Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Detection of Driver Distraction Using Spatio-Temporal Graph Convolutional Networks (ST-GCN) and Attention Mechanism

Mahdi Davari and Razieh Rastgoo*

Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 02 May 2025
Revised 03 June 2025
Accepted 16 September 2025

DOI:10.22044/jadm.2025.16157.2735

Keywords:
Driver Distraction Detection, Deep Learning, Pose Estimation, Keypoints.

**Corresponding author: rrastgoo@semnan.ac.ir (R. Rastgoo).*

Abstract

Detecting driver distraction during driving is of critical importance due to its significant role in increasing road accidents. This study aims to present a hybrid model based on Spatio-Temporal Graph Convolutional Networks (ST-GCN) and the attention mechanism for identifying driver distraction. In this research, skeletal body data of drivers were extracted from the 3D Drive&Act dataset and used as input for the proposed model. The model leverages spatial and temporal graph convolution layers, along with attention layers, to simultaneously analyze the spatiotemporal features of driver movements. Experimental results demonstrate that the proposed model achieves higher accuracy in detecting driver distraction compared to previous models, particularly under complex driving scenarios. Experimental results show that our proposed model achieves an accuracy of 97.47% on the Drive&Act dataset, significantly outperforming previous methods. This system can serve as an intelligent warning tool to reduce road accidents and enhance transportation safety.

1. Introduction

Human activity recognition in videos and motion data represents a fundamental challenge in the fields of computer vision [1-5] and deep learning [6-10], with widespread applications in areas such as surveillance, security, human-computer interaction, and robotics. A particularly critical application in this domain is the detection of driver distraction during driving [1]. Driving is a multifaceted task requiring the simultaneous coordination of cognitive, physical, and sensorimotor skills, including processing visual information, controlling motion, maintaining environmental awareness, and making rapid decisions [11]. However, many drivers engage in secondary activities, such as using mobile phones, adjusting in-car entertainment systems, or conversing with passengers [12]. Collectively, these behaviors are referred to as driver distraction, which significantly diminishes attention to the road and constitutes a serious threat to road safety.

Research has shown that driver distraction is a leading cause of road accidents [1]. In the United States, approximately one-fourth of traffic accidents are attributed to distracted driving [1]. Similarly, in Iran, an estimated 60% of accidents result from driver inattention. The rising prevalence of smartphones and in-car multimedia systems has further exacerbated this issue, elevating the risk of distracted driving [11].

To address this growing concern, the development of intelligent systems for the automatic detection of driver distraction has become imperative. Deep learning, a subset of artificial intelligence [13,14], has emerged as a powerful tool for analyzing driver behavior with high accuracy [12]. These systems leverage video data from in-cabin cameras to detect distractions such as mobile phone usage, looking away from the road, or engaging in physical distractions, providing real-time alerts to drivers [1]. Deep learning models are particularly effective for this task due to their ability to automatically extract hierarchical features from raw input data [15-17], adapt to diverse environments, and generalize across a variety of driving scenarios.

Compared to traditional distraction detection methods, which often rely on handcrafted features such as edge detection, color segmentation, and motion tracking, deep learning offers several advantages. Traditional methods are highly sensitive to variations in lighting, occlusions, and camera angles, and often require extensive manual tuning [18]. These approaches also struggle to generalize across different driving environments. In contrast, deep learning models, particularly those based on convolutional neural networks (CNNs) [19] and graph-based architectures [20], are capable of automatically learning and extracting meaningful features from raw data without the need for manual intervention [21]. Moreover, deep learning methods are more robust to environmental variations, such as changes in illumination and background noise, making them more suitable for real-world applications [13].

Furthermore, advanced architectures such as Spatio-Temporal Graph Convolutional Networks (ST-GCN) [22] and Transformers [23] enable the simultaneous modeling of spatial and temporal dependencies in driver movements, leading to significantly improved accuracy in detecting distractions. Driver behavior classification networks are inspired by architectures commonly used for human activity recognition [1]. These networks can be broadly categorized into RGB image-based networks and human pose-based networks. While RGB image-based models analyze raw pixel data to detect driver behavior, they often face challenges such as lighting variations, background noise, and occlusions [10]. In contrast, human pose-based networks extract skeletal joint positions in 2D or 3D, which makes them more robust to environmental changes [24]. Pose-based models are computationally more efficient and less sensitive to variations in camera angles, backgrounds, and lighting conditions, making them particularly well-suited for real-world applications [25,26].

This study proposes a deep learning-based model aimed at improving the accuracy of driver distraction detection while addressing the challenges mentioned above. The model enables real-time distraction detection, reduces the risk of accidents caused by driver inattention, and provides timely and precise alerts. It performs robustly under varying lighting conditions and adapts to individual behavioral differences. Such a system holds the potential to be integrated into smart vehicles and traffic monitoring frameworks, contributing significantly to road safety.

The proposed model follows a hybrid approach that combines Graph Convolutional Network (GCN) [27,28] and Transformer [29,30] for driver activity recognition. The GCN component extracts spatial features, such as skeletal joint positions, from input motion data. These joints are represented as graph-structured data, allowing the model to efficiently capture spatial relationships between different body parts. After extracting spatial features, the Transformer component models temporal dependencies between video frames, enabling the model to understand motion patterns over time.

To enhance model performance, several key components are incorporated. The spatial attention mechanism determines the relative importance of each body joint in detecting distractions, ensuring that critical movement patterns are effectively captured. Graph convolutional layers refine joint feature representations by updating them based on neighboring relationships. The Transformer encoder further processes sequential data, capturing long-term dependencies between movements. Additionally, a focal loss function is employed to mitigate class imbalance by emphasizing underrepresented distraction categories.

To evaluate the model, the 3D Drive&Act dataset was employed, which contains a diverse set of human activities performed in a driving environment [31]. The 3D Drive&Act dataset is one of the most challenging benchmarks for driver distraction classification, and experimental results on this dataset demonstrate that the proposed model outperforms existing state-of-the-art methods [31].

2. Related Works

Given that driver distraction is widely acknowledged as one of the leading causes of road accidents, a substantial body of research has been dedicated to investigating this issue [1]. The existing literature typically classifies driver distraction into two primary categories: physical distraction and cognitive distraction.

Physical distraction refers to activities that require the driver to engage in actions unrelated to the primary task of driving [32]. Examples include talking to passengers, using a mobile phone, eating or drinking, and adjusting the vehicle's audio system [33]. These activities often result in a temporary loss of vehicle control, which significantly increases the risk of accidents.

In contrast, cognitive distraction is considered a more insidious threat [12]. It occurs when the driver's mental focus is diverted away from driving, even in the absence of visible physical indicators of distraction [34]. Unlike physical distraction, cognitive distraction is more difficult to detect, as it does not necessarily involve observable gestures or movements.

To address physical distraction, some studies have concentrated on detecting and analyzing drivers' body movements. These approaches typically employ video-based or sensor-based systems to monitor critical physical cues, such as hand positions, head orientation, and other bodily signals that may indicate distraction.

2.1. Commercial Systems

In the automotive sector, prominent manufacturers such as Toyota, Ford, and Mercedes-Benz have incorporated driver behavior monitoring systems into their vehicles [11]. These systems are designed to detect indicators of distraction, drowsiness, and diminished attention. Typically, they utilize in-vehicle cameras and computer vision technology to assess the driver's state, issuing timely alerts when signs of distraction or drowsiness are detected [35].

2.2. Scientific Research

In academic research, various approaches have been proposed to detect driver distraction by analyzing facial features, including gaze direction, head movements, and eye distance [6,9,36]. These methods often employ machine learning techniques, with particular emphasis on neural networks, to enhance the accuracy of distraction detection. For example, one study utilized convolutional neural networks (CNNs) to extract facial features and applied clustering techniques to classify instances of driver distraction [34].

Additionally, other studies have leveraged video- and image-based systems, which are particularly effective when combined with image processing techniques. These studies typically focus on analyzing the driver's head and facial movements in video footage, assessing how these movements interact with changing road conditions [35].

2.3. Datasets

Numerous datasets have been curated to facilitate research in the field of driver behavior analysis. Notable examples include datasets such as Ohn et al. [37], Brain4Cars [38], and Drive&Act [31], which serve as valuable resources for training machine learning models aimed at detecting driver distraction. These datasets provide rich visual data about drivers and their surrounding environments, making them instrumental for analyzing driver behaviors, particularly under real-world conditions [39].

In this study, we adopt a driver action recognition approach that leverages skeletal key points. Recent years have witnessed a growing interest in pose-based action recognition, particularly in the context of autonomous vehicles. The goal of this task is to classify driver behavior into predefined categories. These actions may occur while the driver is actively operating the vehicle or when they are a passenger in an autonomous vehicle setting.

Some studies have concentrated on facial pose analysis to determine the driver's gaze direction, while others focus on analyzing the full-body pose. Convolutional neural networks (CNNs) have been extensively utilized for classifying actions based on pose information. For example, one study combined spatial features extracted via CNNs with geometric features to predict the corresponding driver action [12]. Another approach proposed a two-stream recurrent neural network (RNN) [40] architecture to simultaneously model both temporal and spatial dynamics [41].

2.4. Graph Neural Networks and ST-GCN

Graph neural networks (GNNs) are well-suited for modeling driver behavior because they operate directly on graph-structured data, enabling the representation of complex relationships between entities such as body joints or objects in a scene. To further exploit this capability in a spatio-temporal setting, we employ spatio-temporal graph convolutional networks (ST-GCN).

ST-GCN have been widely adopted for video analysis and pose-based activity recognition, particularly in applications related to autonomous driving and human–computer interaction. In these models, human body joints are represented as nodes in a graph, and their spatial relationships are encoded as edges. By extending the graph structure across consecutive frames, ST-GCN can jointly capture spatial dependencies between body parts and temporal dynamics over time.

In practice, ST-GCN employ graph convolutions to extract spatial information from the human skeleton while simultaneously modeling temporal dependencies across video frames [28]. This joint modeling of spatial and temporal features improves the accuracy of recognizing a wide range of human activities, including driver behavior analysis, social interaction understanding, and complex body motion recognition.

A key advantage of ST-GCN for activity recognition is their ability to integrate spatial and temporal information within a unified framework, rather than treating them separately. This makes

ST-GCN particularly suitable for analyzing intricate driver behaviors in realistic and autonomous driving environments. In our work, we use an ST-GCN backbone to model the spatio-temporal dynamics of the driver's skeletal keypoints, as illustrated in Figure 1.

3. Method

In this study, a hybrid approach combining ST-GCN and the attention mechanism has been employed to analyze spatio-temporal skeletal data for the identification and prediction of driver behaviors and distraction activities during driving.

3.2. Model

The model proposed in this study is built upon the ST-GCN, which integrates graph convolutional networks with specialized temporal analysis layers. The primary objective of utilizing this architecture is to effectively analyze and simulate both spatial and temporal interactions, which are critical for accurately identifying and predicting driver behaviors.

This model is specifically designed to detect driver actions that could lead to distractions or pose potential risks. The ST-GCN processes skeletal data as spatio-temporal graphs, where each node represents a body joint, and the connections between nodes model the spatial relationships between these joints. These graphs evolve over time, with data from each frame of driver activity being sequentially fed into the network as input.

3.3. Input

The input to the neural network comprises the 3D coordinates of human body joints. For each joint, the x, y, and z values are recorded across time (frames). For example, in the case of a video featuring 25 joints and 90 frames, the input data is organized into an array with dimensions of 90 × 25 × 3. Our model is intentionally designed to handle situations where keypoint data is missing, such as when lower body parts are occluded due to camera angles or clothing. In such cases, the keypoints corresponding to these occluded regions are assigned a value of zero, as illustrated in Figure 2. Despite these missing data points, the model compensates effectively by utilizing spatial graph convolutions and attention mechanisms, ensuring that high accuracy is maintained even with incomplete or noisy input data.

3.4. Model Architecture

3.4.1. Spatial Attention Layer

To improve the network's performance, a spatial attention layer is incorporated. At this stage of the network, each body joint of the driver is assigned a specific weight, representing its significance in identifying particular activities. The spatial attention layer facilitates the network in extracting motion and spatial features from various body key points with greater accuracy.

When compared to traditional sequential models, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), the attention mechanism presents several distinct advantages. While LSTM and GRU models are effective at capturing temporal dependencies, they face challenges such as vanishing gradients, high computational complexity, and difficulties in processing long-range dependencies. In contrast, attention mechanisms can selectively focus on the most relevant features across all frames, without the constraints imposed by sequential processing. This ability significantly enhances the model's capacity to capture key spatio-temporal patterns, leading to improved performance in driver distraction detection.

Furthermore, the attention mechanism enables the model to dynamically adapt to different distraction scenarios by emphasizing crucial body joints, thereby ensuring greater robustness in varying driving conditions.

The spatial attention module assigns a normalized importance weight to each joint independently, utilizing a learnable parameter vector. These weights are applied to amplify the contribution of critical joints in the input data. It is essential to note that this module does not directly model inter-joint relationships. Instead, the spatial dependencies between joints are captured by the subsequent graph convolution layers, which use a predefined adjacency matrix to model these relationships.

To implement spatial attention, we define a learnable parameter tensor of shape (1, 1, 1, 25), which is applied directly to the input tensor with shape (N, C, T, V), where:

- N: batch size
- C: number of channels
- T: number of time frames
- V: number of joints (25 in our case)

The attention tensor contains 25 scalar weights, each corresponding to one joint. These weights are normalized using a softmax function over the joint dimension (V) and then broadcast-multiplied across the input tensor. This process allows the model to emphasize or suppress specific joints based on their relevance to the target activity. Importantly, the attention is computed independently per joint, without explicitly modeling inter-joint dependencies. Such dependencies are subsequently captured by the

Graph Convolution layers, which process the skeletal topology using an adjacency matrix.

The inclusion of the Transformer enables the model to effectively simulate complex, sequential driver activities that unfold over various time intervals, thereby improving its capacity to recognize behaviors that evolve over extended periods.

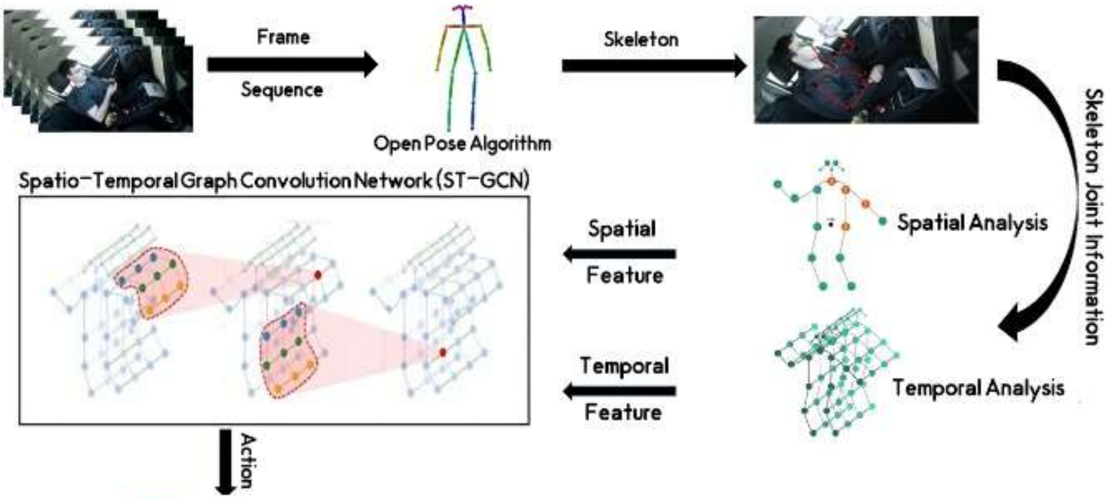

Figure 1: Overview of the model architecture. This diagram provides a comprehensive representation of the proposed system's pipeline. It begins with the input of 3D skeletal data, followed by the processing through the ST-GCN block for extracting spatial and short-term temporal features. Subsequently, the Transformer encoder is applied to capture long-range temporal dependencies.

3.4.2. Spatial Graph Convolution Layer

In this layer, spatial relationships between body joints are processed using predefined graphs, which represent the anatomical connections between these joints. These graphs are crucial for identifying the driver's movements and different postures. This layer enables the model to effectively extract complex spatial features from skeletal data, thereby enhancing its ability to recognize and analyze body movements with high accuracy.

3.4.3. Temporal Graph Convolution Layer

This layer is responsible for processing temporal features. By incorporating this layer, the model is capable of analyzing the driver's motion changes over time, thereby detecting sequential and temporal relationships across frames.

This layer is particularly beneficial for identifying activities that evolve gradually over time, such as drowsiness or distraction, where the signs of these behaviors develop progressively rather than abruptly.

3.4.4. Transformer Encoder

Finally, a Transformer Encoder layer is incorporated into the model to enhance its ability to capture long-term temporal relationships. This component of the model simulates temporal dependencies between frames, enabling the detection of long-term changes in driver movements and behaviors.

3.4.5. Output Layer

After the extraction of spatio-temporal features, the network's output is passed through prediction layers, which transform the extracted features into the corresponding activity classes of the driver. The final outputs of the model are presented as categorized results, indicating the detected driver behavior.

Figure 3 illustrates the complete data processing pipeline implemented in our proposed model. As shown, the process begins with the input of raw video sequences, from which 3D skeletal keypoints are extracted using OpenPose. These keypoints are then structured into a spatio-temporal graph, which serves as the input to the ST-GCN block. After extracting spatial and short-term temporal features, the resulting representations are forwarded to the Transformer encoder to capture long-term dependencies. Finally, classification layers generate the predicted driver activity. This pipeline ensures an end-to-end and efficient workflow from raw video input to final behavior classification, highlighting the modular yet integrated architecture of our model.

3.4.6. Training

The proposed model is trained using a focal loss function [42], which is specifically designed to address classification problems with imbalanced data. In the context of driver activity recognition, certain activities may occur less frequently than others. By employing the focal loss function, the model's accuracy is improved, particularly in recognizing less frequently observed classes, thus enhancing its performance across all activity categories.

3.4.7. Experimental Setup

The model was trained for 300 epochs using a batch size of 64 and a learning rate of 0.01 with the SGD optimizer. The dataset was randomly divided into training and test sets without any subject-based separation. No data augmentation techniques were applied during training. The model was trained on the Drive&Act dataset, utilizing 3D skeletal sequences extracted through OpenPose.

To further assess the model's robustness under challenging conditions, we analyzed frames where keypoint data was partially missing. These conditions include occlusions due to bulky clothing, object interference, and camera angles from behind the driver. We identified such frames across several action classes. For instance:

- Frames showing heavy clothing occlusion appear across multiple action classes.
- Frames with keypoints hidden behind objects, such as hands obscured by a magazine, are particularly common in the "Read/Write Magazine" class, which achieves over 90% accuracy, as shown in Figure 4.
- Frames where the subject faces away from the camera, resulting in missing frontal keypoints, mostly belong to the "Park Exit" class, which achieves close to 90% accuracy.

These examples in Figure 2 demonstrate that even with partially missing keypoints, the model retains strong performance, highlighting the effectiveness of the spatial attention and graph-based design in handling incomplete data.

3.5. Architectural Design and Innovation

Although the individual components used in our model, ST-GCN, spatial attention, and Transformer, are well-known, the innovation lies in how these modules are systematically integrated and adapted for the task of driver distraction detection using 3D skeletal data. Unlike conventional approaches that rely on RGB or depth images, we use graph-based skeletal input, which reduces computational cost and improves generalization.

In our architecture, the ST-GCN blocks are responsible for capturing both spatial and short-term temporal features, using spatio-temporal graph convolutions on joint data. The output of these blocks is reshaped and passed to a Transformer encoder, which models long-range temporal dependencies across the entire motion sequence. This sequential integration, ST-GCN followed by Transformer, is specifically designed to combine local joint-level motion with global temporal understanding.

The use of Focal Loss further enhances performance on rare distraction events by focusing the training process on hard-to-classify samples. This design enables the model to maintain both high accuracy and efficiency, making it suitable for real-time driver monitoring systems.

4. Discussion

The dataset utilized in this research is the Drive&Act dataset, which is specifically designed for human action recognition in the context of autonomous vehicles. This dataset includes data from 15 distinct individuals performing a variety of actions, captured from 6 different viewpoints. This dataset provides two levels of classification for the actions in all sequences, as outlined below:

General Tasks: This represents the highest classification level, consisting of 12 main categories of general tasks. The goal of this study is to achieve the highest accuracy at this level of classification [39].

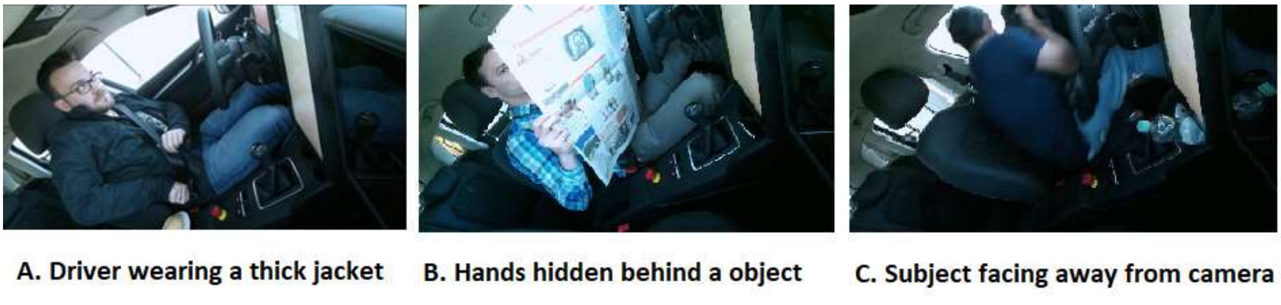
A. Driver wearing a thick jacket
B. Hands hidden behind a object
C. Subject facing away from camera

Figure 2. Sample frames illustrating the model’s robustness to missing keypoints caused by various challenges.

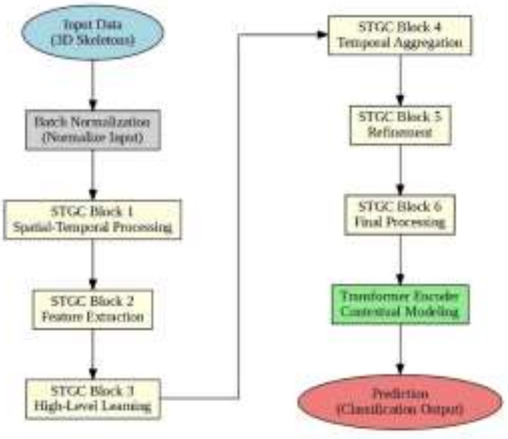

Figure 3. Data processing pipeline from video input to final predictions in the proposed model.

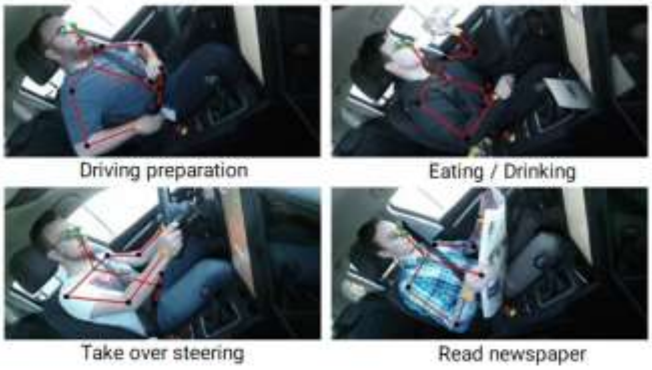

Figure 4. A sample of four frames illustrating the driver's state and the corresponding class in the dataset.

- More Specific Activities: At the second level, activities are further subdivided into 34 distinct classes. Each class corresponds to a specific action that may vary depending on individual driver preferences. These activities include more detailed actions, such as opening and closing a water bottle. Since the primary objective of this study is to classify the general action of the driver, the general task data is most suitable for our analysis. Therefore, this data will be utilized for training and validation, as described in the results section. Our task is to classify the entire sequence of actions, as predefined in the dataset, with the corresponding class label.

To process the input data, we employed a neural network consisting of multiple layers. The model was trained for 300 epochs using the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01 to optimize the model weights. The model utilizes a graph to model the relationships between body joints. This graph is composed of various layers, including SpatialGraphConvolution and STGC_block layers, which are specifically designed to extract both spatial and temporal features from the input data. Each STGC_block layer incorporates spatial attention and graph convolution operations, enabling the model to learn complex relationships between different joints. For processing temporal information, a Transformer encoder layer is used. These layers allow the model to learn intricate temporal dependencies within the data and use them to predict the final activity classes. The model employs Focal Loss for optimization, a loss function specifically designed for imbalanced datasets. This function allows the model to focus more on the less frequent, underrepresented classes, thereby improving performance across all categories.

5. Results

We applied our proposed architecture to the Drive&Act dataset, and Figure 5 illustrates the accuracy for each class in the test set. When compared to methods such as Squeezeformer [39], which also provide accuracy metrics for each class, our model demonstrates a significant improvement in accuracy across all classes. Figure 6 presents the class-wise accuracy of the proposed model. It was observed that the use of focal loss significantly improved performance on low-frequency actions, such as 'take off sunglasses', which are typically challenging to classify. The model's ability to focus on these minority classes underscores the effectiveness of focal loss in addressing issues related to imbalanced data. In particular, several underrepresented classes, such as “Put On Sunglasses,” “Take Off Sunglasses,” and “Put On Jacket”, showed poor performance in earlier configurations without focal loss. These classes suffered from high misclassification rates due to their limited number of training samples. However, after incorporating focal loss, the model demonstrated substantial improvements in these categories, achieving notably higher accuracy. This validates the focal loss function's role in directing the model's attention to minority classes and improving overall classification balance.
To evaluate the contribution of each component within the proposed architecture, we conducted internal experiments comparing various model variants. As shown in Table 1, we started with a baseline model that utilized only the ST-GCN, which achieved an accuracy of 80%. We then progressively incorporated the spatial attention layer and Transformer encoder. The inclusion of the spatial attention layer resulted in a noticeable

improvement, increasing the accuracy to 90%. Finally, adding the Transformer encoder further enhanced the model's performance, achieving a final accuracy of 97.47%. This analysis highlights the significant contributions of each individual component to the overall performance of the model.

When compared to other methods and models that rely solely on keypoint data of the driver's body, the results obtained by our model demonstrate higher accuracy. A detailed comparison of these results is provided in Table 4 and Figure 6. Additionally, the confusion matrix for the final model is presented in Figure 7, and Table 2 shows the results of different hyperparameters.

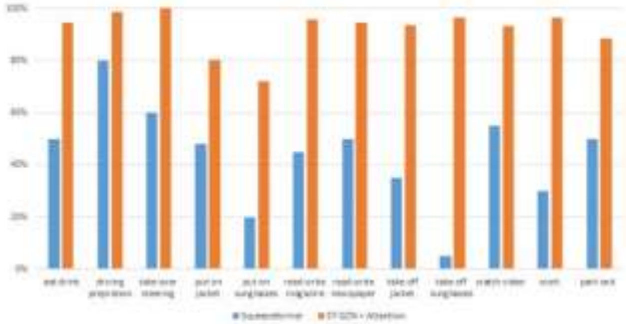

Figure 5. Comparison of the accuracy of the proposed model's results with the Squeezeformer model.

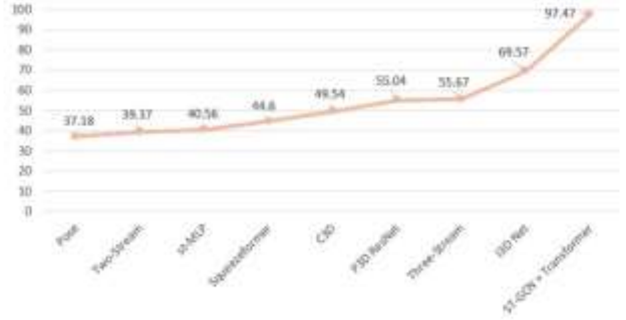

Figure 6. Comparison of different methods and the results obtained with the proposed method.

Table 1. Performance of different model variants.

Model Variant	Accuracy (%)
Baseline (ST-GCN only)	80%
ST-GCN + Spatial Attention Layer	90%
ST-GCN + Transformer Encoder	93.5%
ST-GCN + Spatial Attention + Transformer Encoder	**97.47%**

Due to the efficiency of the graph-based representation, which significantly reduces the data volume, the entire training process was completed in 4 hours, demonstrating that both training and inference times are fast, with a relatively low computational load. As shown in Table 3, we present a detailed breakdown of the model's training time, inference time, GPU memory usage, and overall latency, highlighting the computational efficiency and real-time viability of the proposed system.

Table 2. The results obtained with different hyperparameters.

Epoch	Batch_Size	Dropout	Dim_Feedforward	Lr	Accuracy
80	16	0.3	128	0.01	76
80	16	0.3	256	0.01	73
100	16	0.3	128	0.01	72
100	16	0.5	256	0.01	73
100	32	0.3	128	0.01	78
100	32	0.3	256	0.005	78
120	32	0.5	256	0.005	77
120	32	0.5	512	0.005	79
120	64	0.3	128	0.005	80
150	64	0.3	256	0.005	82
200	256	0.4	128	0.001	38
250	64	0.4	128	0.01	97

Figure 7 presents the confusion matrix for our final model, where we observe notable misclassifications between similar action classes. For example, "eat/drink" and "work" are frequently confused due to overlapping hand and upper-body movements. This overlap is especially pronounced when these actions are viewed from specific angles or when subtle variations in movement occur. The confusion matrix further highlights how activities such as "read/writing newspaper" and "work" share similar postures, leading to frequent misclassifications. Both activities involve seated positions with hand movements, which the model struggles to differentiate in certain contexts, particularly when the movement is subtle or partially obstructed.

This pattern of misclassification can likely be attributed to the contextual and postural similarities between these actions. The model appears to have difficulty distinguishing between activities that involve similar upper-body movements and hand gestures. Additionally, the limited data variety for certain actions in the dataset may exacerbate this issue, making it more challenging for the model to generalize across diverse conditions.

In Section 5, we provide a detailed analysis of the recurring misclassifications observed in the confusion matrix. We discuss how actions such as "eat/drink" and "work" share similar hand and upper-body movement patterns, which often lead to misclassifications. Furthermore, we highlight the contextual ambiguity between activities like "read/writing newspaper" and "work", where the postural similarities complicate the model's ability to distinguish between the two.

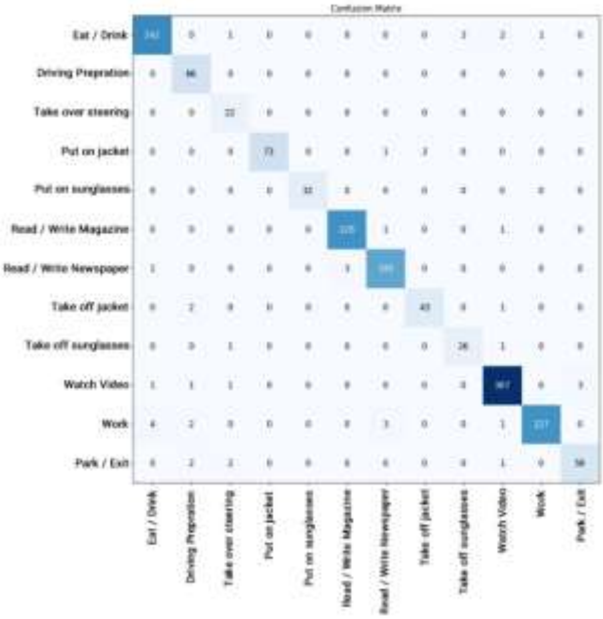

Figure 7. The confusion matrix in the final model.

Table 3. Performance and resource usage summary.

Experiment	Details
Training Time	4 hours
Inference Time per Video	100 ms per 90-frame video (3 seconds at 30 FPS)
GPU Model Used	NVIDIA T4 GPU
Batch Size	64
Video Length	90 frames (approximately 3 seconds at 30 FPS)
3D Skeletal Data (Joints)	25 joints
Average Memory Usage (Training)	4 GB
Average Memory Usage (Inference)	3.5 GB
Overall Latency (Processing one video)	200 ms per 90-frame video (from input to output)
Resource Utilization	Low (due to efficient graph-based representation)

5.1 Proposed Architecture

The proposed architecture demonstrates superior performance compared to other existing approaches, yielding improved results. The key enhancement in this architecture is the integration of ST-GCN with attention mechanisms, which enables the model to more effectively capture and understand the relationships within the data.

6. Discussion and Conclusion

This study introduced an innovative approach for detecting driver distraction during driving by integrating spatio-temporal graph convolutional networks (ST-GCN), Attention Mechanisms, and Transformers. The primary objective of this research was to develop an efficient and accurate model for identifying a wide range of driving activities and distraction behaviors. This objective was successfully achieved, with the model achieving an impressive accuracy of 97.47%, surpassing the performance of other related studies. Although this study primarily focused on architectural innovation and performance validation using the Drive&Act dataset, future work will incorporate statistical testing to further assess the significance of model improvements over baseline methods.

Table 4. Comparison of different methods and the results obtained.

Method	Accuracy
Pose[31]	37.18
Two-Stream[41]	39.37
st-MLP[32]	40.56
Squeezeformer[39]	44.60
C3D[43]	49.54
P3D ResNet[44]	55.04
Three-Stream[45]	55.67
I3D Net[46]	69.57
ST-GCN + Transformer	**97.47**

The model was trained for 300 epochs on a Google Colab environment using an NVIDIA T4 GPU. Each input sequence consisted of 90 frames (3 seconds at 30 FPS), with skeletal data extracted from 25 keypoints. Unlike traditional pixel-based methods, this graph-based input significantly reduces the data size, resulting in faster processing times. While detailed latency benchmarks were not included in this study, the model demonstrated efficient inference during testing, suggesting its potential for real-time deployment. Future work will involve quantitative measurements of latency and computational requirements to better assess the model's viability for practical applications.

The proposed model, which integrates spatial attention layers and the Transformer encoder, effectively extracts spatio-temporal features from the data. Experimental results demonstrated that this approach accurately identified driving behaviors and distractions, achieving excellent performance in distinguishing complex activities. The incorporation of the Focal Loss function further contributed to enhancing the model's accuracy, particularly when handling imbalanced data.

This research makes a significant contribution to the field of driver distraction detection by presenting an effective combination of Graph Neural Networks and Transformers, thereby showcasing its potential for real-world applications. Potential applications of this model include integration into Advanced Driver-Assistance Systems (ADAS) and proactive alert systems in intelligent vehicles, both of which could significantly improve road safety.

For future research, it is recommended to utilize larger and multi-source datasets, such as data from vehicle sensors, thermal videos, and driver's biological data. Additionally, enhancing the model architecture through the use of deeper neural networks and advanced data augmentation techniques could further improve the model's accuracy and reliability.

While the Drive&Act dataset remains the most comprehensive resource for 3D pose-based driver activity recognition, it has limitations in terms of environmental diversity, particularly regarding lighting variations and real-world complexity. Additionally, given the importance of accurate keypoint detection, we utilized the most reliable camera viewpoint available. Future research should focus on experiments with more diverse datasets and explore the potential of transfer learning to assess the model's robustness in uncontrolled, real-world driving conditions.

The current study is based exclusively on the Drive&Act dataset under controlled conditions, which limits the model's generalizability to unseen drivers, varied vehicle interiors, and diverse lighting conditions. Future work will involve cross-subject evaluations and testing on more heterogeneous datasets. Furthermore, integrating multi-modal data (e.g., combining skeletal and RGB data) and applying ensemble or smoothing strategies could significantly enhance both the accuracy and robustness of the model, particularly in real-world driver-assistance systems.

Another promising direction for future research is the application of automated hyperparameter tuning methods, such as grid search or Bayesian optimization, to optimize training configurations and potentially enhance accuracy further.

References

[1] K. Young, M. Regan, and M. Hammer, "Driver distraction: A review of the literature," *Distracted Driving*, 2007, pp. 379–405.

[2] A.M. Ahmadi, K. Kiani, R. Rastgoo, "A Transformer-based model for abnormal activity recognition in video," *Journal of Modeling in Engineering*, vol. 22, no. 76, pp. 213–221, 2024.

[3] R. Rastgoo, K. Kiani, S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools and Applications*, vol. 79, pp. 22965–22987, 2020.

[4] F. Bagherzadeh, R. Rastgoo, "Deepfake image detection using a deep hybrid convolutional neural network," *Journal of Modeling in Engineering*, vol. 21, no. 75, pp. 19–28, 2023.

[5] M. Talebian, K. Kiani, R. Rastgoo, "A Deep Learning-based Model for Fingerprint Verification," *Journal of AI and Data Mining*, vol. 12, no. 2, pp. 241–248, 2024.

[6] H. Zaferani, K. Kiani, R. Rastgoo, "Real-time face verification on mobile devices using margin distillation," *Multimedia Tools and Applications*, vol. 82, no. 28, pp. 44155–44173, 2023.

[7] S. Zarbafi, K. Kiani, R. Rastgoo, "Spoken Persian digits recognition using deep learning," *Journal of Modeling in Engineering*, vol. 21, no. 74, pp. 163–172, 2023.

[8] N. Majidi, K. Kiani, R. Rastgoo, "A deep model for super-resolution enhancement from a single image," *Journal of AI and Data Mining*, vol. 8, no. 4, pp. 451–460, 2020.

[9] R. Rastgoo, K. Kiani, "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning," *Journal of Modeling in Engineering*, vol. 17, no. 58, pp. 103–111, 2019.

[10] F. Alinezhad, K. Kiani, R. Rastgoo, "A Deep Learning-based Model for Gender Recognition in Mobile Devices," *Journal of AI and Data Mining*, vol. 11, no. 2, pp. 229–236, 2023.

[11] T. Stewart, "Overview of motor vehicle crashes in 2020," United States Department of Transportation, *National Highway Traffic Safety*, 2022.

[12] M. Wu, et al., "Pose-aware multi-feature fusion network for driver distraction recognition," in *ICPR*, 2021.

[13] R. Rastgoo, K. Kiani, S. Escalera, "Sign Language Recognition: A Deep Survey," *Expert Systems with Applications*, vol. 164, 113794, 2020.

[14] R. Rastgoo, K. Kiani, S. Escalera, "A transformer model for boundary detection in continuous sign language," *Multimedia Tools and Applications*, vol. 83, pp. 89931–89948, 2024.

[15] R. Rastgoo, K. Kiani, S. Escalera, "Hand pose aware multimodal isolated sign language recognition," *Multimedia Tools and Applications*, vol. 80, pp. 127–163, 2021.

[16] R. Rastgoo, K. Kiani, S. Escalera, M. Sabokrou, "Multi-modal zero-shot dynamic hand gesture recognition," *Expert Systems with Applications*, vol. 247, 123349, 2024.

[17] R. Rastgoo, K. Kiani, S. Escalera, "A deep co-attentive hand-based video question answering framework using multi-view skeleton," *Multimedia Tools and Applications*, vol. 82, pp. 1401–1429, 2023.

[18] R. Rastgoo, K. Kiani, S. Escalera, "ZS-GR: zero-shot gesture recognition from RGB-D videos," *Multimedia Tools and Applications*, vol. 82, pp. 43781–43796, 2023.

[19] R. Rastgoo, K. Kiani, S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Systems with Applications*, vol. 158, 113336, 2020.

[20] R. Rastgoo, K. Kiani, S. Escalera, "A non-anatomical graph structure for boundary detection in continuous sign language," *Scientific Reports*, vol. 15, 25683, 2025.

[21] R. Rastgoo, K. Kiani, S. Escalera, "Real-time isolated hand sign language recognition using deep networks and SVD," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, pp. 591–611, 2022.

[22] A. Holzbock, et al., "A spatio-temporal multilayer perceptron for gesture recognition," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2022.

[23] N. Esfandiari, K. Kiani, R. Rastgoo, "A conditional generative chatbot using transformer model," *Journal of Modeling in Engineering*, vol. 23, no. 82, pp. 99–113, 2025.

[24] R. Rastgoo, K. Kiani, S. Escalera, "Diffusion-Based Continuous Sign Language Generation with Cluster-Specific Fine-Tuning and Motion-Adapted Transformer," in *Proceedings of the Computer Vision and Pattern Recognition Workshop*, pp. 4088–4097, 2025.

[25] R. Rastgoo, K. Kiani, S. Escalera, V. Athitsos, M. Sabokrou, "A survey on recent advances in Sign Language Production," *Expert Systems with Applications*, vol. 243, 122846, 2024.

[26] R. Rastgoo, K. Kiani, S. Escalera, "Sign language production: A review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pp. 3451–3461, 2021.

[27] R. Rastgoo, K. Kiani, S. Escalera, "A Non-Anatomical Graph Structure for isolated hand gesture separation in continuous gesture sequences," *arXiv:2207.07619*, 2022.

[28] Yan, S., Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[29] Vaswani, A., "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[30] N. Esfandiari, K. Kiani, R. Rastgoo, "Development of a Persian Mobile Sales Chatbot based on LLMs and Transformer," *Journal of AI and Data Mining*, vol. 12, no. 4, pp. 465–472, 2024.

[31] M. Martin, et al., "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[32] M. Martin, D. Lerch, and M. Voit, "Viewpoint invariant 3d driver body pose-based activity recognition," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2023.

[33] T.A. Dingus, et al., "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016.

[34] N. Moslemi, M. Soryani, and R. Azmi, "Computer vision-based recognition of driver distraction: A review," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 24, e6475, 2021.

[35] S. Kaplan, et al., "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, 2015.

[36] M.H. Sigari, et al., "A review on driver face monitoring systems for fatigue and distraction detection," *International Journal of Advanced Science and Technology*, vol. 64, pp. 73–100, 2014.

[37] E. Ohn-Bar, et al., "Head, eye, and hand patterns for driver activity recognition," in *2014 22nd International Conference on Pattern Recognition*, IEEE, 2014.

[38] A. Jain, et al., "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[39] P. Pardo-Decimavilla, et al., "Do You Act Like You Talk? Exploring Pose-based Driver Action Classification with Speech Recognition Networks," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2024.

[40] N. Esfandiari, K. Kiani, R. Rastgoo, "A new transformer-based generative chatbot using CycleGAN approach," *Neural Computing and Applications*, vol. 37, no. 31, pp. 26125–26156.

[41] H. Wang, and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[42] T. Lin, "Focal Loss for Dense Object Detection," *arXiv preprint arXiv:1708.02002*, 2017.

[43] D. Tran, et al., "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[44] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[45] M. Martin, et al., "Body pose and context information for driver secondary task detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2018.

[46] J. Carreira, and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

شناسایی حواس‌پرتی راننده با استفاده از شبکه‌های کانولوشنال گراف فضایی-زمانی و مکانیزم توجه

مهدی داوری و راضیه راستگو*

دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.

ارسال ۲۰۲۵/۰۵/۰۲؛ بازنگری ۲۰۲۵/۰۶/۰۳؛ پذیرش ۲۰۲۵/۰۹/۱۶

چکیده: شناسایی حواس‌پرتی راننده هنگام رانندگی از اهمیت بالایی برخوردار است، زیرا نقش قابل توجهی در افزایش تصادفات جاده‌ای دارد. این مقاله با هدف ارائه یک مدل ترکیبی مبتنی بر شبکه‌های کانولوشنال گراف فضایی-زمانی (ST-GCN) و مکانیزم توجه برای شناسایی حواس‌پرتی راننده انجام شده است. در این مقاله، داده‌های اسکلتی بدن رانندگان از مجموعه داده سه‌بعدی Drive&Act استخراج و به‌عنوان ورودی مدل پیشنهادی استفاده شده‌اند. مدل پیشنهادی با بهره‌گیری از لایه‌های کانولوشنال گراف فضایی و زمانی همراه با لایه‌های توجه، ویژگی‌های فضایی-زمانی حرکات راننده را به‌طور هم‌زمان تحلیل می‌کند. نتایج تجربی نشان می‌دهد که مدل پیشنهادی دقت بالاتری در شناسایی حواس‌پرتی راننده، به‌ویژه در سناریوهای رانندگی پیچیده، نسبت به مدل‌های پیشین دارد. نتایج آزمایش‌ها نشان می‌دهد که مدل پیشنهادی ما دقت ۹۷٫۴۷٪ را بر روی مجموعه داده Drive&Act به‌دست آورده و عملکرد آن به‌طور قابل توجهی از روش‌های قبلی برتر است. این سیستم می‌تواند به‌عنوان یک ابزار هشدار هوشمند برای کاهش تصادفات جاده‌ای و افزایش ایمنی حمل‌ونقل مورد استفاده قرار گیرد.

کلمات کلیدی: شناسایی حواس‌پرتی راننده، یادگیری عمیق، برآورد وضعیت بدن، نقاط کلیدی.