Predicting Pedestrian Violations Using Object Detection and Deep Learning: A Comparative Study of LSTM and GRU Model

Stella Roussou
Ph.D. Candidate, Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR15773
Email: s_roussou@mail.ntua.gr

Apostolos Ziakopoulos
Post-Doctoral Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR15773
Email: apziak@central.ntua.gr

Roberto Ventura
Post-Doctoral Research Associate
Department of Civil, Environmental, Architectural Engineering and Mathematics (DICATAM)
University of Brescia - Via Branze, 43 – 25123 Brescia (Italy).
Email: roberto.ventura@unibs.it

George Yannis
Professor
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR15773
Email: geyannis@central.ntua.gr

Word Count: 6,587 words + 3 tables (250 words per table) = 7,337 words

Submitted 14.12.2024

ABSTRACT

Recent advances in machine learning and computer vision have significantly impacted traffic monitoring and safety analysis, particularly in urban environments. The integration of advanced object detection models, such as YOLOv8, with machine learning algorithms like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), offers new opportunities for analyzing pedestrian and vehicle behaviors. This paper presents a comparative analysis of LSTM and GRU models for predicting pedestrian and vehicle movements, focusing on illegal crossings in Panepistimiou road, near Omonoia, in the center of Athens, Greece. The study leverages data from video footage of one peak traffic hour, processed through the YOLOv8 object detection algorithm, integrated with ResNet-50 for feature extraction, Kalman filtering, homography transformations, and object re-identification to ensure high accuracy in detecting pedestrian and vehicle interactions. The primary goal of this study is to evaluate and compare the predictive performance of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models in predicting pedestrian behavior, focusing on illegal crossings during different traffic light phases on Panepistimiou Road, near Omonoia, in Athens, Greece. Both the LSTM and GRU models were trained to predict pedestrian violations, with performance evaluated based on Precision, Recall, and F1-Score. While both models performed well, the LSTM model showed a better balance between precision and recall, while the GRU model demonstrated higher recall, effectively identifying more violations but at the cost of more false positives. The findings suggest that while GRU is better at detecting violations, LSTM provides a more reliable solution in scenarios where minimizing false alarms is important. These results underline the trade-offs between model performance and the specific needs of real-time traffic monitoring systems, where both accuracy and efficiency are critical. In conclusion, this study highlights the strengths and limitations of LSTM and GRU in predicting pedestrian violations and suggests future work in optimizing threshold settings, feature engineering, and model tuning to further improve performance for urban traffic safety systems.

Keywords: Object Detection, Urban Traffic Safety, YOLOv8, LSTM, GRU

INTRODUCTION

Millions of lives are lost and a high number of injuries are recorded every year emphasizing the global impact of road traffic crashes. Human behavior, road characteristics (design and condition of roads), vehicle safety standards, environmental factors, and socioeconomic differences, are some of the complex factors which contribute to road incidents. Road safety is a matter of critical concern and this is why it is very important to address these issues. The European Union and the World Health Organization have set an ambitious goal to reduce by 50% the number of fatal traffic crashes by the year 2030 (6; 21). A crucial role to succeed at this target is the rise of new technologies, especially in the sector of advanced traffic monitoring and analysis.

A dynamic but often hazardous urban environment is created due to the co-existence of vehicles, pedestrians, and vulnerable road users (cyclists, scooters, children, etc.). To mitigate the risks, it is very important to implement effective traffic management and safety interventions (7). The city of Athens, Greece, presents an interesting case study, as it is an urban environment with very high levels of motorization, high traffic density, and urban challenges, and in the meantime, it lacks a comprehensive network of integrated traffic cameras and real-time monitoring systems in all its network. The absence and the dysfunctionality of this infrastructure raise the need for innovative solutions to address this problem and enhance road safety.

To confront this issue, computer vision and video recognition technologies are providing tools and methods to monitor and analyze traffic and in this way becoming fundamental techniques for road safety (8). The main aspect of these technologies is the ability to detect and track objects, and in the case of the urban road environment, vehicles and any transport mode, pedestrians, and other road users. In this way, they offer real-time insights regarding urban traffic conditions and possible dangers, which could affect road safety (14).

Object detection models like YOLO (You Only Look Once) have been extensively studied for their ability to deliver real-time performance, framing object detection as a single regression problem. YOLO models, such as YOLOv3 and YOLOv4, have demonstrated exceptional speed and accuracy, making them well-suited for applications like traffic monitoring and autonomous driving (7; 10). A study highlights the effectiveness of YOLO in pothole detection, demonstrating its potential in various road safety applications (18). The rise of object detection technologies, such as YOLOv8 (You Only Look Once), has revolutionized traffic monitoring by enabling the detection and tracking of objects in real-time from video streams. YOLOv8, in combination with advanced feature extraction models like ResNet-50 and tracking algorithms such as Kalman filters, has been employed in the context of the EU PHOEBE project ("Predictive Approaches for Safer Urban Environment") project (15), which aims to advance the application of traffic simulation tools and road safety assessments to enable transport planners and managers to fully understand and address the safety implications of changes in road conditions, mode choice, new modes, road user behaviors, and other factors.

While object detection and tracking provide essential data, the analysis of dynamic behaviors, such as pedestrian compliance with traffic signals, requires predictive modeling. Deep learning architectures, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, excel in modeling temporal dependencies and dynamic sequences, making them ideal for forecasting pedestrian and vehicle behaviors in urban encironments (21). These models can identify patterns in pedestrian and vehicle movements, predict violations like illegal crossings, and provide actionable insights to improve traffic management and safety. However, the effectiveness of these models depends not only on the quality of the algorithms but also on the availability of accurate and comprehensive data for model training.

This study evaluates and compares the predictive performance of LSTM and GRU models in forecasting pedestrian behavior during different traffic light phases (green, red, and intergreen). By leveraging YOLOv8 (24) for object detection and ResNet-50 for feature extraction, the study focuses on predicting illegal pedestrian crossings, a significant contributor to urban traffic risks. The models are

assessed using accuracy, precision, recall, and F1-score to determine their effectiveness in real-world scenarios. This comparison aims to identify the model that best balances prediction accuracy and false positive rates, thus enhancing the reliability of traffic safety systems.

A significant challenge to this endeavor is the limited availability of video data in Athens. The lack of integrated cameras and comprehensive street view data in most parts of the city of Athens poses a significant gap in the video footage availability and consecutively to the effective function of traffic management systems and safety interventions. Matters are further complicated by the fact that there are significant recording restrictions and overall enforcement imposed by the police as the city center is rife with governmental and public service buildings and facilities while being simultaneously critical for the regulation of traffic throughout the city. As a consequence, transport engineers needing to analyse video data for Athens traffic are very frequently required upon to conduct field research without relying on fixed infrastructure oriented on key locations on the transport network.

The findings of this research contribute to the development of innovative traffic safety solutions, particularly for cities like Athens, where limited monitoring infrastructure and high urban density demand advanced technologies. By demonstrating the utility of LSTM and GRU models in analyzing pedestrian behavior, this study provides valuable insights for urban traffic safety strategies and informs the broader field of machine learning applications in transportation engineering.

Conducted within the framework of the EU PHOEBE project (15), this research aligns with the project's goals to improve urban mobility and road safety through advanced simulation and assessment tools. By addressing the challenges of pedestrian compliance and traffic signal interactions, the study supports efforts to create safer urban environments and meet the ambitious road safety targets set by global initiatives.

The paper is structured as follows. The Introduction provides a comprehensive overview of the background, objectives, and significance of the study, emphasizing the context of the PHOEBE project and the integration of advanced object detection technologies for pedestrian behavior prediction. The Data Description section details the dataset utilized in this study, including the video footage, object detection models (YOLOv8 and ResNet-50), and the tracking methods applied to capture pedestrian and vehicle behavior. The Methodology section outlines the steps involved in data preparation, the training of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, and the evaluation of their performance based on key classification metrics such as precision, recall, F1-score, and confusion matrix. In the Results section, the performance of the LSTM and GRU models is discussed and compared, particularly regarding their ability to predict illegal pedestrian crossings and enhance urban traffic safety. Lastly, the Conclusion summarizes the key findings, discusses their practical implications for improving pedestrian safety, and suggests avenues for future research in the field of urban traffic monitoring and deep learning applications.

METHODS

Data Description

The dataset for this study was collected by strategically setting up smartphone cameras at the key point of Panepistimiou Street in the Omonoia area, a central and heavily trafficked location in Athens, Greece, connecting the two central-most squares of the capital city: Syntagma Square (with parliament and governmental functions) and Omonoia Square (with historic location and business functions). This location of this urban arterial was chosen as part of the EU PHOEBE project to provide a representative sample of urban traffic scenarios, including various types of vehicles, pedestrians, and cyclists (15). The absence of integrated traffic cameras and comprehensive street view data in the area necessitated manual video capture. This was attained by using field researchers who shot video footage using commercially available smartphones and tripods and compiling them in external databases at the end of each workday data collection.

The entire data collection period for a number of key point locations had a duration of 4 days, spread across 2 weeks in order to record two typical days per week (Tuesdays and Thursdays) for each location. The intent was to collect video footage encompassing a peak traffic hour of each workday when the city center is at its busiest (thus presenting the highest challenge to the algorithms), specifically on a weekday from 9 am till 10 am and a non-peak traffic hour later in the same day at 8 pm till 9 pm. Ultimately, more than 8 hours of video were collected for each location. Indicative images depicting one of the recorded locations can be found in Figure 1 below, taken by the authors.

Figure 1 Panepistimiou Urban Arterial in Athens near Omonoia Square

Object Detection and Feature Extraction

The first stage of processing involves detecting objects within the video frames. For this task, the YOLOv8 model is used, which formulates object detection as a single regression problem. Given an input image extracted as a frame from the video, YOLOv8 (24) outputs a set of bounding boxes B=$(B1, B2, \dots, B_n)$, where each bounding box B_i is defined by:

$$B_i = (x_i, y_i, w_i, h_{i,} c_i)$$

Where (x_i, y_i) are the coordinates of the center of the box, $w_i, and\ h_i$ are the width and height of the bounding box, and c_i is the class probability score indicating whether the object in the box is a pedestrian or vehicle.

Once the objects are detected, ResNet-50 is employed for feature extraction. The feature vector F_i corresponding to each detected object B_i is extracted from the output of ResNet-50, represented as:

$$F_i = ResNet - 50(B_i)$$

These features are then used for tracking and re-identification.

To track pedestrians over time, the system utilizes Kalman filters (12). The Kalman filter operates in two phases: Prediction and Update. Firstly, this phase predicts the state estimate at time $t \in T$ based on the estimate from time $(t - \Delta t) \in T$:

$$\hat{s}_m(t|t - \Delta t) = A \cdot s_m(t - \Delta t); \quad \forall t \in T; \ \ \forall m \in M(t); \tag{1}$$

Secondly, the phase predicts the error covariance matrix:

$$P_m(t|t - \Delta t) = A \cdot P_m(t - \Delta t) \cdot A^T + Q_m(t); \quad \forall t \in T; \ \ \forall m \in M(t); \tag{2}$$

Then, for the updating phase the Kalman Gain $K_m(t)$, is computed, which determines how much the predictions are adjusted based on the new measurements:

$$K_m(t) = P_m(t|t-\Delta t) \cdot O^T \cdot \left(O \cdot P_m(t|t-\Delta t) \cdot = O^T + R_m(t)\right)^{-1}; \quad \forall t \in T; \ \forall m \in M(t); \quad (3)$$

The Kalman Gain balances the uncertainties in the prediction and the measurement. Additionally, the Hungarian algorithm is applied for data association. This algorithm optimizes the association between the predicted object locations and new detections, minimizing the cost function for object-to-track association.

A cost matrix C is generated where each element $C_{n,m}$ represents the cost of assigning the detected object $n \in N(t)$ to the tracked object $m \in M(t)$. The total cost $C_{n,m}$ is computed as a weighted sum of the factors as indicated in Eqn. (4). Negative signs are applied to $IoU_{n,m}$, $CosSim_{feat,n,m}$ and $CosSim_{emb,n,m}$ because higher values of these factors indicate a better match, and, thus, a lower cost. Additionally, cosine similarity returns values between -1 and +1, and $d_{n,m}$ represents a distance that can exceed 1. Since the cost matrix involves different metrics, it's common to normalize all metrics to the same range, typically between 0 and 1, to ensure consistency when combining them. Therefore, normalizations are introduced in Eqn. (4) for these factors.

$$C_{n,m} = -w_{IoU} \cdot IoU_{n,m} - w_{feat} \cdot \frac{CosSim_{feat,n,m}+1}{2} - w_{emb} \cdot \frac{CosSim_{emb,n,m}+1}{2} + w_{cent} \cdot \frac{d_{centroid,n,m}}{D_{max}}; \quad \forall t \in T; \ \forall (n,m) \in N(t) \times M(t); \quad (4)$$

Where:

- $IoU_{n,m}$ be the Intersection over Union between the bounding boxes b_n and $\hat{b}_m$, defined as follows (higher values indicates a better spatial match):

$$IoU_{n,m} = \frac{\text{area}(b_n \cap \hat{b}_m)}{\text{area}(b_n \cup \hat{b}_m)}; \quad \forall t \in T; \ \forall (n,m) \in N(t) \times M(t); \quad (5)$$

- $CosSim_{feat,n,m}$ be the Cosine Similarity of feature vectors $r_n(t)$ and $\bar{r}_m(t)$, defined as follows (values close to 1 indicate high visual similarity):

$$CosSim_{feat,n,m} = \frac{r_n(t) \cdot \bar{r}_m(t)}{\|r_n(t)\| \, \|\hat{r}_m(t)\|}; \quad \forall t \in T; \ \forall (n,m) \in N(t) \times M(t); \quad (6)$$

- $CosSim_{emb,n,m}$ be the Cosine Similarity of embedding vectors $e_n(t)$ and $\bar{e}_m(t)$, defined as follows (values close to 1 indicate high visual similarity):

$$CosSim_{emb,n,m} = \frac{e_n(t) \cdot \bar{e}_m(t)}{\|e_n(t)\| \, \|\bar{e}_m(t)\|}; \quad \forall t \in T; \ \forall (n,m) \in N(t) \times M(t); \quad (7)$$

- $d_{n,m}$ be the Euclidean distance between projected ground positions $p_{world,n}$ and $\hat{p}_{world,m}$, defined as follows (smaller distances indicate better alignment):

$$d_{n,m} = \|p_{world,n} - \hat{p}_{world,m}\|; \quad \forall t \in T; \ \forall (n,m) \in N(t) \times M(t); \quad (8)$$

- D_{max} be the maximum expected Euclidean distance $d_{centroid,n,m}$ for normalization.

- $\delta_{n,m}$ be a binary variable, defined as follows:

$$\delta_{n,m} = \begin{cases} 1 & \text{if detected object } n \text{ is assigned to tracked object } m \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Once objects are detected and tracked, the system analyzes pedestrian behavior in relation to traffic signal phases. The pedestrian behavior is classified as either legal or illegal depending on whether they cross the pedestrian line during a red pedestrian light. The behavior status of tracked pedestrian is $S_{ped}(t)$, where:

$$S_{ped}(t) = \begin{cases} illegal & \text{if } S_{light}(t) = red \\ legal & \text{if } S_{light}(t) = green \\ unknown & \text{if } S_{light}(t) = intergreen \end{cases}; \quad (10)$$

Each illegal crossing event is logged, and the number of violations is counted.

After detecting illegal pedestrian crossings, the system generates annotated outputs in the form of video annotations and data files. The video output includes the marked illegal crossings, with timestamps, bounding boxes, and corresponding traffic signal phases for each detected event. The data files contain the count of illegal crossings, along with information about the specific times and locations of the violations.

LSTM Methodology

Long Short-Term Memory (LSTM) networks are a specialized form of Recurrent Neural Networks (RNNs) that are particularly suited for sequential data due to their ability to capture long-term dependencies without suffering from the vanishing gradient problem typical of traditional RNNs (3). In this study, LSTM was employed to predict pedestrian behavior, focusing on their illegal crossings during different traffic light phases. The LSTM model is composed of multiple layers, with the primary advantage being its architecture that processes sequential data efficiently. The network includes a series of repeating cells, each containing three gates:

- Forget Gate: Determines which information from the previous state should be discarded.
- Input Gate: Decides which new information should be added to the memory state.
- Output Gate: Filters the memory and decides which parts of the state will be used as output for the current time step.

These gates allow the LSTM to "remember" important information over long sequences, making it ideal for predicting future behavior based on historical data.

The LSTM model used in this study was designed with a two-layer architecture, with 64 and 32 units in each layer. The Bidirectional LSTM layers enable the model to process the input sequences in both forward and backward directions, capturing both past and future context in the data. ReLU activation is used in the hidden layers to capture complex relationships between input variables, such as the traffic light phases, historical violations, and other environmental factors. A dropout rate of 0.6 is applied after the first LSTM layer to reduce overfitting, and recurrent dropout of 0.5 is applied after the second LSTM layer. These parameters help regularize the model by randomly deactivating some neurons during training, preventing the model from overly relying on specific features and improving its ability to generalize to unseen data. The Adam optimizer is used for model optimization, with a learning rate of 0.0002. The model is trained over 50 epochs with a batch size of 32, and 10% of the training data is reserved for validation. The output layer uses a tanh activation function to constrain the output values between -1 and 1, which correspond to the predicted pedestrian violations (illegal crossings) during different traffic light phases (red, green, and intergreen).

The performance of the LSTM model was evaluated using common classification metrics, including accuracy, precision, recall, and F1-score, as outlined in the following equations:

Accuracy, which measures the proportion of correctly classified observations, is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (11)$$

Precision, which quantifies the number of positive class predictions that actually belong to the positive class, is defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

Recall, also known as True Positive Rate, which measures the proportion of actual positive cases correctly identified by the model, is defined as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

F1-score, which combines precision and recall into a single measure, is defined as follows:

$$\text{f1-score} = \frac{2x\ (Precision)x\ (Recall)}{(Precision)+(Recall)} \quad (14)$$

False alarm rate, which measures the proportion of negative cases incorrectly classified as positive is defined as follows:

$$\text{False Alarm Rate} = \frac{FP}{FP+TN} \quad (15)$$

Where, True Positive (TP) denotes instances belonging to class i that were accurately classified as such. True Negative (TN) refers to instances not belonging to class i and correctly not classified as such. False Positive (FP) indicates instances that do not belong to class i but were erroneously classified as part of it. Lastly, False Negative (FN) signifies instances that belong to class i but were mistakenly not classified as such.

GRU Methodology

The Gated Recurrent Unit (GRU) is a type of Recurrent Neural Network (RNN) designed for sequential data, and it is particularly effective in capturing dependencies over time. GRUs have been widely used for time series prediction tasks, where future behavior is inferred based on prior sequences. Unlike traditional RNNs, GRUs overcome the vanishing gradient problem, which allows them to retain long-term dependencies and memory of past states (16, 2). In this study, the GRU model is employed to predict pedestrian illegal crossings during various traffic light phases, based on historical data and other dynamic features.

The GRU architecture consists of two main components:

- Update Gate: Controls how much of the past information should be carried over to the current state.
- Reset Gate: Determines the amount of previous information that should be discarded.

These gates allow the GRU to capture relevant patterns in sequential data by controlling the flow of information across time steps, enabling the model to make more accurate predictions over longer sequences.

For this study, the GRU model was designed with two Bidirectional GRU layers. The Bidirectional nature allows the model to process data in both forward and backward directions, which captures both past and future context for better prediction accuracy. The GRU layers were followed by a Dense layer with ReLU activation to capture complex relationships between variables, such as traffic light phases,

pedestrian behavior, and other environmental factors. A Dropout layer was added to prevent overfitting, and a tanh activation function in the output layer ensured that the predicted pedestrian violations were constrained within a specified range.

The GRU model was trained using the Adam optimizer, with a learning rate of 0.0002. The training process lasted for 50 epochs with a batch size of 32. EarlyStopping was employed to prevent overfitting, with a patience of 10 epochs for monitoring validation loss. This model was evaluated using the same metrics with the LSTM model.

Compared with Long Short-Term Memory (LSTM) (9), the unit of the GRU is simpler to compute and implement (2). The training efficiency of GRU gets improved with the number of weights reduced. However, LSTM can remember longer sequences than GRU, as LSTM has a more sophisticated memory cell.

RESULTS

Algorithmic outputs

The dataset used in this study was produced through an advanced video analysis algorithm, designed to track pedestrian behavior and violations, particularly illegal crossings, in an urban environment. The video footage was captured at a busy pedestrian crossing in central Athens, where the algorithm was employed to process and analyze the data. Key to this process was the use of YOLOv8 (You Only Look Once) for object detection, ResNet-50 for feature extraction, and Kalman filtering for tracking the movement of pedestrians and vehicles.

A crucial component of the algorithm involved traffic light detection. The system identified the traffic light status (red, green, or intergreen) at the intersection, which was integral to the detection of illegal pedestrian crossings. This was done by leveraging pre-trained models and combining them with the region-based classification of the image frames. The Region of Interest (ROI) for this analysis was defined to focus specifically on the pedestrian crossing zones, which were marked in the video footage. The ROI was crucial for limiting the area under analysis to the relevant pedestrian zones, filtering out irrelevant information, and improving the accuracy of violation detection.

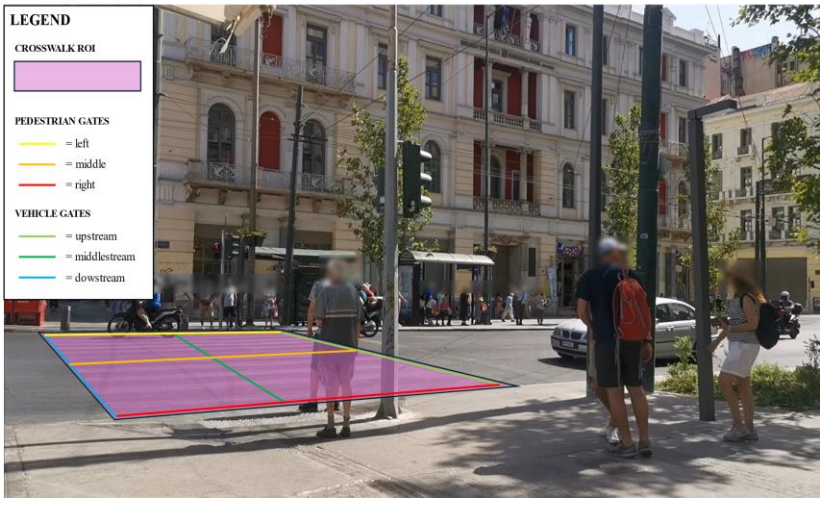

Figure 2: Pedestrian crosswalk ROI, with the pedestrian and vehicle gates defined for the Omonoia spot.

Pedestrian illegal crossings were detected when pedestrians entered the defined ROI during red or intergreen phases. The algorithm tracked the position and movement of pedestrians across the intersection and compared these movements with the current traffic light status to determine whether a violation had occurred.

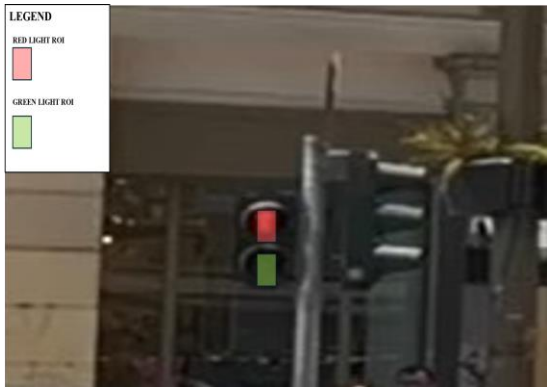

Figure 3: Pedestrian traffic light ROIs defined for the Omonoia Square spot

Once violations were detected, the model recorded the time, location, and traffic light phase in a CSV file, which was used for further analysis. This data was then stored in a CSV file that contains the following columns:

- Time (seconds): The timestamp of when a violation occurred.
- Pedestrian_IllegalMiddleGate: A binary variable indicating whether a pedestrian violated the traffic signal by crossing illegally.
- Predominant Traffic Light Status: The state of the traffic light at the time of the violation (Red, Green, Unknown).
- Vehicle_IllegalMiddlestreamGate: Indicates whether a vehicle also violated traffic rules at the same time.

TABLE 1 Dataset sample with the illegal detection

Time (seconds)	Pedestrian_Illegal MiddleGate	Predominant traffic light status	Vehicle_Illegal Middlestream Gate
428	1	red	0
429	0	green	0
430	0	red	0
431	1	red	0
432	0	red	0
433	2	red	0
434	0	red	0

This dataset was produced using smartphone cameras mounted on tripods at key pedestrian crossing points in Athens, allowing for a flexible and cost-effective solution for data collection. The data collection process ensured a representative sample of pedestrian behavior under varying traffic conditions and light phases, and it avoided the need for fixed infrastructure, making it applicable to urban environments with limited monitoring systems.

By focusing on the ROI and employing traffic light detection, the algorithm was able to provide precise and actionable insights into pedestrian behavior, specifically illegal crossings during different traffic phases. The processed data was used to train and evaluate the performance of the predictive models, providing a comprehensive dataset for assessing pedestrian behavior and violations across urban intersections.

LSTM results

The Long Short-Term Memory (LSTM) model was employed to predict pedestrian behavior and illegal crossings during different traffic light phases. The architecture of the LSTM model consists of two Bidirectional LSTM layers, each with 64 and 32 units respectively. Bidirectional LSTM layers enable the model to capture both past and future context from the input sequences, which is crucial for understanding the temporal dependencies in pedestrian behavior. The network uses ReLU activation in the hidden layers to model complex relationships between input variables such as traffic light phases, historical violations, and environmental factors. Additionally, dropout layers were added to regularize the model and prevent overfitting, with a dropout rate of 0.6 after the first LSTM layer and 0.5 after the second. The LSTM model achieved strong performance in terms of both classification and regression metrics (Tables 2 and 3).

TABLE 2 Confusion Matrix and Classification metrics for the LSTM model

Confusion Matrix		
	Positive	Negative
Positive	449	161
Negative	502	57
Classification Metrics		
Accuracy	0.812	
Precision	0.736	
Recall	0.887	
F1-Score	0.805	

These results indicate that the LSTM model performs well in detecting pedestrian violations, with a high recall of 0.887. This suggests the model's effectiveness in identifying true violations, although its precision of 0.736 indicates that the model also predicts some false positives. The F1-Score of 0.805 highlights the model’s balanced performance between precision and recall, which is essential for applications where minimizing false alarms is a priority.

Below, Figure 4 shows a comparison of true pedestrian violations (represented by the solid line) and the LSTM model’s predicted violations (represented by the dashed line). The LSTM model demonstrates a strong ability to predict illegal pedestrian crossings, particularly highlighting the model's higher recall rate. The LSTM's bidirectional architecture captures both past and future pedestrian behavior, which helps it better generalize over long sequences of data. This is evident in the graph, where the predicted violations closely follow the true violations over time.

It is important to note that, while the LSTM model shows a good alignment with the true violations, there are occasional peaks in the predictions that do not match the actual data, indicating false positives. This behavior emphasizes the trade-off between high recall (detecting more true positives) and precision (minimizing false positives). The higher recall of the LSTM model is beneficial for applications focused on capturing as many violations as possible, even if it results in slightly more false alarms.

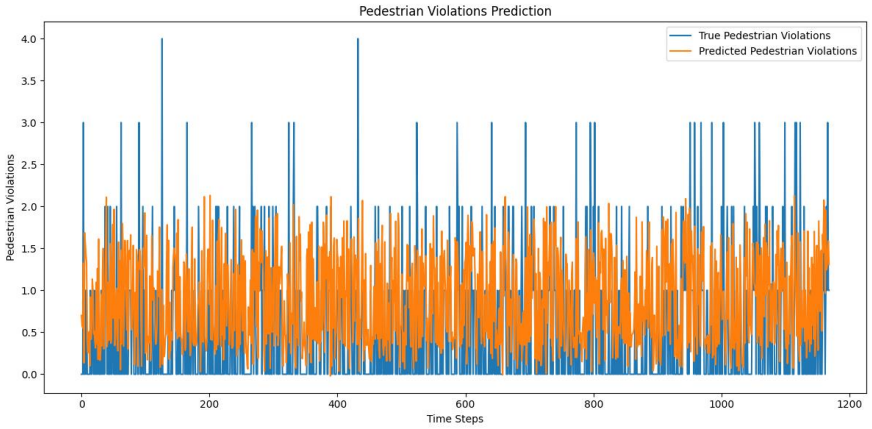

Figure 4: True vs. Predicted Pedestrian Violations Plot for the LSTM model

Figure 5 displays the residuals (errors) of the LSTM model's predictions. A narrow distribution of residuals indicates good model performance, while a wider distribution signals more variability between predicted and true values. The LSTM model's residuals indicate that it is relatively accurate, with some instances showing greater deviation from the true violations.

The LSTM model's residuals show a wider distribution compared to the GRU model as showcased below in figure 8, reflecting a higher occurrence of prediction errors, especially for more complex pedestrian behaviors. This variance can be attributed to the LSTM's sensitivity to different sequences and its trade-off between recall and precision.

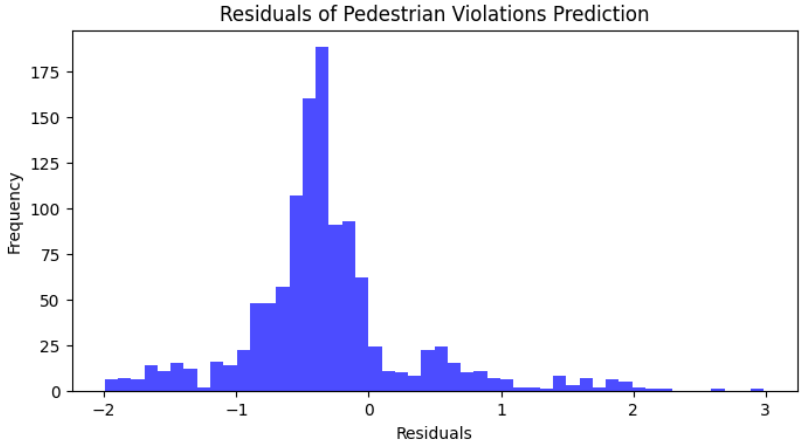

Figure 5: Residuals/Error Distribution Plot for the LSTM model

Regarding Figure 6 the training and validation loss, as well as the training and validation accuracy, over the course of training for the LSTM model are being depicted. The graph indicates that both the training loss and validation loss decrease over time, while accuracy improves. This is a sign that the model is learning effectively from the data and is not overfitting, as there is little divergence between the training and validation curves. The steady decrease in both loss curves and the increase in accuracy indicate that the LSTM model is fitting the data well and generalizing effectively. The slight fluctuations in the validation loss indicate occasional overfitting, but overall, the model is improving with each epoch. The model's learning behavior is typical of LSTM networks, where a longer training process often helps improve generalization for sequential data.

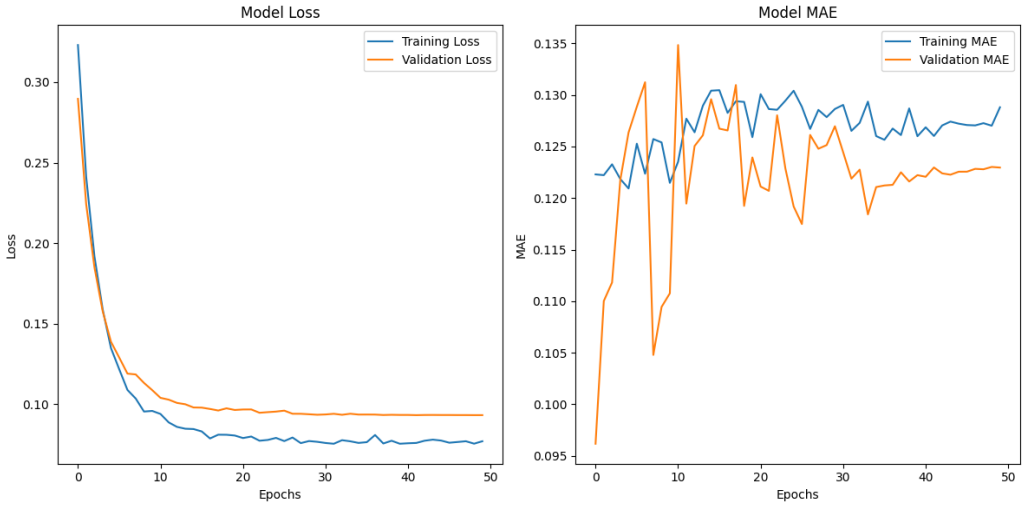

Figure 6: Model Loss and Accuracy Over Epochs

GRU Model Results

The Gated Recurrent Unit (GRU) model was also designed to predict pedestrian behavior, utilizing two Bidirectional GRU layers, similar to the LSTM model, but with 64 and 32 units respectively. The Bidirectional architecture allows the model to process input sequences in both forward and backward directions, capturing long-term dependencies from both the past and the future. The GRU model was optimized using the Adam optimizer with a learning rate of 0.0002, and the model was regularized with dropout layers to prevent overfitting. A ReLU activation was used in the intermediate layers, while the final output layer utilized tanh activation to constrain the predictions within a range suitable for classification. The GRU model exhibited strong recall performance (Tables 4 and 5) but slightly lower precision compared to the LSTM model.

TABLE 3 Confusion Matrix and Classification metrics for the GRU model

Confusion Matrix		
	Positive	Negative
Positive	471	278
Negative	384	36
Classification Metrics		
Accuracy	0.731	
Precision	0.629	
Recall	0.929	
F1-Score	0.749	

The GRU model performed excellently in terms of recall (0.929), indicating its ability to detect a large proportion of true pedestrian violations. However, the precision (0.629) was lower than the LSTM model, indicating a higher rate of false positives. Despite the lower precision, the F1-Score of 0.749 reflects a solid balance between precision and recall.

Figure 7 below presents the comparison between true pedestrian violations (solid line) and predicted pedestrian violations (dashed line) from the GRU model. The GRU model closely tracks the

true violations but with fewer deviations, indicating its higher precision compared to the LSTM model. The GRU model performs well in minimizing false positives, as seen in this plot. The predicted values align more closely with the true violations, demonstrating that the GRU model is more precise, with fewer false alarms. The balance between precision and recall is evident here, where the GRU model captures a significant number of true positives while limiting false positives.

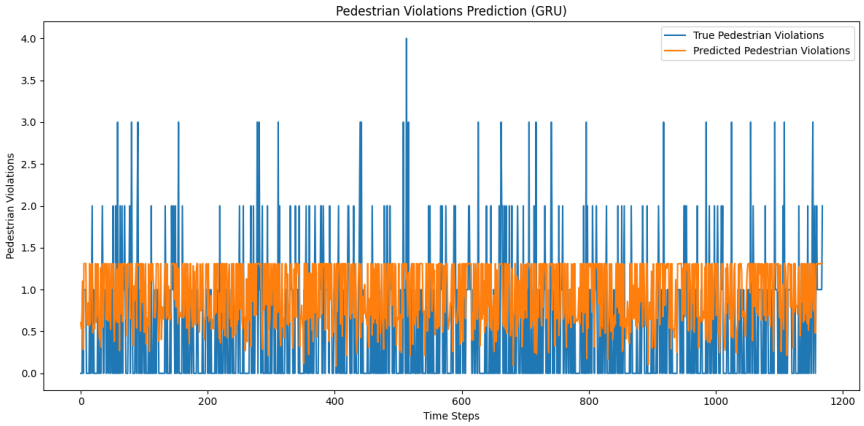

Figure 7: True vs. Predicted Pedestrian Violations Plot for the GRU model

Figure 8 shows the residuals (errors) for the GRU model's predictions. A narrow, centered distribution suggests the GRU model's predictions are accurate, with fewer large deviations from the true violations. The GRU model shows a tighter distribution of residuals compared to the LSTM model, indicating that its predictions are more precise. The fewer large errors imply that the GRU model is more consistent in its predictions, especially in time-series data, where maintaining accuracy is crucial. This makes the GRU model well-suited for applications where reducing false positives is a priority.

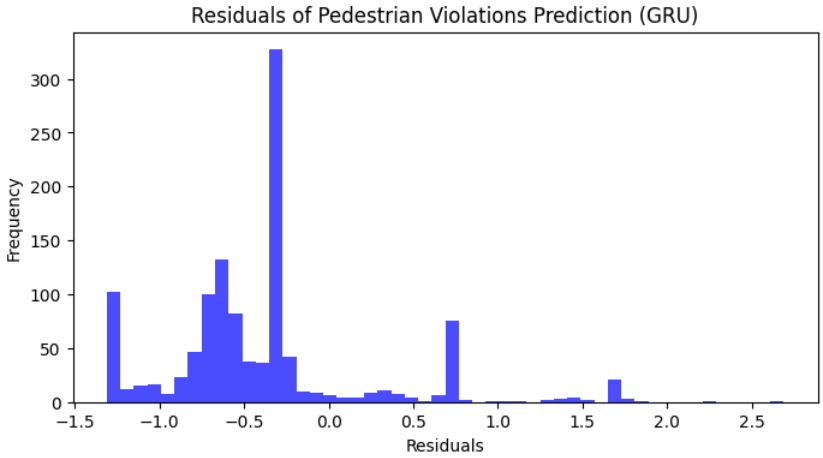

Figure 8: Residuals/Error Distribution Plot for the GRU model

Figure 9 displays the training and validation loss, as well as the training and validation accuracy, over epochs for the GRU model. Like the LSTM model, the GRU model shows a steady decrease in loss and an increase in accuracy over time, suggesting effective learning. The validation loss follows a similar trend to the training loss, indicating that the GRU model is generalizing well without significant overfitting.

The GRU model shows faster convergence compared to the LSTM model, with both training and validation loss decreasing rapidly within the first few epochs. This is indicative of the GRU's ability to

learn more efficiently from sequential data, especially in tasks like pedestrian behavior prediction. The smooth convergence and consistent performance suggest that the GRU model is stable and less prone to overfitting compared to the LSTM.

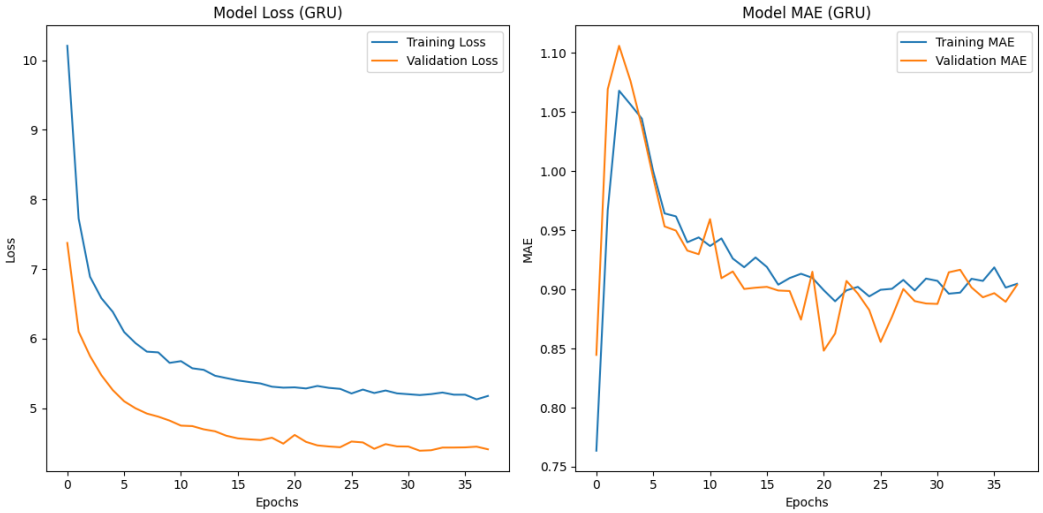

Figure 9: Model Loss and Accuracy Over Epochs for the GRU Model

Comparative Analysis of LSTM and GRU Models

The LSTM and GRU models were evaluated and compared based on their ability to predict pedestrian behavior, specifically illegal crossings during different traffic light phases. Both models were assessed using classification metrics, including precision, recall, and F1-score, which were used to determine their effectiveness in detecting pedestrian violations.

In terms of precision, the LSTM model outperformed the GRU model with a value of 0.736 compared to 0.629 for the GRU model. Precision measures the proportion of true positive predictions among all positive predictions, meaning the LSTM model was more reliable in correctly predicting pedestrian violations. This characteristic is particularly important when false positives are costly or disruptive in real-world applications, where minimizing unnecessary alarms is essential.

On the other hand, the GRU model showed superior recall, achieving a value of 0.929, compared to 0.887 for the LSTM model. Recall measures the proportion of true positives that are correctly identified by the model. The GRU model's higher recall indicates that it is more sensitive in detecting pedestrian violations, capturing a larger proportion of actual violations. However, this comes at the cost of more false positives, as reflected in its lower precision.

When considering the F1-score, which combines precision and recall into a single metric, the LSTM model had a higher score of 0.805, compared to 0.749 for the GRU model. The F1-score reflects the trade-off between precision and recall, and the LSTM model's higher F1-score suggests a better overall balance between minimizing false positives and maximizing true positive detections.

The GRU model excels in recall, making it more effective at detecting pedestrian violations, which is advantageous in applications where detecting as many violations as possible is the primary goal. However, this comes with a drawback of a lower precision, meaning the GRU model generates more false positives.

The LSTM model, with its higher precision and more balanced F1-score, is more reliable for scenarios where minimizing false alarms is critical. Although the LSTM model detects fewer violations compared to the GRU model, it provides more accurate predictions in terms of distinguishing between violations and non-violations.

DISCUSSION

The results from the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models provide valuable insights into the ability to predict pedestrian illegal crossings in complex urban traffic environments. Both models demonstrate strong performance, with notable differences in their capabilities, particularly in terms of recall and precision.

The LSTM model, with its bidirectional architecture, demonstrated higher recall, capturing more true positive instances of illegal pedestrian crossings. This result underscores the model's ability to generalize over longer sequences and leverage both past and future states of pedestrian behavior. The LSTM model's architecture, which included two bidirectional LSTM layers with dropout regularization, effectively mitigated overfitting, a common issue in deep learning models dealing with sequential data. While the LSTM model excelled at detecting more violations, its precision was slightly lower than that of the GRU model, indicating a trade-off between sensitivity and false positives.

On the other hand, the GRU model showed a slightly higher precision and a more balanced performance with respect to precision and recall. This model, utilizing bidirectional GRU layers and regularization techniques, is well-suited for scenarios where minimizing false positives is crucial, while still maintaining a high detection rate. In particular, the GRU model's ability to manage the complexities of sequential data proved effective, especially in time-series prediction tasks, aligning with findings from previous studies (20), who demonstrated the strength of GRU in handling sequential dependencies for traffic predictions. The performance of GRU, especially in the context of its predictive capabilities for illegal crossings, aligns with the expectations of modern urban traffic safety applications.

Both models were trained on a dataset with features such as traffic light phases, pedestrian behavior patterns, combined with temporal sequences. The feature engineering process included encoding traffic light states and introducing lag features to account for previous violations. Additionally, spike detection techniques were employed to handle imbalanced data, addressing the class imbalance often seen in violation detection tasks, as highlighted in a number of studies (13; 4). This allowed the models to adapt to the complexities of real-world traffic scenarios. The combination of GRU, CNN, and LSTM has shown excellent performance in traffic congestion predictions as well, especially in the context of urban road environments (1), where the model integrates these three components for traffic prediction, handling both temporal and spatial complexities in traffic flow data.

In comparison to previous studies (22; 13), this research offers a novel contribution by applying both LSTM and GRU models specifically for pedestrian behavior prediction in urban environments. Zhou et al. (22) emphasized the importance of incorporating posture features and local context, which was also reflected in the feature engineering of this study. Similarly, the use of spatio-temporal information in pedestrian crossing prediction, a method employed by Dofitas et al. (4), further reinforces the relevance of these models in advancing traffic monitoring systems.

Furthermore, studies like Yao and Ye (11), who used LSTM to predict traffic flow, highlight the broad applicability of LSTM models in traffic prediction. Wang et al. (20) explored LSTM model improvements by integrating attention mechanisms to capture high-impact traffic flow values, supporting our findings that enhanced learning techniques can improve the accuracy of pedestrian behavior predictions, especially with complex data inputs.

Additionally, the work of Ullah et al. (19) introduced an attention residual LSTM for anomaly recognition in surveillance videos, which shares conceptual similarities with our approach. Their framework, which incorporates a residual LSTM to handle anomalies efficiently, presents a relevant comparison to the current study's methods of capturing abnormal pedestrian behavior in urban environments. This highlights the potential for combining spatiotemporal learning with attention mechanisms to further improve prediction accuracy in real-time surveillance systems.

CONCLUSIONS

This research contributes to the development of advanced traffic management systems, where deep learning models like LSTM and GRU can play a crucial role in predicting pedestrian behavior and improving urban road safety. The present work demonstrates the efficacy of LSTM and GRU models in predicting pedestrian behavior in urban traffic environments. The LSTM model showed higher recall, making it a more reliable model for capturing illegal pedestrian crossings, while the GRU model excelled in precision, reducing false positives. Both models exhibited strong performance, showing that they can be applied to real-time traffic monitoring systems.

Our results are consistent with previous research that applied deep learning techniques to pedestrian behavior prediction. Pawar et al. (13) applied LSTM to anomaly detection in traffic surveillance, which aligns with our findings regarding LSTM's ability to capture anomalous pedestrian behavior. The integration of YOLOv8 for real-time object detection in our study further enhances the predictive power of the models, similar to the way Dofitas et al. (4) combined CNNs with LSTMs to improve recognition tasks in traffic environments.

This study suggests that while LSTM is beneficial for capturing a broad range of violations, GRU provides an effective approach when minimizing false positives is a priority. The choice between the LSTM and GRU models depends on the specific requirements of the application. If the goal is to detect as many pedestrian violations as possible, even at the cost of more false positives, the GRU model would be preferred. However, if minimizing false alarms and achieving a better balance between precision and recall is essential, the LSTM model is the more suitable choice. In terms of future work, expanding the models to include additional environmental features, such as road conditions or pedestrian density, could further improve prediction capabilities. Integrating multi-modal data from different sensors or video sources would also help create a more robust system that is adaptable to various urban traffic conditions.

Looking ahead, further research could explore incorporating additional environmental factors, such as weather conditions and road types, to refine predictions. Moreover, integrating multi-modal data from different sensors or video feeds could lead to even more robust systems capable of handling various real-world traffic scenarios. As shown by Yao & Ye (11), enhancing feature sets and integrating new learning techniques will contribute to improving traffic monitoring systems' performance.

ACKNOWLEDGMENTS

The present research was carried out within the research project "PHOEBE - Predictive Approaches for Safer Urban Environment", which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101076963.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Stella Roussou, Apostolos Ziakopoulos, George Yannis; data collection: Stella Roussou; Roberto Ventura; analysis and interpretation of results: Stella Roussou, Apostolos Ziakopoulos; draft manuscript preparation: Stella Roussou. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Azhagiri P, R., , M. Deep learning based predicting urban traffic congestion with RGB-coded images using GRU-CNN and LSTM. Multimed Tools Appl 83, 86261–86280 (2024). https://doi.org/10.1007/s11042-024-20376-8

2. Cho, K. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

3. DiPietro, R.; Hager, G.D. Deep Learning: RNNs and LSTM. In Handbook of Medical Image Computing and Computer Assisted Intervention; Elsevier, 2020; pp. 503–519.

4. Dofitas, C., Jr., Gil, J. -M., & Byun, Y. -C. (2024). Multi-Directional Long-Term Recurrent Convolutional Network for Road Situation Recognition. Sensors, 24(14), 4618. https://doi.org/10.3390/s24144618

5. Du, W., Dash, A., Li, J., Wei, H., & Wang, G. (2023). Safety in Traffic Management Systems: A Comprehensive Survey. Designs, 7(4), 100.

6. European Commission. EU Road Safety Policy Framework 2021–2030 - Next Steps towards "Vision Zero", 2019.

7. Gefan, Y., & Yuchi, L. Object Detection in the KITTI Dataset using YOLO and Faster R-CNN.

8. Hammoudeh, M. A. A., Alsaykhan, M., Alsalameh, R., & Althwaibi, N. (2022). Computer Vision: A Review of Detecting Objects in Videos--Challenges and Techniques. International Journal of Online & Biomedical Engineering, 18(1).

9. Hochreiter, S. (1997). Long Short-term Memory. Neural Computation MIT-Press.

10. Islam, S. U., Ferraioli, G., Pascazio, V., Vitale, S., & Amin, M. (2024). Performance Analysis of YOLOv3, YOLOv4 and MobileNet SSD for Real Time Object Detection. The Sciencetech, 5(2), 37-49.

11. Jingxuan Yao, Yuntao Ye, The effect of image recognition traffic prediction method under deep learning and naive Bayes algorithm on freeway traffic safety, Image and Vision Computing, Volume 103, 2020, 103971, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2020.103971.

12. Kalman, R.E. A new approach to linear filtering and prediction problems. J. Basic Eng. 1960, 82, 35–45.

13. Karishma Pawar, Vahida Attar, Deep learning based detection and localization of road accidents from traffic surveillance videos, ICT Express, Volume 8, Issue 3, 2022, Pages 379-387, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2021.11.004.

14. Mammeri, A. J. Siddiqui, Y. Zhao, and B. Pekilis, "Vulnerable Road Users Detection based on Convolutional Neural Networks," 2020 International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 2020, pp. 1-6, doi: 10.1109/ISNCC49221.2020.9297332.

15. PHOEBE Project. PHOEBE Library. https://phoebe-project.eu/phoebe-library/. Accessed July 1, 2024.

16. Pascanu, R. (2013). On the difficulty of training recurrent neural networks. arXiv preprint arXiv:1211.5063.

17. Shile Zhang, Mohamed Abdel-Aty, Yina Wu, Ou Zheng, Modeling pedestrians' near-accident events at signalized intersections using gated recurrent unit (GRU), Accident Analysis & Prevention, Volume 148, 2020, 105844, ISSN0001-4575, https://doi.org/10.1016/j.aap.2020.105844.

18. Sonali Ashok Khude, Nishita Vitthal Patil, Snehal Santosh Darade, Sneha Santosh Galande, Aishwarya Maruti Sawant, Dr. Swati Pawar (2024), Enhancing Object Detection Accuracy Through Custom Dataset Using Yolo, DOI Link: https://doi.org/10.22214/ijraset.2024.59712.

19. Ullah, W., Ullah, A., Hussain, T., Khan, Z. A., & Baik, S. W. (2021). An Efficient Anomaly Recognition Framework Using an Attention Residual LSTM in Surveillance Videos. Sensors, 21(8), 2811. https://doi.org/10.3390/s21082811

20. Wang, J. Zhao, C. Shao, C. Dong and C. Yin, "Truck Traffic Flow Prediction Based on LSTM and GRU Methods With Sampled GPS Data," in IEEE Access, vol. 8, pp. 208158-208169, 2020, doi: 10.1109/ACCESS.2020.3038788.

21. World Health Organization, Global status report on road safety 2023. WHO, 2023.Güney, E., & Bayılmış, C. (2022). An implementation of traffic signs and road objects detection using faster R-CNN. Sakarya University Journal of Computer and Information Sciences, 5(2), 216-224.

22. Zhou, X., Ren, H., Zhang, T., Mou, X., He, Y., & Chan, C. -Y. (2022). Prediction of Pedestrian Crossing Behavior Based on Surveillance Video. Sensors, 22(4), 1467. https://doi.org/10.3390/s22041467

23. Yan Sun, Zheping Yan, Image target detection algorithm compression and pruning based on neural network, Jan. 2021, doi: /10.2298/CSIS200316007S.