ISSN 2007-9737

BiciVR: A Software Engineering Framework for AI-Driven Bicycle Mobility Risk Simulation

Ramón Alejandro Briseño Martínez[1], Wilmer Rodríguez[2], Edgar Cossio[3,*], Carolina Del Valle Soto[1]

[1] Universidad Panamericana, Facultad de Ingeniería, Mexico

[2] Universidad de Pamplona, Colombia

[3] Instituto de Información Estadística y Geográfica de Jalisco, Mexico

me@wrsbyte.com, edgar.cossio@iieg.gob.mx, {rbriseno, cvalle}@up.edu.mx

Abstract. The rise of micromobility in the Guadalajara Metropolitan Area (GMA) introduces new road safety challenges. This study proposes a software engineering framework that integrates virtual reality and data science to identify cyclist behavior and accident risk scenarios. A gamified simulation in Unity 3D, using Oculus Meta Quest headsets, allowed 36 participants to navigate realistic traffic conditions. Data on collisions and violations were analyzed using machine learning and heatmaps. Results show that 4.31% of violations led to collisions, with intersections and traffic density as key risk factors. This framework supports targeted interventions for safer urban mobility.

Keywords. Software engineering, virtual reality simulation, cycling mobility, risk assessment & machine learning analysis.

1 Introduction

According to estimates, the global micromobility market will reach a value of USD 69.32 billion by 2028, with a compound annual growth rate (CAGR) of 13.7% [1]. The evident increase in the use of micromobility vehicles, mainly bicycles and scooters, brings with it greater mobility management problems including the increase in accidents related to these. In the Guadalajara Metropolitan Area (GMA) Jalisco Mexico, the study area of this research, reports presented by groups and government entities of Jalisco registered 331 fatal cycling accidents from 2009 to 2024 [2]. According to IIEG data, between 2015 and 2022, 708 cycling accidents were recorded in GMA: 105 cyclists were killed and 612 injured [3]. In this context, this research proposes the development and implementation of the "BiciVR" framework, which integrates a highly realistic virtual reality simulation of the Guadalajara Metropolitan Area (GMA) with machine learning and geospatial data analysis techniques to identify high-risk scenarios in cycling mobility. The goal of the project is to provide insights for decision-making, such as the development of policies to prevent bicycle mobility accidents.

Within the simulation, participants acting as cyclists navigate a virtual map designed using urban features identified from historical data on cycling mobility accidents in the GMA. Likewise, the BiciVR simulation represents traffic density on three levels (easy, normal, and hard) through artificial intelligence tools that generate vehicles capable of making decisions and strategies that optimize computing resources so that it can be executed on mobile virtual reality devices such as the Oculus Meta Quest 2 or higher.

To understand the factors that contribute to accidents and to design effective solutions that

improve road safety, an experiment was carried out involving 36 test subjects, 12 who interacted with the simulation in easy mode, 12 in normal mode and twelve in hard mode.

The data from the BiciVR sessions was gathered from text log files and analyzed using classification algorithms to later interpret feature importance. Among the results, it It was found that road violations, vehicle turning at intersections while cyclists go straight, and traffic density were among the variables that increased the risk of accidents. These results highlight that the implementation of virtual simulations can offer critical data that includes more detailed characteristics of the scenario in which the incident occurs, as well as the relationship and behavior of the parties involved.

2 Motivation

The central problem addressed in this research is the increase in serious and fatal accidents related to cycling mobility in the GMA and the lack of effective tools to analyze their causes in a preventive manner. Based on this, the follow three key factors are identified as part of the problem.

2.1 High Accident Rate in Micromobility

The GMA has 331 localized deaths of cyclist mobility users in road accidents [2]. To contextualize the problem, according to sources [2] and [3], the cyclist fatality rate in traffic accidents in the Guadalajara Metropolitan Area (AMG) was 8% in 2021. This rate is higher than in any US city with more than 500,000 inhabitants [4].

2.2 Limitations on Existing Data

Official records lack critical details such as maneuvers, speeds, actual interactions between vehicles, and interactions between drivers and the environment. The data is collected post-accident, so it does not allow for preventive actions.

2.3 Lack of Controlled Environments for Analysis

Studying risky behaviors on real streets is dangerous and unfeasible. There is a lack of tools that combine virtual reality, urban simulation, and machine learning to predict critical scenarios.

3 Related Work

Despite advances in infrastructure and the promotion of safe cycling, there remains a significant gap in the ability of cities to ensure the safety of users of sustainable modes of transport, especially in the context of interaction with other vehicles. In this context, various studies have explored the use of innovative technologies such as virtual reality [5-8] to address this challenge, providing a promising avenue to train and sensitize users about risk situations in controlled environments. At works [5,6,8] virtual reality applications are used together with a sensorized physical bicycle to collect information about the reaction of the participants according to what they see in the simulation. On the other hand, the study [7] it only uses simulation, which is a different approach that tries to identify drivers' decisions. The BiciVR simulation aims to identify behavioral patterns when riding a bicycle, so no sensors are used. as in [7] BiciVR plans to analyze mainly the maneuver, speeds and road violations, hence the importance of creating a virtual environment emulating a real one of the GMA similar to what they do in [6] for a city in Belgium.

As in [8] BiciVR incorporates data analysis using machine learning, so a well- structured text log system was considered from the beginning of the simulation's development to enable subsequent information analysis. As in [5,6] the BiciVR simulation was developed in Unity3D and used Oculus Meta Quest virtual reality headsets. In [9], the use of immersive virtual reality (VR) is explored to analyze cyclist behavior at urban intersections—critical environments where accidents frequently occur. The study aims to gain a deeper understanding of cyclists' decision-making processes in complex and potentially hazardous traffic situations, demonstrating that immersive VR is not only valuable for training or awareness purposes, but also serves as an empirical research tool for studying human behavior in intricate urban contexts. Furthermore, it provides a foundation for the design of public policies and improvements in urban infrastructure. As highlighted in [10], although the development of

machine learning (ML) systems is fundamentally a branch of software engineering, there are significant disconnects between AI teams and traditional software development processes.

This fragmentation introduces substantial challenges that hinder the adoption of continuous practices such as integration, delivery, monitoring, and continuous improvement.

Considering this, and in response to the evident gap separating AI from software engineering, this work underscores the critical role of software engineering in the construction of ML-based systems, with particular emphasis on the analysis and design phases.

4 Methodology

As part of the methodology for the design and implementation of the BiciVR framework, four phases are identified: Design of the virtual reality simulation; Experimentation; Data collection; and Data analysis and interpretation of results. Figure 1 shows the methodology process.

4.1 Design of the Simulation

Simulation design has two well-defined stages, modeling, and software development. These two phases are of primary importance for this work because the scenario is developed during them. According to [11], the scenario is not only the virtual world but also includes all the interactions within it. Maria-Eleni Paschali and Ioannis Stamelos mention in [11] that the scenario can be the principal predictor of whether a game will be successful or not, making it the most important part of the development process.

4.1.1 Modeling the Simulation with Software Engineering

For the design of the simulation, it was decided to use a model-driven software engineering approach. This work presents the four key software engineering components of the documentation: functional and non-functional requirements; Deployment diagram; General Use Case Diagram; and diagram of the general states of the simulation. The functional requirements are shown in table 1 and consist of the following.

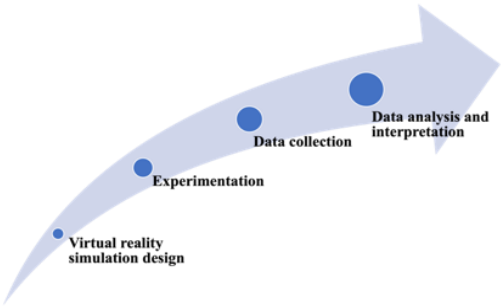

Fig. 1. Methodology process

Table 1. Functional Requirements

ID	Summary Concept
FR001	Interaction between users
FR002	Bike riding
FR003	Smart Traffic
FR004	Difficulty selection
FR005	Automatic traffic light management
FR006	Gamification with Checkpoints
FR007	User Interface with Simulation State Display CSV
FR008	Incident Logging
FR009	Prevention of dizziness due to sudden movements

FR001: The simulation must allow interaction between two types of users, motorized vehicles and the cyclist who will be the player. An incident is defined as interactions of high proximity or collision between a bicycle and a vehicle.

FR002: The simulation must allow the player to ride a bicycle through the virtual world with the possibility of accelerating, braking, turning, and moving forward freely.

FR003: The simulation must have intelligent and autonomous motor vehicle traffic.

FR004: The simulation must allow for selecting between difficulties. Difficulty is expressed as the density of motor vehicles as easy, normal, and hard.

FR005: The simulation must automatically manage the configuration and activation time of traffic lights.

Table 2. Non-Functional Requirements

ID	Summary Concept	Quality Attribute
NFR001	Virtual world based on historical data	Functionality
NFR002	Unity 3D development and VR support	Support
NFR003	Stable execution at 30 FPS	Performance

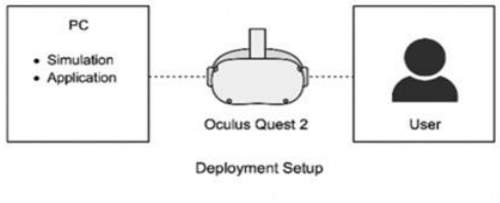

Fig. 2. Deployment diagram of the BiciVR simulation

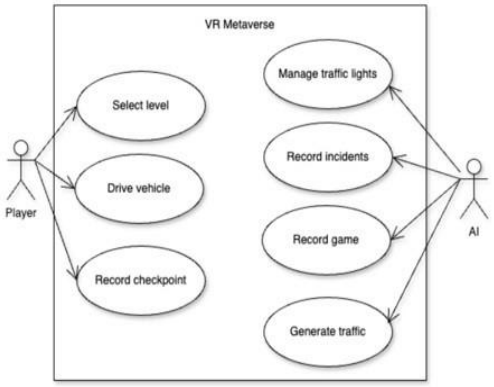

Fig. 3. Use case of the BiciVR simulation

FR006: The simulation must gamify a player's game by providing them with 2 stops as checkpoints and an indication with the direction they must follow to reach each target checkpoint.

FR007: The simulation should provide an interface that indicates the state of the simulation, showing the playtime and the completed checkpoints.

FR008: The simulation must record the incidents that occurred in a game in a CSV file. Among other things, the simulation must record the time of departure where the incident occurred, the distance between the bicycle and the motor vehicle, the speed of the vehicle, the speed of the bicycle, and the coordinates of the bicycle's location on the map.

FR009: If a player wishes to perform maneuvers that include sudden camera movements, the simulation prevents sudden changes of scenery by preventing dizziness in the player.

The non-functional requirements are shown in table 2 and have to do with.

NFR001: The simulation must run on a virtual world that emulates the infrastructure and traffic characteristics obtained from historical data of cycling mobility accidents in the GMA.

NFR002: The simulation must be developed in Unity 3D, in a 2022 editor version or higher, and executable on Oculus Quest 2 devices or later.

NFR003: The simulation must run at least 30 FPS stable.

Figure 2 shows the deployment diagram, where the deployment is done from a PC where the simulation and application are located, which connects through the Oculus Quest 2 to the user. Figure 3 shows the general use case diagram, and figure 4 shows the simulation state diagram.

4.1.2 Simulation Development

As a first step in developing the virtual world, we used a previous study [12] as a reference. This study analyzed the relationship between the infrastructure of the Guadalajara Metropolitan Area (GMA) and fatal cyclist accidents, using historical records of cycling mobility.

Building on the integration of these findings with others reported globally, the world was designed consisting of 4 roundabouts located near the corners of the map, which vary in their design and the number of lanes per entry/exit.

The first roundabout has 2 circulating lanes, with two entries/exits of 2 lanes each, and one 1-lane exit; The second roundabout also has 2 circulating lanes, with two entries/exits of 2 lanes each, and one 1-lane entry; The third roundabout has 3 circulating lanes, with two entries/exits of 2 lanes each, and one entry and one exit of 3 lanes each; The fourth roundabout also has 3 circulating lanes, with two entries/exits of 2 lanes each, and one entry and one exit of 3 lanes each.

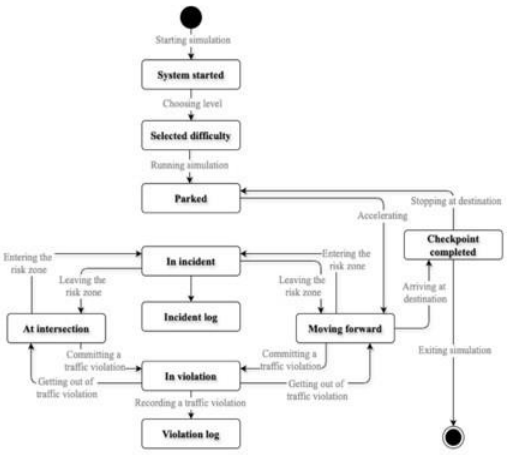

Fig. 4. BiciVR simulation status diagram

Fig. 5. Virtual world seen in panoramic view from the sky

Fig. 6. If the cyclist is in the red dot, the blue square is the area where autonomous vehicles are generated

In addition, the virtual world has 2 main avenues of 6 lanes (3 lines in each direction), intermediate streets of 2 lanes (1 in each direction) and surrounding or connecting streets with 1 lane with only one direction.

The entry and exit angles, as well as the driving directions, were varied to include acute, right, and obtuse angles across the different types of roads: surrounding streets, intermediate streets, and avenues.

The sharpest angle is 45 degrees; the most obtuse, 135 degrees; Right or quasi-right angles are the most common and are found inside the map, between intermediate streets, or connections with avenues. For the creation of this world, the urban designer CiDy 2 was used [13] which allows the creation of streets, roads, sidewalks and include buildings among them. Figure 5 shows the design of the virtual world.

Autonomous traffic is a requirement when implementing a simulation that seeks to create risky situations and generate synthetic incident data. Therefore, an intelligent traffic system was implemented, with the Unity Mobile Traffic System package [14] which uses artificial intelligence techniques so that the vehicles created make their own decisions. As a motor vehicle, a personal sedan- type car was chosen with the following configuration: 1) ”Car” type of vehicle; 2) it has a speed between 15 and 90 km/h; 3) Maximum rotation angle per time of 30 degrees. The vehicle is not limited to turning only 30 degrees in total. Instead, because its direction is checked every frame, it can only change direction by up to 30 degrees at a time; 4) acceleration time of 10 seconds and braking time of 3, and a mass of 1500 kg. The system automatically records the length of the vehicle, the distance between tires, and the size of the collider. In addition, the front and rear lights, stop and turn signals were configured. This ensures that the player understands and foresees the autonomous vehicle’s decisions. The vehicle has a pivot object in front of it, positioned at 1.5 times its length, which allows it to analyze the elements that enter its field of vision and perform appropriate measurements of those elements to decide whether to slow down or search for an alternative route. To optimize computing performance while running the simulation, a player-centric autonomous vehicle generation system was configured, allowing vehicles to appear, and be managed only on streets near the user. For the system’s operation, the terrain, perfectly square in shape, is divided into 54 uniform cells, each measuring 100 meters per side. The traffic algorithm only generates vehicles in the cells adjacent to the user’s current location, including

Fig. 7. Speed of traffic on roads. 30 km/h in blue, 40km/h in cyan, 50 km/h in green, 60 km/h in orange, 70 km/h in brown and 80 km/h in red

Fig. 8. The image shows the 4 intersections that have a traffic light system

Fig. 9. Game capture showing the interface, where the arrow indicates the direction to the next checkpoint, and the user's speed and the number of checkpoints reached are displayed

the cell the user is currently in. Therefore, autonomous vehicles are visible within a 3x3 cell area centered on the user. Figure 6 shows an example of the area considered for traffic generation.

Although the vehicles generated are autonomous and the roads are already created, it is necessary for the traffic system to recognize them as such entities. For this propose, a disconnected graph is created where each edge represents a road or road fragment. Each edge has the following possible configurations: 1) number of lanes, and the width in meters of each one; 2) maximum road speed; 3) For each lane within the road, the direction of vehicular flow, the speed independent of the lane, and the type of vehicles that can travel it are configured.

Once the sections of the graph have been defined, waypoints are generated, which are small fragments within an edge. These allow connections to be generated between lanes in the same direction, allowing an autonomous vehicle to change lanes. For avenues, streets with two or three lanes and of great length, a speed of 80-90 km/h was chosen; for intermediate or connecting streets a speed of 40-60 km/h; and for roundabouts a speed of 30 km/h. Likewise, lane directions were configured, including two-way an one-way streets. The selected lane width was 4 meters. All roads with more than one lane in the same direction allow lane change at any time. Since only the "Car" vehicle type was defined, it was assigned to all lanes of all edges, providing free mobility to autonomous vehicles. Figure 7 shows the representation of the speeds on the map of the virtual world.

Intersections are configured to maintain a logical traffic order. This involves defining the possible maneuvers of a vehicle are configured, and which ones have priority. For the simulation, a standard automatic traffic light system was integrated, based on a special type of intersections, with two variants: 1) at a four-way intersection with a traffic light at each corner, the activation order rotates among the entrances and exits, so only one entry or exit is activated at a time. Then, if the north light is green, the others will remain red until it first turns red, and thus it will rotate its turn to the right (2) at a four-way intersection with a traffic light at each corner, the activation order alternates between opposing entrances and exits, which are positioned at 180-degree angles. So, if the north light is green, the

Fig. 10. Level selection screen

Fig. 11. The yellow dial represents the warning range, and the red dial represents the risk range

south will also be green, and the others will remain red until the first ones turn red. The traffic lights were integrated into intermediate streets of the city, which consist of two-lane roads, one in each direction, as shown in Figure 8.

Finally, the joints of the edges at the in the player's implementation, a standard 26- inch bicycle model was used, which was configured with a maximum speed of (equivalent to 36 km/h), seeking a balance between the standard speed of a bicycle and the perception of speed in the simulation, which tends to be slower in a video game environment. A camera system was implemented that followed the player along the map, from a first-person perspective within the simulation, but allowing the camera to rotate freely, through the Oculus gyroscope, to provide greater immersion and freedom to the player in the virtual reality environment. Likewise, a hybrid control system was developed that allowed interaction both from a computer keyboard, (WASD) and from the controllers of the Oculus Quest 2.

The Oculus' joysticks allowed you to perform the basic movements: accelerating, braking, and turning, thus providing an intuitive and flexible user experience. To create risk scenarios to players through various interactions with the traffic system, a system of goals was implemented that encourages the user to travel through the designed city. In the simulation, the user starts at a fixed neutral point on the map and must move through the city until they complete two checkpoints. Once these points are reached, the game concludes, and the user is redirected to the initial menu.

The objective system was structured as follows: a bidirectional graph was created where all nodes and edges represented the intersections and streets of the designed map with a total of 36 nodes. Additionally, 15 static nodes were located on various sidewalks of the city, called target nodes. At the start of the game, 2 nodes from the target node group are randomly selected as checkpoints to reach, each checkpoint can be identified by an arrow pointing to the ground. The optimal route to the next checkpoint is calculated using the A* algorithm [15]. To make it easier, navigation within the virtual reality environment, a visual arrow was implemented in front of the bicycle, which dynamically indicates the direction towards the next objective, thus improving the orientation and immersion of the player as seen in Figure 9.

Also, as seen in Figure 10, a difficulty selection scene was designed among the options: easy, normal, and hard. The difficulty lies in the number of vehicles generated in the grid close to the user, the minimum generation distance, and their speed. Practically, the level of difficulty was expressed through the volume of traffic the player faced in the game.

For the easy level, 10 vehicles were generated with a minimum distance of 150 meters; for the normal level, 20 vehicles were generated with a minimum distance of 100 meters; and for the hard level, 30 vehicles were generated at a minimum distance of 50 meters.

Finally, a logging system was programmed to automatically records incidents in a CSV file. In

Table 3. Dataset Data Dictionary – Incident Items

Feature	Description	Type/Options
row id	Unique identifier of each record	String
current_date_time_min	Minimum time logged during the incident	DateTime
current_date_time_mean	Average time logged during the incident	DateTime
current_date_time_max	Maximum time logged during the incident	DateTime
time_min	Min. time from departure to incident	Time
time_mean	Avg. time from departure to incident	Time
time_max	Max. time from departure to incident	Time
user id	Unique ID of the subject user	String
lanes	Number of lanes on road	Number
is intersection	Incident at intersection?	Binary
is two-way street	Street is two-way?	Binary
bike maneuver	Cyclist's maneuver	Enum: turn left, reverse, straight
vehicle maneuver	Vehicle's maneuver	Enum: stop, turn right, turn left, reverse, straight
maneuver direction	Relative direction of maneuvers	Enum: same direction, opposite direction, from right, from left
sidewalk climbs	Cyclist on sidewalk?	Binary
running red light	Runs red light?	Binary
drive opposite direction	Drives in opposite direction?	Binary
bad roundabout	Incorrect roundabout maneuver?	Binary
driving between lanes	Drives between lanes?	Binary
crossings without priority	Crosses without priority?	Binary
is bike infringement	Any violation by cyclist?	Binary
real risk	Rated actual risk level	Enum: warning, risk, collision
bike side on collision	Bike part involved in collision	Enum: front, back, left, right
vehicle side on collision	Vehicle part involved in collision	Enum: front, back, left, right
fault of	Alleged culprit in collision	Enum: bicycle, car
vehicle id	Unique vehicle identifier	String
vehicle name	Vehicle model name	String
vehicle position x mean	Avg. vehicle X position	Number
vehicle position y mean	Avg. vehicle Y position	Number
vehicle position z mean	Avg. vehicle Z position	Number
bike position x mean	Avg. bike X position	Number
bike position y mean	Avg. bike Y position	Number
bike position z mean	Avg. bike Z position	Number
vehicle speed mean	Avg. vehicle speed	Number
vehicle speed min	Min. vehicle speed	Number
vehicle speed max	Max. vehicle speed	Number
distance mean	Avg. vehicle- bike distance	Number
distance min	Min. vehicle- bike distance	Number
distance max	Max. vehicle- bike distance	Number
bike speed mean	Avg. bike speed	Number
bike speed min	Min. bike speed	Number
bike speed max	Max. bike speed	Number
level	Game difficulty level	Enum: easy, normal, difficult

that sense, a warning radius of 7.5 meters was established around the bicycle that captures all interactions with any vehicle.

Each log entry automatically includes the following columns: date and time, user ID, distance to vehicle, vehicle ID, vehicle spatial coordinates,

Table 4. Dataset Data Dictionary – User data

Feature	Description	Type/Options
Timestamp	Date and time of the experiment	DateTime
User	User ID in the simulation. Uniquely identifies the game	String
Level played	Difficulty level of the simulation	Easy, Normal, Hard
Age	Age of the test subject	Number
Sex	Sex of the test subject	Hombre, Mujer
Do you ride your bike regularly?	How regularly the test subject rides a bicycle, on a weekly basis	Never, 1-3 times, 3-5 times, 5-10 All the time
Usual mode of transportation	Mode of transport with which the test user usually travels in the city	Macro, Train, Microbus, Private vehicle, Motorcycle, Bicycle, Scooter
How immersive was the experience?	The degree to which the test subject felt fully immersed in the virtual environment during the simulation	Value between 1 and 5
Perception of security	Subjective sense of safety experienced by the test subject during the simulation, with respect to the environment and traffic	Value between 1 and 5
Usability	Ease with which the test subject was able to interact and use the simulation effectively	Value between 1 and 5
Road violation: riding on the sidewalk	Number of times the subject rode onto the sidewalk, per street	Number
Road violation: Running a red light	Number of times the subject crossed at a red light	Number
Road violation: Driving in the wrong direction on the street	Number of times the subject drove in the opposite direction, on the street	Number
Road violation: Misuse of roundabout	Number of times subject entered/exited a roundabout badly	Number
Road Violation: Driving between lanes	Number of times the subject drove in the middle of the lanes	Number
Road violation: Non-priority crossing	Number of times the subject crossed an intersection without having priority	Number
Total road violations	Sum of all the infractions of the subject, regardless of the type	Number
Incident: Total Warnings	Total warning type incidents accounted for by the automatic log	Number
Incident: Total Risk	Total risk incidents accounted for by the automatic log	Number
Incident: Total Collision	Total collision-type incidents by the automatic log	Number
Total incidents	Sum of all the incidents in which the subject was involved, regardless of the type	Number
Playtime	Duration of the subject's departure	Time

vehicle speed, bicycle spatial coordinates, and bicycle speed.

Since incidents were recorded in each frame, the 10- second records are grouped into a single frame, adding the minimum, maximum, and average values for each numerical variable within the time window. Subsequently, with the minimum distance recorded in each interaction, a categorical variable of real risk was calculated.

This variable classifies events according to proximity: collision (distance equal to 0 meters), risk (distance between 0 and 2 meters) and warning (distance greater than 2 meters) as shown in Figure 11.

4.2 Experimentation

The experiment involved 36 test subjects, selected from youth and adults in early stages of life.

Before starting, each participant received a brief introduction to the device's basic controls, which included moving forward, braking, and turning their head to change the visual perspective.

Users were also instructed on the steps required to record their gameplay using Oculus' native camera app. Participants had to select the level of difficulty indicated by the experiment supervisor.

Once inside the simulation, they were instructed that their goal was to complete two checkpoints by riding the bike as they would in real life. It should be noted that each participant completed the simulation only once, which made it possible to guarantee uniformity in data collection and prevented possible biases derived from familiarization with the virtual environment.

A total of 36 experimental sessions were conducted, evenly distributed across the three difficulty levels, with 12 sessions per level. At the end of the experiment, each subject completed a survey designed to gather information about their demographic profile and evaluate their experience during the simulation.

4.3 Data Collection

Once the automatic data was extracted from the game logs, an exhaustive visual review of each of the game video sessions was carried out. This visual analysis aimed to gain a deeper understanding of the risks faced by players during matches, as well as to provide details that could not be captured automatically. For each log entry, detailed information about the road infrastructure and the maneuvers performed by both the bicycle and the vehicles involved were integrated. Regarding infrastructure, the following variables were recorded: the number of lanes (between 1 and 4), the presence of intersections, and whether the road was two-way. These characteristics helped provide better context of the environment in which the incidents occurred.

Regarding the maneuvers, the behavior of both the bicycle and the vehicle at the time of the incident was detailed. The bicycle's maneuvers were classified into five categories: forward, reverse, right turn, left turn, and stop. Similarly, the vehicle's maneuvers were categorized as forward, reverse, right turn, left turn, and stop. In addition, the relative direction of the maneuver between the bicycle and the vehicle was analyzed, specifying whether they were moving in the same direction, in opposite directions, or if the vehicle was approaching from the left or the right, always taking the bicycle as a reference point. This directional analysis provided a clear perspective on the dynamics of the incident and the possible underlying cause.

For the categorization of infractions committed by players, six main types were selected: driving on sidewalks, crossing red lights, driving in the wrong direction, improper entry/exit from a roundabout, driving between lanes, and crossing intersections without having priority. These categories were chosen because of their prevalence among players and their relevance as common causes of accidents in real life. For each log record, Boolean columns were added indicating the presence or absence of one or more of these road violations at the time of the incident. In addition, a Boolean variable was calculated to indicate whether any road infraction had been committed, allowing for a more global analysis of the player's behavior. In cases where a collision was recorded, further analysis was performed to better understand the circumstances surrounding the accident. Specific features such as the side of the bicycle impacted, the side of the vehicle involved in the collision, and a preliminary assessment of whether the vehicle or bicycle was at fault were added. This detailed analysis of collisions enables the identification of patterns and potential intervention points to enhance safety in future simulated scenarios. Finally, two datasets were obtained: a general dataset of all incidents and a dataset per player. Table 3 shows the data dictionary for all incidents and table 4 shows the data dictionary per player.

4.4 Data Analysis and Interpretation of Results

This stage consists of 3 steps, 1) an exploratory analysis using descriptive statistics. This analysis was based on the use of data groupings and frequency tables, using the Python Pandas library; 2) a spatial analysis was conducted to examine how users navigated the map and interacted with vehicles during the simulation. To support this, a visual representation system of the incidents was implemented using heatmaps on the game's road map; 3) A classification analysis using machine learning was conducted to identify the most relevant variables associated with incidents. The

Table 5. AUC and Recall scores by model for collision class

Model	AUC	Recall
Neutral Network	85.1	83.3
Random Forest	86.3	66.7
Logistic Regression	75.7	66.7
Naive Bayes	82.1	50.0
Gradient Boosting	87.3	50.0

Fig. 12. Position of incidents (variable real risk) by type: warnings in green, risks in yellow, and collisions in red

Fig. 13. Position of the incidents Risk and collision: Risk are shown in yellow, and incidents in red.

analysis used the real risk variable as the target, as the goal was to determine which characteristics are most likely to lead to collisions or risks. The analysis was a binary classification that considered risk incidents and collision incidents. For the classification, it was necessary to balance the incident data set using SMOTE since the number of risk incidents was much higher than collision incidents.

5 Results and Discussions

From the descriptive statistical analysis, it was found that dataset 1 of incidents contains 1904 records with the following distribution: 70.64% of the incidents corresponded to warnings (distances greater than 2 meters), 26.42% to risk (distances less than 2 meters) and 2.94% to collisions, accounting for a total of 56 collisions.

The difficulty of the simulations showed a significant impact: 52.57% of the incidents occurred in hard games, 33.56% in normal games, and 13.86% in easy games. The road violation analysis revealed that 52.36% of the incidents involved at least one violation, accumulating a total of 997 violations. Among these, the most frequent were driving in the wrong direction (30.56%) and crossing intersections without priority (21.42%). The 4.31% of the road violations ended in collisions, 24.57% in dangers and 71.11% in warnings. The most dangerous combination of road violations, accounting for 12% of collisions, involved simultaneously crossing a red light, driving the wrong way, and crossing intersections without priority. The results, consistent with those in the study [16], show that a higher number of road violations committed by cyclists is associated with an increased occurrence of accidents.

The average speed of vehicles was 23.22 km/h while that of bicycles was 16.35 km/h in risk- type incidents. In the case of collisions, the average speeds were slightly lower for both vehicles (14.10 km/h) and bicycles (10.78 km/h). The collisions showed specific patterns depending on the point of impact and the culprit. For bicycles, the left side was the most affected (40%), while for vehicles it was the right side (43.63%). In 30.90% of the collisions, the left side of the bicycle impacted the right side of the vehicle, being mostly the fault of the bicycles (82.35%). Overall, 83.63% of collisions were attributed to bicycles, compared to 16.36% for vehicles.

Speaking about infrastructure, 85.71% of collisions occurred on two-way streets and 71.42% of collisions took place at intersections,

which is consistent with studies [17,18] which argue that intersections are the place with the highest probability of accidents for cyclists.

The level of the simulation (traffic density) is related to the number of incidents, a finding that has been found in previous studies where they mention that the higher the traffic density, the greater the probability of encounters between cyclists and vehicles [19].

Of the incidents classified as warnings, the hard level concentrates 52.94% of the cases, followed by the normal level with 34.65%, while the easy level has a significantly lower contribution of 12.42%. In risk type incidents, the hard level predominates with 51.89%, the normal level has 30.02%, and the easy level has 18.09%. Collisions are more evenly distributed between the hard level (50%) and the normal level (39.29%), while the easy level concentrates only 10.71%.

In Dataset 2, where data is organized per player, there are 36 participants with an average age of 22 years, the youngest player is 18 years old and the old is being 32. The distribution by gender shows a predominance of men (68.5%) compared to women (31.43%). Regarding previous experience with bicycles, most (62. 86%) indicated that they rarely or never used this mode of transportation, while only 11.42% reported using it regularly. In addition, the immersion experience received a score of 4.1 on a scale of 1 to 5. This means that the virtual world of the simulation achieved the goal of being realistic and engaging for the players.

Spatial analysis using heatmaps reveals that the simulation map has been uniformly traversed across all sections, with incidents recorded in most types of planned infrastructure. As can be seen in figures 12 and 13, collisions are concentrated in certain specific areas, where internal two-way streets are the most dangerous infrastructure. These zones and intersections have the highest frequency of collisions, which could be attributed to the inherent complexity of this type of infrastructure, where bidirectional flow increases the chances of conflicts between users. In the machine learning analysis of the incident dataset, the time and location variables were dropped. Each record in the incident dataset represents 10 seconds of interaction between cyclists and vehicles. Therefore, we decided to use the maximum speed as the most representative measure for both bike and vehicle speeds. Consequently, we dropped the variables bike speed min, bike speed mean, vehicle speed min, and vehicle speed -mean from the analysis. We also dropped the variables distance max, distance mean, and distance min because the target variable real risk was calculated using distance min, introducing a clear dependency.

Thus, the dataset was composed of a total of 17 features, 16 independent variables and 1 target variable.

The goal of the machine learning analysis is to explore feature importance in classification, where the real risk variable was taken as the target variable. The real risk variable originally contains three classes: warning, risk, and collision. However, for the classification task, we selected only the instances labeled as risk and collision to perform a binary classification. The goal is to distinguish between actual collisions and near-collision incidents.

Before applying SMOTE, the data were divided into a training set (80%) and a test set (20%) using stratified random sampling, ensuring that the proportion of classes was maintained in both divisions. This strategy allows oversampling to be applied only to the training set, preserving the validity of the test set to evaluate the model.

The models used for the classification analysis were: Random Forest, Gradient Boosting, XGBoost, Neural Networks, Naive Bayes, and Logistic Regression. Since the dataset was imbalanced, with 56 records labeled as collision and 503 as risk, we used AUC and recall for the collision class as evaluation metrics. These metrics focus on the model's ability to distinguish between classes and to correctly identify collision instances.

The best performance was presented by neural networks with an UAC of 85.1% and a Recall of 83.3% for the minority class (see table 5).

To train and evaluate the models, we used the Orange data mining software. We also used its Feature Importance and Explain Model functionalities to gain insights into the impact of the features on the classification. By analyzing the interpretation of the Feature Importance and the Explain Model on the neural network, we obtained the following insights:

– The ten most important variables in descending order for the UAC model is bike - infringement, drive opposite direction, bike maneuver, maneuver direction, vehicle - maneuver, is intersection, lanes, bike speed max, driver between lanes, and level.

– The classes that most often lead to a classification as a collision are: is bike - infringement in the positive class, drive opposite direction in the negative class, is intersection in the positive class, bike maneuver as straight, maneuver direction as from right, and level as hard.

As in the study [20], vehicle turns at intersections are identified as the most dangerous situations. It is important to mention that this maneuver was not the most frequent in the experiment; the most common maneuver for vehicles was crossing the intersection without turning.

Global interpretation Intersections and two- way streets represent the critical points with the highest concentration of collisions. Combining more than one violation at a time increases the risk of having an accident. For example, crossing a red light, driving in the wrong direction, and entering intersections without priority simultaneously accounted for 12% of collisions in the experiment. On the other hand, the most dangerous maneuver combination involves vehicles turning right while cyclists go straight at an intersection, both traveling in the same direction.

6 Conclusions

The main contribution of this work is an open-source tool called BiciVR, which provides enriched data to support the development of public policies aimed at promoting cycling mobility, such as the redesign of intersections.

Also, the study showed why high-difficulty scenarios, intersections and two-way streets are the riskiest for cyclist. Collisions are strongly associated with multiple violations and risky behaviors. The BiciVR framework turned out to be a useful tool for simulating and analyzing critical traffic situations in a safe and controlled way. The objective of the work was achieved, as it was0 possible to identify maneuvers and driver behaviors that lead to traffic accidents.

References

1. **Global Market Insights Inc. (2025).** Micro-mobility market size & share, growth trends 2023–2032. https://www.gminsights.com/industry - analysis/micro-mobility-market.
2. **White Bicycle Fatal database (2025).** White bicycle. https://bicicletablanca.org/.
3. **IIEG (2025).** Siniestralimap. https://iieg.gob.mx/siniestralimap/index.html.
4. **IIEG (2021).** 2021 Data- Bicyclists and Other Cyclists.
5. **Tsuboi, H., Toyama, S., Nakajima, T. (2018).** Enhancing bicycle safety through immersive experiences using virtual reality technologies. Augmented Cognition: Smart Technologies, **Schmorrow, D.D., Fidopiastis, C.M. Eds.** Cham: Springer International Publishing, pp. 444–456.
6. **Zeuwts, L.H.R.H., Vanhuele, R., Vansteenkiste, P., Deconinck, F.J.A., Lenoir, M. (2023).** Using an immersive virtual reality bicycle simulator to assess hazard detection and anticipation of overt and covert traffic situations in young cyclists. Virtual Reality, Vol. 27, No. 2, pp. 1507–1527.
7. **Sasaki, Y., Fujiwara, K., Mitobe, K. (2024).** Bicycle accident induced risks: Measuring and analyzing cyclists' behavior when going straight and turning right using a bicycle simulator. Accident Analysis & Prevention, Vol. 194, pp. 107338.
8. **Losada, F.J., Páez, F., Luque, F., Piovano, L., Sánchez, N. et al. (2024).** Models for predicting collisions between vehicles and cyclists through the application of machine learning techniques to data from virtual reality bicycle simulators. Applied Sciences, Vol. 14, No. 9.
9. **Chuang, T.Y., Chang, Y.S. (2022).** Using immersive virtual reality to investigate cyclist behavior in urban intersections. Accident Analysis & Prevention, Vol. 170, pp. 106659.
10. **Vänskä, S., Kemell, K.K., Mikkonen, T., Abrahamsson, P. (2024).** Continuous software engineering practicesin AI/ML development

past the narrow lens of MLOps: Adoption challenges. e-Informatica Software Engineering Journal, Vol. 18, pp. 240102. https://www.e-informatyka.pl/index.php/einform atica/volume s/volume-2024/issue-1/article-2/.

11. **M.E. Paschali and I. Stamelos. (2023).** Computer Game Scenario Representation: A Systematic Mapping Study. e-Informatica Software Engineering Journal, Vol. 17. https://www.e-informatyka.pl/index.php/einfor matica/volumes/volume-2023/issue-1/article-3/
12. **Briseño, R.A., Arellano, R.M., Franco, E.G.C., Larios, V.M., Beltrán, R.J. et al. (2022).** Fatal cyclist-car accidents scenarios at intersections from the Guadalajara Metropolitan Area. International Journal of Combinatorial Optimization Problems and Informatics, Vol. 13, No. 4, pp. 10–25.
13. **Unity Asset Store (2023).** CiDy 2 | Level Design. https://assetstore.unity.com/packages/ tools/le vel-design/cidy-2-55298.
14. **Unity Asset Store (2023).** Mobile Traffic System v3 | Behavior AI. https://assetstore. unity.com/packages/tools/behavior-ai/mobile-traffic-system-v3-305800.
15. **Hart, P.E., Nilsson, N.J., Raphael, B. (1968).** A formal basis for the heuristic determination of least-cost trajectories. IEEE Transactions on Systems Science and Cybernetics, Vol. 4, No. 2, pp. 100–107.
16. **O'Hern, S., Estgfaeller, N., Stephens, A.N., Useche, S.A. (2021).** Bicycle rider behavior and crash involvement in Australia. International Journal of Environmental Research and Public Health, vol. 18, no. 5.
17. **Aldred, R., Kapousizis, G., Goodman, A. (2021).** Association of infrastructure and route environment factors with bicycle injury risk at intersection and non- intersection locations: A crossover study of cases from Great Britain. International Journal of Environmental Research and Public Health, Vol. 18, No. 6.
18. **Lin, Z., Fan, W.D. (2021).** Analysis of the severity of cyclist injuries with mixed logit models at intersections and non-intersectional locations. Journal of Transportation Safety & Security, Vol. 13, No. 2, pp. 223–245.
19. **Johnsson, C., Laureshyn, A., D'Agostino, C., Ceunynck, T.D. (2021).** The effect of 'safety in density' for cyclists and motor vehicles in Scandinavia: An observational study. IATSS Research, Vol. 45, No. 2, pp. 169–175.
20. **Schröter, B., Hantschel, S., Huber, S., Gerike, R. (2023).** Determinants of bicycle accidents at signalled urban intersections. Journal of Safety Research, Vol. 87, pp. 132–142.

Article received on03/08/2025; accepted on 11/10/2025.
**Corresponding author is Edgar Cossio.*