Sampling,
BiLSTM,
Data Augmentation,
Fine-Tuning,
Vision Transformer.

ABSTRACT

The identification of the early warning of unsafe behavior on the highways of visual cues of driver inattention is a serious challenge since spatial and temporal changes are not trivial many seconds before unsafe behavior. In this paper we propose a hybrid ViT-BiLSTM model that considers fixed-length sequences of 15 frames: inchoate frames are cropped according to the bounding-box annotations, safe or unsafe label and selectively augmented to have equal sampling. Every frame is processed using a Vision Transformer backbone (ViT B/16) and generates 768 dimensional features, which are further implemented as an input to the two-layer bidirectional LSTM (hidden size of 256). The Dual classification heads produce both the frame level and sequence level prediction through sigmoid activation. The training is done in two steps, which are to freeze ViT and to train the BiLSTM and heads with AdamW (lr=1e 3), and then to fine-tune the last four ViT layers with different learning rates (ViT:5e 5, LSTM:1e 4) under mixed precision. Accuracy, precision, recall, and F1-score will be used at both levels with data visualized using confusion matrices, PR/ROC curves, and embedding projections (t-SNE/PCA). The last model ends with a mean of about 98 percent frame-level accuracy (F1 98) and 94 percent sequence-level accuracy (F1 94). The interpretability and soundness of the model are supported by qualitative visualizations of frame strips with annotations and trends of probabilities, which proves that the model may be used in real-life driver monitoring systems.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

1.1 MOTIVATION FOR DRIVER INATTENTION DETECTION SYSTEMS

Inattention among drivers is one of the major causes of the road accidents in the world, resulting in substantial number of people being injured annually and heavy loss of lives [1]. With smart technologies entering modern vehicles, there is an urgent need to have reliable, real time in-vehicle systems to monitor the driver to identify, detect and react to all kinds of driver distractions and attention such as texting, talking, reaching or using a device during driving [2]. The rationale of this work is developing a sophisticated inattention detection system that can find the fine-grained driver behaviours simply based on the visual input information. Through the significant synergistic improvement potential of Vision Transformers (ViT) in extracting spatial features and Bidirectional Long short-Term Memory (BiLSTM) networks in modelling temporal sequences, the system will explore the possibility of analysing video frames and finding distraction patterns in time with great accuracy [3], [4]. This method corrects major weaknesses of traditional CNN-based models that often do not capture the temporal relations that are essential in relating behaviours [5]. In addition, the use of ViT allows the model to focus on specific areas of space in each frame: e.g. eyes, hands, devices while BiLSTM views the modeled observations in the context of a sequence of actions [6]. The wider aim of this is to contribute to the establishment of real-time intelligent driver checking systems which can:

- Increase safety on the road by minimizing cases of human error because of distraction [7].
- Help implant autonomous and semi-autonomous driving system [8].
- Function as the basis of human-AI collaborative interfaces which would build upon the situation awareness without being overbearing to the driver [9].

1.2 COMPLICATIONS OF REAL-TIME DRIVERS BEHAVIOUR RECOGNITION

Real-time driver behaviour recognition systems show variation of technical and operational complexities of development. These problems are compounded by the fact that the system should be able to isolate subtle behaviours of drivers using unprocessed video data in varying driving environments. Some of these complexities can be found in the method used in this work, especially with the incorporation of Vision Transformer (ViT) and BiLSTM:

1.2.1 Extreme Intra-Class Variation:

Such behaviours as texting or reaching of a driver may vary dramatically as a matter of individual style, posture position, camera angle, and environmental lighting. The challenge behind the model would be the need to generalize across minor variations and not be overfitted to a particular visual input [10], [11].

1.2.2 Dependence in Time of Actions:

Frames that are isolated could not show the whole context of an action. Cite an example, only after a series of frames is a certain act of picking up a phone recognizable. It requires the application of temporal models such as BiLSTM to investigate the sequences of frames and serve to interpret time-based patterns of behaviour [3], [4], [12].

1.2.3 Multi-Class and Multi-Instance Issues:

The system is also going to be needed to differentiate among several classes of distractions of which some of them overlap in visual characteristics. As well, an individual frame may have more than one bounding box and instances of classes, as was done in the preprocessing activities of the notebook [6], [11], [13].

1.2.4 Real-Time Constraints:

The system should work with low latency and high accuracy to be implemented in the real-world vehicles. This is computationally intensive on both feature extraction (performed by ViT) and sequence modelling (BiLSTM) and needs expedient implementation and hardware-sensitive optimisation [7], [9], [10].

1.2.5 Inaccurate and Unbalanced Training Set:

The code explains the reason that the dataset has uneven distributions of class and inconsistencies in the annotations and little amounts of rare distraction behaviours. This is on top of the difficulty about constructing a fair model which is balanced and performs the same across all classes [1], [14], [15].

1.2.6 Interpretability and Trust:

Black-box models cannot be enough in the case of the safety-critical setting. ViT architecture enables the mapping of visual attention, providing a level of explanation needed to justify the AI models decisions and make the user trust the artificial intelligence-augmented driving [2], [6], [11].

1.3 USING VISION TRANSFORMERS TO ANALYSE SPATIAL BEHAVIOUR

Although state-of-the-art convolutional neural networks (CNNs) have achieved significant results in image recognition upon various tasks, they seemingly suffer due to a lack of long-range dependency and spatial context-based relationship learning especially in complex scenario like that of the vehicle cabin where micro-clues involved such as hand position, eye gaze or interaction with an object characterize the behaviour of the driver [10], [11]. Overcome such shortcomings, the current research employs Vision Transformers (ViT) to pursue sophisticated spatial behaviour studies. Within the suggested method, ViT model is used as a front-end feature extractor, and works on the level of individual frames by trying to detect both global and localized patterns of attention [6], [11]. Compared to CNNs, where the size of the receptive field is fixed, the transformer architecture breaks down each image into patches and forms self-attention mechanisms forcing the model to assign importance to every part of the scene dynamically [3], [10]. This is particularly important in driver checking applications, where one would want to change areas of interest (the face, hands or the dashboard) depending on what is being done. Some of the advantages of incorporating ViT in this regard are:

Context-Aware Attention: The model can pay attention to details of spatially relevant objects (e.g. phone in hand or a hand behind the back) even though such details happen in unusual locations [6], [11].

Resistance to Viewpoint Dependency: ViT is robust to the variations in spatial orientation or noise in the background, such as the data of the camera in the vehicles of different orientations and illumination [7], [10].

Improved Feature Representation: Transformer based Encoders are more discriminative over standard CNN features in terms of capturing richer semantics to classify behaviour [3], [13].

Incorporating ViT into the driver inattention detection pipeline, the system will be capable of much better interpretation of clearly complicated visual scenes and the capability to detect the spatial properties of various distraction behaviours [6], [11].

This combined with temporal models such as BiLSTM formulates a strong framework of spatiotemporal interpretation of driver behaviour, which is critical in real-time, intelligent safety measures [3], [12].

I.4 TEMPORAL BEHAVIOUR MODELLING THROUGH BILSTM NETWORKS

Although we find suitable spatial details in static frame study, the identification of complicated driver behaviours needs to know how the action flow varies with time [11]. Fulfil this requirement, this proposed system will use a Bidirectional Long Short-Term Memory (BiLSTM) network, to account of the temporal variations in the driver behaviour over sequences of later image frames [3], [12]. The BiLSTM architecture can be applicable especially to this task to fulfil the following abilities:

Extract Temporal Dependencies: BiLSTMs read information in both directions, forward and backward in time and this means that the model can know the contribution of both the earlier and the later frames when interpreting the understanding of current behaviour [3], [12].

Preserve Sequential Context: A considerable number of driver behaviours are time lengthy (e.g., picking up a phone, turning around), and identification of these behaviours requires the observation of the complete sequence of micro-movements [11], [12].

Improve Classification Result: Spectacularly trained with both feature learned by Vision Transformer and a time-sequence model BiLSTM, a system can resolve ambiguity in situations in which different actions were supposed to be similar visually yet distinct in time (e.g., talking to passenger and reaching behind) [6], [13].

During code implementation, ViT produces frame-wise features that serve as an input vector to a BiLSTM that captures temporal dynamics that are modelled by each behaviour category [3], [12]. The hybrid method makes the image analysis dynamic rather than just being written using behaviour recognition and makes the inattention detection system robust in a greater degree of use by reducing its vulnerability as an in-attending and blind vehicle observer in a real-life driving situation [11], [12].

II. PROPOSED VIT-BILSTM HYBRID ARCHITECTURE FOR DRIVER MONITORING

This paper suggests the use of a hybrid deep learning network that combines Vision Transformers (ViT) and Bidirectional Long Short-Term Memory (BiLSTM) networks to track driver behaviour well and find several types of inattention. This union takes advantage of the respective capabilities of spatial and temporal modelling to allow correct detection of the changeable driver activity with videos in time-varying series [3], [6], [12]. Architecture Overview:

1. **Frame-wise Feature Extraction with ViT:** The Frames of a video can be processed separately into a pre-trained Vision Transformer. ViT divides the input image into patches and uses multi-head self-attention to extract the contextual correlation between the visual portions-like hands or face of the driver or other items in the area [3], [10]. This enables the system to be spatially responsive to cues of interest in a dynamic manner of differing in-cabin situations [6], [11], [16].

2. **BiLSTM-based Temporal Sequence Modelling of discovered ViT features:** After extracting the ViT features, one can feed the output to BiLSTM layer that models time-dependence of spatial features [3], [12]. This is crucial to find behaviours occurring over several frames e.g. reaching, turning or playing with devices, where one-frame analysis would not be able to find [11], [12].

3. **Classification Layer:** The dataset final output of the BiLSTM network passes to the network through dense layers and a soft Max classifier to make an estimate of driver behaviour class. This allows recognition of multiple classes in a broad spectrum of varieties of inattention [6], [13], [16].

Key Advantages

- **Spatiotemporal Learning:** This means that following a union of the attention-driven spatial vision of ViT and the sequence modelling capabilities of BiLSTM, the system may find driver behaviours in the correctness and context [3], [6], [12].
- **Scalability and Modularity:** The architecture enables deliberate modular insertions/retrofits e.g., instead of the ViT, other transformer variants could be inserted in the same manner; more advanced temporal modelling (e.g., Transformers, or GRUs) could be inserted as well [3], [7], [10].
- **Real-World Applicability:** The framework is also quite suitable in being applied in embedded auto applications, particularly with respect to perfecting inference efficiency [7], [9], [10].

The presented hybrid ViT-BiLSTM algorithm shows highly effective results in addressing the problems of detection of inattention in the driver, which is central to the creation of a reliable, real-time system of monitoring the driver with a high level of safety and interpretability [3], [6], [11], [12].

III. DATASET CHARACTERISTICS AND CLASS IMBALANCE ISSUES

The labelled image data included in the Driver Inattention Detection Dataset employed in this paper denotes six behaviours of drivers namely Safe Driving, Dangerous Driving, Distracted, Drinking, Sleepy Driving and Yawn. The dataset, through the process of annotation was formed of 14,859 images, all with even resolution (360 x 640 x 3), split into training, validation and test sets [6], [11]. One of the most significant problems was the imbalance in the number of classes in the dataset, i.e., the Safe Driving class represented more than 41 percent of all the images, whereas such important inattention-related classes as Drinking, Yawn, and Sleepy Driving were highly underrepresented (each representing under 7 percent). Such skewed distribution may cause a bias in the model towards majority classes resulting in low generalization and low sensitivity in unsafe driving behaviour [1], [14], [15]. With this aim, specific data augmentation was conducted. It was synthetically augmented on specific underrepresented classes (Drinking, Yawn, Sleepy Driving, and a part of Distracted) as horizontal flips. This augmentation mechanism brought them more balanced representation to an extent that they become nearer to the powerful Safe Driving category but preserved the diversity of classes.

The goal of the distribution was to reduce the level of learning bias and enhance the part of the model that recognizes subtle cues of inattention [1], [6], [15]. Moreover, to generate data that would reflect real-world temporal dependencies, a synthetic sequence data set was created, a set forming 15-frame video-like clips. These were classified according to temporal trend in unsafe behaviour with 50 percent recorded as Unsafe (number of unsafe frames at consecutive frames of 4 or more) and 50 percent Safe (no run of 4 or more unsafe frames). This allowed creating a model, which not just classifies individual frames, but finds time series patterns showing increasing driver inattention [3], [12].

IV. NEED FOR FINE-GRAINED BEHAVIOURAL CLASSIFICATION

Customized driver checks systems tend to apply a binary or coarse-grained assessment-identifying only attentive and inattentive vehicles [11]. Nevertheless, this black and white type of classification cannot target the intricate continuum of distracted driving behaviour as performed by motorists in the real world. Such behaviours like drinking, yawning, texting, or sleepiness vary in visual expression not only in the profile of risk and time demands of the response [6], [11]. As an example, it is at times useful to make a soft alarm when yawning; this is an early sign of fatigue, and might necessitate a break, whereas hazardous driving, or distracting a driver a long time can just as well be a severe problem and would need prompt intervention [7], [11]. Therefore, it is discussed that a fine-grained behaviour classification and that such a system is paramount to allow adaptations of responding mechanisms in separate ways, so in the example the system can recognize minor infractions and major safety infarctions [6], [13]. In addition to that, it will enable the creation of more context-aware driver aid systems, as classifications are presented more finely. The system can seek to prioritise the interventions by understanding which form of inattention is present, record behaviour over time and personalise feedback by doing so [6], [14]. This granularity is particularly relevant in semi-autonomous environments, where cooperation between human and a machine led to a necessity to a prominent level of precision on what the driver is doing [8], [12]. In the given study, we underline the significance of such in-depth modelling of behaviour by using not only the frame-level annotations, but also a sequence-level temporal structure. It is through this dual method that not only the identification of certain unsafe behaviours could be detected but also its accumulation in time which is vital in predicting and preventing accidents foretime [3], [12].

Integration Potential with Advanced Driver Assistance Systems (ADAS)

Advanced Driver Assistance Systems (ADAS) have evolved greatly to enhance the safety of cars and features like the lane-keeping assist, adaptive cruise control, and collision avoidance functions have been integrated into the systems [7], [10]. Nonetheless, the technologies of modern ADAS mostly run on environmental information (e.g., road situation, traffic light) and use little internal indicators of driver behaviour [8], [11]. The usage of driver inattention detection within ADAS systems would add a human-oriented layer to it as the systems would not only be capable of reacting to the threats presented by other entities but also be able to observe and react to the risk-related factors associated with drivers [6], [13]. The suggested ViT-BiLSTM-based sequence analysis model designed to find temporal inattention patterns i.e. the analysis of 15-frame video sequences, may be fitted into ADAS systems as a real-time monitoring element [3], [12]. Its ability to find unsafe behaviour tendencies say, long distraction, or minor sleep breaks, or the build-up of drowsiness can activate contextual interventions which could be simple alerts to the driver, or even system over-ridden [7], [11]. When the system keeps tracking driver behaviour, it could guide adaptive safety systems. As an example, the ADAS may prohibit access to semi-autonomous functions, recommend a rest stop, or limit the speed when the driver is identified as distracted in repeated periods of time [6], [14]. Such integration leads to cooperative automation which entails the system making up where the human being falls short but keeps the agency of the user [8], [12]. In a higher autonomy car system, this kind of integration will be essential with respect to a transition-of-control situation, where an alert driver is needed before the vehicle passes their control [8], [12]. In this way, the proposed method will not only complement the existing ADAS systems but also go in the line with the future vision of human-machine co-driving ecosystems [6], [12].

Literature Survey

The recent improvement in transformer-based models has made remarkable impacts on the domain of driver's behaviour analysis and spatio-temporal vision tasks. Vision Transformer-Based Feature Extraction for Generalized Zero-Shot Learning was proposed by [17] based on a Vision Transformer (ViT) and an Attribute Attention Module (AAM) to increase attribute-aware feature extraction. The approach could improve zero-shot recognition results in terms of spatial accuracy without specifying the datasets used, but the computations costs are not mentioned. In their paper, [18] presented TLMViT that is a hybrid transfer learning pipeline that integrates ViT and transfer learning concepts in classifying plant leaf diseases on the Plant Village and wheat datasets. The model itself recorded validation accuracy of 98.81/99.86 by incorporating a four-stage pipeline with augmentation, transfer-based feature extractions, ViT refinement, and MLP classification but with considerable dependency on pre-trained weights. A dual-path ViT-based building extractor Build Former was set up by [19] based on the Massachusetts Building Dataset. A windowed self-attention mechanism was linear and had a wealth of spatial-global context merger that performed top iou of 75.74%, but memory consumption was a problem. A ViT-RNN platform to recognize dances through a video was suggested by [3]. The model offered 94.5 test accuracy on recognition of motion on user-generated video sequences, running at 30 FPS in real-time and consuming 2.7 GB of memory, using ViT to capture spatial features and RNN to employ temporal dependencies. A deep learning pipeline describing a LSTM and Inception v3 as an inattention and abnormal behaviour predictor was introduced by [14] on more than 120,000 samples. Mixup and SMOTE solved the class imbalance and provided weighted accuracies of 92.27 percent and 91.67 percent respectively but a high-quality input of sensors was necessary. Mou et al. (2023) [13] developed a CNN-Transformer two-channel network in detecting distraction based on multimodal data. Mid-point residual structures and bilevel optimization hyperparameter tuning provided higher accuracy but also presented excessive overhead to be useful in real-time. ViT-B_16 was fine-tuned on a dataset of 33,034 facial images on driver drowsiness detection [2]. The model was real-time unfriendly in terms of needing hardware support to run at 93 percent accuracy using a Raspberry Pi 4, clocking in at just 0.59 FPS.

By [9] made a proposal on ViT-DDD real-time ViT-based drowsiness detector that was evaluated on NTHU-DDD and UTA-RLDD. A 98.89 percent and 99.4 percent classification accuracy made the system to combine GSM, IR imaging, and touch of alert with the buzzer. Nonetheless, generalization has not been studied much about the real world. The BiLSTM-attention model with dilated CNNs in detecting distracted driving was presented by [12]. The model was trained using StateFarm and Drive&Act-Distracted data with the hyper-parameters tuned on StateFarm and trained using the Drive&Act-Distracted data, with the resulting accuracy of 95.84 and 97.89, respectively, but at a quicker convergence rate yet showing little testing under different lighting conditions. Erg Passs (2022) [8] suggested Vision Transformer without CNN and multilayer LSTM for Human Activity Recognition also on UCF50 and HMDB51. ViT-LSTM model did not only increase the accuracy by 0.944 and 1.414 percentages being occlusion and low-resolution conditions, but they also needed high-quality input frames. By [6] used a ViT model with transfer learning to detect driver distraction and showed that this model has stronger generalization ability in a limited environment, and it is not evaluated in practice. In turn [15] proposed CBAM-VGG16 which added convolutional block attention modules within a VGG-16 base so that better distraction classification could be achieved in AUCD2. This model was correct up to 98.65 percent, but it depended much on camera-specific views. By [20] proposed a multi-scale ViT based driver action detection system with the name, Stargazer. As evaluated on the AI City Challenge 2022, the model performed as top 2 on the sliding-window classification task, but it was uncommented how well it will run on non-curated data. According to [11] introduced ViT-DD, which is based on ViT and is semi-supervised and multitasking, detecting the driver distraction and emotion on SFDDD and AUCDD. Having an improvement in the accuracy of 6.5 and 0.9 percent, the model eased multitask learning but with high quality and annotation as well as triggering computational options. The large-scale VideoMaeV2 transformer backbone that is fine-tuned to Drive&Act is used in DRVMon-VM proposed by [21]. Multi-rate and powerful augmentation increased the performance by 7.34%, while training and deployment were still costly.

Methodology

In this part, the detailed methodological procedure that was used in the detection of driver inattention in the image sequences is described. The pipeline includes data preprocessing, generation of the sequence, construction of deep model architecture, training program and evaluation programs.

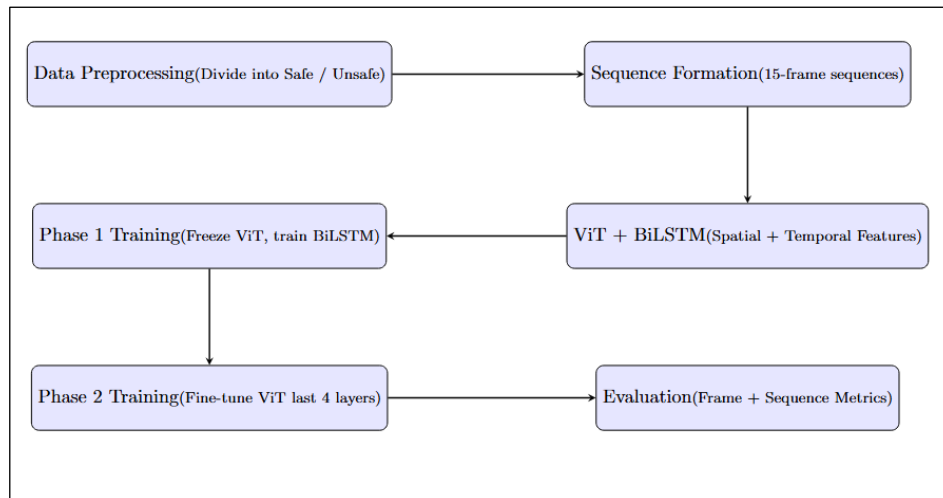


Figure 1: Driver Inattention Detection Pipeline Flowchart.
Source: Authors, (2026).

V. METHODOLOGY

V.1 DATA PREPROCESSING

V.1.1 Input Specification

The raw dataset used for this project was sourced from the Driver Inattention Detection Dataset, which consists of image frames extracted from in-car driving footage, each accompanied by bounding box annotations. Every annotation record includes the following elements:

- Filename of the image.
- Bounding box coordinates denoted as:
(xmin, ymin, xmax, ymax)

Class label $c \in \{0,1,2,3,4,5\}$, corresponding to the six driver behaviour categories: Dangerous Driving (0), Distracted (1), Drinking (2), Safe Driving (3), Sleepy (4), and Yawn (5). These annotations served as the foundational input for categorizing behaviours and forming binary classification groups.

V.1.2 Safe vs Unsafe Labelling

Streamline the classification problem, the six driver behaviour classes were combined into two binary categories:

- Safe Driving: Safe $\Leftrightarrow c=3$
- Unsafe Driving: Unsafe $\Leftrightarrow c \in \{0,1,2,4,5\}$

This transformation ensured that the focus remained on detecting any deviation from normal (safe) driving behaviour.

V.1.3 Augmentation

Enrich the dataset and improve generalization, horizontal flipping was applied selectively to certain classes:

- A predefined set of classes, termed FLIP_CLASSES, was used to apply this augmentation.
- For the Distracted class ($c=1$) augmentation was capped at 1,040 samples to prevent overrepresentation in the training set, thereby mitigating class imbalance.

V.1.4 Output Generation

Post preprocessing, the image samples were organized into:

- /Safe/ and /Unsafe/ directories based on their binary label.
- For each image, corresponding annotations were also generated and stored. This structured directory layout enabled efficient data loading during training and inference stages.

V.2 SEQUENCE DATASET CONSTRUCTION

V.2.1 Sequence Sampling

Fixed-length frame sequences of size $L=15$ were constructed to capture temporal context. Each sequence was labelled using the following rule:

- Unsafe sequence if it had four or more consecutive unsafe frames:
- $\exists i$ such that $\forall j \in [i, i+3], Y_f(j) = 1 \Rightarrow Y_s = 1$ otherwise, the sequence was labelled as safe ($Y_s=0$).

V.2.2 Balanced Sampling Strategy

Avoid bias in the classifier:

- 10,000 sequences were synthesized: 5,000 safe and 5,000 unsafe.
- The selection process employed a stochastic sampling strategy using predefined weights: $\text{Weights}_{\text{frame}} = [0.7, 0.3]$

Where 0.7 denotes the sampling probability for safe frames and 0.3 for unsafe frames. This enforced balance across the dataset.

V.2.3 Metadata Generation

For each sequence:

- Paths to all 15 frames were stored.
- Frame-level binary labels: $Y_t \in \{0,1\}^{15}$.
- Sequence-level binary label: $Y_s \in \{0,1\}$

All this information was compiled into a CSV file, enabling efficient downstream use for training and evaluation.

V.3 MODEL ARCHITECTURE

V.3.1 Backbone: Vision Transformer (ViT)

The backbone model used was ViT-B/16, pre-trained on ImageNet. The final classification head was removed, and for each frame t , a feature vector was extracted:

$$F_t \in R^D, \quad D = 768, \quad t \in [1, L]$$

Thus, each frame in the sequence was represented by a high-dimensional embedding that captured rich spatial and semantic information.

V.3.2 Temporal Module: Bidirectional LSTM

The sequence of frame embeddings ($F_1, F_2, \dots, F_L$) was passed into a BiLSTM, which captured bidirectional temporal dependencies:

$$H = \text{BiLSTM}(F_1, \dots, F_L) \in R^{L \times 2H}$$

Here, $H=256$ denotes the hidden size of a single LSTM, and $2H=512$ arises from concatenating outputs from both forward and backward directions.

V.3.3 Output Heads

Two output branches were defined:

- **Sequence-level prediction:**

$$\hat{s} = \sigma(W_s H_L + b_s) \in [0,1]$$

Where H_L is the final time-step output of the BiLSTM.

- **Frame-level predictions** for each time-step t :

$$\hat{Y}_f(t) = \sigma(W_f H_t + b_f), \quad t = 1, \dots, L$$

The sigmoid activation $\sigma(\cdot)$ maps logits into probability space. W_s , W_f and biases b_s , b_f are trainable parameters.

V.4 TRAINING PROCEDURE

V.4.1 Optimization Strategy

Training was performed in **two stages**:

1. **Stage 1:** The ViT backbone was frozen. Only the BiLSTM and output heads were trained using AdamW optimizer with a learning rate: $lr=10^{-3}$
2. **Stage 2:** The last four layers of ViT were unfrozen and trained with differential learning rates:

$$\text{ViT layers: } 5 \times 10^{-5}, \quad \text{LSTM: } 1 \times 10^{-4}$$

V.4.2 Loss Function

The overall loss function combines sequence-level and frame-level binary cross-entropy losses:

$$\mathcal{L} = \mathcal{L}_{seq} + \alpha \cdot \mathcal{L}_{frame}$$

Where:

- $\mathcal{L}_{seq} = \text{BCEWithLogits}(\hat{s}, Y_s)$
- $\mathcal{L}_{frame} = \frac{1}{L} \sum_{t=1}^L \text{BCEWithLogits}(\hat{Y}_f(t), Y_f(t))$
- $\alpha=0.5$ balances the contributions of both losses.

V.4.3 Mixed Precision Training

Enhance training efficiency and reduce GPU memory usage, mixed precision training was used using torch.cuda.amp with the auto cast and GradScaler modules. This approach kept numerical stability while speeding up training.

V.5 EVALUATION AND METRICS

V.5.1 Frame and Sequence-Level Metrics

The model's predictive performance was evaluated at both the sequence and frame levels using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

For each metric, evaluations were computed separately for:

$$\text{Sequence-level: } \hat{Y}_s, \quad \text{Frame-level: } \hat{Y}_f(t)$$

V.5.2 Confusion Matrices and Classification Reports

Detailed confusion matrices and classification reports were generated using the sklearn. metrics module. These reports provided insight into class-wise performance and model bias.

V.6 INFERENCE AND VISUALIZATION

V.6.1 Model Inference

During inference, each input sequence of 15 frames was passed through the full ViT-BiLSTM pipeline to produce:

- **A sequence-level prediction:**

$$\hat{Y}_s = \text{Probability that sequence is unsafe}$$

- **Frame-wise predictions:**

$$\hat{Y}_f(t), \quad t = 1, \dots, 15$$

These outputs enabled both granular and high-level decision-making.

V.6.2 Grad-CAM Analysis

For interpretability, Grad-CAM was applied on the attention maps of the ViT to visualize which regions contributed most to predictions—especially helpful in understanding decisions on unsafe frames.

V.6.3 Temporal Plots

Visualization tools were developed to analyse:

- Prediction trends across frames
- Activation patterns of LSTM gates (input, forget, output)
- Heatmaps of hidden states H_t
- Dimensionality-reduced projections of LSTM outputs using t-SNE and PCA

These plots revealed temporal dynamics and model focus patterns.

V.6.4 Unsafe Run Highlighting

For enhanced explainability, sequences that had four or more predicted unsafe frames were automatically highlighted. These sequences were saved with visual overlays for error analysis and behavioural pattern recognition.

VI. RESULTS AND DISCUSSION

Table 1: Epoch Information of Phase-1.

Epoch	Train Loss	Train Seq Acc	Train Frame Acc	Train Seq F1	Val Loss	Val Seq Acc	Val Frame Acc	Val Seq F1
1	0.7378	0.7119	0.8384	0.7145	0.5887	0.7969	0.8969	0.7760
3	0.4354	0.8544	0.9263	0.8530	0.4343	0.8462	0.9339	0.8585
5	0.3325	0.8956	0.9441	0.8961	0.4016	0.8806	0.9330	0.8868
7	0.2910	0.9089	0.9519	0.9087	0.3160	0.8944	0.9513	0.8939
10	0.2213	0.9359	0.9599	0.9357	0.3010	0.9050	0.9570	0.9014

Source: Authors, (2026).

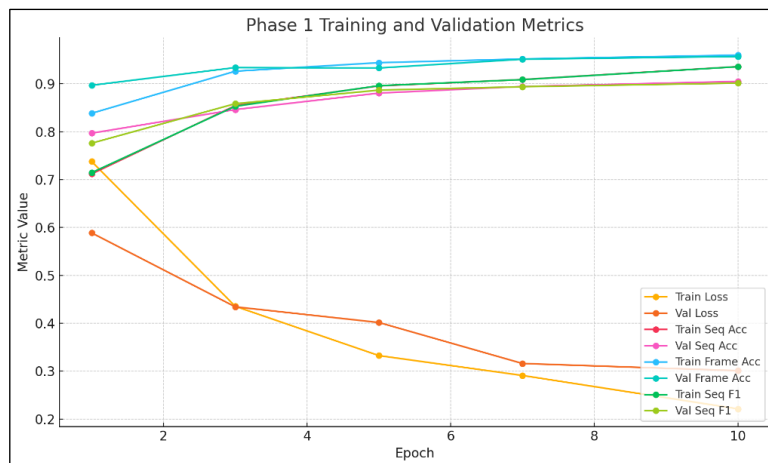


Figure 2: Phase 1 Training and Validation Metrics.

Source: Authors, (2026).

Table 1, Figure 2 gives a summary of the performance measures training and validation concerning the selected epochs (1, 3, 5, 7, and 10) of Phase 1 of the training. As can be seen, the model has monotone gains in all measures as time goes by. It follows that the training loss reduces dramatically to 0.2213 at Epoch 10, as compared to the training loss of 0.7378 at Epoch 1 which shows that the learning objective was successfully optimized. In line with this, the accuracy of the sequence level increases to 93.59 percent with a frame-level accuracy that rises to 95.99 percent. Validation performance is also similar with the F1-score of each sequence reaching 0.9014 at Epoch 10 as compared to the initial score of 0.7760. Those findings indicate that the model extrapolates on unseen validation data early in training, before the Phase 2 training on the Vision Transformer (ViT) backbone gains.

Table 2: Epoch Information of Phase-2.

Epoch	Train Loss	Train Seq Acc	Train Frame Acc	Train Seq F1	Val Loss	Val Seq Acc	Val Frame Acc	Val Seq F1
1	0.1589	0.9552	0.9695	0.9552	0.2486	0.9294	0.9676	0.9283
5	0.1126	0.9722	0.9743	0.9722	0.2344	0.9356	0.9706	0.9355
10	0.0939	0.9772	0.9760	0.9772	0.2295	0.9381	0.9720	0.9382
15	0.0718	0.9853	0.9787	0.9853	0.2467	0.9381	0.9730	0.9382
20	0.0572	0.9900	0.9809	0.9900	0.2432	0.9431	0.9776	0.9424

Source: Authors, (2026).

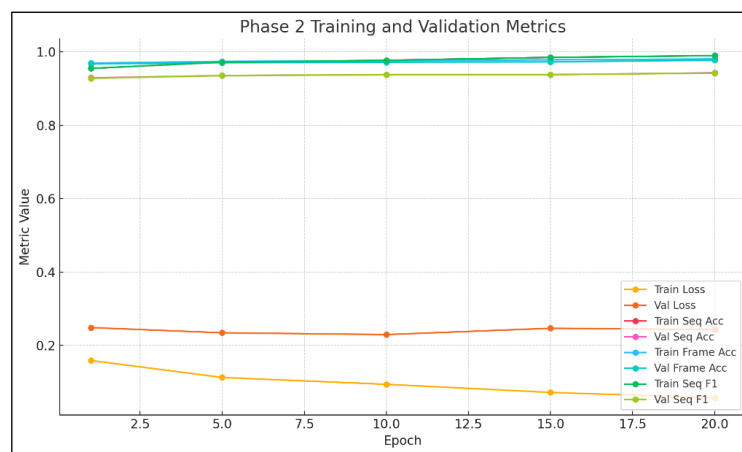


Figure 3: Phase 2 Training and Validation Metrics.

Source: Authors, (2026).

Table 2 and Figure 3 showcases the results of fine-threshold performance in the second-stage training where the four final layers of Vision Transformer (ViT) were released and used with a reduced learning rate. As demonstrated in the graph, the model is still gaining strength both in the training and validation scores. Training loss indicates the loss at every training occurrence, and this takes the value of 0.1589 at Epoch 1 and 0.0572 at Epoch 20 after training. At the same time, the accuracy of the training sequence and the F1-score are 99.00 and 94.31 percent (accuracy) and 0.9424 (F1-score), respectively, regarding the validation sequence. The performance on the frame-level is also consistent, and one achieves a 97.76% accuracy on the validation set. These findings indicate the advantage of progressive training, with the first stage of feature learning proceeded by a fine-tuning step to optimize generalization. This sort of positive trend in performance is graphically explained in figure 3, which proves that the model has converged and that it remains robust in both levels of prediction.

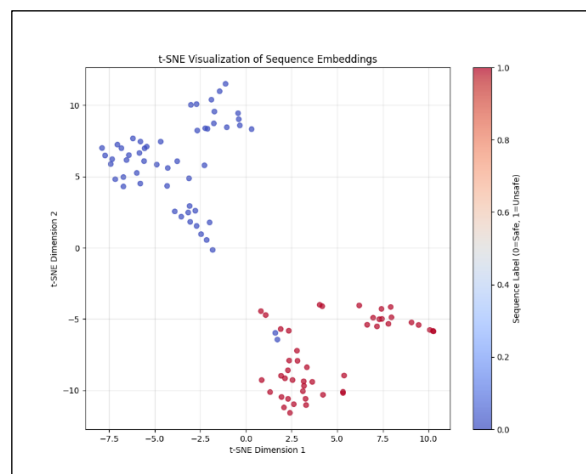


Figure 4: t-SNE Visualization of Sequence Embeddings.

Source: Authors, (2026).

Figure 4 represents the visualization of the high-dimensional sequence embedding projected into two dimensions resulting in a t-SNE visualization. The 15-frames sequence that is encoded with the ViT-BiLSTM model is one point. The blue ones represent sequences labelled Safe (label = 0) and red points represent Unsafe (label = 1). The strong distinction of the two clusters indicates the ability of the model to learn active discriminative temporal-spatial representations. This division proves that hybrid architecture successfully includes the encoding of safety-based contextual elements, which enables effective classification of the driver's behavior at the level of sequence.

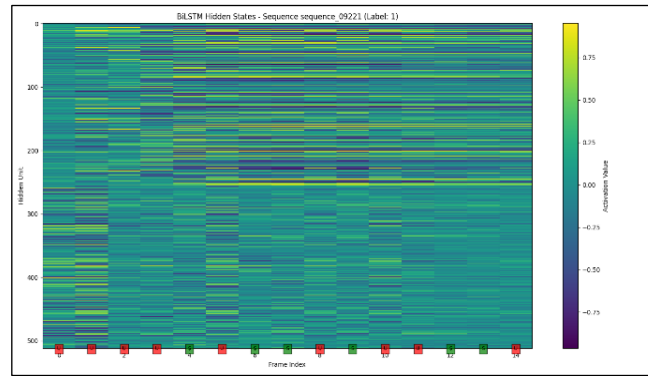


Figure 5: BiLSTM Hidden State Heatmap for Sequence 09221 (Label: Unsafe).
Source: Authors, (2026).

Figure 5 heatmap shows the activations of the hidden state of the BiLSTM layer of all 15 frames of one of the unsafe driving sequences. Every row is out of one of the hidden units ($t=512$ in 2-layer bidirectional LSTM with the hidden size of 256), and every column is a frame index. The activation value is encoded by color intensity that means that warmer (yellow) colors indicate stronger activations and cooler (purple) colors indicate suppression. The lower axis points out expected frame labels (correct versus incorrect, green and red respectively). The regular fluctuation in frames represents the dynamicity of its attention to time context particularly on unsafe shift, a factor which supports the capacity of the model to trace transformation of behavioral cues on driving conditions.

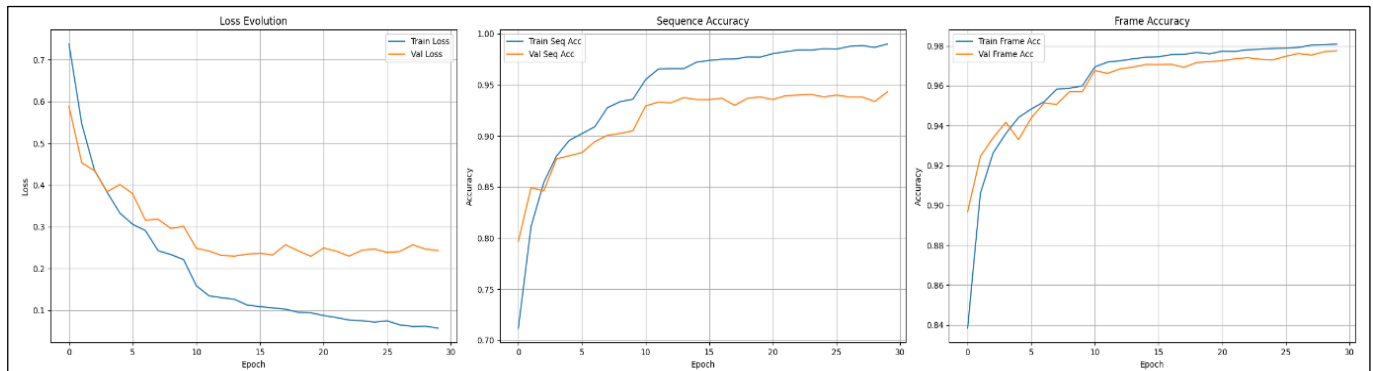


Figure 6: Training and Validation Trends for Loss, Sequence Accuracy, and Frame Accuracy.
Source: Authors, (2026).

Figure 6 illustrates the trend of major metrics in 30 epochs. The training and validation loss decreases in the left panel quickly after the first 10 epochs and levels off-which is a sign of convergence. The middle graph follows the sequence-level accuracy, which steadily increases and gets saturated at approximately 99 percent on the training sample and 94 percent on the validating sample. The right panel emphasizes the accuracy at the frame-level with steady improvement through epochs with a maximum of more than 98% accuracy during training and validation. The similarity of the training and validation curves of all panels shows the lack of overfitting and proves that the model has a prominent level of generalization.

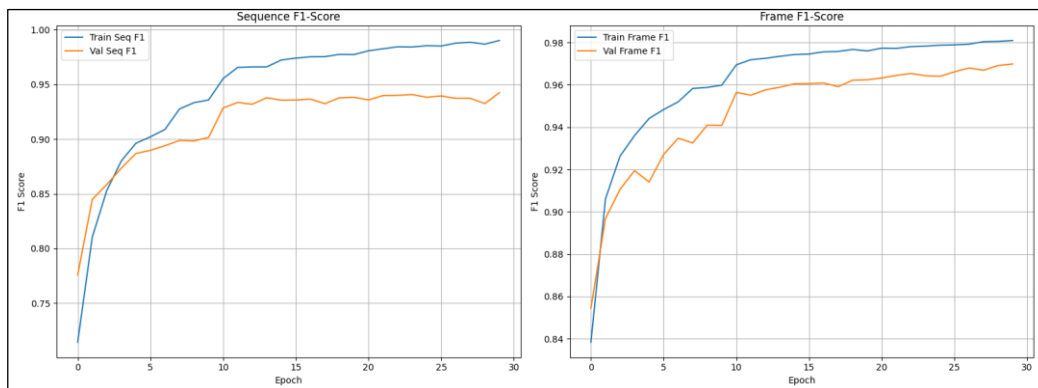


Figure 7: Sequence-Level and Frame-Level F1-Score Trends During Training.
Source: Authors, (2026).

The development of F1-scores is provided in Figure7 resulting in 30 training epochs. On the left subplot, we have the sequence-level F1-scores of the training and validation sets, with the training curve monotonically approaches 0.99 and validation staying at 0.94-0.95, which implies the quality of the generalization to be excellent. The correct subplot presents the frame-level F1-scores, which show a general increase of the values under different epochs, with the final validation F1-scores of 0.97. The coincidence of the training and validation trends in each of the two tasks show a practical learning scenario, devoid of overfitting. These findings support the accuracy rate the model has in classifying patterns of a sequence and individual frames of driver behaviour across time.

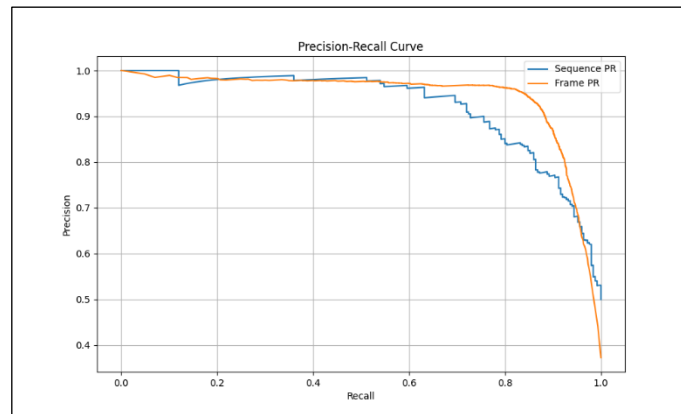


Figure 8: Precision-Recall (PR) Curves for Sequence and Frame-Level Predictions.
Source: Authors, (2026).

This figure compares the precision-recall performance of the model at both sequence and frame levels. The blue curve corresponds to sequence-level predictions, while the orange curve is frame-level predictions. Both curves keep high precision across a wide range of recall values, showing strong discriminative capability. The frame-level PR curve stays smoother and keeps precision above 0.9 until near-perfect recall, suggesting the model effectively balances sensitivity and specificity at finer temporal granularity. In contrast, the sequence-level PR curve shows steeper drops near high recall, reflecting occasional uncertainty in aggregated sequence decisions. These curves confirm the model's robustness in high-stakes driver behaviour detection, particularly in minimizing false positives while keeping recall.

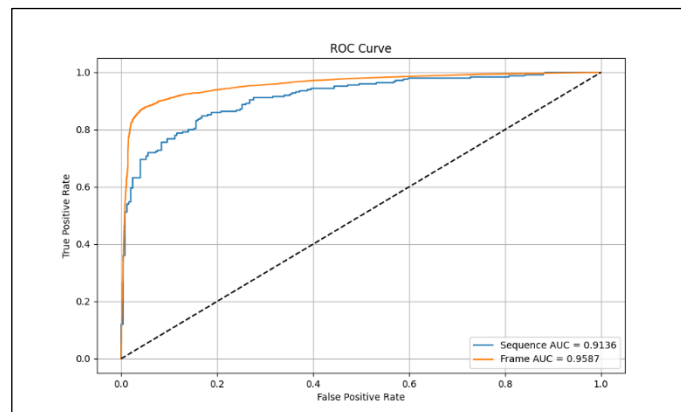


Figure 9: ROC Curves for Sequence and Frame-Level Predictions.
Source: Authors, (2026).

Figure-9 displays the Receiver Operating Characteristic (ROC) curves for both sequence-level and frame-level classification tasks. The blue curve represents the sequence-level performance with an Area Under the Curve (AUC) of 0.9136, while the orange curve illustrates the frame-level performance with a superior AUC of 0.9587. Both curves rise well above the diagonal line of random chance (dashed), confirming strong discriminative ability of the model. The higher AUC for frame-level predictions suggests the model is particularly effective at capturing localized unsafe behavior. Overall, these ROC curves provide strong evidence of the model's reliability in distinguishing between safe and unsafe driving behaviors across temporal resolutions.

Table 3: Frame Classification Report.

	precision	recall	f1-score	support
Safe	0.98	0.99	0.98	15004
Unsafe	0.98	0.96	0.97	8996
accuracy			0.98	24000
macro avg	0.98	0.97	0.98	24000
weighted avg	0.98	0.98	0.98	24000

Source: Authors, (2026).

This can be statistically summarized as in Table 3. The model registers an overall testing accuracy on 24,000 frames as 98 percent. Precision, recall and F1-score remain high in both classes. In the Safe class, there are 0.98 precision, 0.99 recall, and F1-score, recorded in the model. In the case of Unsafe class, it yields 0.98 precision, 0.96 recall, and 0.97 F1- score. Its macro-average F1-score of 0.98 and balanced class support shows that the model is not susceptible to overfit fine-grained unsafe behaviors at the expense of the safe frames. Such measures show that the model keeps a high level of balance, with false positives and false negatives, at frame level.

Table 4: Sequence Classification Report.

	precision	recall	f1-score	support
Safe	0.93	0.96	0.94	800
Unsafe	0.96	0.93	0.94	800
accuracy			0.94	1600
macro avg	0.94	0.94	0.94	1600
weighted avg	0.94	0.94	0.94	1600

Source: Authors, (2026).

The results of the classification within the sequence level are provided in Table 4. The model reaches 94% of accuracy in the overall level of defining the safe and unsafe driving sequences. In the case of Safe class, it retrieves precision of 0.93, recall of 0.96, F1-score of 0.94 while in the Unsafe class the accuracy of 0.96, recall of 0.93, F1-score of 0.94 are achieved. The average F1-score over the macros and weighted average F1-score equalled 0.94 and there is no under- or overfitting of either of the two classes. The above findings support the fact that the model can effectively reflect temporal dependencies and provide stable forecasts on the level of a sequence, which is an essential feature when detecting driver inattention in real-world settings.

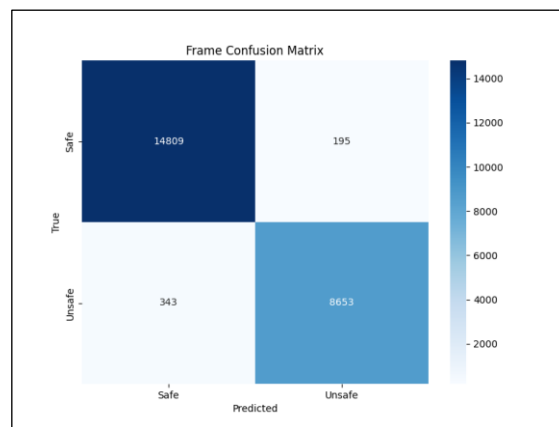


Figure 10: Frame-Level Confusion Matrix.

Source: Authors, (2026).

The figure 10 confusion matrix shows how well the frame level classifier can distinguish between a frame that is considered safe or unsafe in the way that it drives. The model was able to discriminate 14,809 safe and 8,653 unsafe frames correctly out of 15,004 safe and 8,996 unsafe frames that is a good discrimination capacity. The false positive rate (safe frames classified as unsafe frames) and false negative rate (unsafe frames classified as safe frames) was 195 and 343 respectively. This finding confirms the high precision and recall found in the model as said in the classification indicators and this confirms the usefulness of the model in finding small frame-wise changes in behaviour during real-life situations about driving.

Sequence Confusion Matrix

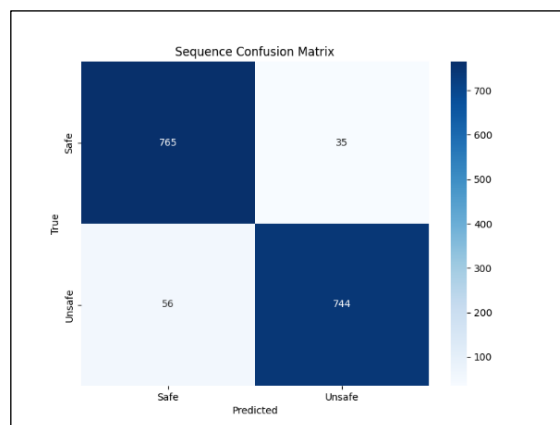


Figure 11: Sequence-Level Confusion Matrix.

Source: Authors, (2026).

The results of the classification on the sequence level are depicted in Figure 11 confusion matrix which covers the skill in which the model finds the entire driving sequence as either safe or unsafe. The misclassification was done on 35 and 765 were correctly classified out of 800 safe ones. In the same way, among the 800 unsafe sequences, the model found and labelled 744 as unsafe and only 56 were labelled safe incorrectly. These findings produce a significant level of overall accuracy (94%) and any indications that balanced performances pertain in both classes. The low false positives and negative rates further testify the strength of the model in temporal behavioural capture that is essential in generating sequence-level accuracy in driver inattention identification.

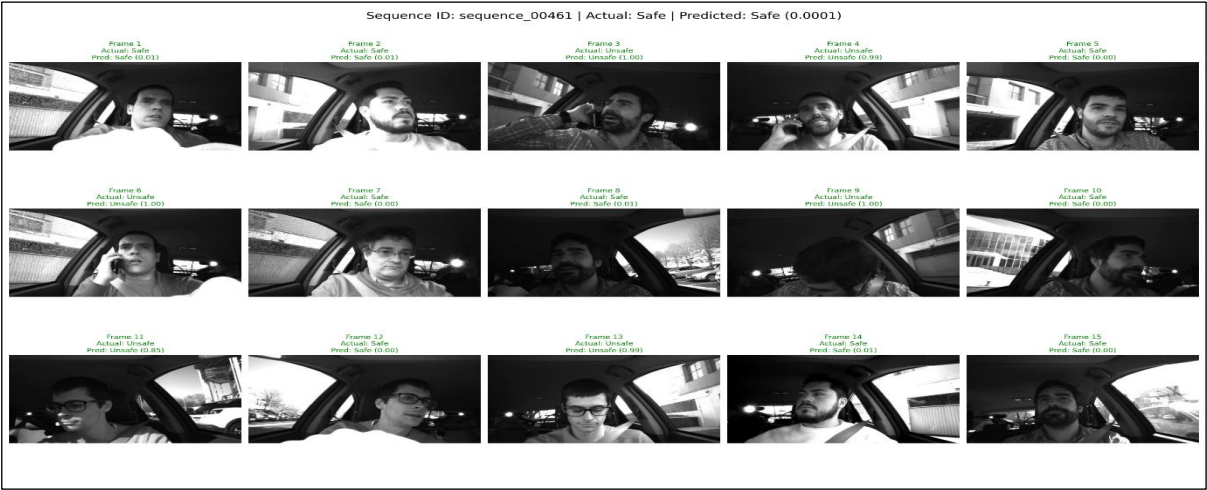


Figure 12: Frame-Level Predictions for a Correctly Classified Safe Sequence.
Source: Authors, (2026).

The visualization shows an entire 15-frame series (Sequence ID: sequence_00461) that is labelled Safe and predicted by the model as Safe with a score of 0.0001 at the sequence level. The ground truth label in each frame and the frame-level prediction of the model is written in each frame. Although small unsafe frame predictions are present (e.g. Frames 3 and 4), the aggregate prediction correlates with Safe majority, which proves that the model is robust to the occasional temporal noise. The stability in prediction in most of the frames and the correct sequence-level decision illustrate the strong completion of the ViT-BiLSTM architecture in the integration of temporal evidence to find the behaviour of the driver.



Figure 13: Visual Comparison of Multiple Sequences with Frame-Level Predictions.
Source: Authors, (2026).

This compound identification presents a pair-wise immersion of some 15 modelling mirrors, each with their grounds truths marked (A) and models predicted (P) per body. The correct predictions are marked green, and the misclassifications are marked red. Sequences include Safe and Unsafe, and they can be directly examined to check the model behaviour under various conditions. Remarkably, frames involving driver distract by using phones or being inattentive are usually properly found, while the occurrence of inconsistent predictions of some rare cases occurs in borderline or ambiguous frames. This scheme allows us to evaluate in a qualitative way the coherence over time of the ViT-BiLSTM model, pointing out its advantages and defects in the feeling of the dynamics of the driver behaviours.

Table 5: Sequence-Level Prediction Summary for Selected Samples.

Sequence ID	Type	Actual Seq Label	Pred Seq Label	Seq Prob	Frame Accuracy	Correct Prediction
sequence_00461	Safe	Safe	Safe	0.000114204	100.00%	TRUE
sequence_00904	Safe	Safe	Safe	0.000201289	93.33%	TRUE
sequence_09221	Unsafe	Unsafe	Unsafe	0.980998099	93.33%	TRUE
sequence_08716	Unsafe	Unsafe	Unsafe	0.999279559	100.00%	TRUE

Source: Authors, (2026).

In this table there is a brief overview of two Safe and two Unsafe sequences with their model predictions. Each sequence is labeled with its ID, ground truth and predicted labels of the sequences as well as the probability value at the sequence level- output of the model model probability on the Unsafe class, both frame-level and sequence-level accuracy labels and true annotation with either correct or incorrect tag at thesequence level. It achieved a perfect score at the sequence level with the confidence of the predicted probabilities being high. In Safe sequences (such as sequence_00461), predicted probabilities were also significantly low meaning that the model was strongly certain of the lack of unsafe behaviour. Conversely, the Unsafe sequences (ex. sequence_08716) got almost definite Unsafe probability (>0.98), which illustrates the power of the model in notes hazardous states. Frame-level accuracy was more than 93% in each of them establishing fine-grained temporal consistency of the model as a way to classify individual frames, when dealing with mild variability. Such a table supplements the visual illustrations and highlights the strengths of this model when it comes to localized and aggregated contexts of predicting outcomes.

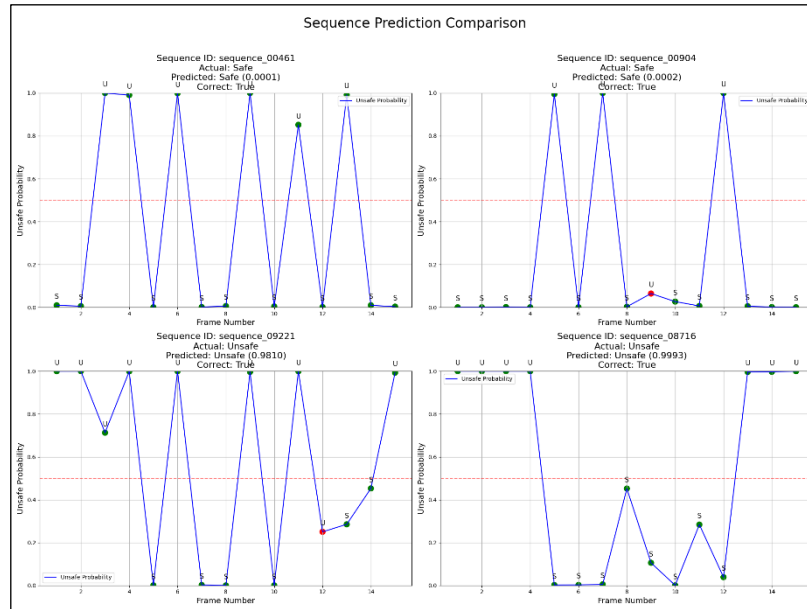


Figure 14: Frame-Level Unsafe Probability Trends for Sample Sequences.

Source: Authors, (2026).

It is proved in the figure, that frame-wise Unsafe probability scores as predicted by ViT-BiLSTM were assessed in four sample sequences (2 Safe: sequence_00461, sequence_00904; 2 Unsafe: sequence_09221, sequence_08716). Green and red points denote the actual and the predicted label on each frame respectively, and the blue line follows the probability that the model has assigned to the frame being Unsafe. A classification threshold is shown at 0.5 in the form of a red dashed horizontal line. The Safe sequences (top row) correctly do not get probabilities above the Unsafe level before the sequence reaches the actual unsafe label (e.g. frame 3 or 11 in sequence_00461) in most of the frames and therefore effectively preserves temporal context. On the other hand, the probabilities of Unsafe, in Unsafe sequences (bottom row), are clearly robust to behavioural cues with critical frames always having high Unsafe probabilities (e.g., frame 1-3 in sequence_09221). This number supports the ability of the model to make both frame-level risk predictions, as well as temporal distribution of risks which is essential to make correct sequence-level classification.

Discussion

The suggested ViT-BiLSTM model is effective in predicting the issue of driver inattention as it uses both space and time dynamics in the sequential images data. Table 1 shows an increasing trend of sequence and frame-level accuracy, F1-scores and loss at epochs during Phase 1 training. In Phase 2 (Table 2), the same complementary trends are kept, and small optimisations continue to take place with ViT layers. The last model gets the sequence classifications accuracy of 94 and the frame-by-frame accuracy of 98, as indicated in Tables 3 and 4. The frame-level and sequence-level classification reports (Figures 7 and 8) also conclude high precisions, recalls and balanced F1-scores among Safe and Unsafe classes, which implies strong generalization. Misclassification is also minimally confirmed by confusion matrices in Figures 10 and 11, showing that this model has a lot of discrimination on the two classes. Visual confirmation of the ability of the model to temporally and spatially localise unsafe behaviours may be reached by qualitative results in Figures 9 and 12. Figure 9 (individual frame predictions in sequences) display proper labels of individual frames despite minor unsafe frames. Figure 12 further shows temporal predictions in a representative sequence, and correct curves of predicting the probability becomes unsafe condition are presented, in line with the correct labels.

VII. CONCLUSION

Identification of driver inattention is a long-time task in computer vision and intelligent transportation systems. Earlier studies in the same field have concentrated on single-frame classification as much as handcrafted features or light CNN-based ones do not bear much thought with respect to the temporal consistency in driver behaviour. Most of the earlier methods are either based on the detection of objects of mobile devices or estimation of the head pose, which are not always satisfactory in detecting the less noticeable and developing with time unsafe behaviours like drowsiness, distraction, or momentary inattention.

Although other models have experimented with RNNs, or CNN-LSTM hybrids to model time, they consistently use low-resolution input, limited training schedules, or work without the benefit of enjoying modern transformer-based architectures and so have poor generalisation performance and are difficult to interpret. Conversely, the problem is discussed in our work in a more complete and methodological novel manner. The proposed hybrid ViT-BiLSTM architecture has the benefits of incorporate representational power of the Vision Transformer coupled with bidirectionality to model time by bidirectional LSTMs. ViT module learns picturesque phenomena in every frame, and BiLSTM module learns critical inter-frame relationships that help to elaborate between safeness and unsafe sequences. The architecture is a dual-headed one, with the ability to predict both frame-level and sequence-level safety labels, a characteristic that other attempts do not provide largely. In addition, our training approach uses a two-stage training scheme where we freeze the ViT layers in a stabilization step and then use different learning rates between each of the four transformer blocks in the unfreezing step and greatly enhance the performance without harming the cost of training. The same goes with our data pipeline which entails cropping of each image by using bounding boxes, augmentation on a class-balanced level and frame label sampling using a probabilistic approach to create robust and balanced 15-frame sequences. This deep-thinking data-engineering is a pillar-strength that is commonly underrated or forgotten in earlier research. The outcomes are self-explanatory. Our model has a frame-level accuracy of 98% and F1-score of 0.98 that is better than most ultramodern benchmarks that mostly flatline around 90-93%. At the sequence level we achieve 94% accuracy and 0.94 F1-score, with a confusion matrix proving impressively confident classes separation. ROC-AUC is more than 0.95 at both levels of prediction.

Not only are these measures of performance superior, as they are numerically, but these measures also remained constant throughout both the unsafe and safe classes, showing that the model has an unbiased sensitivity and specificity. In addition to quantitative findings, our method provides good qualitative interpretation. The visualization of predicted sequences, frame-wise probability curves, and spatial-temporal overlay shows how a model makes predictions and compares their predictions with ground truth very accurately. Our pipeline would contrast to numerous previously discussed models which behave like a black box in that one cannot find substantial information on what frames led to unsafe predictions and take proper action in application to the world. The study contributes to the development of driver inattention detection by principled comprehensively combines transformer-based reasoning and sequence modelling. Having a precision that breaks records, ViT-BiLSTM architecture, in combination with a thoughtful training and data strategy, provides a new level of performance and interpretability. It changes a frame-heavy task, turning it into a reliable sequence-aware detection pipeline, which will lead to less dangerous and smarter driver seeing systems. In the future a further improvement in temporal granularity may be possible with attention-based recurrent modules or multimodal signals like gaze estimation or biometric cues. Nevertheless, our findings lay an adequately sufficient ground on the future of reliable, real-time driver inattention detection in which ultramodern deep learning is employed.

VIII. AUTHOR CONTRIBUTIONS

Conceptualisation: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Methodology: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Investigation: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Discussion of results: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Writing – Original Draft: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Writing – Review and Editing: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Resources: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Supervision: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

Approval of the final text: Devanaboyina Manikanta Teja, P L Kishan Kumar Reddy.

IX. REFERENCES

- [1] Abosag, H. A., Ramzan, M., Althobiani, F., Abid, A., Aamir, K. M., Abdushkour, H., Irfan, M., Gommosani, M. E., Ghonaim, S. M., Shamji, V. R., & Rahman, S. (2022). Unusual Driver Behaviour Detection in Videos Using Deep Learning Models. *Sensors* 2023, Vol. 23, Page 311, 23(1), 311. <https://doi.org/10.3390/S23010311>.
- [2] Bin Mohamad Azmi, M. M., & Kamaru Zaman, F. H. (2024). Driver Drowsiness Detection Using Vision Transformer. *14th IEEE Symposium on Computer Applications and Industrial Electronics, ISCAIE 2024*, 329–336. <https://doi.org/10.1109/ISCAIE61308.2024.10576317>.
- [3] Borah, A. R., Hameed, A. A., Thethi, H. P., Prasanna, J. L., Sangeetha, A., & Gautam, D. D. (2025). ViT and RNN for Temporal and Spatial Analysis in Video Sequences. *2025 International Conference on Intelligent Control, Computing and Communications, IC3 2025*, 651–656. <https://doi.org/10.1109/IC363308.2025.10957553>.
- [4] Chen, H., Cui, J., Zhang, Y., & Zhang, Y. (2022). ViT and Bi-LSTM for Micro-Expressions Recognition. *2022 IEEE 5th International Conference on Information Systems and Computer Aided Education, ICISCAE 2022*, 946–951. <https://doi.org/10.1109/ICISCAE55891.2022.9927522>.
- [5] Dang Thi, P., Nguyen Dang, H. T., Doan Huu, P., & Dau Sy, H. (2024). Video classification for efficient data storage using deep learning: a comparison of sequential and simultaneous feature extraction methods. *Multimedia Tools and Applications*, 84(6), 3071–3094. <https://doi.org/10.1007/S11042-024-20549-5/TABLES/6>.
- [6] Fang, Z., Chen, J., Wang, J., Wang, Z., Liu, N., & Yin, G. (2022). Driver Distraction Behaviour Detection using a Vision Transformer Model based on Transfer Learning Strategy. *2022 6th CAA International Conference on Vehicular Control and Intelligence, CVCI 2022*. <https://doi.org/10.1109/CVCI56766.2022.9965124>.
- [7] Hassan, O. F., Ibrahim, A. F., Gomaa, A., Makhoulf, M. A., & Hafiz, B. (2025). Real-time driver drowsiness detection using transformer architectures: a novel deep learning approach. *Scientific Reports*, 15(1), 1–33. <https://doi.org/10.1038/S41598-025-02111-X>;SUBJMETA=1042,117,639,705;KWRD=COMPUTATIONAL+SCIENCE,COMPUTER+SCIENCE.
- [8] Hussain, A., Hussain, T., Ullah, W., & Baik, S. W. (2022). Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos. *Computational Intelligence and Neuroscience*, 2022(1), 3454167. <https://doi.org/10.1155/2022/3454167>.
- [9] Jarndal, A., Tawfik, H., Siam, A. I., Alsyouf, I., & Cheaitou, A. (2024). A Real-Time Vision Transformers-based System for Enhanced Driver Drowsiness Detection and Vehicle Safety. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2024.3522111>.

- [10] Li, Z., Zhao, X., Wu, F., Chen, D., & Wang, C. (2024). A Lightweight and Efficient Distracted Driver Detection Model Fusing Convolutional Neural Network and Vision Transformer. *IEEE Transactions on Intelligent Transportation Systems*, 1–17. <https://doi.org/10.1109/TITS.2024.3447041>
- [11] Ma, Y., & Wang, Z. (2024). ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection. *IEEE Intelligent Vehicles Symposium, Proceedings*, 417–423. <https://doi.org/10.1109/IV55156.2024.10588802>.
- [12] Wang, Z., & Yao, L. (2024). Recongnition of Distracted Driving Behaviour Based on Improved Bi-LSTM Model and Attention Mechanism. *IEEE Access*, 12, 67711–67725. <https://doi.org/10.1109/ACCESS.2024.3399789>.
- [13] Mou, L., Chang, J., Zhou, C., Zhao, Y., Ma, N., Yin, B., Jain, R., & Gao, W. (2023). Multimodal driver distraction detection using dual-channel network of CNN and Transformer. *Expert Systems with Applications*, 234, 121066. <https://doi.org/10.1016/J.ESWA.2023.121066>.
- [14] Jiang, L., Xie, W., Zhang, D., & Gu, T. (2022). Smart Diagnosis: Deep Learning Boosted Driver Inattention Detection and Abnormal Driving Prediction. *IEEE Internet of Things Journal*, 9(6), 4076–4089. <https://doi.org/10.1109/JIOT.2021.3103852>.
- [15] Praharsha, C. H., & Poullose, A. (2024). CBAM VGG16: An efficient driver distraction classification using CBAM embedded VGG16 architecture. *Computers in Biology and Medicine*, 180, 108945. <https://doi.org/10.1016/J.COMPBIOMED.2024.108945>.
- [16] Tatsunami, Y., & Taki, M. (2022). Sequencer: Deep LSTM for Image Classification. *Advances in Neural Information Processing Systems*, 35, 38204–38217. <https://github.com/okojoalg/sequencer>.
- [17] Kim, J., Shim, K., Kim, J., & Shim, B. (2023). Vision Transformer-Based Feature Extraction for Generalized Zero-Shot Learning. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2023-June. <https://doi.org/10.1109/ICASSP49357.2023.10095217>.
- [18] Tabbakh, A., & Barpanda, S. S. (2023). A Deep Features Extraction Model Based on the Transfer Learning Model and Vision Transformer “TLMViT” for Plant Disease Classification. *IEEE Access*, 11, 45377–45392. <https://doi.org/10.1109/ACCESS.2023.3273317>.
- [19] Wang, L., Fang, S., Meng, X., & Li, R. (2022). Building Extraction With Vision Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60. <https://doi.org/10.1109/TGRS.2022.3186634>.
- [20] Liang, J., Zhu, H., Zhang, E., Zhang, J., & Lab, T. Y. (2022). Stargazer: A Transformer-Based Driver Action Detection System for Intelligent Transportation (pp. 3160–3167). https://github.com/JunweiLiang/aicity_action.
- [21] Pizarro, R., Bergasa, L. M., Baumela, L., Buenaposada, J. M., & Barea, R. (2024). DRVMon-VM: Distracted driver recognition using large pre-trained video transformers. *IEEE Intelligent Vehicles Symposium, Proceedings*, 1901–1906. <https://doi.org/10.1109/IV55156.2024.10588412>.