Learning-based Dynamic Risk Indicators (LADRI) for Risk Prediction in Automated Driving Systems

Vom Fachbereich Informatik der
Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau
zur Verleihung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation
von

M.Sc. Anil Ranjitbhai Patel

Datum der wissenschaftlichen Aussprache: 06.12.2024

Dekan:	Prof. Dr. Christoph Garth
Berichterstatter:	Prof. Dr.-Ing. habil. Peter Liggesmeyer
Berichterstatter:	Prof. Dr. Karsten Berns

DE - 386

“Should dynamic risk contexts shape our risk assessment models, or should risk assessment models redefine our understanding of dynamic risks?”

Abstract

As Automated Driving Systems (ADSs) revolutionize the intelligent transportation landscape, ensuring unparalleled safety becomes increasingly essential. Moreover, as automation levels rise, the responsibility for safety shifts from the driver to the system developers. This transition necessitates a reevaluation of safety assumptions and the creation of a risk assessment framework to assess system failures within an ADS.

The challenges of continuous risk assessment and early hazard identification for ADS is multifaceted. Firstly, it is crucial to achieve comprehensive scenario coverage, as current standards often focus on a limited set of predefined conditions and may not encompass the wide range of potential real-world scenarios an ADS might encounter. Secondly, integrating complex system interactions is challenging because traditional approaches may not fully account for the detailed interactions among various ADS subsystems or effectively analyze the associated risks. Thirdly, integrating and analyzing diverse data streams is essential but challenging, and current frameworks may not adequately address this aspect. Lastly, validation under various operational conditions and continuous model improvement is vital, with standard validation processes potentially not accounting adequately for the wide variability in operational conditions faced by ADS and failing to specify measures for open traffic scenarios.

The LeArning-based Dynamic Risk Indicators (LADRI) framework addresses these challenges, aiding safety engineers in early hazard detection for ADS development. It employs a Plan-Do-Train-Adjust-Assess cyclic process, enabling continuous improvement in risk assessment across diverse dynamic driving conditions. This approach leverages advanced learning algorithms and incorporates risk-specific context information from both the operational environment and the ADS itself, providing runtime contextual insights for the model to accurately predict severity and controllability indicators. This strategy bridges traditional gaps in risk assessment by allowing the evaluation of severity and controllability to adapt to different dynamic environments, instead of relying on subjective judgment. By monitoring the risk spectrum during driving operations, safety engineers can identify specific risk profiles, thereby enhancing safety mechanisms for subsequent iterations of risk assessment. Iteratively applying this framework allows developers to transition from unknown-unsafe regions to known-safe regions, thus progressively enhancing safety.

A use case of Highway-Lane Following with Adaptive Cruise Control (ACC) functionality is utilized, creating extensive testing across a variety of driving scenarios, on different road shapes under diverse conditions, which has demonstrated the framework's robust capability for accurate risk prediction. The analysis underscores the importance of integrating time-based, distance-based, and impact-based risk features for a comprehensive risk assessment. These findings, supported by comprehensive model performance metrics and evaluations, position the LADRI framework as an advanced tool to enhance the risk assessment process.

Zusammenfassung

Mit der Revolutionierung des intelligenten Verkehrswesens durch Automatisierte Fahrsysteme (AFS) wird die Gewährleistung beispielloser Sicherheit zunehmend unerlässlich. Mit steigendem Automatisierungsgrad verlagert sich die Sicherheitsverantwortung von den Fahrern auf die Systementwickler. Dies erfordert eine Neubewertung der Sicherheitsannahmen und die Entwicklung eines Rahmens zur Risikobewertung von Systemausfällen innerhalb eines AFS.

Die kontinuierliche Risikobewertung und frühzeitige Gefahrenerkennung für AFS bringt vielschichtige Herausforderungen mit sich. Erstens fehlt eine umfassende Abdeckung potenzieller Szenarien. Aktuelle Standards berücksichtigen oft nur begrenzte, vordefinierte Bedingungen, was reale Herausforderungen unzureichend abdeckt. Zweitens sind komplexe Interaktionen zwischen AFS-Teilsystemen schwer zu integrieren, da herkömmliche Ansätze nicht die ganzheitliche Analyse dieser Interaktionen ermöglichen. Drittens bleibt die Nutzung und Analyse diverser Datenströme unzureichend, obwohl diese für eine präzise Risikobewertung essenziell ist. Viertens mangelt es an standardisierten Validierungsverfahren, die die Variabilität der AFS-Betriebsbedingungen oder offene Verkehrsszenarien berücksichtigen.

Das LeArning-basierte Dynamische Risikoindikatoren (LADRI)-Rahmenkonzept adressiert diese Herausforderungen und unterstützt die frühzeitige Gefahrenerkennung in AFS. Mithilfe eines Plan-Do-Train-Adjust-Assess-Zyklus wird eine kontinuierliche Verbesserung der Risikobewertung unter verschiedenen dynamischen Fahrbedingungen ermöglicht. Dieser Ansatz nutzt fortschrittliche Lernalgorithmen und integriert risikospezifische Kontextinformationen aus der Umgebung sowie dem Fahrzeug selbst, um Laufzeitkontextdaten zu liefern. So können Schweregrad- und Kontrollindikatoren präzise vorhergesagt werden. Im Gegensatz zu herkömmlichen Methoden überbrückt LADRI bestehende Lücken, indem es die Anpassung der Risikobewertung an dynamische Umgebungen statt subjektiver Urteile ermöglicht. Die Risikoüberwachung während des Fahrzeugbetriebs erlaubt die Identifikation spezifischer Risikoprofile, was die Weiterentwicklung von Sicherheitsmechanismen in iterativen Prozessen unterstützt. Dadurch können unsichere Szenarien sukzessive in sichere überführt werden.

Ein Anwendungsfall von Highway-Lane Following mit adaptiver Tempomat-Funktionalität wurde genutzt, um die Fähigkeiten des LADRI-Konzepts zu demonstrieren. Tests unter vielfältigen Fahrszenarien und Bedingungen verdeutlichen die robuste Fähigkeit zur genauen Risikovorhersage. Die Analyse unterstreicht die Relevanz der Integration zeit-, distanz- und wirkungsbasierter Merkmale für eine umfassende Risikobewertung. Diese Ergebnisse, gestützt auf umfassende Metriken und Evaluationen, etablieren das LADRI-Rahmenkonzept als fortschrittliches Werkzeug zur Verbesserung des Risikobewertungsprozesses.

Acknowledgments

I am immensely grateful to Prof. Peter Liggesmeyer for his expertise and guidance throughout my PhD journey. His unwavering support and insightful feedback have not only shaped this academic endeavor but also inspired my personal and professional growth.

I would also like to express my deepest gratitude to Prof. Karsten Berns for his thorough assessment of my research work.

I want to take this moment to acknowledge the invaluable support of my colleagues, who have generously offered their assistance without any expectations in return. A very special thanks to Dr. Felix Möhrle, Nikita Haupt, Nishanth Laxman, Alexander Günther, Mengisti Berihu, Caroline Frey, Thomas Schneider, and Claudia Schwager for their unwavering support throughout this journey.

Additionally, I would like to thanks to my colleagues from the other side of the crossroad, Dr. Daniel Schneider, Dr. Rasmus Adler, and Jan Reich. Their engaging and insightful discussions have not only shaped some of my research ideas but have also broadened my perspectives and deepened my understanding of safety research.

A PhD journey is indeed a transformative experience, yet, all the skills and hypotheses we develop during this time would fall short without the foundational support of our families. The environment our parents create allows us to explore and realize our potential. I am at a loss for words to fully express my deep love and gratitude towards each member of my family, my parents, my brother Umesh, and my sister-in-law Heny, for their unwavering support and belief in me.

Throughout this challenging and rewarding PhD journey, there has been one person whose silent support has been my cornerstone, the love of my life, my wife, Sweta. She has made countless sacrifices to ensure that I could give my best to my research. Her unconditional love and care have been instrumental in navigating the ups and downs of this demanding process. She generously proofread my thesis, providing valuable insights from a non-background perspective. Additionally, I cannot overlook the joy brought into my life by our son, Avinash. His arrival into this world and his remarkable resilience over these months have inspired me greatly. His tiny hands, too, have provided support, adding an extraordinary dimension to our family's journey together.

Table of Contents

List of Figures

List of Tables

Acronyms

ACC	Adaptive Cruise Control
ADS	Automated Driving System
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASIL	Automotive Safety Integrity Level
AV	Autonomous Vehicle
CV	Conventional Vehicle
DDT	Dynamic Driving Task
DRA	Dynamic Risk Assessment
DRAC	Deceleration Rate to Avoid Crash
ETA	Event Tree Analysis
FMEA	Failure Modes and Effects Analysis
FTA	Fault Tree Analysis
GBDT	Gradient Boosting Decision Tree
GUI	Graphical User Interface
HARA	Hazard Analysis and Risk Assessment
HAZOP	Hazard and Operability Analysis
KE	Kinetic Energy
LADRI	LeArning-based Dynamic Risk Indicators
MDAC	Minimum Distance to Avoid Crash
ML	Machine Learning
ODD	Operational Design Domain

PDTAA	Plan-Do-Train-Adjust-Assess
RF	Random Forest
SAE	Society of Automotive Engineers
SFD	Safety Distance
STD	Stopping Distance
SVM	Support Vector Machine
TTC	Time to Collision
TTE	Time to Escape
TTS	Time to Stop

1 Introduction

As the transportation sector evolves with the advent of new technologies, the safety-critical behaviors of autonomous modes of transport become a focal point for research and development. These autonomous systems, whether they operate on the rails, through the skies, or along our urban roadways, are not only expected to revolutionize how we travel but also to ensure the highest standards of safety and reliability. Consequently, the following table provides a comparative overview of these behaviors across three major autonomous transport modes: trains, airplanes, and cars. It outlines the distinct challenges and operational parameters that each mode must navigate to maintain safety and reliability in their respective domains.

Table 1.1: Comparison of Safety-Critical Behaviors of Autonomous Transport Modes [45]

Mode	their behaviors.
	Expected Operation
Trains	Adherence to schedules and signal compliance.
Airplanes	Flight plan and emergency response readiness, in-flight monitoring.
Cars	Road rule adherence and responsive to dynamic conditions.
	Freedom of Movement
Trains	Restricted to rails.
Airplanes	Operate in three-dimensional space with set flight paths.
Cars	Can move in any direction on roadways.
	Avoidable Traffic Encounters
Trains	Other trains, predictable paths.
Airplanes	Other aircraft, monitored by air traffic controller, relatively predictable.
Cars	Diverse traffic including pedestrians and cyclists, unpredictable.
	Environmental Monitoring
Trains	Signals, weather affecting tracks.
Airplanes	Weather patterns, air traffic over long distances and high altitudes.
Cars	Road conditions, traffic signals, and weather.
	Collision Risk
Trains	Collisions with other trains, derailments.
Airplanes	Risk of mid-air collisions, highest control during takeoff/landing.
Cars	Collisions with all road users, high variability.

In light of this context, Table 1.1 sheds light on the fact that the complexities and challenges for autonomous cars are more pronounced due to their operation in a highly dynamic and unpredictable environment. While trains and airplanes follow predetermined paths with strict scheduling and air traffic control oversight, respectively, autonomous cars must navigate a myriad of unpredictable factors such as diverse traffic conditions, variable road rules, and unanticipated actions from pedestrians and other road users. The operational demands on autonomous cars necessitate a level of adaptive behavior and runtime decision-making that is far more complex than that required for the other two modes of transport, thus presenting greater challenges for safety and standardization in the domain of automotive.

Autonomous cars are composed of a complex network of sensors, actuators, and processing units. This allows vehicles to independently perceive, interpret, and navigate their surroundings. When delving into the domain of autonomous cars, confusion often arises regarding the terminology used, such as autonomous cars, automated driving, self-driving vehicles, and driverless vehicles. Although these terms are frequently used interchangeably, they possess almost same meanings. "Autonomous" and "self-driving" vehicles are closely related, as are "automated" and "driverless" vehicles [47, 54, 99, 139, 143]. A truly autonomous vehicle would independently determine its destination and route without any human input.

The Society of Automotive Engineers (SAE) has established a classification system for such system, which ranges from Level 0 (no automation) to Level 5 (full automation) [60]. At Level 0, the human driver maintains complete control over the vehicle at all times. Conversely, at Level 5, the vehicle operates entirely without human intervention. For the sake of consistency and clarity, this thesis follows the term Automated Driving System (ADS) specifically referring to Level 5 automation, where no human intervention is considered necessary. Nevertheless, it is recognized that some research has used "Autonomous Vehicles (AVs)" as case studies. In these scenarios, this thesis aligns the terminology by treating AV as synonymous with ADS. This way thesis work ensures uniformity in the examination and discussion of automated driving technologies.

ADS emerge as a promising solution for significantly reducing the vast number of road accidents, and enhancing overall driving experiences. They offer clear advantages, such as eliminating the need for manual vehicle operation, reducing delays from congestion, and creating safer environments by minimizing human error. The push for developing ADS is primarily driven by the fact that human error or misconduct causes the majority of road accidents, with more than 90% attributed to such factors including distracted driving, aggressive maneuvers, driving under the influence, speeding, and lack of attention due to fatigue or drowsiness [123, 125]. This underscores a critical need for ADS to mitigate driving misconduct and enhance road safety.

However, as ADS operate in dynamic, safety-critical environments, it is imperative that they maintain safety for all traffic participants, including their users, in the event of any hazard or system failure. By comprehensively understanding their own functional capabilities and the environment, ADS can not only reduce the severity but also diminish the likelihood of accidents. Furthermore, it is crucial that they do not introduce any additional hazards that could potentially increase the number of road accidents.

Therefore, in this thesis, the research is centered on evaluating ADS across diverse scenarios to determine whether their behavior is safe or unsafe. The structure of the remainder of this chapter is outlined as follows:

In Section 1.1, the thesis focuses on motivation by highlighting the post-deployment safety challenges faced by ADS and discusses how much risk is considered safe enough. This section also explores the gap in current risk assessment techniques.

In Section 1.2, the thesis explains the problems that ADS face with current risk assessment techniques, where the thesis work can provide solutions, and why dynamic risk assessment and a learning-based approach is necessary.

In Section 1.3, the thesis outlines its contributions towards addressing the issues highlighted in Sections 1.1 and 1.2.

In Section 1.4, the thesis discusses the key assumption made during the development of its contributions.

Finally, an overview of the structure of the thesis is provided in Section 1.5

Chapter Content

1.1 Motivation

Ensuring the safety of ADS demands a comprehensive and interdisciplinary approach that spans various areas, including safety engineering, computing hardware, robotics, human-machine interaction, software, security, and testing [77]. A key challenge is validating ADS against unexpected inputs and achieving the reliability necessary for widespread deployment, especially with the driver out of the loop [76]. Developing an integrated design and deployment process that consolidates safety considerations across multiple technical domains into a cohesive strategy presents a significant hurdle [15].

As depicted in Fig. 1.1, with automation levels increasing to Level 5, the responsibility for safety shifts from the driver to system developers. This change compels developers to incorporate safety measures throughout ADS to effectively address the complex challenges it may face during its operation. System developers bear the responsibility of adhering to and potentially exceeding existing safety engineering standards, which may not fully address the unique challenges presented by ADS. This entails a comprehensive reevaluation of safety assumptions and the creation of dependable safety mechanisms to manage rare but inevitable system failures across an entire fleet, accommodating the absence of a human driver in the operational loop.

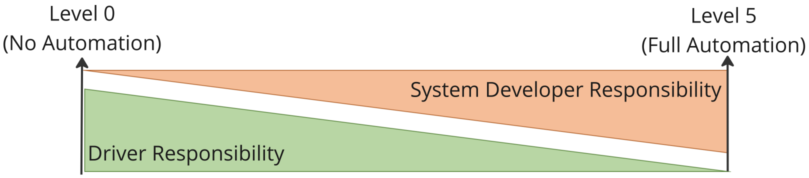


Figure 1.1: Safety Responsibility: Level of Automation [60]

The quantification of risk in ADS presents multiple challenges, particularly in validating the complex and interconnected systems essential to ADS functionality. These systems often rely on inductive reasoning, complicating the assurance of their behavior in unforeseen conditions. Additionally, the complexity and variability of real-world data present significant obstacles to achieving system safety.

To address these challenges, a comprehensive validation strategy that goes beyond traditional methods is necessary. Systems like ADS, capable of learning and adapting over time, require dynamic testing approaches to ensure thorough validation. These strategies must accommodate the dynamic nature of ADS operations, demanding rigorous and innovative approaches to ensure that ADS operate safely under all conditions.

1.1.1 Post-Deployment Safety Challenges

To further illustrate the ongoing efforts to monitor and enhance the safety of ADS post-deployment, the National Highway Traffic Safety Administration (NHTSA, USA) has issued a Standing General Order requiring manufacturers and operators to report crashes involving vehicles equipped with ADS [2]. This order seeks to gather detailed, real time crash data to address the primary safety challenge of understanding ADS behavior in a wide range of real-world scenarios. If an ADS-equipped vehicle crashes and the ADS was active within 30 seconds before the crash, causing property damage or injury, it must be reported. Although the report does not specifically discuss how safety challenges differ across various environments or conditions, it implies that the Operational Design Domains (ODDs) of ADS vehicles, which specify the conditions under which a given ADS can function, are crucial. The report mentions that among the reported crashes, several involved vulnerable road users, including cyclists, motorcyclists, buses, heavy trucks, pickup trucks, vans, SUVs, and passenger cars. According to the NHTSA, as depicted in Fig. 1.2, there have been 244 crashes involving vehicles with Level 5 autonomy over the last 12 months.

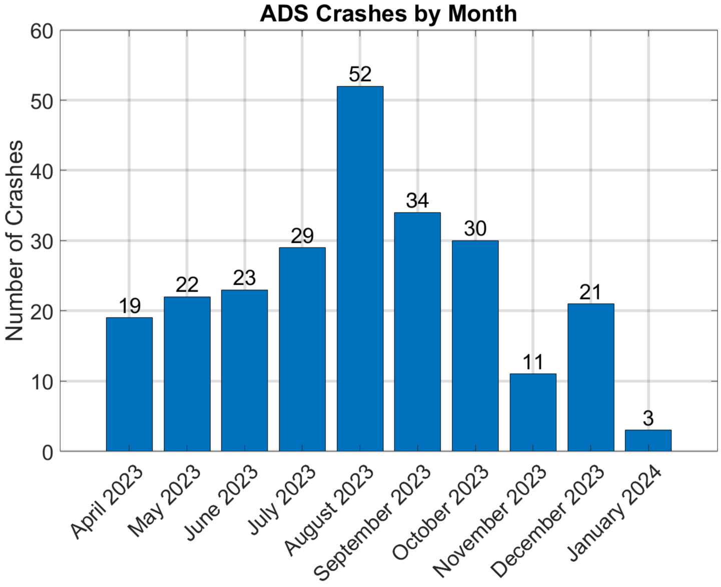


Figure 1.2: ADS equipped Vehicle Crashes in Last 12 Months (NHTSA Dashboard) [2]

The primary safety challenges faced by ADS after deployment include [77]:

- Self-Monitoring and Fail-Operational Capabilities: ADS must be capable of self-monitoring to reliably detect system degradation or failures.
- Validating Inductive Learning and Novel Environmental Inputs: A significant challenge involves validating Artificial Intelligence (AI) algorithms

against novel environmental inputs not included in the training or testing datasets. The complexity of real-world scenarios complicates ensuring that ADS can handle unexpected situations with the necessary safety.

- Achieving High Levels of Dependability: ADS must attain ultra-dependability, a safety level akin to commercial aviation, to be considered safe for large-scale deployment. This requires systems to be robust against diverse environmental conditions, including adverse weather, sensor noise, and unpredictable road hazards.
- Managing Infrequent Failures at Scale: With large-scale ADS deployment, managing rare failures becomes challenging. Even infrequent failures can pose unacceptable risks when scaled to a fleet of millions of vehicles. This requires a comprehensive approach to ensuring the safety of computer-based automotive systems that goes beyond traditional functional safety approaches.
- Social Acceptance and Human Factors: Gaining public trust in ADS technology is critical for its adoption. This requires demonstrating the safety and reliability of ADS and addressing challenges related to human factors, including interactions with ADS occupants, pedestrians, and drivers of non-autonomous vehicles [89].

1.1.2 How Much Risk is Safe Enough?

Transitioning from the specific challenges to a broader regulatory perspective, the question of "How Much Risk is Safe Enough?" emerges as a pivotal concern. The concept of "Acceptable Risk" is balanced with public safety and regulatory standards through various means. Firstly, the safety of ADS is recognized not solely as a technical issue but also as a matter of public and regulatory importance, necessitating engagement with broader societal values and standards. Secondly, risk quantification uses metrics like "miles per disengagement" or "fatalities per million miles driven," which are useful but limited. These need to be complemented by qualitative assessments that consider the broader societal impacts of the technology. Lastly, trust in ADS encompasses their performance (how safely they operate), the processes underpinning their development and deployment, and their intended purposes. This multifaceted approach to technological risk suggests that governance should extend beyond performance metrics to also consider development processes and the technology's goals within society [126].

The comparative analysis in Fig. 1.3 indicates a notable disparity in the types of collisions involving ADS compared to Conventional Vehicles (CVs), underscores the necessity of integrating both quantitative metrics and qualitative assessments in determining acceptable risk levels. Rear-end collisions and proceeding straight incidents are significantly more common in accidents involving ADS, which can be attributed to the aggressive driving style of these systems. Additionally, factors such as unsafe speed, following too closely, side swipes, violations of traffic signals and signs, and passing other vehicles represent a combination of maneuvers and errors by ADS that lead to a higher rate of accidents compared to CVs. Unlike broadside collisions, which is typically more severe but less common in ADS accidents. This suggests that ADSs are good at avoiding some high-risk crashes but might struggle with unusual situations that human drivers rarely face, leading to more rear-end collisions because of their unexpected way of driving [115]. This indicates that the overarching question of acceptable risk reinforces the argument for a multifaceted evaluation of ADS before deployment.

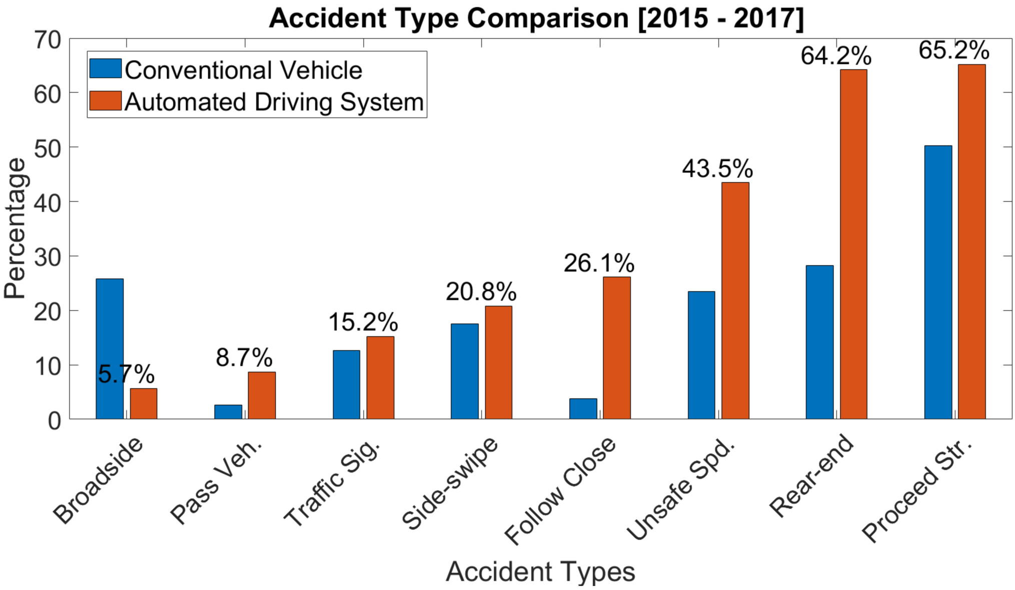


Figure 1.3: Comparative Analysis of Accident Types between ADS and CV [115]

Measuring the safety of ADS or quantifying risk is both essential and difficult. There are still some open questions and challenges to define how safe is safe enough [17, 75] or in other words, how much risk is safe enough. The accident state in the context of ADS is not binary, it encompasses a spectrum that ranges from a safe (non-catastrophic) state to an unsafe (catastrophic) state [1]. As shown in Fig. 1.4, accidents involving ADS range from safe state to catastrophic state, each presenting unique insights for system improvement. For instance, near-misses (S1C1), where accidents are averted at the last moment, highlight the importance of ADS or human intervention, serving as crucial data points for analysis and enhancement of the system. Minor incidents (S1C2,S1C3), involving

[1] Machin et. al in [92] proposed partitions between catastrophic states and non-catastrophic states for safety invariant

low-speed collisions or slight contacts, though typically resulting in minimal damage, reveal areas where the ADS could be improved. Moderate accidents (S2C2, S2C1), which involve significant vehicle damage or minor injuries, indicate failures in the ADS's decision-making or navigation capabilities. These issues may lead to stricter regulations and safety enhancements. Serious accidents (S2C3, S3C1), which cause major injuries or fatalities, point to significant failures in the ADS's operational abilities, requiring in-depth investigations and major system modifications. Catastrophic events (S3C2,S3C3), the most severe category, involving multiple vehicles and casualties, impact public trust and regulatory policies, possibly pausing ADS deployment for extensive system reviews. This gradation of incidents underscores the continuous learning and adaptation process, leveraging data from each event to refine algorithms, reassess design principles, and enhance sensor functionalities for improved safety [2].

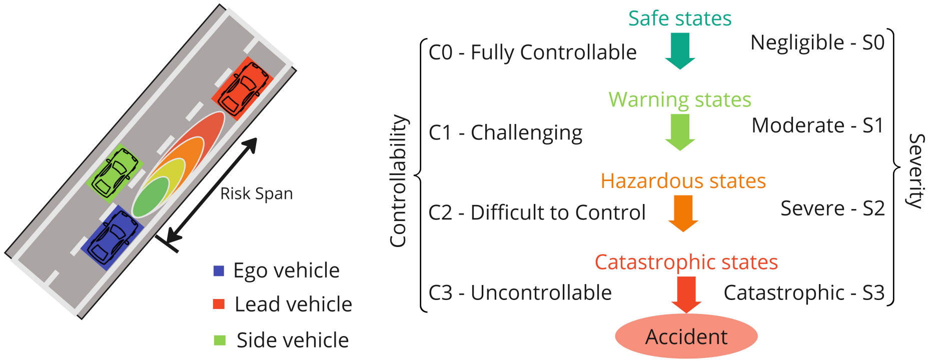


Figure 1.4: The Classification of Severity and Controllability as System-Dependent Risk Levels

As discussed earlier, traditional metrics like "miles per disengagement" and "fatalities per million miles driven" have limitations; not all disengagements equate in risk level, and such metrics don't account for non-fatal incidents or near misses [126]. Therefore, the RAND Corporation report [17, 44] emphasizes the importance of identifying effective safety measures and recommends both leading and lagging indicators for measuring safety. Additionally, the report advises that safety measurement methods must be valid, feasible, reliable, and non-manipulatable. Leading measures can serve as proxy measures of driving behaviors correlated to safety outcomes, while lagging measures involve actual safety outcomes that include harm.

In this context, controllability levels (C0-C3) serve as leading indicators, offering runtime insights into the ADS's ability to manage different driving scenarios and suggesting areas for technological improvement or risk mitigation. Conversely, severity levels (S0-S3) act as lagging indicators, mea-

[2] The classification of controllability and severity as system attributes underscores their inherent relation to the system's design, operation, and potential failure modes [71].

suring the outcomes of ADS interactions in terms of incident severity, from no injury to fatalities. This approach highlights the complexity of quantifying risk in ADS and stresses the need for a comprehensive set of metrics, including both predictive and outcome-based indicators, to ensure safety.

1.1.3 The Gap in Current Risk Assessment Techniques

The ISO 26262 standard [41], while comprehensive for Electrical/Electronic (E/E) system failures, does not fully cover the broader risks associated with ADS, including challenges outside of E/E problems, such as decision-making by software algorithms under uncertain conditions [133]. Traditional risk assessment technique, such as Hazard Analysis and Risk Assessment (HARA), relying on predefined scenarios and static hazard analysis, struggle with the dynamic and unpredictable nature of real-world driving environments [95]. This creates a gap in identifying hidden connections and operational states that are not recognized as failure modes, highlighting the need for functional safety in complex automotive systems.

The complexity of ADS, marked by their interconnected nature and the requirement to interact with their environment, poses significant challenges. This complexity leads to a system-of-systems scenario, increasing the risk of malfunctions that could spread across system boundaries. The absence of standardized practices for assigning Automotive Safety Integrity Levels (ASILs) and subjective assessments complicates early hazard identification in the ADS development process [57]. Additionally, functional insufficiencies, where system sensors, algorithms, or overall functionality fail due to environmental constraints, highlight another gap in current risk assessment approach. This indicates that even if a vehicle is safe from hardware malfunctions, it might still be risky in complex driving situations for which it was not designed or tested [35].

Additionally, the use of AI Algorithms in ADS introduces two major obstacles: the lack of complete specification for AI-based components due to the complexity of defining all potential environmental interactions, and the non-interpretability of AI models, especially deep neural networks, which hinders the ability to manually verify the correctness of these systems. The lack of complete specifications and the issue of non-interpretability significantly affect the ability to verify and test AI-based components in ADS. Approximately half of the verification and testing methods outlined in ISO 26262 become inapplicable under these conditions, leading to challenges in assuring the safety and reliability of ADS [120].

Traditional safety standards and methodologies, such as ISO 26262 and ISO 21448 [43], do not adequately consider environmental impacts over the vehicle lifecycle, struggle to obtain statistically valid failure probabilities, and lack established quantitative metrics for ADS. The inability to perform complete vehicle-level testing and the challenge of handling complex and uncertain failure modes can result in overlooked potential hazards.

Without clear risk metrics for ADS, making decisions under uncertainty becomes difficult, potentially compromising safety in various operational situations when ADS are deployed [130].

This situation necessitates an enhanced risk assessment methodology that includes a wider spectrum of operational scenarios beyond typical testing environments. Developing a deeper understanding of the ODDs, including environmental, geographical, and temporal limitations for safe operation, is essential. Additionally, integrating robustness and resilience in system design is crucial, allowing ADS to maintain safety or degrade gracefully under conditions beyond its nominal operational capabilities.

1.2 Problem Statement

This section focuses on identifying and analyzing the risks involved in developing and deploying ADS as part of the problem statement on ADS risk assessment. The importance of considering dynamic factors in risk assessment is highlighted, acknowledging that risks evolve in real-world scenarios where conditions, human interactions, and ADS behaviors can change unpredictably. Traditional HARA methods, as outlined in ISO 26262, struggle to adapt dynamically and may not fully understand the complex nature of ADS environments. Additionally, subjective risk assessment ratings can be problematic because they depend on personal judgment, which can vary greatly between people and may not always consider dynamic factors. Subjective risk ratings might find it hard to accurately measure risks and adjust to new or changing information.

Therefore, this thesis proposes a new, more objective, and dynamic framework for evaluating the risks associated with ADS. It aims to improve current methods by incorporating specific information about risks and using learning-based algorithms to better adapt to dynamic conditions. The goal is to thoroughly examine ADS under various conditions to confirm their safety, reliability, and functionality before they deployed. This includes identifying potential hazards and evaluating their effects using dynamic risk indicators. While this work aims to provide a basis for taking safety measures to reduce risks to acceptable levels, the detailed implementation of such measures is beyond its scope. This method seeks to increase the precision of risk evaluations and aid in making more informed decision-making for the development and deployment of ADS.

In this thesis, given the evolving risks in ADS environments and the limitations of HARA, the following research questions and objectives are proposed:

1. Research Question: What are the specific dynamic factors that significantly influence ADS risk, and where do current models fall short in capturing these dynamics?

Research Objective: Conduct a thorough investigation into dynamic factors affecting ADS risks, such as environmental variability and ADS behavior, and pinpoint the deficiencies in current risk assessment models regarding these factors.

2. Research Question: How can a more effective risk assessment methodology be developed to integrate dynamic factors into ADS evaluations?

   Research Objective: Design and validate a comprehensive and iterative framework that incorporates dynamic factors (e.g., failure modes, environmental changes, ADS behavior) into the risk assessment process for ADS, utilizing runtime data and predictive analysis. Study how the dynamic risk indicators can vary in diverse safety-critical scenarios.

3. Research Question: How can supervised Machine Learning (ML) algorithms enhance HARA to increase the objectivity and effectiveness of ADS risk assessment?

   Research Objective: Develop and test ML algorithms to dynamically assess risks, aiming to automate the process, enhance accuracy, and improve adaptability to new situations compared to the current HARA method. Study how correlations between risk factors affects risk assessment accuracy. Evaluate the quality of empirical evidence that supports the use of dynamic risk indicators.

1.2.1 Dynamic Driving Task and Operational Environment

The Dynamic Driving Task (DDT) of ADS is composed of multiple system components integrated together, as depicted in Fig. 1.5. For instance, sensors focus on data acquisition, underscoring their role in environmental perception and obstacle detection. The Driver-Human Machine Interface (HMI) monitors ADS functions and provides safety alerts when necessary. Vehicle actuators are essential for implementing ADS decisions, influencing steering, braking, and throttle. The Decision and Control unit makes tactical decisions such as speed adjustment and obstacle navigation, while Vehicle Dynamics Management coordinates actuators and monitors the vehicle's state to ensure safe and intended movements.

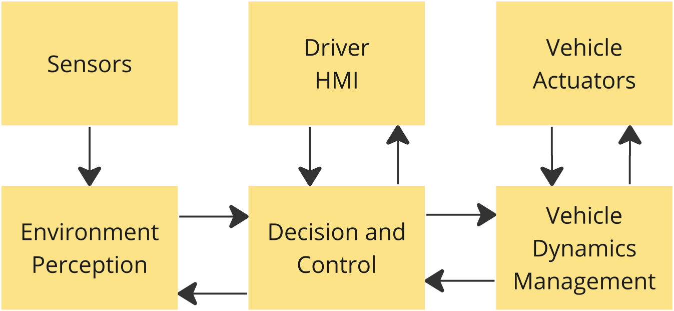


Figure 1.5: Conceptual Architecture of ADS [65]

A DDT for ADS encompasses a broad spectrum of activities that require runtime processing and decision-making in response to ever-changing environmental conditions, traffic patterns, and sensor data. Unlike conventional driving tasks where human drivers use their senses and cognitive skills to respond road situations, ADS rely on sensors and AI algorithms to perceive their environment and make decisions. For instance, ADS must continuously adjust driving strategies to maintain safety amidst varying traffic densities, weather patterns, and road conditions. The dynamic nature of these tasks is further complicated by the need for tactical decision-making, such as speed adjustments and lane changes based on sensor inputs, which are affected by factors like sensor reliability, environmental variability, and sensor degradation over time.

Not only this, the operational environment plays a critical role in influencing the performance and risks associated with ADS. Variability in weather conditions such as rain, fog, snow, and changes in lighting can significantly affect sensor performance and the decision-making algorithms of ADS. For instance, poor visibility or sensor obstruction can impede the ADS's ability to accurately perceive its surroundings, increasing the risk of misjudgment and accidents. Traffic congestion and fluctuating traffic patterns demand continuous adjustments in driving strategy, posing challenges in maintaining safety. Additionally, changes in road conditions, obstacles, and unexpected events like accidents or roadworks require ADS to make runtime adjustments, thereby influencing their operational performance and associated risks.

Modeling the DDT for risk assessment presents several challenges. First, the variability in environmental conditions, and operational environment necessitates sophisticated models that can accurately simulate the myriad of possible scenarios ADS may encounter. Second, the performance and reliability of sensors and detection systems, which can be affected by environmental factors and wear and tear, add another layer of complexity in modeling the ADS's ability to perceive its surroundings accurately. Third, the inclusion of adaptive and AI algorithms in ADS introduces unpredictability, as these systems shape their responses based on accumulated data, potentially leading to unforeseen behavior. Last, the need to account for the dynamic interaction with other road users complicate the modeling process, making it challenging to assess risks accurately.

1.2.2 The Need for Dynamic Risk Assessment

Dynamic Risk Assessment (DRA) offers a significant advancement over static HARA by enabling runtime evaluation of risks in accordance with the actual operational context and the system's current state. Trapp et al. in [132] highlighted the shift towards Dynamic Safety Management to address the uncertainties in autonomous systems' behaviors and operational contexts, allowing for runtime safety and performance optimization (as shown in Fig. 1.6). To implement the adaptation process and understand its roles, answers to different aspects of the questions (5W+1H), coined by [122] and used in Table 1 of [111], are essential. The central question in this framework, "WHY system need to perform adaptation?", triggers the adaptation process based on runtime risk assessments.

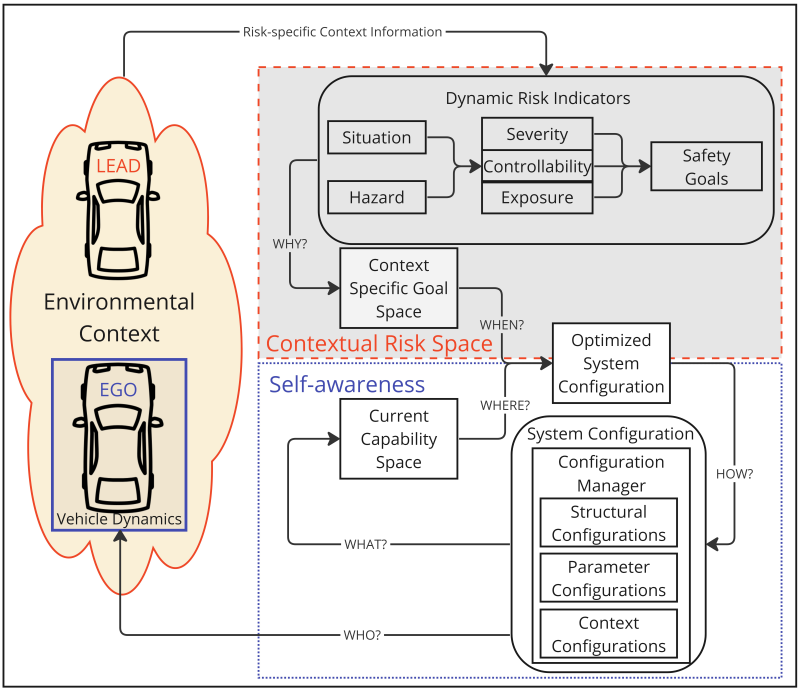


Figure 1.6: Basic Idea of Dynamic Safety Management (inspired from [132] and adapted from [111]).

The inherent non-linear and non-deterministic behaviors of AI algorithms that drive ADS pose a challenge to foresee all potential operational scenarios during the design phase. Due to the specific characteristics of ADS, there are many uncertainties making it difficult or even impossible to predict the system's behavior necessitates a shift from static, worst-case scenario-based assessments to more flexible, DRA. Additionally, it provides a more realistic and flexible basis for the development of safe and cost-efficient ADS. DRA enables ADS to dynamically assess risk and it can adjust (using Why? as risk reasoning for) safety measures in response to actual risks related to the current context [3] [111, 112]. This adaptability ensures more efficient and effective operations compared to static assessments' constraints.

[3] The focus on safety measures based on DRA is beyond the scope of this thesis. However, DRA could be coupled with self-adaptive systems that offer "Risk Reasoning" to explain why the system must perform adaptation operations.

However, testing ADS with DRA in real-world driving is not a practical solution. Additionally, testing ADS over extensive driving distances is physically costly and not a feasible approach. Therefore, DRA using simulation-based approach could be beneficial for the testing of ADS across a vast array of driving scenarios, including rare or hazardous situations that are difficult or dangerous to reproduce in real-world testing.

1.2.3 Learning from Risk-Specific Context Information

The HARA process in the automotive domain evaluates three critical components for each identified hazard to provide insights into risk decomposition (R) shown in Eqn. 1.1:

$$R = f(S, E, C) \tag{1.1}$$

1. Severity (S) measures the potential impact of a hazard occurring, indicating the seriousness of injury or damage that could result. Influenced by factors such as impact energy and environment characteristics, severity aids in understanding the hazard's potential consequences [64].
2. Exposure (E) assesses the likelihood or frequency of encountering a situation where a hazardous event might occur, reflecting the system or its components' susceptibility to conditions that could lead to a hazard. Exposure assessment can vary subjectively based on expert knowledge, affecting the exposure ratings [10].
3. Controllability (C) evaluates the vehicle occupants' (or users') ability to avoid harm when facing a potential hazard (i.e. ADS capabilities in absence of driver). It measures the manageability of a hazard by the driver or user under different conditions, influenced primarily by the vehicle's capability to alter its trajectory and environmental factors impacting this ability [100].

By assessing severity, exposure, and controllability, developers can prioritize hazards and implement suitable safety measures. The standard categorizes severity into four classes: S0 (negligible), S1 (moderate injuries), S2 (severe injuries), and S3 (catastrophic injuries). Exposure is divided into: E1 (very low), E2 (low), E3 (medium), and E4 (highly probable). Controllability is classified as: C0 (fully controllable), C1 (challenging), C2 (difficult to control), and C3 (uncontrollable) [4]. As shown in Fig. 1.7, the two-dimensional risk assessment is depicted through a risk matrix that considers two primary factors: the ability to avoid harm (controllability) and the potential impact of a hazard (severity). This matrix is commonly used to categorize risks into different levels (e.g., low, medium, high) based on these two dimensions. This two-dimensional model is good for broadly categorizing

[4] In this thesis, the research focuses solely on severity and controllability for the specified hazardous events, with the exposure rating considered as E4 (highly probable).

risks but might miss the details of complex risk scenarios, especially those that change over time or are influenced by various risk factors.

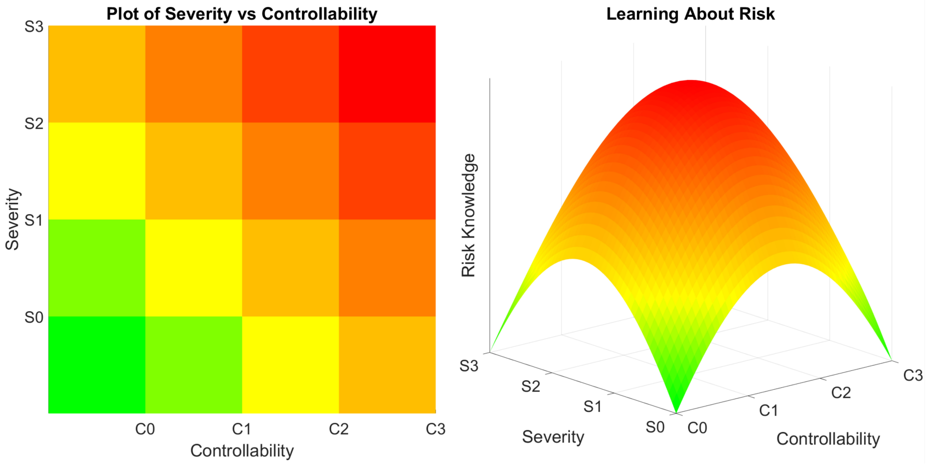


Figure 1.7: Learning about Risk: Incorporating the Risk Knowledge Dimension (inspired from [11, 108])

The introduction of the risk knowledge dimension transforms the traditional two-dimensional model into a three-dimensional model, as shown in Fig. 1.7. This illustrates that as risk knowledge increases, the perception and categorization of risk can shift, emphasizing the need for continuous learning and adaptation in risk management. The importance of risk knowledge (i.e. risk-specific context information) for DRA is important due to the evolving nature of risk. As ADS technologies and operational environments change, so do the risks associated with them. This evolution necessitates a continuous reassessment of risks, incorporating new knowledge to refine our understanding and management of these risks.

Risk knowledge enables a deeper understanding of the performance and limitations of ADS within high-risk environments. An approach presented in [11, 108, 110], emphasize the importance of continuous information systematization and the inclusion of new risk evidence through continuous monitoring (i.e. through iterative process). This approach is centered around the knowledge dimension of risk, advocating for an iterative process that adapts to changing conditions and improves risk management based on evolving knowledge.

Truth tables and rule-based approach, while useful in certain contexts, confront significant challenges when applied to DRA in complex and uncertain environments. Their limitations stem from issues with complexity, scalability, adaptability, and a static nature that does not readily accommodate updates without manual intervention. Both struggle with ambiguity and lack the capability to learn from new data, often oversimplifying real-world scenarios and potentially overlooking critical risk factors.

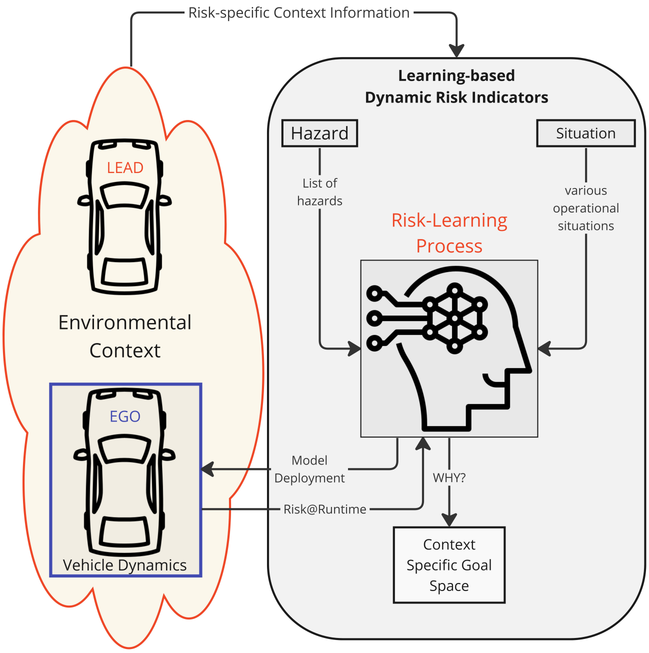


Figure 1.8: Idea of Risk-Learning Process [113]

To address these limitations, the application of a learning-based approach introduces a dynamic method to continuously refine risk assessment strategies [113]. By treating risk knowledge as an evolving variable, supervised ML-based algorithms can automate the process of updating risk knowledge based on new data, insights from past incidents, and ongoing operational feedback (as shown in Fig. 1.8). This approach enables the analysis of vast datasets to uncover hidden patterns, trends, and correlations that may elude human analysts (i.e. Safety engineer), thus enriching the depth and breadth of risk-related knowledge. The iterative process that a learning-based approach provides establish a feedback loop for the continuous integration of new information into risk assessments. This iterative process ensures the risk knowledge base remains current, reflecting the latest in operational conditions, and ADS advancements.

Furthermore, to extend the risk assessment coverage from unknown-unsafe region to known-safe region (Table XII from [21]) within ADS, Zio et al. in [147] suggest that a vast, combinatorial set of possible scenarios, events, and conditions needs consideration, with only a few rare ones leading to critical, unsafe situations. Through this proactive exploration of both accidents and incidents [4], ADS developers can identify potential risks and evaluate the effectiveness of mitigation strategies. The integration of a learning-based approach thereby equips decision-makers with the most current and comprehensive risk information, enhancing decision-making in the face of uncertainty.

1.3 Contributions

This thesis introduces the LeArning-based Dynamic Risk Indicators (LADRI) framework, a novel approach designed to enhance risk assessment for ADS by leveraging runtime risk-specific context information. The LADRI framework implements a cyclic process, named Plan-Do-Train-Adjust-Assess (PDTAA), as depicted in Fig. 1.9, to dynamically update and refine risk indicators. This allows for a more accurate and comprehensive assessment of risks, ensuring the framework's adaptability to new insights and shifts in ADS operational conditions. The LADRI framework is centered around the PDTAA cyclic process, designed to iteratively refine risk assessments by integrating both design and non-design parameters along with potential failure conditions, as shown in Fig. 1.10.

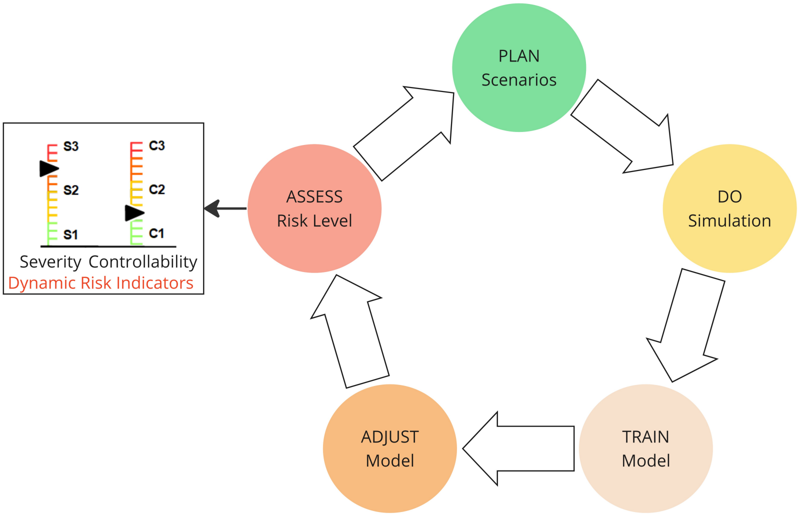


Figure 1.9: Concept of Enhancing Continuous Risk Assessment (Inspired from [39])

This thesis seeks to address the limitations mentioned earlier by using a learning-based approach within the LADRI framework. This method helps discover hidden patterns and connections related to specific risks, improving the precision of risk assessments. Leveraging supervised ML algorithms, the LADRI framework seeks to provide a more objective and dynamic tool for risk assessment. The task of each phase of the LADRI framework contribute to the next phase of the cyclic process:

1. PLAN: This initial phase defines the framework for simulation by identifying Design parameters (directly controllable aspects like sensor configurations and vehicle control strategies), Non-Design parameters (uncontrollable factors like environmental conditions and traffic behavior), and Failure conditions (potential system failures or errors, identified using Hazard and Operability Analysis (HAZOP) Guide words).

Study objective:

- 1a: What are the critical design and non-design parameters, along with failure conditions, that must be considered to effectively simulate real-world driving conditions for ADS?
- 1b: How can these parameters be quantitatively defined to maximize the relevance and comprehensiveness of the LADRI framework?

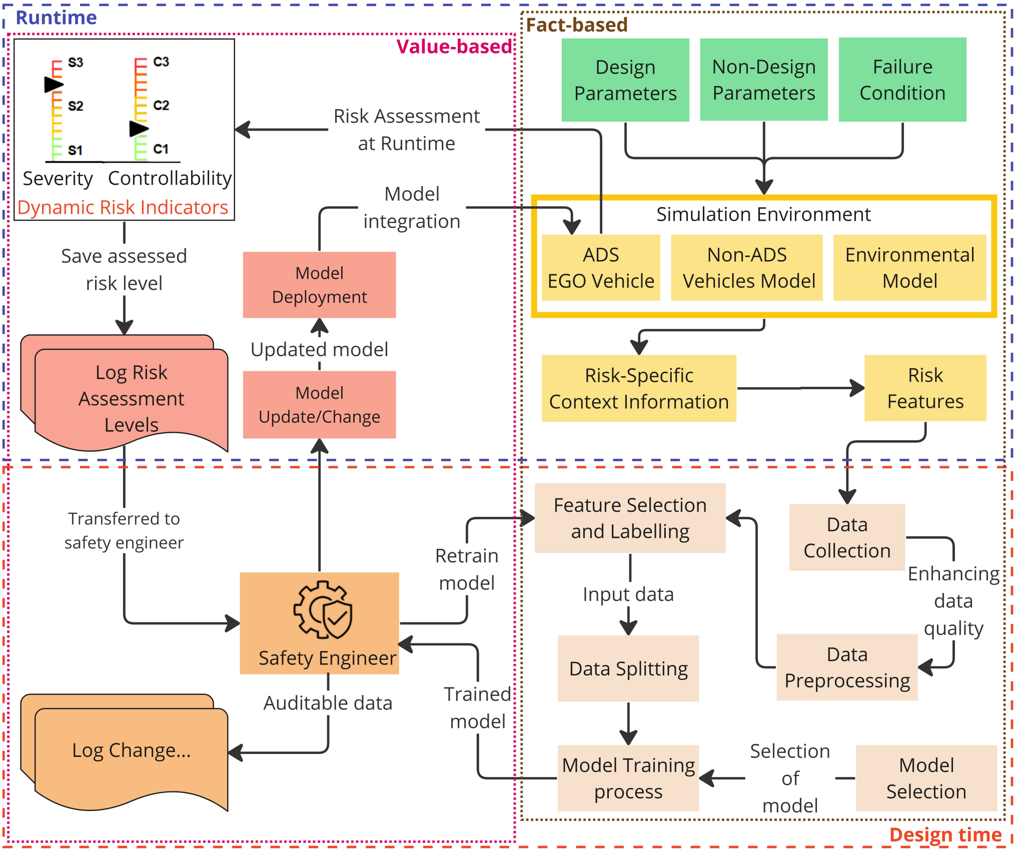


Figure 1.10: Schematic Representation of the LADRI Framework

2. DO: Following the Plan, this phase involves the execution of simulations that replicate the interactions between the ADS, the environment, and surrounding traffic behavior across a spectrum of driving conditions. The aim is to assess how these interactions influence the ADS's decision-making algorithm in different scenarios, generating risk-specific context information and transformed them into risk features for further analysis.

   Study objective:

   - 2a: How can interactions between the ego vehicle model, the environmental model, and the surrounding traffic behavior model be accurately simulated to reflect a wide range of driving conditions?
   - 2b: How do these interactions can generate a reach risk-specific context information in various scenarios?

3. TRAIN: Utilizing the risk features generated in the Do phase, this phase focuses on training and testing supervised ML models to predict dynamic risk indicators. The process involves analyzing the data to identify underlying patterns and correlations that inform the risk assessment.

   Study objective:

   - 3a: Which ML algorithms learn most effectively from risk features extracted from risk-specific context information?
   - 3b: How ML model can be optimized (e.g., by changing hyperparameters) to better manage the complexities and variations of ever-changing road environments?

4. ADJUST: In this phase, safety engineer review the performance of the trained models. If the ML models' predictions do not satisfactorily reflect or fail to meet predefined criteria, adjustments or retraining occurs. This way safety engineer ensures the models' accuracy and performance in predicting risk indicators in runtime.

   Study objective:

   - 4a: What are the roles of a safety engineer in intervening to refine a ML model to ensure it neither overfits nor underfits?
   - 4b: How can ML model be adjusted to accurately reflect the risk thresholds and acceptance criteria for risk prediction?

5. ASSESS: With the ML model fine-tuned, it is deployed within the ADS to assess risk levels in operational settings, identifying severity and controllability indicators for scenarios not encountered during the training phase. Safety engineers analyze these predictions to determine necessary improvements or adjustments in the ADS, such as enhancing braking performance or modifying sensor configurations.

   Study objective:

   - 5a: Upon deploying the ML model into ADS operations, how can its performance and accuracy in risk assessment be continuously monitored and logged for further refinement of the tool?

The process repeats cyclically once safety updates are made, with each iteration building on the accumulated risk knowledge and risk indicators from previous cycles. The cyclic process continues until all unknown-unsafe situations within a particular ODD are explored. As depicted in Fig. 1.11,

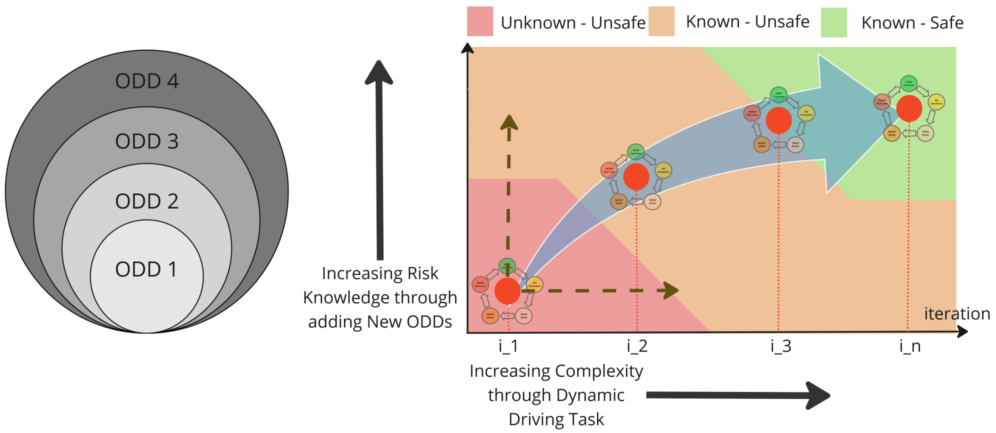

Unknown - Unsafe
Known - Unsafe
Known - Safe
ODD 4
ODD 3
ODD 2
ODD 1
Increasing Risk Knowledge through adding New ODDs
iteration
i_1
i_2
i_3
i_n
Increasing Complexity through Dynamic Driving Task

Figure 1.11: Advancement of Risk Knowledge Through Iterative Cycle

the process begins with ODD 1 in the first iteration, encountering the first failure, followed by subsequent iterations for each addressing additional failures. Once all failures are attempted, a new ODD included. For instance, if highway lane following straight road driving is considered in ODD 1, then ODD 2 might be a curved road, ODD 3 an uphill scenario, ODD 4 a downhill scenario, and so forth. The iterations on the X-axis correspond to increasing complexity in DDT; for example, the first iteration might cover basic acceleration/deceleration and braking operations, the second iteration could introduce steering maneuvers, the third might involve overtaking maneuvers, and the fourth could include exiting and entering the highway.

As mentioned earlier, advanced simulation tools can generate a wide range of hypothetical scenarios based on the existing knowledge base. By exploring these scenarios, ADS developers can proactively identify potential risks and assess the effectiveness of various mitigation safety strategies.

1.4 Assumptions

The research focuses solely on severity and controllability risk indicators for the specified hazardous events, with the exposure rating considered as E4 (highly probable). The thesis employs a simulation environment for validating its hypothesis, as articulated in the motivation section and reiterated throughout the thesis when relevant. Within this environment, three vehicles are utilized: one ego vehicle equipped with ADS functionality and two non-ADS vehicles, positioned at the side and in the leading side, respectively. To observe the ADS's behavior under the influence of failure, a Graphical User Interface (GUI) is employed for injecting runtime failure conditions during the simulation. Failures are introduced via a switch mechanism, halting specific behaviors to assess the risk. This thesis examines two types of failures: the first involving the throttle pedal position sensors and the second involving the brake pedal position sensors. The study focuses on the HAZOP guide word "Wrong Value" to analyze these failures. The primary focus of this research is on highway-following driving scenarios. Although the principles developed may have broader applicability, the thesis specifically addresses the unique challenges and dynamics of highway driving, excluding lateral maneuvers from its scope of investigation. Acknowledging that absolute safety assurance is unattainable, the thesis sets a threshold whereby ML model performance above 99% is considered safe for deployment in ADS for risk assessment.

1.5 Thesis Structure

The thesis is structured into five chapters, each methodically contributing to the exploration and development of the LADRI framework for ADS.

Chapter 2 sets the stage by defining key terminologies and reviewing the state of the art, identifying gaps in current research.

The development of the LADRI framework is detailed in Chapter 3, where the PDTAA cyclic process is introduced.

Chapter 4 evaluates the framework across different scenarios using varied risk feature combinations, discussing the optimization of ML models and the framework's applications and limitations.

The thesis concludes in Chapter 5, summarizing the findings, discussing the implications for future research, and highlighting the LADRI framework's potential to advance ADS risk assessment.

2 State of the Art

In this chapter, the thesis systematically unfolds the foundation and state-of-the-art approaches relevant to research on risk assessment.

Section 2.1: Terminology - key terms and definitions are established in this section, ensuring a common language and understanding for the discourse ahead.

Section 2.2: Background Work - This section delves into the evolution and current methodologies of risk assessment, from its theoretical underpinnings to practical applications, and how it is connected to LADRI framework.

Section 2.3: Related Work - An exploration of contemporary studies and methodologies that share goals with the research being performed in this thesis.

Section 2.4: Bridging the Gap - this section argues the need for a new approach in risk assessment, one that transcends current state-of-the-art practices to address unresolved challenges.

Chapter Content

2.1 Terminology

The purpose of this section is to clarify essential terms and establish a terminology for exploring the depths of this thesis and risk assessment in particular.

Automated Driving System

An Automated Driving System (ADS) is defined as a Level 3 to 5 driving automation system according to the Society of Automotive Engineers (SAE) Levels of driving automation [139]. This thesis considers a vehicle equipped with Level 5 automation.

Item

An item is a system or combination of systems, to which ISO 26262 is applied, that implements a function or part of a function at the vehicle level [41].

Hazard

A hazard is a potential source of harm (i.e., physical injury or damage) caused by malfunctioning behavior (i.e., failure or unintended behavior) of the item [41].

Hazardous Events

A hazardous event is a combination of a hazard and an operational situation [41].

Functional Insufficiency

Functional insufficiency refers to the inadequacy in a system's specification or performance that contributes to hazardous behavior or an inability to prevent, detect, and mitigate a reasonably foreseeable indirect misuse when activated by one or more triggering conditions [43].

Hazard Analysis and Risk Assessment

Hazard Analysis and Risk Assessment (HARA) is a method to identify and categorize hazardous event of items and to specify safety goals and Automotive Safety Integrity Levels (ASILs) related to the prevention or mitigation of the associated hazards in order to avoid unreasonable risk [41].

Dynamic Driving Task

The Dynamic Driving Task (DDT) refers to the runtime operational and tactical functions necessary for operating a vehicle in traffic. It encompasses various functions, including lateral and longitudinal vehicle motion control (operational), monitoring the driving environment and responding to objects and events (operational and tactical), maneuver planning (tactical), and enhancing vehicle conspicuity through lighting, signaling, or gesturing (tactical). These functions are critical for safely navigating a vehicle in diverse traffic conditions [43].

Operational Design Domain

SAE J3016 defines an Operational Design Domain (ODD) as "Operating conditions under which a given driving automation system, or feature thereof, is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics [119]."

Safety

Safety is absence of catastrophic consequences on the user(s) and the environment [13].

Safety Measure

A safety measure is an activity or technical solution to avoid or control systematic failures and to detect or control random hardware failures, or mitigate their harmful effects [41].

Safety Goal

A safety goal is a top-level safety requirement as a result of the HARA at the vehicle level [41]. One safety goal can be related to several hazards and several safety goals can be related to a single hazard.

Design Parameters

Design parameters refer to the elements or variables within a system that can be adjusted or configured to influence the system's performance. In the context of ADS, design parameters are the directly controllable aspects that affect how the system operates [50]. These parameters include:

- **ADS Vehicle Dynamics**: In this thesis, to assess physical characteristics of the ADS vehicle, vehicle´s engine, brakes, and aerodynamics, acceleration, and deceleration capabilities are used to assess the ADS algorithms.
- **Sensor Configurations**: The arrangement and types of sensors used in the ADS, such as Cameras, Radar, and Lidar. These sensors gather data about the vehicle's surroundings.
- **Algorithm Parameters**: The settings within the algorithms that process data from the sensors and use it to make decisions. These parameters can affect how the system interprets data and chooses actions (e.g. ACC Controller).

Non-Design Parameters

Non-design parameters refer to external factors or conditions that impact a system's performance but are beyond the control of the system's designers or operators. In the context of ADS, non-design parameters are the variables related to the environment, traffic behaviors, and road condition that the ADS must adapt to, even though it cannot alter these conditions [50]. These parameters include:

- **Other traffic participants**: The behavior of other vehicles (e.g., their speed, direction, lane changes, and braking actions) is external to the ADS, and it considered as non-design parameters.
- **Environmental Conditions**: Weather patterns such as rain, snow, fog, or varying light conditions that can affect sensor performance and vehicle handling.
- **Infrastructure Elements**: The condition and layout of the road infrastructure, including signs, signals, road markings, and the presence of construction zones or obstacles.

Failure Conditions

A system failure occurs when a system's performance deviates from its intended function, transitioning from delivering correct to incorrect service [12]. It is an inability of a system, component, or process to perform a required function according to its intended operation or design specifications. Failures can be caused by various factors, including design flaws, hardware malfunctions, software bugs, operational errors, or external disturbances. In the computer-based systems, it is divided as follows [97]:

- **Omission**: Refers to the absence of a response or action that should have been provided by the system. It represents a failure condition where the system fails to perform a function or produce an output that is expected under given circumstances.

- **Wrong Value**: Refers to the occurrence of incorrect values being produced or utilized by a system, which may initially seem valid, making detection challenging. Even slight inaccuracies in these values can result in substantial adverse outcomes.
- **Delay**: Refers to an action or operation is performed later than expected. This could impact system performance or functionality when timing is critical, such as in runtime systems where late execution might lead to missed deadlines or failure to synchronize with other system components.

Risk-Specific Context Information

Risk-specific context information refers to the broader environmental, operational, and situational data that characterizes the setting in which risks occur. This includes information about the physical environment, operational conditions, technological systems, and any external influences that might affect risk.

Risk Features

Risk features are specific, quantifiable attributes or variables derived from the risk-specific context information that are directly used in Machine Learning (ML) models to identify, assess, and predict risks. These can include measurable parameters or derived metrics that are indicative of risk levels.

Risk Indicators

Risk indicators are data points (i.e risk levels) that provide insights into the current state of risk or the potential for future risks within a system. For instance, severity, as a risk indicator, refers to the potential impact or consequences of a risk event if it were to occur, whereas controllability refers to the degree to which a ADS system can effectively manage or mitigate a risk event once it has been detected or as it is occurring.

Supervised Learning

Supervised learning is a ML approach that uses labeled data, involving a training phase and a testing phase. In supervised learning, the model learns to categorize data into predefined labels based on the features of the training dataset [14]. This categorization capability is then tested on the testing dataset to measure the model's ability to accurately predict the labels of unseen data. Once the model has been trained and tested, the learned model can be deployed to predict unseen and unlabeled data, extending its utility beyond the initial dataset to real-world applications.

2.2 Background

This section provides the technical background essential for understanding the LeArning-based Dynamic Risk Indicators (LADRI) framework discussed in this thesis. Initially, an overview of risk assessment and the phases integral to its process are explained, detailing the systematic approach adopted within this field (Section 2.2.1). This serves as a foundation for the subsequent exploration of HARA method, as outlined in ISO 26262 (Section 2.2.2), while acknowledging the limitations and structural considerations previously discussed in Section 1.1.3.

Further, the nature of risk assessment is examined from both qualitative and quantitative perspectives (Section 2.2.3), emphasizing the importance of dynamic risk assessment in responding to the evolving nature of risks (Section 2.2.4).The discussion includes identifying various risk features (Section 2.2.5) and exploring the importance of risk assessment indicators, especially leading and lagging indicators (Section 2.2.6). The advancement towards simulation and scenario-driven approaches represents the latest trend in risk assessment methodologies, reflecting a shift towards more nuanced and adaptable strategies (Section 2.2.7).

Finally, the role of supervised learning algorithms in enhancing risk assessment practices is explored (Section 2.2.8), highlighting the intersection of advanced algorithms and traditional risk assessment techniques.

2.2.1 Overview of Risk Assessment

The landscape of risk assessment is changing due to technological advancements and new safety challenges in systems like ADS. It emphasizes the need to integrate simulation techniques, focus on resilience, and adopt dynamic approaches to assess risks. The complexity of cyber-physical systems, climate change impacts, and emerging risks/threats are underscored as critical areas needing attention. The use of computational capabilities and data to enhance risk assessment methodologies suggests a shift towards more comprehensive and adaptive strategies for managing risks in complex environments [147].

Defined by ISO 31010 [42], risk assessment is a systematic process that includes Risk identification, Risk analysis, and Risk evaluation:

Risk Idenfitication: Techniques for identifying risks include a wide array of methods designed to pinpoint potential hazards. From Failure Modes and Effects Analysis (FMEA) to Hazard and Operability Analysis (HAZOP) studies and beyond, these methodologies extend to perception surveys, what-if scenarios, scenario analysis, and the utilization of checklists or taxonomies. The advantages of these structured approaches lie in their comprehensiveness and the thoroughness achieved through the application of multiple techniques, showcasing effective due diligence. However, these methods also face limitations, particularly in adapting to technological advances and the dynamic risks associated with the rapidly evolving field of ADS. The inherent complexity and variability of such systems may diminish the efficacy of some traditional risk identification techniques.

Risk Analysis: Within the broader context of risk identification, risk analysis methods delve into the sources of risk, potential consequences, control effectiveness, and event likelihoods. These methods, including Cause-Consequences analysis, Fault Tree Analysis (FTA), Decision Tree Analysis, and others, offer significant benefits by facilitating an in-depth understanding of risks and their potential implications, which in turn supports the development of targeted mitigation strategies. However, traditional risk analysis methods may not adequately capture the dynamic interactions and interdependencies between various system components and external factors in real-time. This can lead to underestimations of risk in scenarios where system behavior changes in response to external variables, such as weather conditions, road types, or unpredictable human behavior. Additionally, these methods often rely heavily on historical data and expert input, which may not always reflect future conditions or novel threats introduced by advancing technologies. This discrepancy necessitates the adoption of more system-specific approaches to accurately assess and address these challenges.

Risk Evaluation: The Consequence/Likelihood Matrix is a notable risk evaluation method, allowing for the effective comparison and communication of risks based on potential outcomes and occurrence probabilities (e.g. ALARP, Bayesian Networks, Risk Indices etc.). This method, known for its simplicity and the clarity of its visual representations, facilitates the quick ranking and comparison of varied risks. Despite its advantages, certain challenges persist; crafting a valid matrix demands specific expertise, and the inherent subjectivity in determining a single indicative value for consequences can complicate the accurate depiction of the complex, multifaceted risks associated with ADS. Moreover, the subjective basis of this evaluation tool may result in inconsistent risk assessments, especially within the intricate contexts typical of dynamic behavior of the system.

2.2.2 Hazard Analysis and Risk Assessment

In automotive domain, ISO 26262 [41] outlines a structured approach to hazard analysis within automotive electrical/electronic control systems, addressing different levels of system decomposition. At the item level, the highest level of decomposition, it recommends methods like brainstorming, checklists, quality history, FMEA, and field studies, which rely heavily on past experiences and expert knowledge. For more detailed safety analysis at lower levels, it employs, FTA, Event Tree Analysis (ETA), and HAZOP, among others, as common practices in the automotive industry.

However, these methods face limitations when applied to modern, complex electronic control systems. Techniques like FMEA, FTA, and ETA, while effective for addressing random hardware failures qualitatively and quantitatively, fall short in guiding the identification of unsafe system interactions and are based on oversimplified linear chain-of-event models that do not always apply to complex systems or software, given software's lack of random failures [84].

FMEA's bottom-up, inductive approach and FTA's top-down, cause-focused method highlights the analytical challenges in complex systems, emphasizing the labor-intensive process of identifying potential failure combinations that could lead to system hazards [37]. HAZOP, with its focus on guidewords and actual process modeling, offers a nuanced examination of component failures and system interactions, yet its application to automotive electronic control systems is not straightforward and may require modifications for effectiveness.

Furthermore, ISO 26262 suggests the applicability of these safety analysis methods to software development, despite the majority of software-related accidents stemming from requirements flaws rather than coding errors [84]. This distinction highlights the inadequacy of traditional hazard analysis methods for software, underscoring the need for tailored approaches that account for software's unique failure modes and the complex interdependencies within modern electronic control systems [133].

2.2.3 Nature of Risk Assessment: Qualitative Vs Quantitative

Qualitative assessments use expert judgment to categorize risks without precise numerical values, which is useful in initial stages or when data is scarce. Quantitative assessments, in contrast, use mathematical models for a detailed evaluation of risk probabilities and impacts. ISO 26262's traditional qualitative HARA may not suit ADS due to complex operational situations and the dynamic nature of ADS exposure to these situations. Unlike manually driven vehicles, where hazards are identified based on human-operated functionalities, ADS requires continuous adjustment of tactical decisions and performance capabilities, challenging the completeness of identified hazards and hazardous events.

The conservative design approach of traditional HARA, assuming global validity of situational frequencies, may lead to overly cautious ADS designs, failing to account for ADS's inherent variability and adaptability [138]. Established methods often rely on ordinal scales for qualitative assessments, which suffer from limitations such as poor resolution and neglecting correlations, as highlighted in [8, 27, 59].

The quantitative risk assessment approach, on the other hand, evaluates risks by assigning numerical values to both the likelihood of hazard occurrences and their potential consequences. Unlike qualitative risk assessment methods that rely on descriptive analysis and expert judgment, quantitative risk assessment utilizes data and statistical methods to quantify risks. It involves calculating the frequencies of failures, probabilities of different outcomes, and the expected severity of consequences in measurable terms. The output of a quantitative risk assessment is typically a numerical value that represents the risk, facilitating a comparison against established risk acceptance criteria to determine if the risk is acceptable or if further mitigation measures are needed. Through the quantification of risks, this approach emphasizes actual outcomes over theoretical risks, enhancing the completeness and efficiency of safety goals by classifying incidents into predefined categories [138].

The integration of qualitative and quantitative methods for ADS risk assessment is not only feasible but advantageous. This approach yields safety goals with a quantitative integrity attribute, such as a numeric value (e.g. S3 or C3) for the maximum frequency of each incident type or the impact of a hazard, diverging from ISO 26262's qualitative norms on ASIL inheritance and decomposition. Instead, it adopts traditional mathematical quantitative rules to refine these safety goals into allocated safety requirements during the ADS development phase. This synthesis enables a thorough safety argumentation encompassing all potential safety risk causes, whether systematic faults in software or hardware design, random hardware faults, or performance limitations of sensors or actuators.

Furthermore, it supports an integrated HARA refinement strategy, Verification & Validation (V&V) approach, and safety case structure, addressing issues traditionally separated in ISO 26262 and ISO 21448. The customized HARA emphasizes that a quantitative framework still allows for the inclusion of qualitative evidence. For instance, criteria like ASIL from ISO 26262 for ensuring freedom from systematic faults still apply in this mixed approach. This means that safety goals based on quantitative data can also include qualitative evidence, leading to a thorough and robust risk assessment for ADS.

Connection to LADRI: LADRI acknowledges the limitations of traditional qualitative HARA in addressing the complex operational situations of ADS, mirroring concerns about the dynamic nature of ADS exposure. By incorporating a learning-based approach, LADRI facilitates the continuous adjustment of tactical decisions and performance capabilities of ADS, addressing the need for a more nuanced risk assessment method. Recognizing the conservative design approach of traditional HARA, LADRI integrates both qualitative judgments and quantitative data. This dual approach allows LADRI to navigate the inherent variability and adaptability of ADS, moving beyond overly cautious designs to reflect real-world operational variability. Even within a predominantly quantitative framework, LADRI acknowledges the value of qualitative evidence in supporting safety goals derived from quantitative assessments. This mixed approach ensures a robust and comprehensive risk assessment for ADS, highlighting the framework's flexibility in incorporating qualitative insights to support quantitative findings.

2.2.4 Dynamic Risk Assessment

In order to facilitate the safety assurance of autonomous systems, Trapp et al. [131] introduced the concept of context awareness, which allows a system to monitor and analyze the operational situation from a safety perspective. Here is a summary of key points regarding the performance and necessity of Dynamic Risk Assessment (DRA):

– Integration of Perception Information and HARA Models: DRA leverages real-time perception information about the system's environment and operational context, in conjunction with HARA models that are accessible at runtime. This integration enables the system to evaluate current risks based on the actual operational situation, moving away from static worst-case assumptions.
– Operational Situation Impact on Risk: The assessment takes into account hazardous events, which are combinations of potential hazards and specific operational situations. The risk associated with a hazardous event is significantly influenced by the operational context. The HARA process assesses the exposure (likelihood of encountering a particular operational situation), controllability (the ability to mitigate a hazardous event), and severity (expected harm) of these events.
– Dynamic HARA at Runtime: Implementing HARA at runtime offers a balance between flexibility and assurability. It allows the system to become aware of its current context, using parameters monitored in real-time such as speed, weather conditions, and traffic context. This context awareness enables the system to dynamically assess risks. By describing situations as vectors of characteristic parameters and using continuous functions to determine these parameters, the system achieves a high degree of flexibility in DRA.

Additionally, DRA can be part of a broader dynamic risk management framework that enables systems to assess and manage risks in real-time, considering the actual operational context and internal state (as explained in [112] and shown in Fig. 4.16). This approach marks a shift from static, worst-case-based safety assessments to a more flexible, context-aware strategy capable of adapting to the complexities and uncertainties inherent in dynamic operational environments.

2.2.5 Risk Features

Risk features, criticality metrics, and risk indicators, though related and often used interchangeably, have distinct focuses and applications within risk assessment. Understanding these differences is crucial for precise communication and effective implementation of risk assessment methodologies.

Risk Features: are identifiable characteristics or factors that can influence the likelihood or impact of a risk. In the context of ADS, risk features might include vehicle speed, weather conditions, traffic density, or the proximity of objects. The identification of risk features is a foundational step in risk assessment, serving to characterize the operational environment and potential hazards. Risk features provide the raw data necessary for further analysis and are often used as input variables in models predicting risk metrics or assessing criticality.

Criticality Metrics: are quantitative measures that evaluate the urgency or severity of a potential hazard or situation. They are often used to prioritize risks based on their immediate importance or potential impact (e.g. Time to Collision (TTC), Deceleration Rate to Avoid Crash (DRAC), etc). Criticality metrics focus specifically on the aspect of urgency and severity, providing a means to assess how critical a situation is for immediate attention or response. These metrics are crucial for decision-making processes, especially in dynamic systems like ADS, where prioritizing responses to potential hazards is essential for safety [141]. However, criticality metrics do not necessarily provide information about risk or the level of risk [68].

Risk Indicators: are quantitatively assess the potential impact of risks (i.e. level of risk), often combining multiple risk features to provide an overall risk evaluation. Risk indicators in ADS might encompass collision probability, potential injury severity, or likelihood of loss (e.g. severity, exposure, and controllability). Risk indicators offer a comprehensive assessment of risk, integrating various dimensions of risk features and criticality metrics to estimate overall risk levels. They are instrumental in broader risk assessment to classify the critical and non-critical situations.

Although all three concepts contribute to understanding and managing risks, they serve different purposes in the risk assessment process. Risk features identify factors contributing to risk; criticality metrics evaluate the significance of those factors in specific scenarios; and risk indicators provide an overall assessment of risk likelihood and impact.

The distinction lies in the scope and application of each term. Identifying risk features is about recognizing potential sources of risk. Assessing criticality is about understanding the immediate implications of those risks. Calculating risk indicators is about evaluating the overall level of risk. This tiered approach enables detailed analysis of specific hazards and a comprehensive overview of risk levels, facilitating targeted and effective risk assessment strategies.

Connection to LADRI: In this thesis, the risk assessment approach is refined by converting risk features into criticality metrics, then incorporating them as key risk features within the methodology. This transformation facilitates the calculation of risk indicators, specifically severity and controllability ratings, tailored to the operational dynamics of ADS. By adopting this refined classification, these risk features become instrumental in runtime risk assessment, enabling a proactive and dynamic evaluation of potential hazards. The categorization of these risk features falls into three distinct but interconnected domains: Vehicle Dynamics, Temporal, and Environmental [1].

Table 2.1: Summary of vehicle Dynamics Risk Features

Risk Features	Impact On Runtime Risk Assessment
Required Lateral Acceleration [62]	Measures the lateral force necessary to maintain trajectory in a curve. Critical for evaluating vehicle stability and handling.
Required Longitudinal Acceleration [62]	Indicates the needed acceleration or deceleration to match traffic flow or to stop. Affects vehicle's ability to adapt speed based on dynamic conditions.
Deceleration Rate to Avoid Crash [9]	Measures vehicle's ability to prevent crashes, affecting controllability. Evaluates the vehicle's capability to safely decelerate under emergency conditions.
Lateral Jerk [38]	Sudden changes in lateral acceleration. Important for assessing comfort and control during maneuvers.
Longitudinal Jerk [40]	Indicates abrupt changes in acceleration or deceleration. Affects both risk of rear-end collisions and passenger comfort.
Change in Velocity [82]	Reflects impact severity, informing on potential injuries or damage. Assists in understanding the potential severity of an accident.

Vehicle Dynamics Risk Features: This category encompasses features that directly relate to the physical capabilities and limitations of the vehicle itself, as shown in Table. 2.1. By focusing on fundamental aspects of vehicle behavior such as lateral and longitudinal, and the ability to decelerate effectively, this category assesses the intrinsic risk features of the vehicle. These features are crucial for understanding how well a vehicle can respond to and avoid potential hazards, underpinning the basic premise of ADS safety by evaluating the vehicle's stability, handling, and crash prevention capabilities.

Temporal Risk Features: This category prioritize the dimension of time in assessing risk, focusing on the critical time windows available for preventative actions against potential hazards (as shown in Table. 2.2). Metrics such as time to collision, provide insights into the immediacy of threat and the available response time, respectively. This category highlights the importance of timing in executing maneuvers and the adequacy of response

[1] While the original literature identifies over 30 criticality metrics [141], this thesis focuses on a selected few to underscore the risk assessment perspective and combined them into categories.

Table 2.2: Summary of Temporal Risk Features

Risk Features	Impact On Runtime Risk Assessment
Time To Collision [128]	Indicates imminent collision risk, crucial for severity assessment. Helps in predicting the urgency of avoiding actions.
Time To Brake [55]	Assesses the vehicle's response window for avoiding obstacles. Indicates the critical time window for initiating braking to prevent a collision.
Time To Maneuver [134]	Determines flexibility in maneuvering to mitigate risks. Measures the adaptability of the vehicle in responding to unforeseen obstacles.
Time To Steer [56]	Assesses the immediate time available for steering to prevent collisions and gauges the maneuverability to reduce risks.

time in avoiding collisions. By quantifying the urgency of situations, temporal features offer a direct measure of the severity of risk and the necessity for timely interventions, enhancing the predictive and responsive capabilities of ADS.

Table 2.3: Summary of Environmental Risk Features

Risk Features	Impact On Runtime Risk Assessment
Conflict Index [7]	Evaluates interaction risks, important for dynamic environments. Helps in predicting potential conflict points between multiple agents, facilitating proactive safety measures.
Time To React [128]	Quantifies crash likelihood in multi-agent scenarios, impacting severity. Assesses the time window for reaction to prevent crashes in complex environments involving multiple agents.
Responsibility Sensitive Safety-Dangerous Situation [63]	Defines situations that violate safety models, indicating an imminent threat. Key for identifying scenarios where intervention is necessary to maintain safety margins.

Environmental Risk Features: This category account for the complex interactions between the vehicle and its surroundings, including other vehicles, pedestrians, and infrastructure. This category recognizes that the operational environment of ADS extends beyond the vehicle itself to include dynamic and potentially unpredictable elements, as shown in Table. 2.3. Metrics such as critical index, assess how external factors influence risk levels, emphasizing the need for ADS to adapt to changing conditions and interactions. The inclusion of the responsibility sensitive safety-dangerous situation further underscores the system's need to identify and navigate scenarios that could compromise safety. This category bolsters the holistic assessment of ADS risk by incorporating the multifaceted nature of dynamic driving environments.

2.2.6 Risk Indicators

Risk indicators play a crucial role in assessing the potential impact of risks in dynamic environments, combining various factors to provide a comprehensive risk evaluation [53, 66, 83, 106]. These indicators, derived from risk models using available data, can be categorized into leading and lagging indicators, each serving a distinct purpose in risk assessment.

Leading indicators actively monitor key events or activities essential for achieving safety outcomes, identifying early deviations that could lead to negative consequences. For instance, in a highway lane-following scenario, indicators such as unusual deceleration patterns of lead vehicles may signal potential emergency braking events, enabling proactive risk control measures.

Conversely, lagging indicators focus on reactive monitoring, involving the reporting and investigation of incidents to pinpoint system weaknesses. They indicate when safety outcomes have not been met, providing essential feedback for corrective actions. A balanced risk assessment includes both leading and lagging indicators, offering a forward-looking perspective while considering historical performance data [81].

Landucci et al. in [81] further classify risk indicators into three types, emphasizing the monitoring of technical, human, and organizational factors. Retrospective indicators draw from historical incident data to evaluate safety performance over time, whereas predictive indicators use models to forecast future risks. Aggregated indicators compile expert judgments, accident analyses, and risk modeling, with a distinction between general aggregation and those specifically aimed at proactive risk assessment.

Integrating leading and lagging indicators within the LADRI framework creates a DRA strategy, combining the predictive power of leading indicators with the empirical insights of lagging indicators. This dual approach allows for a nuanced understanding of risk, combining the immediate assessment of potential hazards with a reflection on past safety performance. However, it is essential to recognize the limitations of both types of indicators, such as the potential for false positives in leading indicators and the reliance on historical data in lagging indicators, which may not always accurately predict future scenarios in rapidly changing ADS operational environment.

Connection to LADRI: In this thesis, severity and controllability indicators are adopted as key risk indicators within the LADRI framework, positioned to leverage the distinctions between leading and lagging indicators for a nuanced approach to risk assessment in ADS. Severity ratings, which gauge the potential impact of a risk event, align with the anticipatory nature of leading indicators. They enable the identification of conditions that may escalate into more severe outcomes, allowing for proactive risk mitigation strategies. Conversely, controllability ratings assess the ADS's ability to manage a risk event, reflecting the retrospective essence of lagging indicators by evaluating past incidents.

By incorporating severity and controllability indicators, this thesis establishes a comprehensive risk assessment methodology that dynamically integrates both predictive and reflective risk indicators. This integration ensures that the LADRI framework can not only forecast potential risks through severity indicators but also draw on historical data through controllability indicators to enhance system robustness and resilience. Thus, the classification of severity and controllability indicators under leading and lagging indicators respectively, enriches the risk assessment capability, allowing for a more adaptable and informed approach.

2.2.7 Risk Assessment Approach: Simulation and Scenario-Driven

The challenges inherent in risk assessment approaches necessitate exploring diverse methodologies to forge a pathway toward a comprehensive understanding of risks, delving into the strengths of simulation-based analyses and scenario-based evaluations. This section aims to articulate the distinctive aspects of each approach while setting the stage for their integration within the LADRI framework.

Simulation-Driven Risk Assessment: Simulation-driven risk assessment represents a cornerstone in the evolution of ADS and cooperative ADS. Given the intricate and dynamic nature of these technologies, a robust and comprehensive approach is indispensable for validating their safety and effectiveness. Recent advancements underscore the importance of this method, shedding light on various dimensions that enhance the development and validation processes [6, 31, 51, 67, 96, 107].

A pivotal development in this field is the creation of simulation toolchains, equipped to identify and evaluate critical scenarios that cooperative ADS might encounter in real-world operations [49]. Traditional validation methods, constrained by the impracticality of extensive real-traffic testing, are complemented by the innovation of digital twins. These advanced digital prototypes replicate real ADS with high fidelity, allowing for precise testing and refinement of automated driving functionalities.

Furthermore, the adaptability of simulation toolchains through exchangeable driving functions, evaluation metrics, and parameter spaces broadens their applicability across different scenarios. This versatility is instrumental in identifying critical behaviors, utilizing safety and traffic quality metrics to fine-tune ADS functionalities. The introduction of Prototype-in-the-Loop and other X-in-the-Loop methods bridges the gap between simulation and real-world testing. By incorporating actual ADS into traffic simulations on proving grounds, these approaches significantly enhance the realism and reliability of simulation results. This integration ensures a more authentic validation of simulation outcomes with real vehicle tests.

Emerging methodologies such as criticality assessment for ADS highlight the efficacy of simulation in safety approval processes. Employing risk features, this approach identifies and evaluates safety-critical scenarios based on their likelihood and potential consequences. The innovation of variable criticality thresholds further refines this process, allowing for a more nuanced assessment that accounts for the dynamic interactions within traffic[104].

Lastly, the SAHARA methodology exemplifies how simulation aids in hazard analysis, streamlining the assessment process through automation [116]. By generating simulation templates and integrating fault injection algorithms, SAHARA supports a systematic classification of scenarios, aligning with the ISO 26262 framework for functional safety engineering.

Connection to LADRI: Collectively, these advancements illustrate the transformative impact of simulation-based risk assessment in advancing the safety and functionality of ADS. The LADRI framework incorporates simulation-based risk assessment as a strategy to enhance the risk assessment process. By leveraging advanced simulation toolchains, LADRI aligns with recent advancements in the field, emphasizing the importance of identifying and evaluating critical scenarios that ADS might encounter. This approach mirrors the development and utilization of digital twins and Prototype-in-the-Loop methodologies, enabling the precise testing and refinement of ADS functionalities within a controlled, simulated environment. These features allow for a nuanced examination of safety-critical scenarios, accounting for the dynamic interactions within traffic systems.

Furthermore, by integrating risk features and diverse scenario generation methods, LADRI efficiently identifies potential hazards, enhancing its ability to predict in runtime. By embodying these simulation-based advancements, LADRI advances the validation process of ADS, aligning with the ISO 26262 framework for functional safety engineering, and underscoring the transformative potential of simulation in addressing the complex safety challenges of ADS.

Scenario-Driven Risk Assessment: The evolution of risk assessment methodologies, driven by recent research, marks a significant shift towards dynamic, scenario-based approaches [5, 24, 25, 73, 85, 93, 101, 140, 146]. Central to this evolution is the integration of Systems Theoretic Process Analysis within scenario-driven testing frameworks, illustrating a move away from traditional risk evaluation strategies and emphasizing the replication of real-world conditions for effective hazard identification and ADS safety evaluation [72].

The development of an optimization algorithm based on vehicle risk assessment aimed at improving autonomous driving capabilities, particularly in lane-keeping and collision avoidance scenarios, adds depth to this evolution [33]. By incorporating mathematical modeling and simulation trials, the algorithm ensures vehicle safety and stability by dynamically adjusting steering and braking based on environmental conditions.

Further complexity is introduced in scenario-driven study that merges traditional hazard analysis techniques with innovative enhancements tailored for ADS [79]. This approach offers an integrated safety assessment framework, combining scenario identification and risk quantification, to facilitate a thorough evaluation of potential risks and implement necessary safety measures.

The convergence of a manageable set of scenario classes from which relevant test cases can be derived is highlighted by structured methodologies for verifying and validating ADS [105]. The structured process involves scenario elicitation, requirement gathering, test derivation, execution, and evaluation, showcasing a detailed approach to safety validation.

This evolution culminates with the proposal of an adaptive testing scenario library generation method for ADS, addressing performance gaps between surrogate models and actual vehicles [34]. Utilizing Bayesian optimization to refine testing scenario libraries, this method accelerates the evaluation process and increases precision. By dynamically updating the testing library to better match ADS performance, this innovative approach reduces the occurrence of suboptimal testing scenarios, marking a pivotal advancement in the field.

Connection to LADRI: Collectively, these contributions weave a coherent theme, the imperative transition towards scenario-based, dynamic analyses that more accurately reflect the intricacies of automated driving environments. Despite each methodology offering a distinct approach, including systems theory integration, optimization algorithms, and adaptive testing frameworks, they all agree that scenario-driven approaches are crucial for improving the risk assessment process of ADS.

The LADRI framework marks a shift towards dynamic, scenario-based risk assessment approach, reflecting current research advancements in this area. By employing scenario-based analysis, LADRI utilizes adaptive testing principles to accurately replicate real-world conditions, thus improving HARA for ADS. This strategy is enhanced by optimization algorithms designed to modify vehicle behavior for essential functions like lane-following, considering the real-time environmental conditions and outcomes of simulation trials. The framework focuses on identifying scenarios and quantifying risks while continually refining testing scenarios using a vehicle dynamics mathematical model. This ensures a thorough and nuanced risk assessment framework that adapts dynamically, accurately capturing the complex interactions characteristic of automated driving operational environments.

2.2.8 Supervised Learning Algorithms: Applications in Risk Assessment

Artificial Intelligence (AI), a subset of computer science, is dedicated to enabling machines to learn from data and make predictions or decisions without being explicitly programmed. ML, a crucial component of AI, allows machines to learn from data through three types of algorithms: supervised, unsupervised, and reinforcement learning, each tailored for specific data types and problems (as shown in Fig. 2.1). Among the supervised ML algorithms, notable ones include Artificial Neural Networkss (ANNs), Support Vector Machine (SVM), Decision Trees, Logistic Regression, Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Naïve Bayes, k-Nearest Neighbors, and Linear Regression. This thesis primarily focuses on four techniques[2]: SVM, ANN, RF, and GBDT.

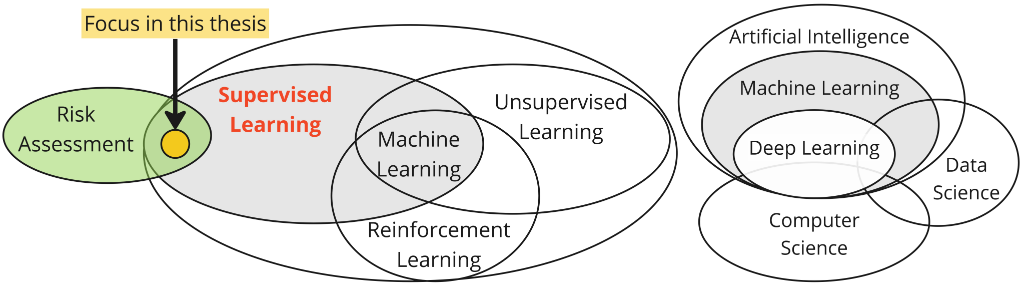


Figure 2.1: Venn Diagram Representing the Components of AI and ML with a Focus on Supervised Learning for Data-Driven Risk Assessment. (Right: Adapted from [46])

Hegde et al. have noted that ADS are equipped with a wide array of sensors generating data that can be processed using ML algorithms for risk assessment [52]. They also highlighted that while IEC 31010 [42] outlines approximately thirty different risk assessment techniques, their capability for performing runtime risk assessment remains limited. Thus, engineering risk assessment could greatly benefit from the adoption of ML algorithms. A comprehensive review addressed research questions such as "Which ML algorithm is adopted the most? Which ML algorithm has been implemented and verified as suitable for risk assessment? What kind of data is used to develop ML algorithms for risk assessment?" Therefore, this section aims to explore the implementation of the four selected models in risk assessment tasks across various aspects, highlighting their potential to enhance engineering risk assessment.

Support Vector Machines: Support Vector Machines (SVM) are supervised learning models extensively used for classification, regression, and outlier detection. Their significance in risk assessment is largely due to their ability to model complex, non-linear relationships and maintain accuracy

[2] The rationale behind selecting these four models is presented in Table. 3.6

with high-dimensional data, making them ideal for identifying potential risks [61, 74, 80, 86, 88, 91, 109, 135]. SVM combats overfitting and ensures robustness, essential in risk scenarios with limited data or diverse inputs.

One application of SVM in risk assessment includes its use in an Adaptive Cruise Control (ACC) system to evaluate collision risks and adjust vehicle distances, utilizing a Radial Basis Function (RBF) kernel for multiclass classification [144]. This study used traffic flow data from over 6,000 vehicles in Germany, focusing on lane-keeping and changing behaviors. The system, which achieved a 99.2% F1 score and 0.065m average root mean squared error in performance metrics, showed a notable reduction in collision risks compared to human drivers and traditional ACC systems.

Another study applied SVM to analyze driver injury severity in rollover crashes, employing polynomial and gaussian RBF kernels [20]. With data from 3,158 vehicle/driver records in New Mexico, this approach highlighted variables like seatbelt use and vehicle damage as significant. Polynomial kernels showed superior classification accuracy, suggesting SVM's utility in traffic safety research despite the challenges of overfitting and model applicability for severe injuries.

SVM was also used to predict driver drowsiness using eyelid-related parameters from EOG data, achieving high detection accuracy, particularly for severe drowsiness [58]. This study highlights SVM's potential in reducing accidents caused by drowsiness, but applying these findings to real-world driving conditions still presents challenges.

These diverse applications of SVM, from enhancing driving safety and analyzing crash injury severity to predicting driver drowsiness, illustrate the adaptability of SVM algorithms to various challenges in risk assessment. They highlight the need for further research to improve data comprehensiveness and model robustness. As SVM continues to improve, its potential for developing reliable risk mitigation strategies becomes increasingly clear, offering advancements in risk assessment across various domains.

Artificial Neural Network: Artificial Neural Networks (ANN) are inspired by biological neural networks and consist of interconnected units or nodes that process data to learn tasks without being explicitly programmed. ANNs are crucial in ML for modeling complex patterns in vast data sets, applicable in areas ranging from road safety to data mining related to traffic patterns [19, 121, 127].

One study focused on enhancing road safety through risk-aware route planning with a hybrid ANN model [87]. It clustered data using a fuzzy C-means algorithm, training separate networks for each cluster to predict road risk indices for safer route planning. The hybrid ANN showed superior performance in prediction accuracy over traditional models.

Another study applied ANN to improve autonomous vehicles' safety, focusing on lane detection and behavior prediction. It employed Convolutional Neural Networks, demonstrating significant advancements in lane detection and behavior prediction, thus contributing to vehicular safety [26]. A different study utilized Recursive Neural Networks for risk assessment at road intersections, employing a non-linear model and information encoding to enhance safety [22]. It showed promising results in recognizing unforeseen patterns, suggesting future applications in real-world scenarios. Additionally, ANNs were integrated with probabilistic analysis for dynamic failure assessment in chemical processes, predicting accident probabilities based on process variables [1]. This approach highlighted ANN's efficacy in managing complex industrial systems' dynamic failures.

These studies illustrate ANN's broad applicability, showing how its unique architectures and strategies tackle complex challenges across various fields. ANNs' ability to handle nonlinear relationships and vast datasets significantly enhances predictive accuracy, underscoring their importance in addressing evolving challenges across different domains.

Random Forest: Random Forest (RF), an ensemble learning method that builds multiple decision trees for training and produces either the mode of the classes (for classification) or the mean prediction (for regression) of the trees, is pivotal in risk assessment. Its ability to handle large datasets with high dimensionality and to provide robustness against overfitting makes it highly suitable for complex risk scenarios, particularly in traffic safety and autonomous driving.

A study examining road traffic accidents in the UK in 2020 utilized RF, alongside Decision Tree, LightGBM, and XGBoost, to select features from a dataset of 135,453 records [98]. Despite challenges like data imbalance, the study underscored the impact of vehicle characteristics and driver age on accident severity, suggesting directions for enhancing road safety. Another study in Zarqa City, Jordan, forecasted traffic crash severity using RF among other models [3]. With data on 97,900 accidents from 2014-2018, RF excelled in performance, indicating its effectiveness in predicting crash severity levels better than models like AdaBoost. A study from Gauteng, South Africa, aimed to model road traffic accidents using RF, which, after optimization, showed superior prediction capability over other models [18]. The study faced challenges with dataset scope but suggested RF's potential in transportation safety.

Assessing collision risks for ADS in dynamic environments, another study utilized RF to analyze autonomous driving activity scenes [103]. Despite challenges in predicting human-driven vehicle behavior, significant predictive accuracy was noted, especially with features like relative distance and time to collision. Finally, leveraging vehicle-to-vehicle communication data for traffic accident detection, RF demonstrated a 92% accuracy and 94% sensitivity in a traffic simulation environment, outperforming SVM and ANN [28]. This study points to RF's potential in real-time traffic management, despite challenges in data processing.

Collectively, these studies highlight RF's versatility and effectiveness across various risk assessment contexts. Its strength in managing imbalanced data, resisting noise, and providing superior predictive accuracy emphasizes its potential in enhancing traffic safety with efficient feature selection and accurate risk predictions.

Gradient Boost Decision Tree: Gradient Boosted Decision Trees (GBDT) excel as an ensemble learning technique, merging multiple decision trees to refine predictions and manage data inconsistencies and complexities. Its versatility across data types (numerical, categorical), ability to model nonlinear relationships without extensive data preprocessing, and reduction of overfitting through gradient boosting make it a standout method in risk assessment. GBDT addresses challenges such as data heterogeneity, imbalanced datasets, and the necessity for precise prediction accuracy.

A study investigating risky driving behaviors with GBDT focused on driver-dependent vehicle features and achieved 100% accuracy on selected top features, highlighting GBDT's potential for enhancing driving safety [48]. Another research aimed at predicting the severity of single-vehicle crashes with GBDT, analyzing 33,327 crash records from California [145].

Similar to this, another framework assessing driving behavior and predicting risk levels utilized GBDT, particularly XGBoost, extracting around 1300 driving behavior features from vehicle trajectory data [124]. With an overall accuracy of 89% for predicting risk levels, the study demonstrated GBDT's efficacy in risk assessment, despite challenges like class imbalance. Furthermore, a study analyzing Maryland State Police crash data from 2015 to 2019 with GBDT highlighted its superior accuracy in crash severity prediction, outperforming other models and suggesting future research directions for broadening predictive capabilities by incorporating driver behavior data [32].

These studies collectively underscore GBDT's high performance, adaptability, and robustness in risk assessment scenarios, showcasing its effectiveness in handling complex relationships and diverse datasets.

Connection to LADRI:The LADRI framework integrates a comprehensive approach to risk assessment, utilizing the strengths of diverse ML models such as SVM, ANN, RF, and GBDT, to meet the specific demands of learning-based risk assessment. SVM is selected for its ability to define precise boundaries in high-dimensional spaces, essential for nuanced risk identification. ANN is chosen for its ability to model the complex dynamics and interactions within automated driving scenarios, owing to its flexibility and predictive accuracy. RF is valued for its versatility in managing imbalanced data and providing accurate risk predictions, making it suitable for dynamic environments. GBDT is included for its adaptability and effectiveness in refining risk assessments through iterative improvements. Collectively, the integration of these models within the LADRI framework leverages their unique strengths to address the complex challenges of ADS risk assessment effectively.

2.3 Related Work

This section reviews related work in the field of risk assessment, highlighting significant advancements across diverse methodologies. It begins with an exploration of advanced risk assessment approaches, focusing on structural and situation-aware improvements (Section 2.3.1), followed by discussions on iterative hazard analysis and function refinement (Section 2.3.2). The section further delves into dynamic risk ratings via deep learning (Section 2.3.3), integrated collision risk assessment strategies (Section 2.3.4), the implementation of the KnowGo framework for adaptive learning-based risk assessment (Section 2.3.5), and concludes with insights into scenario-based collision detection using ML (Section 2.3.6). This overview frames the context for understanding the current state and directions of risk assessment research and practice.

2.3.1 Advancing Risk Assessment: Structured and Situation-Aware Approach

The Structured Approach for Hazard Analysis and Risk Assessment (SAHARA) framework aims to address existing methodological shortcomings in automotive systems [70]. The approach provides a model-based representation. It includes components like GOBI (Gradation of Baneful Influence) to formalize understanding of HARA. OASIS (Ontology-based Analysis of Situation Influences on Safety) is used to formalize operational situations and their safety impacts. HEAT (Handy Exposure Assessment Technique) is employed for modeling and assessing situation exposure. Utilizing Parnas' Four-Variable Model, SAHARA extends interactions to include system, human, and environmental factors and employs the ARID (Analysis of Risk through In-system Degradation) algorithm to manage the complexities of multiple service failures. The framework's development and validation involved over 300 operational situations from 10 industrial case studies, enhancing its ontology and models. SAHARA's effectiveness is demonstrated through feasibility studies, especially ARID's capability for automatic hazard analysis, which ensures thorough coverage of critical system parts. This implementation highlights SAHARA's advantages in improving consistency, efficiency, and correctness in HARA processes over traditional methods, significantly reducing reliance on expert judgment. Its efficiency in managing multiple service failures marks SAHARA's distinctiveness. The framework's comprehensive, automated approach offers a significant leap forward in handling complex scenarios with multiple service failures, distinguishing it from conventional manual HARA methods.

However, the study did mentions some criticisms and limitations of SAHARA, particularly regarding the algorithm's complexity and the extensive domain knowledge required for its effective application. It highlights the challenge of integrating SAHARA into existing safety processes due to its novel approach and the potential resistance from industry practitioners accustomed to traditional methods. Additionally, while SAHARA aims to reduce dependency on expert judgment, the initial setup and customization for specific automotive contexts still demand significant expertise, which could limit its accessibility and scalability for broader adoption without further simplification and automation enhancements.

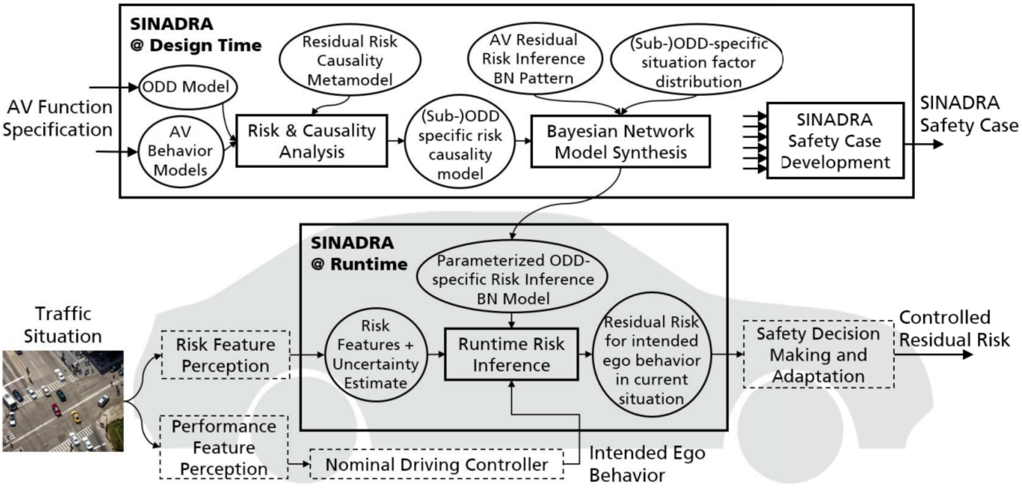

SINADRA @ Design Time
AV Function Specification
ODD Model
AV Behavior Models
Residual Risk Causality Metamodel
Risk & Causality Analysis
(Sub-)ODD specific risk causality model
AV Residual Risk Inference BN Pattern
(Sub-)ODD-specific situation factor distribution
Bayesian Network Model Synthesis
SINADRA Safety Case Development
SINADRA Safety Case
SINADRA @ Runtime
Parameterized ODD-specific Risk Inference BN Model
Traffic Situation
Risk Feature Perception
Risk Features + Uncertainty Estimate
Runtime Risk Inference
Residual Risk for intended ego behavior in current situation
Safety Decision Making and Adaptation
Controlled Residual Risk
Performance Feature Perception
Nominal Driving Controller
Intended Ego Behavior

Figure 2.2: Situation-aware Dynamic Risk Assessment Framework Overview [117]

The SItuatioN-Aware Dynamic Risk Assessment (SINADRA) framework, on the other hand, addresses the challenge of assuring safety for Autonomous Vehicles (AVs) by shifting traditional risk assessment from design time to runtime, employing a model-based approach to mimic human-like risk reasoning under uncertainty [117]. This approach enhances safety assurance in AVs by enabling DRA, allowing AVs to adapt to changing conditions instead of relying on predefined worst-case scenarios. Utilizing a Bayesian network synthesis and assurance process, SINADRA integrates tactical situational knowledge for probabilistic runtime risk monitoring within an adaptive safety management system. Key components of SINADRA include situation class detection, Bayesian network-based behavior intent prediction, trajectory distribution generation, and risk assessment, as demonstrated by implementation in the CARLA open-source simulator [29]. SINADRA employs probabilistic environmental knowledge and datasets specific to the operational design domain for learning Bayesian networks, as shown in Fig. 2.2.

The SINADRA framework tackles the challenge of dynamically controlling driving functions based on residual risk assessments. It utilizes a comprehensive collision risk framework that incorporates probabilistic risk metrics. This approach enhances situational awareness and improves the accuracy of behavior intent predictions, demonstrating superior risk assessment capabilities over simpler DRA models [118]. The ongoing development of SINADRA focuses on quantitatively evaluating its performance gains compared to traditional DRA methods, tackling challenges like the complexity of real-time risk assessment under uncertainty. Systematic design-time methods and runtime risk inference modeling are employed to manage the integration of comprehensive environmental knowledge effectively.

However, SINADRA faces limitations in real-time data processing that are crucial for immediate decision-making in dynamic driving scenarios. Its current scope may not adequately capture complex interactions at intersections or predict pedestrian intents with high fidelity. Future improvements are suggested to extend SINADRA's applicability to a broader range of scenarios and to conduct sensitivity analyses on environmental factors influencing collision risk, which are vital for enhancing the framework's sensitivity and overall effectiveness in diverse traffic conditions.

Both methodologies leverage concepts such as Bayesian networks, ontological analysis, and situation exposure modeling to address the challenges of DRA in complex, uncertain environments. By aiming to overcome the limitations of traditional methods, both frameworks propose advanced solutions for runtime adaptation and systematic assessment of risks in automotive systems.

2.3.2 Iterative Hazard Analysis and Function Refinement

The proposed risk assessment method addresses the challenge of ensuring safety of ADS by identifying and quantifying hazardous scenarios, enhancing both functional safety and safety of the intended functionality [79]. The method integrates established HARA from the automotive domain, with enhancements for automated driving. It employs scenario-based identification with a keyword-based HAZOP approach and causal chain analysis through extended FTA. Data characteristics were not explicitly detailed in the sections reviewed but involve the use of real-world data, expert judgment, and scenario modeling for hazard identification and quantification. Risk is assessed by combining the probability of occurrence and potential severity of identified hazards, with an emphasis on reducing risk to tolerable levels through iterative refinement and risk mitigating measures. The method, tested on the PEGASUS Highway-Chauffeur project, demonstrated its utility for identifying and quantifying risks in automated driving, indicating its effectiveness in supporting development and risk assessment processes. Suggestions include further refinement of the method to enhance scenario specification and combining expert analysis with data-driven approaches for comprehensive risk assessments in more complex environments, such as urban areas.

The limitations related to the complexity and dynamic nature of automated driving environments present challenges for exhaustively identifying and accurately quantifying hazardous scenarios. It emphasizes the difficulty in creating comprehensive models that accurately represent all potential real-world conditions and interactions. Further, it highlights the reliance on expert opinion and data availability, which may introduce subjectivity and limitations in risk assessment accuracy. These aspects underline the need for ongoing refinement of methodologies and the incorporation of more sophisticated, data-driven approaches to enhance reliability and comprehensiveness in risk assessments for ADS.

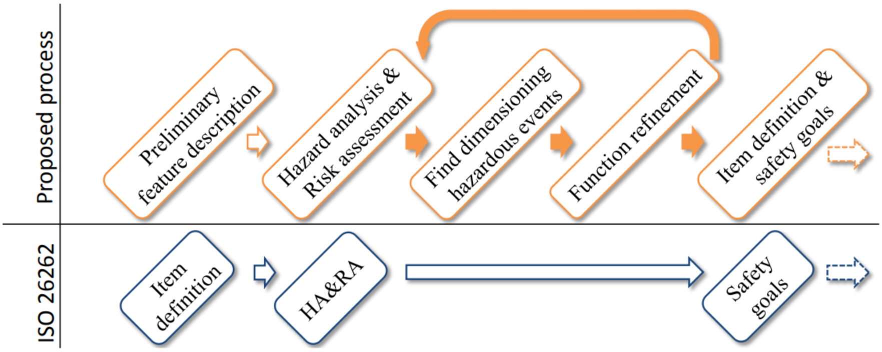

Proposed process
Preliminary feature description
Hazard analysis & Risk assessment
Find dimensioning hazardous events
Function refinement
Item definition & safety goals
ISO 26262
Item definition
HA&RA
Safety goals

Figure 2.3: Proposed hazard analysis and function refinement process compared to ISO 26262 [137]

Building on the iterative process previously discussed, a method proposed in [137] aims to enhance the safety of ADS through a structured, iterative hazard analysis and function refinement process. This method, designed to ensure comprehensive identification and mitigation of hazards in ADS operations, emphasizes the dynamic nature of automotive risk assessment. Unlike traditional approaches, this method considers the item definition as an outcome of the HARA, rather than an input. This adjustment allows for a more dynamic and thorough exploration of safety requirements, as depicted in Fig. 2.3. The iterative hazard analysis process is configured with steps including preliminary feature description, systematic hazard analysis, risk assessment, and function refinement, culminating in the definition of safety goals and further refinement based on identified hazardous events. Challenges arise from managing the complexity of autonomous functions and ensuring the completeness of safety goals. The method provides a structured approach to improving safety but faces limitations in foreseeing and accurately simulating all potential real-world scenarios. Future work suggests developing tool support to automate and refine the process further, thereby enhancing its efficiency and applicability across a wider range of autonomous vehicle functions.

2.3.3 Dynamic Risk Ratings using Deep Learning

A study highlighted in [36] introduces a deep learning-based method for DRA in highly automated vehicles, utilizing images from front stereo cameras. This method contrasts with traditional automotive approaches by harnessing deep learning's potential for enhancing automotive safety and situational awareness. The methodology involves creating a dataset from simulated driving scenarios, annotated with a straightforward, objective risk metric, and employing a Convolutional Neural Network for risk prediction from images. With a 72.87% accuracy in risk assessment, this approach shows promise as a supplement to traditional systems, but it has limitations for use in safety-critical applications. A major reason for the network's sub-optimal performance is identified as the discrepancy between the camera's view and the area considered by the risk metric calculator in the risk assessment. Critical situations occurring outside the camera's view could not be captured or assessed accurately, leading to a mismatch between the predicted risk and the actual risk assessed by the risk metric calculator. This limitation highlights the difficulty in meeting the stringent reliability and accuracy requirements for automotive safety applications, where errors can have severe consequences.

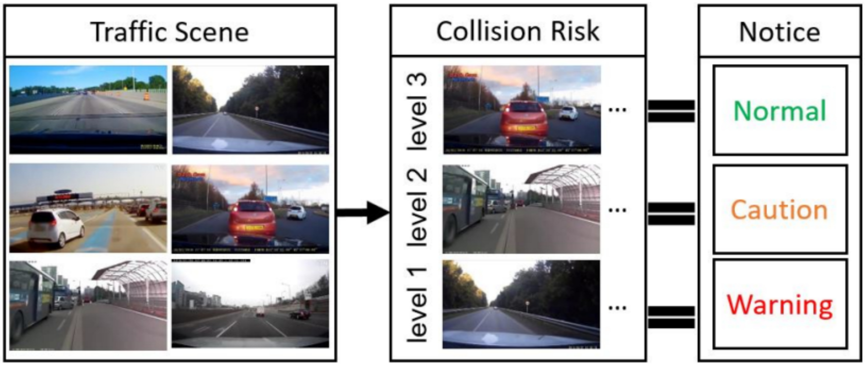


Figure 2.4: Risk Ratings Estimation of Traffic Scene using Learning-based Approach [136]

Expanding on the concept of DRA, [136] introduces a collision risk rating system designed to assess the probability of collisions using dashboard camera videos, depicted in Fig. 2.4. This system leverages two-stream Convolutional Neural Networks and rank learning for predicting collision risk levels, supported by the creation of a novel traffic collision dataset. The findings highlight the advantages of integrating spatial and temporal features for accurate collision risk prediction, outperforming existing video classification methods. The study points towards future improvements, such as incorporating pedestrian detection and semantic segmentation, to enhance the collision risk rating's effectiveness. Further expansion of the dataset is also recommended for more refined risk rate definitions. This research enhances driving assistant systems and autonomous vehicle technologies by providing a comprehensive insight into collision risks in real-time traffic scenarios. However, the explicit rank prediction approach performed poorer than classification approaches for collision risk rating. This limitation arises from two factors. First, optimizing rank learning is inherently more complex than optimizing classifier learning. Second, ranking machines have a smaller model capacity compared to multi-class linear SVMs or fully connected layers in Convolutional Neural Networks. This limitation suggests a potential area for improvement in developing more effective ML models for collision risk assessment that can accurately capture the ordered nature of risk levels (as shown in Fig. 1.4).

2.3.4 Integrated Collision Risk Assessment

For integrated collision risk assessment, the study introduces dynamic bayesian network model that combines network-level collision prediction with vehicle-level risk assessment for autonomous vehicles, aiming to enhance real-time safety perception [69]. The dynamic bayesian network configuration includes a new layer for network-level collision risk (CRN), complementing traditional layers for vehicle-level collision risk (CRV), sensor measurements (Z), and vehicle kinematics (K), as shown in Fig. 2.5. The data used encompasses traffic simulations on a 4.52 km motorway section, yielding 7800 conflict events and 23400 non-conflict cases. Performance metrics focus on classification accuracy, recall, and specificity, with results indicating enhanced model performance in identifying collision-prone conditions. This approach offers a comprehensive risk assessment tool, applying across traffic safety and autonomous driving domains.

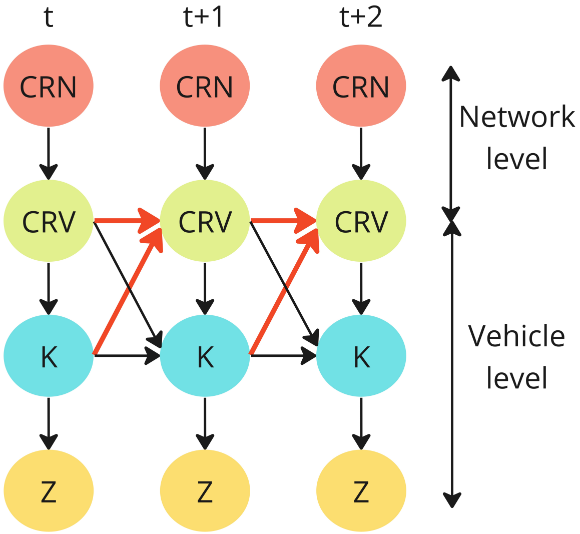


Figure 2.5: Integrated Risk Assessment Network [69]

The study acknowledges challenges in ensuring the accuracy and reliability of collision risk assessment models, particularly in dynamically changing environments like autonomous driving. One critical limitation is the model's dependency on accurate and comprehensive data for training and validation, which might not always encapsulate the complexity of real-world driving scenarios. Moreover, the balance between false positives and false negatives in risk prediction is highlighted as an area needing improvement to avoid unnecessary interventions or missed hazard detection. These limitations suggest a need for ongoing refinement of risk assessment methodologies, incorporating broader data sources and advanced computational techniques to enhance model performance and applicability in diverse driving conditions.

2.3.5 KnowGO: Adaptive Learning-based Dynamic Risk Assessment

The KnowGo score framework incorporates an architecture for DRA. It features a modular setup with components like risk scorers, a data ingestion module, a scorer selector, and a scorer aggregator, all designed to dynamically enact, manage, and fuse risk predictions from multiple models based on real-time data and vehicle states (as shown in Fig. 2.6). The framework employs both ML and rule-based methods to dynamically select and tune risk scoring models according to the changing conditions. The data for the study was generated using a simulation environment, provided a wide range of driving conditions, events, and automation levels, creating a comprehensive dataset for training and evaluating the risk scoring models implemented within the KnowGo framework.

The framework's effectiveness was evaluated through predictive accuracy and the ability of the auto-tuning feature to adapt the risk scoring based on changes in automation levels. The models showed varied accuracies, with non-ML models for night driving and weather conditions reaching 100% accuracy, while ML-based models such as linear regression for journey duration and a multi-model approach for driver alertness had lower accuracies. The study found that KnowGo could accurately predict automotive risks, with performance significantly influenced by the selection and tuning of risk scorers. Auto-tuning enhanced prediction accuracy across varying levels of automation, showcasing the adaptability of the framework. A mix of ML and non-ML models was used, ensuring high confidence in the system's overall risk assessments. This application is critical for enhancing safety and reliability as vehicles transition between different levels of automation.

The KnowGo framework, designed for DRA, inherently faces several limitations from a risk assessment perspective. First limitation is the variability and unpredictability of real-world scenarios pose significant challenges to ensuring consistent accuracy and reliability. Second, the effectiveness of KnowGo is heavily dependent on the quality and availability of input data. Inconsistent, heterogeneous, and volatile data, as often encountered in

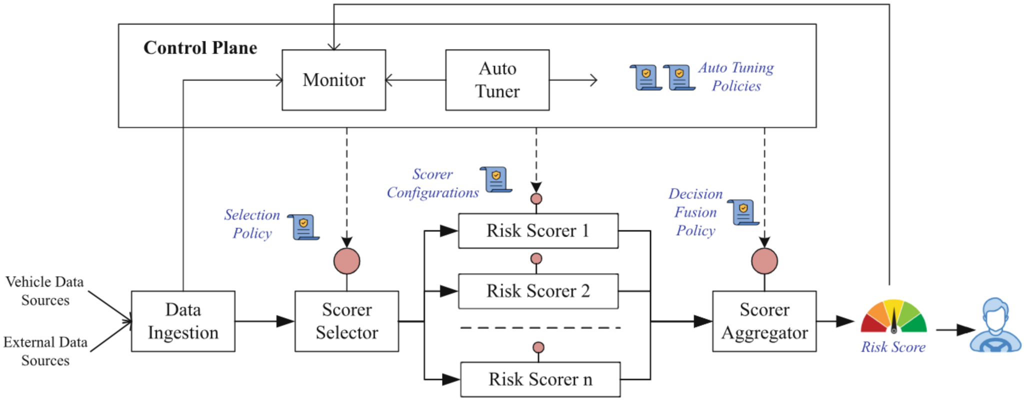


Figure 2.6: KnowGo: Dynamic Risk Assessment Framework [102]

dynamic driving environments, can significantly impact the system's ability to make accurate risk assessments. Third, the framework's capability to dynamically select, tune, and integrate multiple risk scoring models adds a layer of complexity, potentially impacting its operational efficiency and scalability. Managing the interaction between different models, especially in real-time, requires sophisticated coordination and optimization strategies to prevent latency and ensure timely risk assessments. Lastly, the use of multiple models and dynamic decision-making processes could hinder the transparency and explainability of risk assessments. For stakeholders, including drivers, manufacturers, and regulators, understanding the basis of risk scores and the rationale behind model selection and tuning decisions is crucial for trust and acceptance.

2.3.6 Scenario-based Collision Detection Using Machine Learning

The study aimed to enhance ADS safety by identifying potential collisions through scenario-based hazard analysis, as shown in Fig. 2.7. It utilized multilayer perceptron to detect collisions in the safety-related concept phase, thus contributing to reducing the number of scenarios needed for HARA. The multilayer perceptron configuration was chosen for its relevance to the datasets used for collision detection. The model was evaluated across various datasets to support safety argumentation. The multilayer perceptron model had three hidden layers, and experiments were performed to optimize its performance in terms of accuracy and loss. Two types of simulation-based scenario datasets were examined: knowledge-based and data-driven scenarios. The knowledge-based scenarios were derived from expert knowledge and parameter boundaries estimated using Monte Carlo Simulation. The data-driven scenarios were optimized through parameter-based analysis and sensitivity inspection. These datasets emphasized vehicle safety-related parameters like vehicle speed, distance between vehicles, and lane change duration. Performance evaluation was conducted using accuracy, loss, and cross-validation methods.

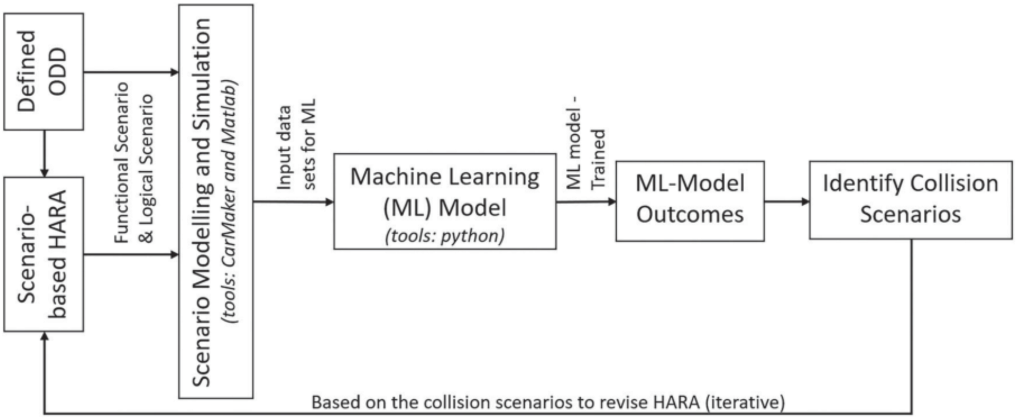

Defined ODD
Scenario-based HARA
Functional Scenario & Logical Scenario
Scenario Modelling and Simulation
(tools: CarMaker and Matlab)
Input data sets for ML
Machine Learning (ML) Model
(tools: python)
ML model - Trained
ML-Model Outcomes
Identify Collision Scenarios
Based on the collision scenarios to revise HARA (iterative)

Figure 2.7: Scenario-based Hazard Analysis using Machine Learning [73]

A systematic approach was used to collect, generate, and optimize input data, showcasing the model's high accuracy in detecting actual collisions. Notably, parameter-based optimized dataset, showed higher accuracy in model predictions. The study also highlighted the importance of quality input datasets, revealing that realistically optimized datasets could significantly enhance model prediction accuracy. The model performed better with the parameter-based optimized dataset than with the Monte Carlo simulation-based dataset.

The study recognizes ML's potential in improving ADS with scenario-based HARA but emphasizes the need for further research to effectively integrate ML techniques into hazard identification and risk assessment. It raises concerns about the assurance of ML models' safety, given their inherent unpredictability and the difficulty in validating their performance across all possible operational scenarios. The critical perspective here emphasizes the need for a careful, scrutinized approach to adopting ML in safety-critical applications, underscoring the balance between innovation and safety.

A significant challenge in applying ML for ADS is ensuring the completeness and quality of scenario data used for training ML models. The current state of scenario-based testing and data collection methodologies, indicating the scope and depth of scenarios might not be comprehensive enough to cover all potential hazardous situations. This limitation could affect the model's ability to accurately predict or detect collision scenarios, suggesting a gap in the data acquisition process that needs addressing.

The exponential increase in the number of scenarios for HARA poses a challenge for training ML models effectively. The study points out the difficulty in managing large datasets and the complexity of scenarios, which could hinder the ML model's performance and its application in real-world situations. This limitation underscores the need for advanced data management and model training strategies that can accommodate the vast and complex nature of driving scenarios.

2.4 Bridging the Gap: Need for a New Approach in Risk Assessment

The landscape of HARA in ADS development is replete with diverse methodologies, each with its unique approach to ensuring safety. Despite the advancements, the limitations highlighted in existing research underscore the necessity for a novel framework capable of overcoming these challenges. The LADRI framework emerges as a solution, integrating learning-based dynamic risk indicators (e.g. severity and controllability), iterative processes, and simulation-based scenarios to address the shortcomings of current practices. Through a comprehensive examination of related work, this thesis elucidate the imperative need for LADRI, drawing connections between its capabilities and the gaps identified in prior studies.

Addressing Complexity and Expertise Demand: Frameworks such as SAHARA have made significant strides in structuring HARA through model-based representations and formalized understandings of safety impacts. However, the complexity and extensive domain knowledge required for effective application limit their scalability and accessibility [70]. LADRI mitigates these challenges by incorporating advanced learning algorithms that adapt to dynamic driving conditions without the need for extensive manual configuration, thereby reducing the barrier to entry and reliance on expert judgment.

Dynamic Risk Assessment: The shift towards DRA, as explored in SINADRA and other dynamic HARA models [69, 117], highlights the critical need for frameworks that can adapt to changing conditions while driving in dynamic environment. LADRI's employment of learning-based dynamic risk indicators and its Plan-Do-Train-Adjust-Assess cyclic process provide a mechanism for continuous improvement and real-time hazard detection, surpassing the static nature of traditional approaches.

Incorporating Learning-based Methods for Enhanced Predictability: The reliance on scenario and simulation-based methods across multiple frameworks underscores the importance of accurate risk prediction under varied operational conditions. While works like those of Feth et al. [36] and Wang et al. [136] demonstrate the potential of learning-based methods in DRA, LADRI advances this approach by integrating ML with traditional HAZOP and iterative HARA processes. This integration allows for an objective assessment of severity and controllability, leveraging risk-specific context information to provide runtime contextual insights, a capability not fully explored in previous models.

Table 2.4: Summary of Related Work Vs LADRI

Related Work	Scenario-driven	Learning-based	Simulation-based	Use of HAZOP	Iterative HARA	Dynamic HARA
[70]	✓					✓
[117]	✓		✓			✓
[79]	✓			✓	✓	✓
[137]	✓				✓	
[36]	✓	✓	✓			✓
[136]	✓	✓				✓
[69]	✓		✓			✓
[102]	✓	✓	✓			✓
[73]	✓		✓			✓
LADRI	✓	✓	✓	✓	✓	✓

Iterative HARA: The iterative approach to hazard analysis and function refinement, as detailed in research by Kramer et al. [79] and Warg et al. [137], resonates with LADRI's cyclic process. However, LADRI distinguishes itself by embedding this process within a learning-based framework that not only identifies hazards but also evolves with each iteration, enhancing safety measures in subsequent cycles of ADS development.

Simulation-based Scenario Integration: Simulation-based scenarios play a pivotal role in risk assessment, offering a controlled environment for testing and validation. LADRI's use of simulation-based scenarios goes beyond testing to include training its models within these environments. This ensures that the framework's predictions and assessments are based on thorough and realistic driving conditions.

The comparative analysis provided in Table 2.4 illustrates LADRI's comprehensive inclusion of scenario-driven, learning-based, simulation-based approaches, use of HAZOP, iterative HARA, and dynamic HARA processes. This integration positions LADRI as a holistic framework that addresses the noted limitations of existing research, offering advancements in DRA and risk quantification. The development of LADRI is not just another HARA methodology; it is an essential evolution that fills existing gaps in ADS risk assessment. By leveraging the strengths of existing approaches and addressing their limitations, LADRI introduces a framework that enhances traditional HARA.

3 Development of a Learning-based Dynamic Risk Indicator

The focus of this chapter is to present a complete process of the development of a specialized tool known as the "LeArning-based Dynamic Risk Indicators (LADRI)." Initially, the chapter commences with an overview of the solution. Following this, the thesis delves into the meticulous steps involved in creating a simulation environment. The text then transitions to the core of the research, discussing the training and testing phases of the ML model. The chapter concludes with the model adjustment and deployment processes. The text demonstrates how the ML model, once trained and refined, is deployed back into the simulated environment. This completes the research cycle, showcasing an iterative process that enables continuous learning from the data.

Chapter Content

3.1 Solution Overview

This thesis introduces the LADRI framework, a novel approach aimed at enhancing continuous risk assessment and early hazard identification for ADS by leveraging runtime risk-specific context information. The LADRI framework utilizes a supervised ML model to provide a more dynamic and evidence-based tool for risk assessment. This approach aims to reduce subjectivity in risk ratings by enhancing the precision of assessments through the identification of hidden patterns and correlations associated with specific risks. The framework employs a cyclic process named Plan-Do-Train-Adjust-Assess (PDTAA), as depicted in Fig. 3.1, to dynamically update and refine risk indicators. Each phase of the PDTAA process contributes to the next, creating a comprehensive approach to risk assessment:

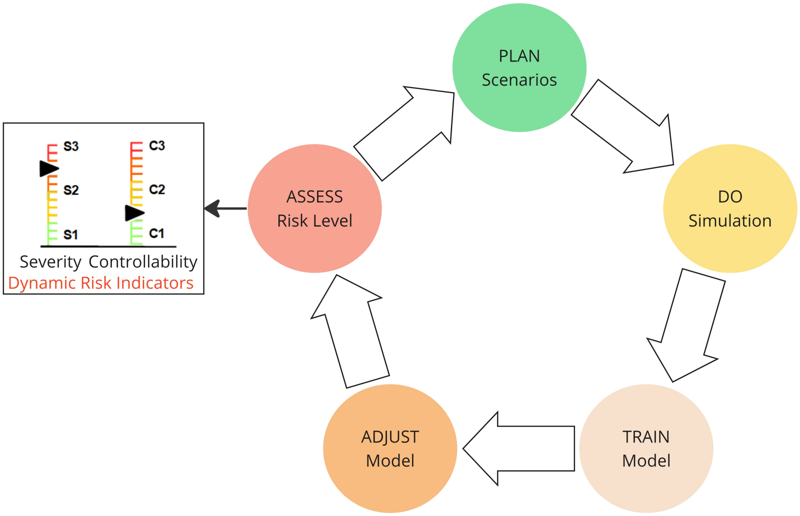


Figure 3.1: The Cyclic Process of the Learning-Based Dynamic Risk Indicator Tool

PLAN-Scenarios: In the Plan phase, the framework identifies three key categories of parameters: design parameters, non-design parameters, and failure conditions. Design parameters, which are directly controllable, include aspects such as sensor configurations, and vehicle capabilities like acceleration and deceleration, alongside vehicle control strategies. Non-design parameters encompass uncontrollable factors, including environmental conditions, the behavior of traffic participants, and road conditions. Failure conditions refer to potential system failures, identified using HAZOP guiding words, such as the omission of a sensor signal or incorrect values considered by actuators.

The systematic variation of parameters and failure conditions within the framework reveals hidden vulnerabilities, with insights from failure injection being critical for guiding design improvements. This way, Plan scenarios can create a wide range of conditions, even rare and extreme ones, ensuring coverage of all foreseeable hazardous events, including low probability but high impact situations. This phase sets the foundation for accurately simulating driving scenarios, incorporating both vehicle dynamics and environmental influences for precise risk assessment.

DO-Simulation: In the Do phase, the previously identified parameters are simulated to replicate ADS interactions with the environment and surrounding traffic (e.g. Non-ADS) across diverse conditions. Simulating interactions between the ADS-equipped ego vehicle and surrounding traffic through comprehensive behavioral algorithms and dynamic environmental factors, such as rain, snow, and fog, enables realistic and complex driving scenarios. This phase tests the ego vehicle's adaptability and decision-making capabilities against unpredictable scenario conditions, including variable weather and aggressive maneuvers by other vehicles. By exposing the ego vehicle to a range of hazards and operational challenges, the simulation provides valuable risk-specific context information and data on its performance.

This comprehensive risk-specific context information forms the basis for identifying critical elements influencing risk scenarios. From this broad context, risk features are extracted and meticulously selected based on their relevance and potential to indicate risk levels. Risk features are specific, quantifiable attributes, meaning they can be numerically expressed to identify, assess, and predict risks. These quantifiable attributes, used as inputs by ML models, allow for learning from historical patterns, recognizing risk factors, and making predictions about potential future risks. The selection, quality, and relevance of risk features significantly affect the accuracy and effectiveness of ML models in risk prediction.

TRAIN Model: In this phase, the focus is on utilizing risk features generated during the Do phase to predict dynamic risk indicators through the training and testing of supervised ML models. Initially, data labeling is conducted to enable the ML model to distinguish between various risk levels accurately. This process involves tagging each data point with an output label that represents different levels of severity and controllability, which the model aims to predict based on the inputs. Such supervised learning ensures that ML models can learn from input-output pairs and make informed predictions.

Following data labeling, the dataset undergoes normalization and balancing. Normalization addresses datasets with risk features on vastly different scales, which can hinder the training process. Normalization uniformly scales the feature space, which helps models converge more smoothly and quickly to optimal parameters, reducing numerical instability and speeding up the learning process. On the other hand, balancing the dataset is crucial for ensuring the model learns equally from all risk classes, regardless of their occurrence frequency. This balanced learning approach improves the model's generalization to unseen data, enhancing its predictive accuracy across different risk categories.

With the data preprocessed through labeling, balancing, and normalization, the ML model proceeds to training. During this stage, the model learns the patterns and correlations within the prepared dataset that indicate different levels of risk severity and controllability.

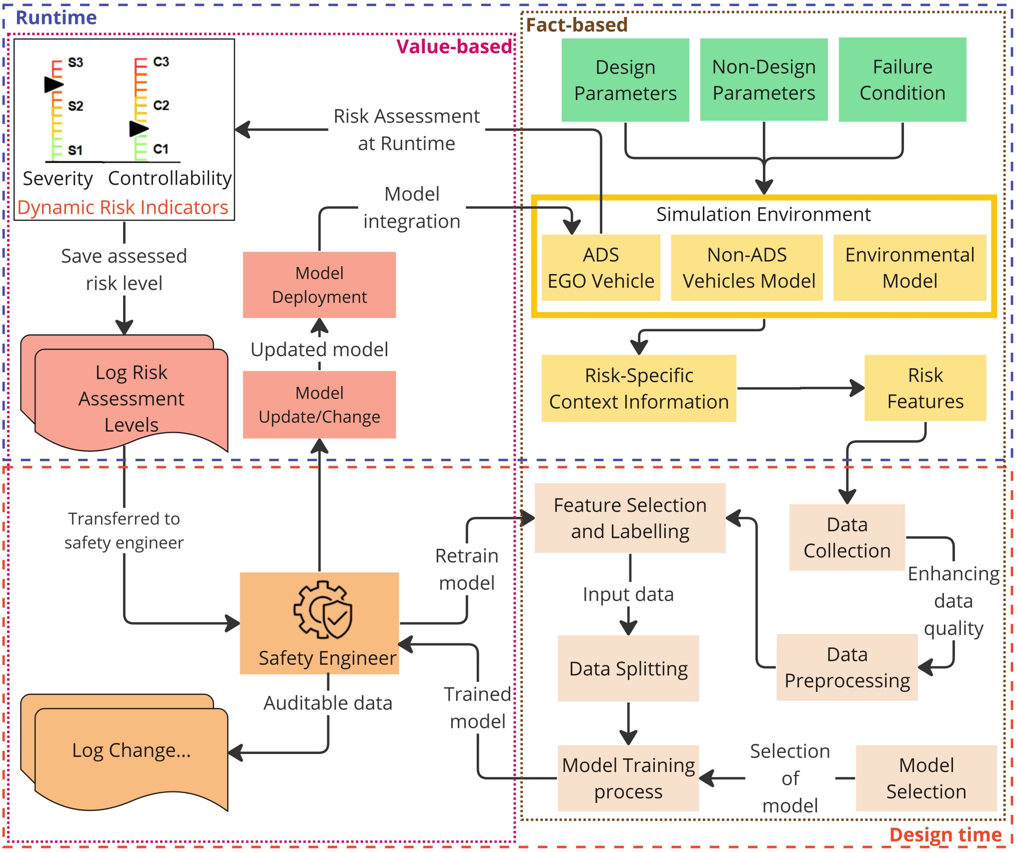


Figure 3.2: Comprehensive Overview of LADRI Process: Intersections of Design Time, Runtime, Value-Based, and Fact-Based Considerations

ADJUST Model: During this phase, the safety engineer rigorously evaluates the ML model's performance using metrics like accuracy, precision, recall, specificity, and F1 score, focusing on its runtime risk indicator prediction capabilities. If the model's predictions don't meet the predefined accuracy criteria (>= 99%), it undergoes adjustments or retraining process. This may involve sessions with modified parameters or datasets and thorough validation to ensure the model's compliance with performance criteria. The safety engineer examines potential overfitting or underfitting issues, employing cross-validation and analyzing learning curves for insights. Additionally, the risk features utilized by the model may be refined or adjusted for relevance and comprehensive representation of operational complexities. Modifications are made to the rules or thresholds for classifying risk labels, aiming to improve the accuracy of predictions and prevent biases in the model. Hyperparameters are also fine-tuned to achieve an optimal balance between bias and variance, aiming for a model that generalizes effectively to new unseen and unlabeled data. Continuous post-deployment monitoring and periodic reassessment using feedback loops ensure the model's sustained performance and adaptability to evolving risks, maintaining strict accuracy and performance standards. Safety engineers also maintain comprehensive documentation of all changes, updates, and performance logs related to the ML model for future reference.

ASSESS Risk Level: In this phase, the refined ML model is integrated into the ADS for runtime risk assessment, focusing on detecting severity and controllability indicators in scenarios not previously encountered. These indicators allow safety engineers to establish safety goals for the necessary enhancements or modifications to ADS, including improvements to braking systems or alterations to sensor configurations. Monitoring the ML model's performance during runtime risk assessments introduces unique challenges, particularly when it encounters new, unlabeled data, and ground truths are unknown to the safety engineer.

To navigate these challenges, safety engineers employ data visualization tools to scrutinize the model's predictions against new data, facilitating the discovery of emerging patterns, trends, or anomalies indicative of previously unrecognized risk indicators. Moreover, the application of model explainability and interpretability tools, like Shapley values or Brier score, sheds light on the rationale behind the model's predictions in unfamiliar scenarios. This understanding aids safety engineers in pinpointing unseen risk indicators and grasping the model's decision-making logic.

The LADRI framework delineates operational dynamics into runtime and design time activities, as depicted in Fig. 3.2. This structured division aligns with ADS risk assessment's specific objectives, facilitating a continuous, iterative loop crucial for ADS development. During runtime, the Plan and Do phases dynamically simulate ADS behavior across varied operational contexts, while the Assess phase translates simulation outcomes into actionable risk indicators like severity and controllability.

Conversely, the Train and Adjust phases, classified as design time activities, focus on leveraging simulation data to develop and refine predictive ML models. These models are trained to identify risk indicators from complex datasets, and safety engineers adjust them to accurately reflect ADS operations based on performance.

Incorporating both fact-based and value-based analyses is pivotal within this framework. Fact-based analysis, which draws on empirical data and objective evidence, supports the Plan, Do, and Train phases by generating, executing, and learning from simulated scenarios that mirror real-world conditions. This ensures the models are empirically grounded, enhancing their relevance and accuracy.

Meanwhile, value-based analysis, which relies on professional judgment and expertise, is key to the Adjust and Assess phases. These phases involve evaluating and modifying predictive models with a nuanced understanding of safety principles and assessing risk indicators against operational safety goals. This blend of fact-based and value-based approaches ensures a comprehensive risk assessment, leveraging empirical evidence and expert judgment to enhance HARA.

3.2 Simulation Environment

In this thesis, to simulate an ADS-equipped ego vehicle, the example "Highway Lane Following with Intelligent Vehicles" from Matlab/Simulink is utilized and modified as per the thesis requirements. This example allows the ego vehicle to travel within a marked lane [129].

The simulation environment comprises several key modules designed to simulate and analyze the behavior of ADS within a 3D scenario (as shown in Fig. 3.3). The Simulation 3D Scenario subsystem sets the foundation by defining the road, ADS-equipped ego vehicle, surrounding non-ADS vehicles, and synthesizing sensors for the simulation. To understand the vehicle's surroundings, the lane detector algorithm model detects lane boundaries using data captured by the camera sensor, while the vehicle detector algorithm model identifies vehicles within the frame. Enhancing the vehicle's perceptual capabilities, the forward vehicle sensor fusion algorithm model amalgamates detection of vehicles ahead of the ego vehicle from both vision and radar sensors.

The lane following decision logic algorithm model determine the vehicle's movement, providing lane center information and the most important object related data to the controller for lateral and longitudinal decision-making. Complementing this, the ego Adaptive Cruise Control (ACC) algorithm governs the vehicle's steering, braking and acceleration/deceleration. Lastly, the vehicle dynamics module outlines the model for the ego vehicle, completing the suite of tools designed to simulate and refine ADS functionalities.

The simulation environment for ADS serves as a dynamic "Dashboard," integrating a blend of design and non-design parameters along with a variety of potential failure conditions. This integration create a scenario-rich platform that produces extensive and contextually diverse data. Such data is important for accurately reflecting the complexities and unpredictability of real-world dynamic driving conditions. This environment is exceptionally conducive to the development, training, and testing of the LADRI frame-

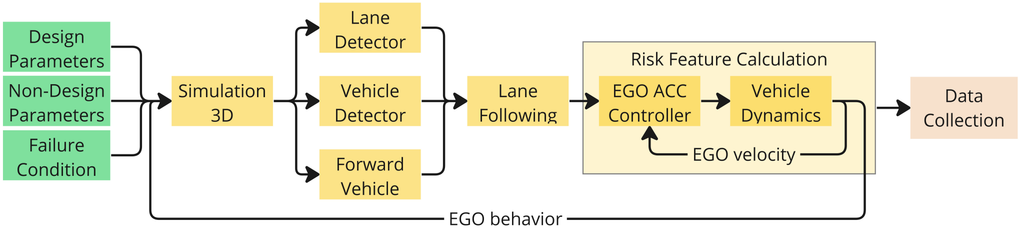


Figure 3.3: Integrated Simulation Environment: Combining Design and Non-Design Parameters with Failure Conditions

work. It encompasses an array of driving scenarios, ranging from standard operational conditions to rare and extreme situations, ensuring the data collected is thorough. This comprehensive data collection is essential for developing a LADRI framework that is both adaptable and efficient.

The simulation environment is structured through a Plan-scenario phase, which involves the identification and categorization of various parameters and conditions. This phase guides the subsequent Do-Simulation phase effectively as shown in Fig. 3.3. The simulation environment varies parameters based on three main pillars defined in the Plan-Scenario phase: Design Parameters, Non-Design Parameters, and Failure Conditions. Each of these elements plays a role in creating detailed and variety of real-world scenarios, providing an environment for evaluating LADRI framework performance.

3.2.1 Design Parameters

Design parameters are those that have a direct influence on a vehicle's dynamic behavior and the decision-making capabilities of the ADS. These parameters include the ego vehicle's set speed, maximum and minimum longitudinal acceleration, the reaction time of the ADS, default safe distance, vehicle mass, runtime longitudinal acceleration, vehicle speed, and ACC control gain parameters (as shown in Table. 3.1). The rationale behind the selection of these parameters lies in their direct impact on how the ADS operates under various traffic conditions and environmental settings, which in turn, shapes the risk assessment process.

For example, maximum acceleration and deceleration rates determine the ADS's emergency response capability, while reaction time and ACCC control gains dictate the system's responsiveness and adaptability. The inclusion of vehicle mass and default safe distance further allows for a nuanced understanding of how physical characteristics and predetermined safety protocols affect risk scenarios. Together, these parameters provide enriched scenarios for LADRI framework to analyze and predict potential hazards, offering insights into the effectiveness of the ADS in safety-critical situations. In addition, the simulation environment offers opportunities to modify sensor specifications, including adjustments to sensor angles, their position, or altering their range and sensitivity. However, to maintain focus and conciseness, these adjustments were not explored in this study.

Table 3.1: Design and Non-Design Parameters

Type	Parameters	Relevant Source	Units
Design Parameter	Set Speed of Ego ($SetSpeed_{ego}$)	Static Input	m/s
	Max. Longitudinal Acceleration ($a_{max,ego}$)	Static Input	m/s^2
	Min. Longitudinal Acceleration ($a_{min,ego}$)	Static Input	m/s^2
	Max. Longitudinal Deceleration ($d_{max,ego}$)	Static Input	m/s^2
	Min. Longitudinal Deceleration ($d_{min,ego}$)	Static Input	m/s^2
	ADS Response Time ($\mathcal{P}$)	Static Input	seconds
	Default Safe Distance ($d_{default}$)	Static Input	meter
	Ego Vehicle Mass (m)	Static Input	kg
	Longitudinal Acceleration (a_{ego})	Accelerometer	m/s^2
	Longitudinal Vehicle Speed (v_{ego})	Speed sensor	m/s
	Position of Ego vehicle (x_{ego})	GPS sensor	Coordinates
	ACC Control Gain Parameters	Software Algorithm	-
Non-Design Parameter	Set Speed of Lead ($SetSpeed_{lead}$)	Static Input	m/s
	Max. Longitudinal Acceleration ($a_{max,lead}$)	Static Input	m/s^2
	Min. Longitudinal Acceleration ($a_{min,lead}$)	Static Input	m/s^2
	Max. Longitudinal Deceleration ($d_{max,lead}$)	Static Input	m/s^2
	Min. Longitudinal Deceleration ($d_{min,lead}$)	Static Input	m/s^2
	Length of the Lead (L_{lead})	Static Input	meter
	Lane Width (l_{width})	Static Input	meter
	Longitudinal Acceleration (a_{lead})	Accelerometer	m/s^2
	Longitudinal Vehicle Speed (v_{lead})	Speed sensor	m/s
	Road friction coefficient (μ)	Wheel slip sensors	-

3.2.2 Non-Design Parameters

As shown in Table 3.1, non-design parameters such as the lead vehicle's maximum longitudinal acceleration, length, speed, lane width, traffic density, weather conditions, and road friction coefficient are crucial for capturing the variability and complexity of real-world driving conditions, impacting the behavior of ADS. Although these parameters are not directly controllable, they have a significant impact on ADS behavior.

For instance, the maximum longitudinal acceleration and speed of the lead vehicle, along with its length, directly impact the spacing and timing decisions of the ego vehicle, affecting how it adjusts its speed or changes lanes in response to the lead vehicle's actions. Similarly, lane width influences maneuverability and safety margins, especially important in scenarios involving lane changes or avoiding obstacles (e.g., during Cut-in and Cut-out scenarios).

Furthermore, traffic density metrics from GPS data provide critical context for the LADRI framework, enabling it to assess risk ratings based on vehicle proximity and potential for congestion. Weather conditions and road friction coefficients, gathered from weather sensors and wheel slip sensors respectively, are crucial for understanding environmental impacts on ADS performance, particularly in terms of braking distances and traction.

By integrating these non-design parameters into simulation environments, the LADRI framework enhances its ability to evaluate ADS behavior and risk ratings in a detailed manner. This approach allows the LADRI framework to adjust its risk ratings dynamically across a spectrum of scenarios, from ideal to challenging conditions, including heavy traffic, adverse weather, varying road surfaces, ensuring more precise and context-aware predictions.

3.2.3 Failure Conditions

To ensure the safety and reliability of ADS, it is crucial to rigorously test these systems under a wide range of conditions, including those that are unlikely but potentially catastrophic. The rationale behind synthetically generating specific failure scenarios is rooted in the HAZOP methodology [97], which is designed to identify and evaluate potential hazards in a system by examining possible deviations from normal operations. By applying HAZOP guide words to simulate failure conditions, engineers can systematically explore the effects of various faults and malfunctions, thereby gaining insights into the ADS's behavior under adverse conditions. This approach enables the identification of weaknesses in the system design and the development of mitigation strategies to enhance overall safety.

Various failure conditions are considered to rigorously evaluate the effectiveness and robustness of the LADRI framework under safety-critical scenarios that ADS may encounter during driving operation on highway lane following situation. The chosen failure conditions include unintended acceleration, unintended braking, insufficient engine/brake power, jerky and fluctuating acceleration/deceleration, and object detection delay by radar sensors as shown in Table. 3.2.

For example, unintended acceleration scenarios evaluate the system's response to abnormal speed increases, which could potentially lead to front-end collisions. Similarly, scenarios involving unintended braking aim to assess rear-end collisions caused by unexpected deceleration of the vehicle. The ultimate objective of these simulations is to mimic both common occurrences and rare, yet potentially safety-critical scenarios. This approach not only highlights real-world issues that could lead to accidents if not

Table 3.2: Potential Failure Conditions Leading to Hazardous Scenarios

No	Failure conditions	Associating HAZOP Guide Words with Faults and Simulation Methods
1	Unintended acceleration leading to front-end collision	**HAZOP:** Wrong Value; **Faults:** Software glitches in acceleration control; **Injection:** Simulate erroneous throttle command inputs in the control algorithm.
2	Unintended braking leading to rear-end collision	**HAZOP:** Wrong Value; **Faults:** Hardware failure in braking system; **Injection:** Introduce random actuator malfunctions into braking system.
3	Insufficient engine/brake power despite higher demand, creating potential risk scenario	**HAZOP:** Delay; **Faults:** Communication delays affecting power delivery; **Injection:** Simulate message loss impacting engine or brake commands.
4	Jerky and fluctuating acceleration/deceleration leading to possibility of rear or front-end collisions	**HAZOP:** Omission; **Faults:** Intermittent software or hardware faults in throttle/brake control; **Injection:** Introduce random fluctuations in throttle/brake actuation commands.
5	Object detection delay by radar sensors leading to unnecessary acceleration/deceleration	**HAZOP:** Delay; **Faults:** Radar sensor drift or processing errors; **Injection:** Manipulate radar sensor data to create delayed in object detection scenarios.

properly addressed but also test the prediction capability of LADRI framework in variety of challenges.

3.2.4 Integrated Overview of ADS

An integrated ADS behavior, particularly provide large amount of data for diverse driving scenarios to serve as an input for the LADRI framework. In the simulated environment, vehicle behaviors are meticulously modeled to reflect real-world dynamics. For instance, as shown in Fig. 3.4, the behavior of an ego vehicle, equipped with an ACC system, is simulated to adapt its speed in response to a lead vehicle's actions. This is achieved by configuring the ego vehicle to automatically adjust its speed to maintain a safe following distance, thereby emulating the ACC controller's role in runtime ADS operation. The incorporation of design parameters, such as the ego vehicle's maximum acceleration and deceleration rates, alongside non-design parameters like lead vehicle speed and weather conditions, facilitates the creation of diverse driving scenarios.

Failure conditions introduce an additional layer of complexity, simulating potential system malfunctions or sensor inaccuracies that might lead to unintended vehicle behaviors. For example, a failure condition introduced

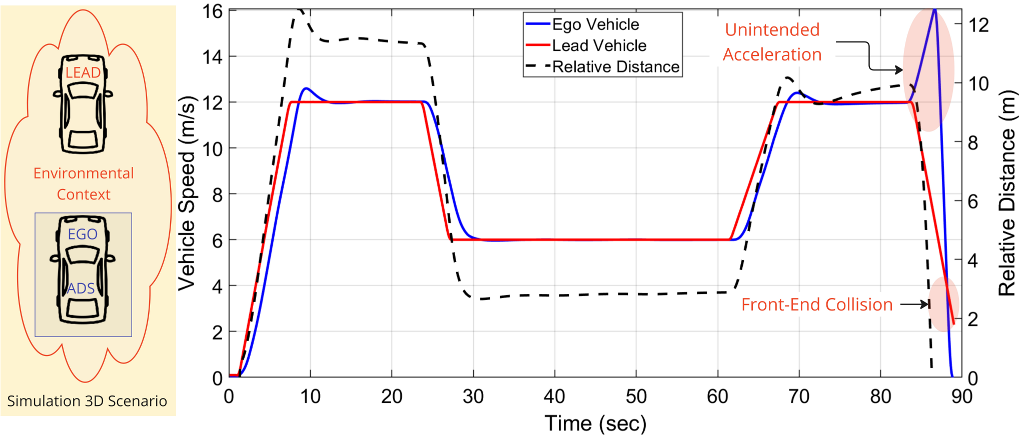


Figure 3.4: Integrated ADS behavior

during a braking operation could artificially impair the performance of the ego vehicle, causing it to reduce the distance between itself and the lead vehicle much more quickly than intended, resulting in a front-end collision (as shown in Fig. 3.4). This scenario not only tests the ACC controller's responsiveness but also examines the LADRI framework's ability to identify and evaluate the emerging risk, assigning a risk indicator based on the severity and controllability of the hazardous scenario. By varying the speed, environmental conditions, and applying different sets of parameters, the simulation environment enables a nuanced exploration of how LADRI framework can assess risks across a spectrum of scenarios. The risk-specific context information collected (at every 0.1 second) from sensors during these simulations, such as vehicle speed, distance to the lead vehicle, and acceleration/deceleration patterns, are critical for risk feature extraction.

3.2.5 Risk Feature Extraction

In the LADRI framework, a "Risk Feature" is defined as a quantifiable attribute or metric derived from risk-specific context information, indicative of potential hazards or conditions that may lead to safety-critical events within an autonomous driving context. These features are crucial for ML algorithms, enabling the model to identify, assess, and predict risk ratings by analyzing complex data patterns that traditional risk assessment methods might not discern. These risk features can be used to identify intricate hidden patterns and correlations within their data points, offering a nuanced and comprehensive understanding of the risk landscape.

Traditional safety metrics such as Time to Collision (TTC), often termed Surrogate Indicators [94] or Criticality Metrics [141], serve as the foundation of driver safety assessment. Their incorporation as risk feature within an ML-driven algorithm enhances their utility. Embedding these metrics enables the LADRI framework to dynamically interpret these features in light of a broad spectrum of data inputs, such as environmental conditions, vehicle dynamics, and traffic behavior patterns. This enriched analysis capability enables LADRI framework to discern that unintended acceleration decreases TTC, indicating an imminent front-end collision risk, whereas unintended braking might paradoxically increase TTC with respect to the lead vehicle but still poses a high-risk scenario due to the potential for a rear-end collision. LADRI framework evaluates these dynamics, understanding how different failure conditions affect the overall risk level, recognizing that metrics like TTC require contextual interpretation; an increase in TTC is not universally safe, nor is a decrease always risky without considering the specific failure condition and surrounding traffic dynamics.

By simulating various design and non-design parameters as static inputs for diverse driving scenarios in the Matlab/Simulink environment, LADRI framework acquires both quantitative and qualitative risk features. To gather these risk features, common static inputs were initially introduced into the simulation environment, as detailed in Table 3.3. Dynamic inputs were then collected through the sensor fusion capabilities provided by Matlab/Simulink. As a result, three types of feature sets: Time-based, Distance-based, and Impact-based risk features, were derived to represent the comprehensive behavior of ADS, thereby facilitating a thorough risk assessment.

Time-based Risk Features: Time-based risk features encompass metrics related to the timing of potential safety-critical incidents, focusing on the duration until a critical event, such as a collision or the need for evasive action, might occur. These are essential for understanding the urgency and timing of safety-related decisions.

TTC, Time to Escape (TTE), Time to Stop (TTS) each provide unique insights into the critical time windows available for action before a potential incident. While TTC offers a direct measure of the time until a collision at current speeds, TTE gives an indication of how quickly a vehicle can avoid an obstacle, emphasizing maneuverability. TTS, on the other hand, focuses on the vehicle's capability to come to a complete stop, highlighting braking efficiency. As shown in Fig. 3.5, understanding these temporal risk features is essential for designing ADS that can make informed decisions about when to initiate braking to avoid collisions.

TTC: can be defined as the time available for the ego vehicle to collide with a lead vehicle, given the prevailing speeds, distances, and trajectory of the ego vehicle. In a scenario where two vehicles (e.g. ego and lead) driving at same speed, the denominator diminishes, and TTC value turn to infinite. Therefore, to avoid this issue, enhanced TTC is used [142].

Table 3.3: Simulation Parameters: Static and Dynamic

Type	Parameters	Values
Dynamic	Velocity of EGO vehicle (Longitudinal)	v_{ego}
	Velocity of LEAD vehicle (Longitudinal)	v_{lead}
	Relative Velocity	$v_{rel} = v_{lead} - v_{ego}$
	Position of EGO vehicle (Longitudinal)	x_{ego}
	Position of LEAD vehicle (Longitudinal)	x_{lead}
	Relative Distance	$d_{rel} = x_{lead} - x_{ego}$
	Acceleration of EGO vehicle (Longitudinal)	a_{ego}
	Acceleration of LEAD vehicle (Longitudinal)	a_{lead}
	Relative Deceleration	$D_{rel} = a_{lead} - a_{ego}$
	Velocity of EGO vehicle (Lateral)	$v_{ego,lat}$
	Velocity of Object vehicle (Lateral)	$v_{obj,lat}$
	Relative Lateral Speed	$v_{rel,lat} = v_{obj,lat} - v_{ego,lat}$
	Position of EGO vehicle (Lateral)	$x_{ego,lat}$
	Position of Object vehicle (Lateral)	$x_{obj,lat}$
	Relative Lateral Distance	$d_{rel,lat} = x_{obj,lat} - x_{ego,lat}$
Static	Max. Acceleration of EGO vehicle	$a_{max,ego} = 2m/s^2$
	Min. Acceleration of EGO vehicle	$a_{min,ego} = -2m/s^2$
	Max. Deceleration of EGO vehicle (brake)	$d_{max,ego} = 9.8m/s^2$
	Min. Deceleration of EGO vehicle (brake)	$d_{min,ego} = 3.8m/s^2$
	Max. Acceleration of LEAD vehicle	$a_{max,lead} = 2m/s^2$
	Min. Acceleration of LEAD vehicle	$a_{min,lead} = -2m/s^2$
	Max. Deceleration of LEAD vehicle (brake)	$d_{max,lead} = 9.8m/s^2$
	Min. Deceleration of LEAD vehicle (brake)	$d_{min,lead} = 2m/s^2$
	Lane Width	$l_{width} = 3.5$ meter
	Default Safe Distance	$d_{default} = 2$ meter
	Response Time	$\mathcal{P} = 1.4$ seconds
	Vehicle Mass	$m = 1575kg$
	Length of the LEAD	$L_{lead} = 4.848$ meter

$$\text{TTC} = \frac{\sqrt{v_{rel}^2 + 2D_{rel}d_{rel}} - v_{rel}}{D_{rel}} \quad ; \quad v_{rel}^2 > 2D_{rel}d_{rel} \tag{3.1}$$

TTE: By quantifying the risk of lateral movement, TTE offers a more comprehensive risk assessment, going beyond TTC, which mainly focuses on longitudinal risks. Integrating TTE into the risk assessment process aids in identifying and implementing collision avoidance maneuvers that are not solely reliant on braking, but also take into account lateral movements.

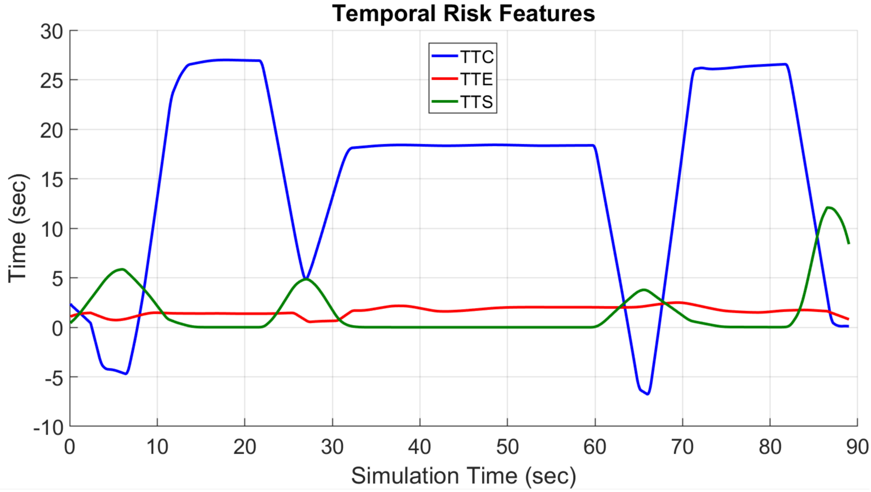


Figure 3.5: Temporal Risk Features for Model Training

This becomes particularly crucial in scenarios where braking alone may not suffice to avoid a collision, such as when a vehicle from an adjacent lane unexpectedly merges into the lane of the ego vehicle [85]. Moreover, TTE offers insights into the amount of time the ego vehicle has to avoid a front-end collision in cases of sudden deceleration and braking by the lead vehicle.

(3.2) $$\mathrm{TTE} = \frac{-l_{\mathrm{width}} + d_{\mathrm{rel,lat}}}{v_{\mathrm{rel,lat}}}$$

TTS: is an essential metric for risk assessment in the domain of ADS and vehicular control. It represents the crucial interval needed for an ego vehicle to come to a complete halt from its current speed, post-initiation of braking. In the context of risk assessment, TTS provides valuable insights into the immediate actionable window available to prevent potential collisions or mitigate their severity. Moreover, the behavior of TTS, akin to other time-based safety metrics, indicates that longer stopping times are generally associated with less severe outcomes and greater controllability over the vehicle’s state. This correlation underscores the importance of maintaining adequate stopping distances and highlights the role of TTS in facilitating proactive risk indicator [142].

(3.3) $$\mathrm{TTS} = \frac{v_{\mathrm{rel}}}{d_{\mathrm{max,ego}}} \quad ; \quad d_{\mathrm{rel}}, v_{\mathrm{rel}} > 0$$

Distance-based Risk Features: Distance-based risk features measure the necessary safety distances around the vehicle, including stopping distances under various conditions. These risk features are essential for spatial awareness and maintaining safe buffers between vehicles and obstacles.

Minimum Distance to Avoid Crash (MDAC), Safety Distance (SFD), Stopping Distance (STD) each contribute to understanding the spatial requirements for safety, as shown in Fig. 3.6. MDAC focuses on the critical minimum buffer needed to prevent collisions, directly informing safety margins. SFD is more prescriptive, indicating the ideal following distance under normal conditions, while STD calculates the total distance needed to halt the vehicle, including both reaction and braking distances. These spatial metrics collectively provide a multidimensional view of the vehicle's spatial requirements for safety, essential for maintaining proper positioning on the road.

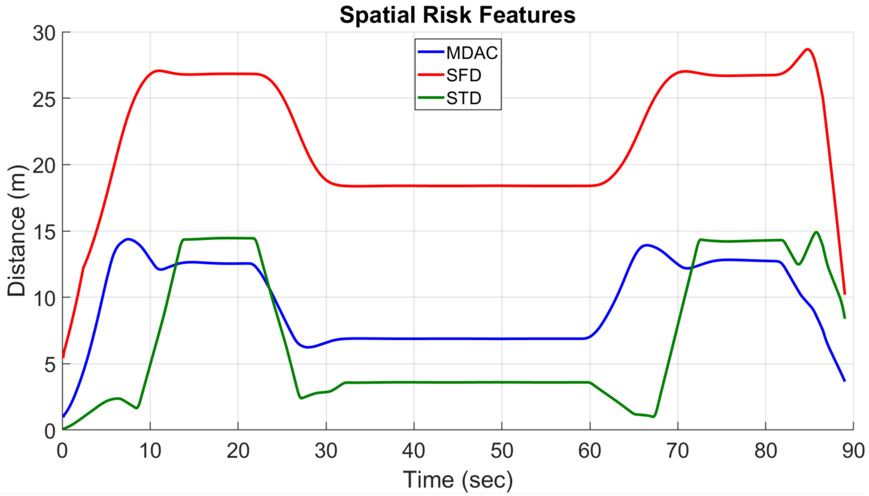


Figure 3.6: Transformation of Raw Sensor Data into Features for Model Training

MDAC: The minimum distance required to avoid a crash at any given time between the ego and lead vehicles is defined as the shortest distance necessary to prevent a collision. This concept, also known as the Responsibility-Sensitive Safety model [78], is derived from a scenario in which the lead vehicle performs a panic stop using the maximum possible braking force. In this scenario, the ego vehicle, initially traveling at a distance greater than or equal to the MDAC, detects this action and responds by panic braking with a deceleration matching at least its minimum braking capability. Given the challenges posed by variability in real-world driving conditions, such as road surface conditions, vehicle braking capabilities, and unexpected maneuvers by other road users, the MDAC risk feature aids in quantifying the impact of these variables on safety margins. This, in turn, allows for a more robust and accurate risk assessment.

$$\text{MDAC} = \left[v_{ego} \cdot \mathcal{P} + 0.5 \cdot a_{max,ego} \cdot (\mathcal{P})^2 + \frac{(v_{ego} + \mathcal{P} \cdot a_{max,ego})^2}{2 \cdot d_{min,ego}} - \frac{(v_{lead})^2}{2 \cdot d_{max,lead}} \right] \tag{3.4}$$

SFD: Safety distance refers to the minimum distance that should be maintained between two vehicles to ensure safety while driving. This distance allows an ego vehicle sufficient time and space to react and stop to avoid a collision with the vehicle ahead in case of sudden braking or any unexpected event. The safety distance varies based on the ego vehicle's speed, the acceleration capabilities of the ego, and the reaction time of the ACC system [30]. It accounts for the reaction time of the system's response time, providing a buffer period to initiate a braking maneuver. Maintaining an appropriate safety distance reduces the risk of rear-end collisions, which are among the most common types of vehicle accidents.

$$\text{SFD} = v_{ego} \cdot \mathcal{P} + \frac{v_{ego}^2}{2 \cdot a_{max,ego}} + d_{default} \quad ; \quad a_{lead} \neq 0 \tag{3.5}$$

The concepts of MDAC and SFD, while related, serve different roles in the context of risk assessment for vehicle safety, particularly when evaluating the severity of potential accidents and the controllability of vehicles to prevent those accidents. MDAC is a precise measure of the minimal distance needed between two vehicles to avoid a collision under specific conditions, focusing on reactive scenarios. SFD, however, is a broader concept recommending a minimum distance for safe driving under normal conditions, accounting for factors such as vehicle speed, braking capabilities, road conditions, and the reaction time of the ADS. While MDAC addresses the minimal distance to prevent crashes in emergency situations, SFD provides extra margins for safety, influencing the evaluation of crash severity potential and vehicle controllability. A vehicle maintaining more than the SFD under varied conditions shows good controllability, while one closer to the MDAC limit has a higher crash risk due to reduced error margin.

STD: refers to the total distance a vehicle travels from the moment a system perceives a need to stop (including the reaction time) to the point where the vehicle comes to a complete stop. This includes both the reaction distance (distance covered during the system's reaction time) and the braking distance (distance covered from the start of braking to a complete stop) [142].

$$\text{STD} = \left(v_{rel} + D_{rel} \cdot \frac{\mathcal{P}}{2} \right) \cdot \mathcal{P} + \frac{(v_{rel} + D_{rel} \cdot \mathcal{P})^2}{2 \cdot D_{max,rel}} \quad ; \quad a_{lead} \neq 0 \tag{3.6}$$

Impact-based Risk Features: By concentrating on the dynamics of deceleration and the kinetic energy involved in potential collisions, impact-based risk features offer a quantitative foundation for assessing the forces of impact, as illustrated in Fig. 3.7. These features evaluate the severity of possible impacts and the vehicle's ability to avoid them. Deceleration Rate to Avoid Crash (DRAC) considers the vehicle's ability to decelerate to prevent collisions, focusing on the physical capabilities and limitations of the vehicles involved. Kinetic Energy (KE) adds another layer by quantifying the energy that would be involved in a collision, offering a measure of the potential severity of impacts. These features provide a foundation for evaluating the potential outcomes of scenarios and the effectiveness of possible interventions, focusing on the physics of collision avoidance and impact mitigation.

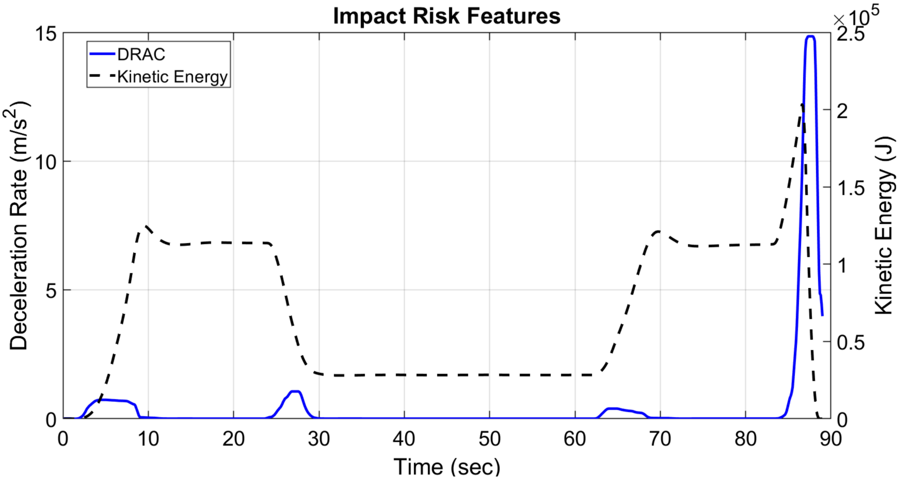


Figure 3.7: Transformation of Raw Sensor Data into Features for Model Training

DRAC: is defined as the rate of deceleration that the ego vehicle must achieve at any given time to avoid a crash, assuming the lead vehicle continues moving at the same speed and trajectory [23]. A key aspect of using deceleration rates effectively is the recognition of different driving conditions, such as wet or icy roads, which significantly affect the actual deceleration a vehicle can achieve. Safety engineers and ADS developers use standard deceleration rates under normal conditions as a baseline but must also consider adverse conditions when planning and developing vehicle safety features.

$$\text{DRAC} = a_{\text{lead}} + \frac{v_{\text{rel}}^2}{2 \cdot d_{\text{rel}}} \quad ; \quad a_{\text{lead}} \neq 0 \tag{3.7}$$

KE: is proportional to the vehicle's mass and the square of its speed, determining the severity of a collision. The KE of a moving vehicle directly affects its stopping distance when brakes are applied. A vehicle with higher KE will require a longer distance to come to a complete stop, assuming the same braking force. For ADS, understanding the KE dynamics of the vehicle and its surroundings is necessary for effective risk assessment.

$$KE = \frac{1}{2} m \cdot v_{ego}^2 \tag{3.8}$$

The delineation into temporal, spatial, and impact-based feature sets underpins a comprehensive LADRI framework, ensuring that every facet of risk assessment is meticulously accounted for. Each category of risk features provides a layer of granularity that contributes to a more detailed risk profile. Temporal features offer insights into the timing aspects of potential hazards, spatial features provide context regarding the physical positioning and distances involved, and impact-based features assess the potential severity of outcomes. This categorization ensures that the Train model phase is equipped with a multidimensional understanding of risk, crucial for accurate predictions of severity and controllability indicators across various highway lane-following scenarios. The integration of these diverse sets of features into the Train model phase provides a holistic and nuanced approach to risk assessment.

3.3 Model Training and Testing

In this phase, the focus is on utilizing risk features generated in previous section to train and test of supervised ML models, as shown in Fig. 3.8. Risk features provide the necessary input that allows ML models to learn the relationship between input data and the corresponding output labels (e.g., severity and controllability)[1].

However, risk features cannot be directly used as inputs for training ML models. There are certain preparatory steps, as highlighted in an established study [109, 110], such as data preprocessing and feature selection. Both steps enhances the learning and prediction capabilities of ML models. This section elaborates in detail on each step of the Train phase, as depicted in Figure 3.8.

[1] In this thesis, the research focuses solely on severity and controllability for the specified hazardous events, with the exposure rating considered as E4 (highly probable). The classification of controllability and severity as system attributes underscores their inherent relation to the system's design, operation, and potential failure modes [71].

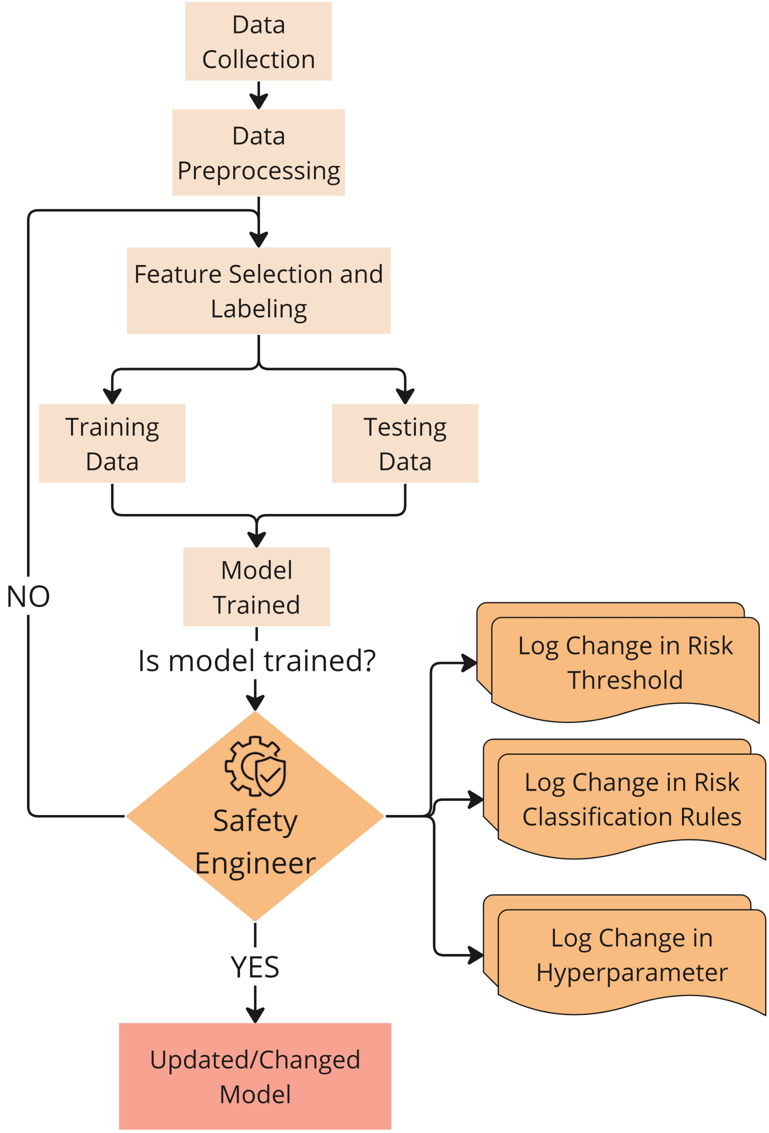


Figure 3.8: The Role of Safety Engineer

3.3.1 Data Preprocessing

This stage lays the groundwork for a robust predictive analysis by implementing a series of essential preprocessing techniques aimed at refining the input data. It encompasses three key processes: Labeling, Normalization, and Class balancing, each tailored to enhance the quality and relevance of the data before it enters the training pipeline.

Labeling Rules: Severity and Controllability: Labeling is an essential for supervised learning, serving as the foundation for models to learn the mapping between inputs and outputs from example pairs. It provides the necessary guidance for algorithms to understand and predict outcomes based on input features, similar to learning animal categories with specific examples. Unlike rule-based systems, which depend on predefined rules, labeling enables models to learn and infer patterns from the data, allowing for generalization and prediction on new, unseen data. This process enables models to identify patterns and relationships within the data, allowing them to manage new scenarios or minor deviations from the training data. For example, they can predict the risk associated with a new type of sensor failure under specific conditions by leveraging learned patterns of the environmental impact on vehicle behavior.

Table 3.4: Thresholds for Classification Rule-Set

Category	Level	Condition
TTC	1	$\geq$ 15s
	2	10s $\leq$ TTC < 15s
	3	5s $\leq$ TTC < 10s
	4	< 5s
MDAC	1	$\geq$ 30m
	2	20m $\leq$ MDAC < 30m
	3	10m $\leq$ MDAC < 20m
	4	< 10m
DRAC	1	$\leq 1\,m/s^2$
	2	1 < DRAC $\leq 3\,m/s^2$
	3	3 < DRAC $\leq 5\,m/s^2$
	4	$> 5\,m/s^2$

In this thesis, the labeling of severity and controllability risk indicators for LADRI framework is meticulously designed around three critical parameters: TTC, MDAC, DRAC. The selection of TTC, MDAC, and DRAC for labeling severity and controllability is grounded in their direct relevance to the physical and dynamic aspects of driving scenarios that critically influence the outcome of potential collision events. By quantifying the temporal urgency (TTC), spatial constraints (MDAC), and the vehicular response demands (DRAC), these parameters provide a comprehensive foundation for assessing and categorizing risks.

From a learning-based risk assessment perspective, training models with data labeled according to these parameters equips the system with the nuanced understanding needed to evaluate the severity and controllability of unseen scenarios. As shown in Table. 3.4 and Table. 3.5, this rule-set and threshold ensures that the assessment model can make informed predictions about the risk indicators, considering the intrinsic dynamics

Table 3.5: Combined Severity and Controllability Ratings Based on Threshold Names

Feature	**Severity Ratings**				**Controllability Ratings**			
	TTC 1	**TTC 2**	**TTC 3**	**TTC 4**	**DRAC 1**	**DRAC 2**	**DRAC 3**	**DRAC 4**
MDAC 1	S0	S1	S2	S3	C0	C1	C2	C3
MDAC 2	S1	S1	S2	S3	C1	C1	C2	C3
MDAC 3	S2	S2	S2	S3	C2	C2	C2	C3
MDAC 4	S3	S3	S3	S3	C3	C3	C3	C3

of vehicular motion and the operational limits of ADS. For example, the model can understand how the combination of high kinetic energy and low safety distance under unintended acceleration increases collision risk.

Normalization: Normalization is a crucial preprocessing step in ML, particularly for supervised learning algorithms. It involves scaling the feature data to a specific range, typically 0 to 1 (by using eqn. 3.9). This process is essential for several reasons, especially in the context of risk assessment, where features might have different scales of measurement. Normalization ensures that each feature contributes equally to the model's learning process, preventing any single feature from dominating the model's predictions due to its scale. This equality is crucial for models that rely on gradient-based optimization methods, as features (**x**) on different scales can distort the optimization landscape, making it more challenging for the model to converge to an optimal solution.

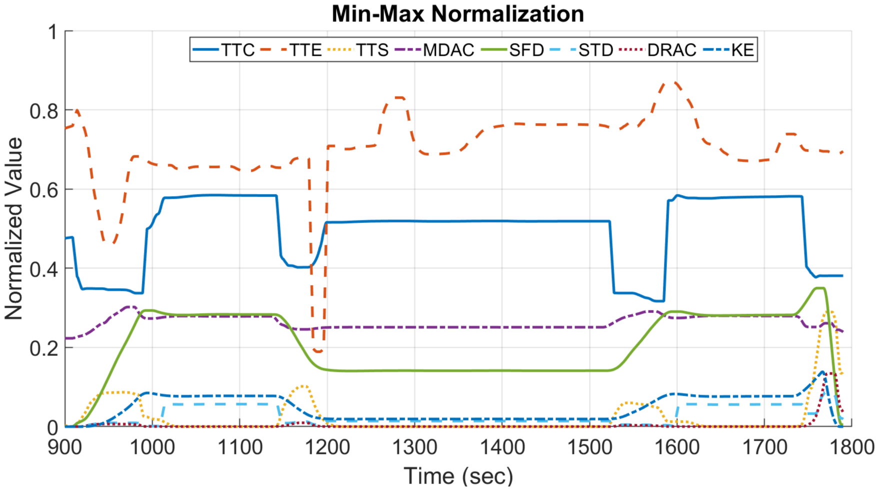


Figure 3.9: Normalization of Risk Features: Min - Max

(3.9) $$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

In risk assessment, features such as TTC (temporal), MDAC (spatial), and DRAC (impact-based) are measured on entirely different scales. Normalization ensures that these features are on a uniform scale, facilitating a balanced learning process where each feature's importance is based on its contribution to the outcome rather than its scale. As shown in Fig. 3.9, by scaling the features to a common range, normalization can speed up the training process. It helps in achieving faster convergence during the model's optimization phase, as the gradient descent paths are smoother and more straightforward when all features are normalized. Normalized data can lead to better model generalization on unseen data. This is because the model learns in a more balanced environment, where the influence of each feature on the prediction is proportional to its relevance, not its scale.

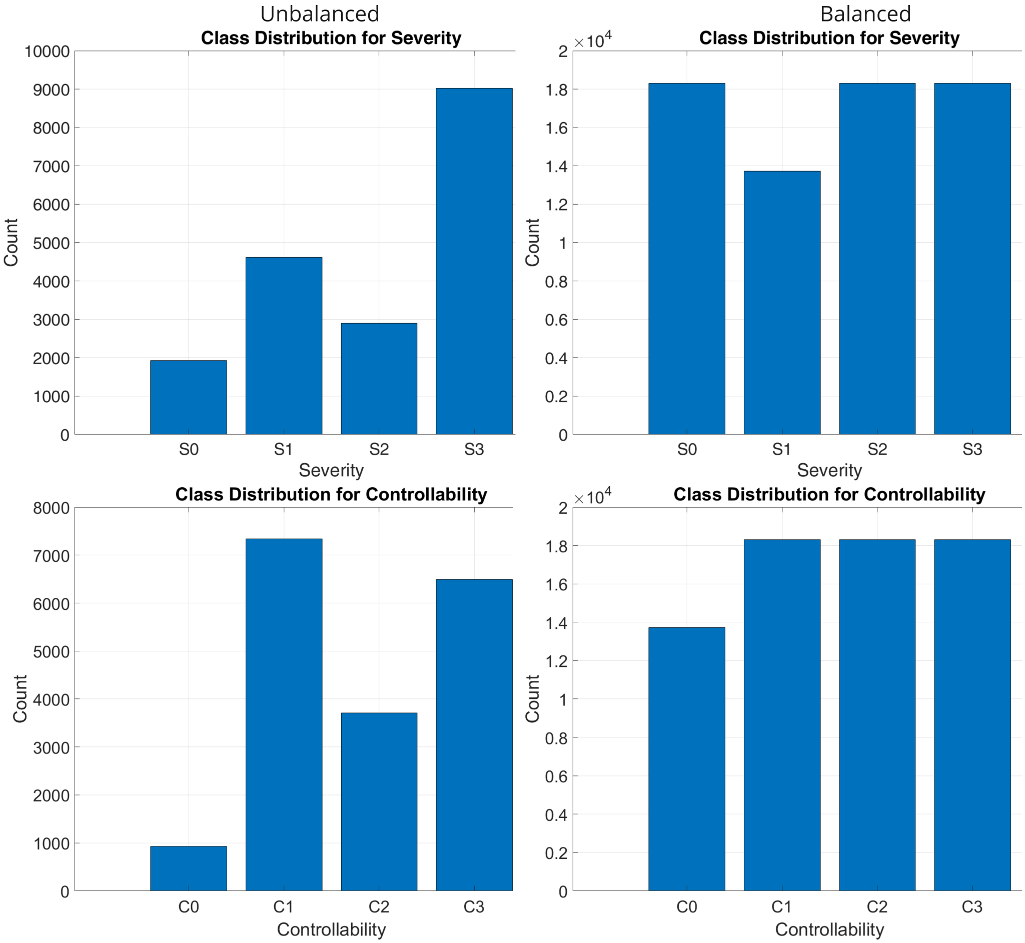


Figure 3.10: Comparative Plots of Unbalanced vs. Balanced Class Distributions for Severity and Controllability Ratings

Class Balancing: In the context of risk assessment, having a dataset with balanced representations of each class is crucial for developing ML models that can accurately predict a range of outcomes. In experiments with risk indicators for severity (S0, S1, S2, S3) and controllability (C0, C1, C2, C3), an unbalanced dataset where lower risk indicators (S0, C0) are less represented than higher ones (C3, S3), as shown in Fig. 3.10, can skew model performance. This imbalance affects the model's ability to learn from less frequent classes, often resulting in a bias towards predicting the more common classes. Models trained on unbalanced datasets may not generalize well to unseen data, particularly for underrepresented classes. This limits their practical applicability as they fail to identify critical but less frequent risk scenarios.

Class balancing can be achieved using techniques like Synthetic Minority Over-sampling Technique.It works by identifying k-nearest neighbors for instances in the minority class and generating synthetic samples along the line connecting each instance with its neighbors. This process continues until class distribution becomes balanced (as shown in Fig. 3.10), adding diversity to the dataset and enhancing model generalization. This technique benefits risk assessment by improving model sensitivity to rare risk indicators, enhancing predictive performance across classes, and preventing overfitting through the generation of varied synthetic examples, thus making models more reliable for identifying a wide range of risk scenarios.

3.3.2 Model Selection

In developing an effective LADRI framework, it is crucial to recognize that no single ML model is universally best for all types of data or complexity levels. Different models have unique strengths and weaknesses, making some more suitable for certain tasks than others. Given this variability, a safety engineer must adopt a flexible approach, selecting and comparing various models based on the specific characteristics of the dataset at hand and the complexity of the risk assessment task [2]. This section explores the strategic selection and comparison of models by safety engineers to optimize performance in diverse contexts, as shown in Table. 3.6.

[2] This section focuses on the characteristics of ML models in the context of risk assessment for ADS, deliberately omitting detailed mathematical explanations. This approach aims to maintain focus on the practical implications of these models within ADS environments, as the mathematical foundations are well-documented in many scientific literatures.

Table 3.6: Comparison of Machine Learning Models for Risk Assessment

Model	Description
Support Vector Machine (SVM)	Handles high-dimensional data well, suitable for scenarios with a clear margin of separation. Moderate interpretability and computational efficiency. Good performance on imbalanced data.
Logistic Regression	Well-suited for binary outcome predictions and serves as a reliable baseline model. High interpretability and computational efficiency but moderate performance on imbalanced data.
Artificial Neural Networks (ANN)	Excels in identifying complex, non-linear relationships within very high-dimensional data. Lower interpretability and computational efficiency dependent on network size. Effective with proper tuning, even on imbalanced datasets.
K-Nearest Neighbors	Ideal for simple classification tasks on small datasets. Lower interpretability and computational efficiency, with poor performance on imbalanced data.
Random Forest (RF)	High data complexity handling, excellent interpretability, and computational efficiency. Performs very well on large datasets and in determining feature importance. Particularly effective on imbalanced data.
Naive Bayes	Offers a baseline probabilistic approach with high interpretability and computational efficiency. Good performance on imbalanced data.
Gradient Boosting Decision Tree (GBDT)	Good at handling unbalanced data and ensuring predictive accuracy. High data complexity handling with moderate interpretability and computational efficiency.
Decision Trees	Provides very high interpretability, making it excellent for analyzing simple relationships and understanding decision paths. Moderate computational efficiency and good performance on imbalanced data.
Long Short-Term Memory	Specialized for time series and sequential data, handling high data complexity with sequence-dependent performance. Lower interpretability and computational efficiency varying with sequence length and model complexity.

Recommendation: For further analysis, models highlighted in gray (SVM, ANN, RF, GBDT) are selected based on their robust data handling, adaptability to complex patterns, and balance between interpretability and computational efficiency.

3.3.3 Support Vector Machine

SVM is a powerful classification method that works by finding the best hyperplane that separates different classes in the feature space. Imagine plotting data points in a multi-dimensional space, where each dimension represents a feature of the data. SVM finds the "hyperplane" that best divides data into classes, aiming to maximize the margin between different categories. For risk assessment, SVM can effectively segregate scenarios into different risk levels by finding the optimal boundary that separates them based on their features.

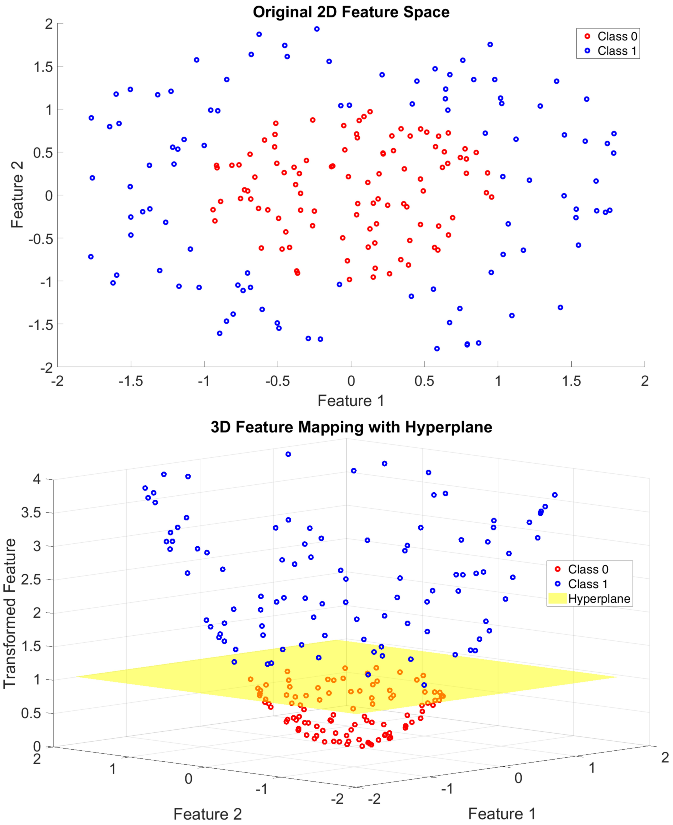


Figure 3.11: Feature Space Transformation: From 2D non-separability (left) to 3D linear separability with SVM hyperplane (right).

SVM is particularly useful when the data is clearly non-separable and utilizes the feature map transformation. It can transform low-dimensional feature space to high-dimensional feature space (many features), making it suitable for complex risk assessment scenarios where multiple factors determine the risk level.

As shown in Fig. 3.11, the top plot shows a synthetic dataset in a 2D feature space, where the two classes (Class 0 and Class 1) are not linearly separable. This is a common scenario in many real-world problems, where the relationship between classes and features is not straightforward. The bottom plot, on the other hand, illustrates the SVM transformation of the same dataset into a 3D feature space, using a simple feature mapping. In this transformed space, the two classes become more clearly separable, allowing for the construction of a hyperplane (shown in yellow plane) that can effectively classify the data points into two different classes.

3.3.4 Artificial Neural Networks

ANNs are inspired by the biological neural networks that constitute animal brains. An ANN is composed of layers of nodes (neurons), with each node connecting to several other nodes in the next layer, and weights (W) assigned to these connections as shown in Fig. 3.12. The network processes inputs (data features) through multiple layers, where each neuron computes a weighted sum of its inputs, applies a non-linear activation function (e.g., sigmoid, relu, tanh), and produces the output, classifying the data into different categories.

ANNs are highly flexible and can model complex non-linear relationships, making them ideal for risk assessment scenarios where the relationship between the input features and the risk levels is not straightforward. They can learn to identify subtle patterns and interactions between features that may indicate different levels of risk.

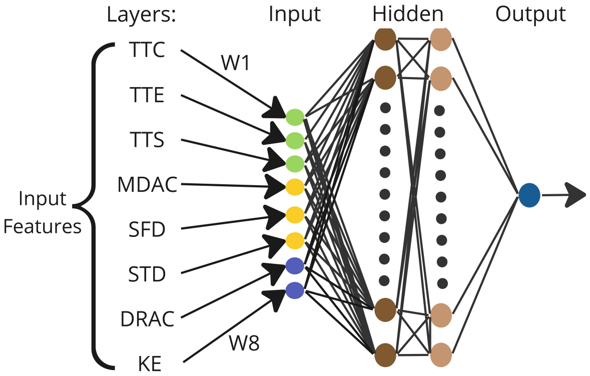


Figure 3.12: Basic Structure of ANN Model

A carefully chosen configuration is utilized for the classification task in order to ensure the effectiveness of the ANN model in predicting unseen data outcomes. This configuration includes a batch size, a learning rate, a Multi-Layer Perceptron classifier with hidden layers structured as 8-9-9-1, and the sigmoid activation function. The model processes large batches of data simultaneously, facilitating efficient training. The small learning rate aids in achieving precise weight updates, and the hidden layers capture complex relationships within the data. The sigmoid function introduces non-linearity, enabling intricate decision boundary learning.

3.3.5 Random Forest

RF uses an ensemble of decision trees to make predictions, leveraging the strength of multiple trees to produce a more accurate and stable model than a single decision tree could provide. For classification tasks, each decision tree in the forest outputs a class prediction, and the final output prediction of the RF is determined by a majority vote. For instance, if model has generated 500 trees, and 351 trees predict class A while 149 trees predict class B for a particular input, then class A will be the final output of the model for that input (as shown in Fig. 3.13). RF correct for decision trees' habit of overfitting to their training set, making the ensemble more robust and accurate.

RF is particularly useful for risk assessment because it can handle a large number of input variables without variable deletion. It is robust against overfitting and can model complex interactions between features. RF can provide insights into the importance of each feature in predicting risk levels, which is valuable for understanding risk factors.

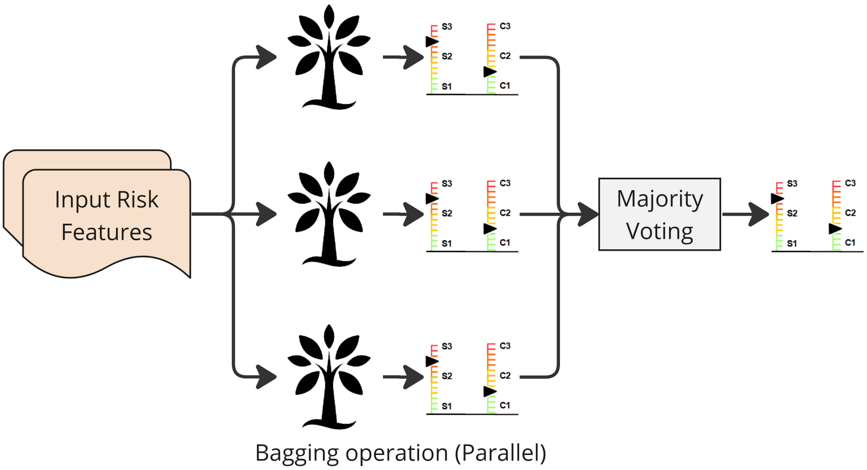


Figure 3.13: Basic Structure of RF Model

The configuration of RF algorithms plays a key role in predicting risk in unseen scenarios. Key hyperparameters include the number of trees in the forest, which enhances model robustness and accuracy by introducing randomness. The maximum number of splits helps control tree depth to prevent overfitting and reduce computation time, while the minimum leaf size ensures the model does not become overly precise on the training set, thereby maintaining generalizability. RF algorithms are particularly effective in risk assessment due to their resistance to overfitting, ability to highlight important features for targeted model tuning, and efficient parallel processing capabilities. These features, along with their capacity to handle non-linear relationships between features, make RF well-suited for the dynamic and complex nature of autonomous driving environments, where it is essential to predict risks accurately in real-time.

3.3.6 Gradient Boosting Decision Tree

GBDT is an ensemble method that builds decision trees in a sequential manner, where each subsequent tree is specifically designed to correct the residual errors made by the previous trees. This model synergizes multiple weak predictors, primarily decision trees, to construct a robust predictive model. The focus of GBDT is on minimizing loss, which is quantified as the difference between the actual and predicted risk levels. It does this through an iterative process that refines its predictions by continuously adjusting the sum of log odds, enhancing the accuracy of the final model output. Each tree adds its log odds prediction to the ensemble's cumulative prediction, progressively reducing the residual error and boosting the model's accuracy with each iteration, as illustrated in Fig. 3.14.

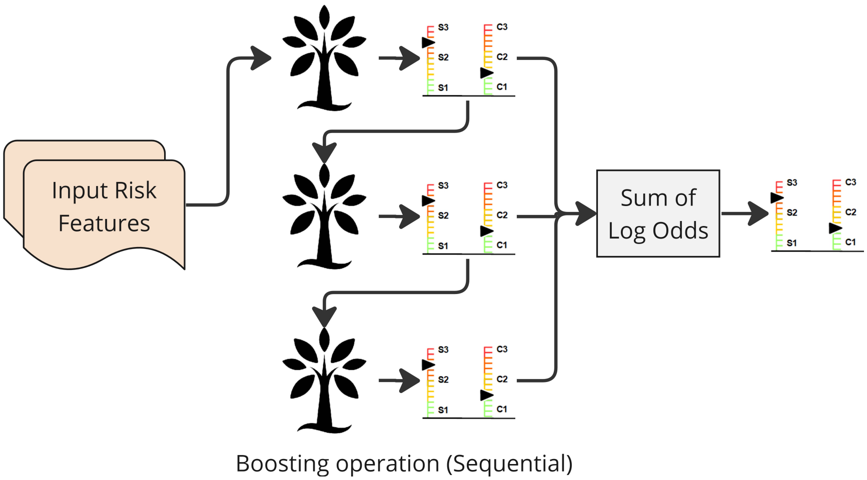


Figure 3.14: Basic Structure of GBDT Model

GBDT is highly effective for handling unbalanced data, which is typical in risk assessment where high-risk scenarios may be rare. It excels at capturing complex non-linear relationships and interactions between features, making it adept at accurately classifying scenarios into different severity and controllability levels. Key parameters such as the number of learning cycles dictate how well the model can capture intricate patterns in dynamic environments, enhancing its ability to adapt to new data. The learning rate determines the impact of each iteration on the final model, facilitating rapid adaptation to changing risk factors, which is critical in dynamic environments. Sub-sampling adds randomness, reducing bias and increasing adaptability. GBDT's robustness comes from its capacity to discern complex feature relationships and its sequential training approach, which allows it to continuously refine its predictions. This makes GBDT especially effective in environments where risk features evolve, ensuring it remains relevant and effective over time.

Model Selection Criteria: The safety engineer plays a pivotal role in selecting the most suitable model or combination of models. They evaluate the performance of various models, considering factors such as accuracy, interpretability, and the model's responsiveness to dynamic conditions. Important considerations also include computational efficiency, the model's ability to generalize from training data to new situations, and its robustness against overfitting. This thorough evaluation ensures that the selected model not only meets the technical needs but also fulfills the operational demands of the ADS risk assessment process.

3.4 Model Performance

After the model selection phase, the model training process begins, where the model is meticulously tuned to recognize patterns and correlations indicative of potential risks. This tuning process is not merely about adjusting the model to fit the training data but is focused on enhancing the model's ability to accurately identify risk ratings. It emphasizes the model's generalization capabilities, ensuring that it can perform well on new, unseen data. This fine-tuning process is necessary for developing a model that is not only sensitive to the nuances of risk indicators but also robust against the variability inherent in diverse driving scenarios.

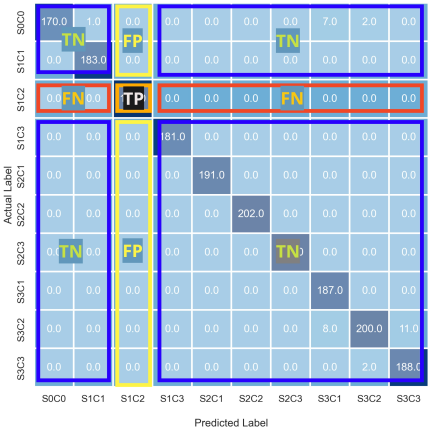


Figure 3.15: Confusion Matrix illustrating the classification outcomes with areas designated for True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)

Following the training phase, the model undergoes a critical testing phase. This phase employs a separate dataset that has not been used during the training process, providing an unbiased evaluation of the model's predictive performance. The use of this distinct test dataset is essential for assessing the model's effectiveness across diverse scenarios, thereby validating its operational accuracy. Performance evaluation during this phase employs a suite of metrics, encapsulated in a confusion matrix, which offers a detailed view of the model's predictive True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as shown in Fig. 3.15.

Central to this process is the role of the safety engineer, as depicted in Fig. 3.8, who checks the model's verification process. The safety engineer conducts a critical analysis of key performance metrics: Accuracy, Precision, Recall, F1-Score, and Specificity. Each metric provides insight into different aspects of the model's predictive performance. The safety engineer is also responsible for identifying any instances of overfitting or underfitting to ensure the model performs optimally.

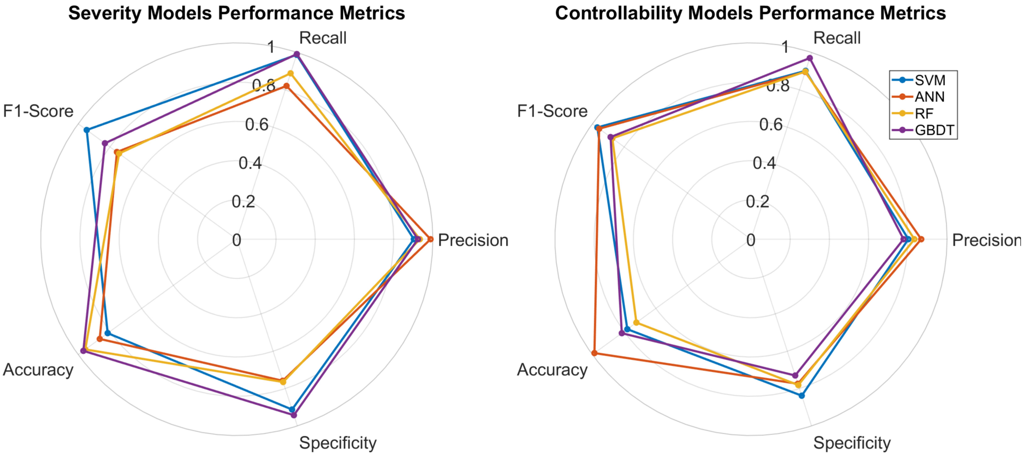


Figure 3.16: Various Model Performance Metrics: Severity and Controllability

Accuracy: Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. It provides a high-level view of the model's overall performance across all classes, making it a straightforward and intuitive metric. However, accuracy might not be reliable in cases of imbalanced datasets where one class significantly outnumbers another. Accuracy is determined by the ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

Precision: Precision, also known as positive predictive value, measures the proportion of true positive results in all positive predictions made by the model. It is crucial for scenarios where the cost of a false positive is high, indicating how reliable a model's positive classifications are. Precision is particularly important in applications where the goal is to minimize incorrect positive identifications. Precision is determined by the ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.11}$$

Recall: Recall, or sensitivity, measures the proportion of actual positives that are correctly identified by the model. It is essential in situations where missing a positive instance (false negative) carries a greater risk than incorrectly identifying a negative instance as positive. High recall is critical in ADS, particularly in detecting hazardous scenarios such as potential collisions or system failures. Failing to identify these risks can lead to severe consequences on the road. Recall is determined by the ratio of correctly predicted positive observations to all observations in the actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.12}$$

F1-Score: The F1-Score is the harmonic mean of precision and recall, offering a balance between the two by taking both false positives and false negatives into account. It is a better measure than accuracy for cases with imbalanced classes or when false positives and negatives have a different cost. The F1-Score is particularly useful for comparing the performance of models across datasets that may not share the same distribution of classes. F1-Score is determined by the weighted average of Precision and Recall.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.13}$$

Specificity: Specificity, or true negative rate, measures the proportion of actual negatives that are correctly identified by the model. High specificity indicates that the model is effective at identifying negative instances, making it complementary to recall, which focuses on positive instances. Specificity is crucial in fields like autonomous vehicle safety, where falsely identifying potential hazards can lead to unnecessary braking or evasive maneuvers. This can disrupt traffic flow or cause undue stress to the vehicle's occupants. Specificity is determined by the ratio of correctly predicted negative observations to all observations in the actual negative class.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.14}$$

Upon concluding the training and testing phases, and with a comprehensive understanding of the model's performance through the calculated metrics, the safety engineer decides whether to retrain the model with adjustments to increase its performance. The following section delves into strategies for refining the model further, addressing any identified deficiencies. It encompasses adjustments in hyperparameters, which could enhance the model's accuracy and its ability to generalize. Moreover, it may involve updating classification rules and adjusting thresholds to optimize performance, ensuring that the model meets the required performance criteria. This seamless transition from evaluating the model's current state to implementing adjustments underscores a dynamic, iterative process of continuous improvement, emphasizing the adaptability and precision required in the risk assessment process.

3.5 Model Adjustment

In this Adjust phase, the role of the safety engineer in the development and implementation of ML models for LADRI framework is crucial. They are primarily responsible for calibrating risk thresholds within the ML model. Their rigorous analysis of the model's performance in various simulated driving scenarios is essential. This analysis aids in identifying potential weaknesses and enhancing the model's robustness, ensuring it can effectively handle diverse and unpredictable driving conditions.

In addition, the safety engineer is pivotal in facilitating cross-disciplinary coordination. They act as a bridge between data scientists, software developers, and domain experts, fostering a holistic approach to risk assessment. This collaboration is essential for developing a comprehensive risk assessment tool that accurately represents the complexities of ADS.

Model retraining involves updating a deployed ML model with new information, such as changes in classification rules or threshold modifications. When a model falls short of meeting the predefined performance criteria (>= 99%), it undergoes retraining with adjusted parameters. This cycle of training, testing, and retraining fosters continuous improvement, refining and optimizing the model. The model's ongoing refinement is achieved through rigorous training and comprehensive testing to ensure enhanced accuracy and effectiveness in predicting runtime risks.

3.5.1 Updating Classification Rules and Thresholds

Calibrating the prediction model is essential for improving its performance in unknown scenarios. The process of updating risk thresholds and selecting specific features for severity and controllability classification is fundamental to enhancing the model's performance and adaptability across various driving conditions. For the sake of brevity, the classification rules for severity and controllability ratings incorporate three distinct features, as outlined in Table.3.4, offering a comprehensive view of both immediate and potential risks. This approach ensures an unbiased assessment across different severity or controllability ratings.

Thresholds for these risk features are determined through extensive simulation testing, utilizing a trial-and-error approach to encompass a wide range of potential driving scenarios comprehensively. As indicated in Table 3.5, combining MDAC with TTC provides a solid foundation for assessing severity, capturing both spatial and temporal safety margins. This combination emphasizes the urgency and potential impact of a scenario. Conversely, coupling MDAC with DRAC proves effective for evaluating controllability. The determination of optimal thresholds for these risk features is a dynamic, ongoing process that evolves through continuous simulation testing. By employing a trial-and-error approach, safety engineers meticulously explore various combinations of risk features and their thresholds. This iterative exploration ensures that the ML models are finely tuned to the complexities of driving conditions.

This meticulous selection process guarantees that the model accurately responds to a wide spectrum of critical driving dynamics, including aspects such as distance, time, and the vehicle's capability in critical situations. Should model verification reveal areas of improvement, a safety engineer undertakes refinement efforts, as depicted in Fig. 3.8. This iterative method, embedded within a transparent and traceable development cycle, might involve retraining the model with revised features or altering classification thresholds based on the testing outcomes. Additionally, sensitivity analysis might be conducted to understand the impact of each risk feature on the model's predictions, further guiding the optimization of feature selection and threshold setting. This systematic and data-driven approach ensures that the model maintains relevance in complex driving environments.

Such iterative refinement is pivotal as new data emerges or when prevailing patterns change, necessitating updates to preserve the model's precision. This ongoing learning strategy assures the model's sustained relevance and efficacy in evolving conditions. Each cycle of testing and adjustment is meticulously documented, charting a clear path for subsequent hyperparameters tuning phases. Every adjustment is made with intention, aimed at optimizing the LADRI framework to provide accurate and precise risk assessments.

3.5.2 Hyperparameter Selection

In the domain of ML, optimizing models and their hyperparameters plays a key role in enhancing model performance. Hyperparameters, predefined settings configured prior to training, significantly influence the learning process, guiding it in a way that differs from the adaptation of model parameters, which are learned during training itself. The selection of the most effective model and hyperparameter values presents a complex and resource-intensive challenge.

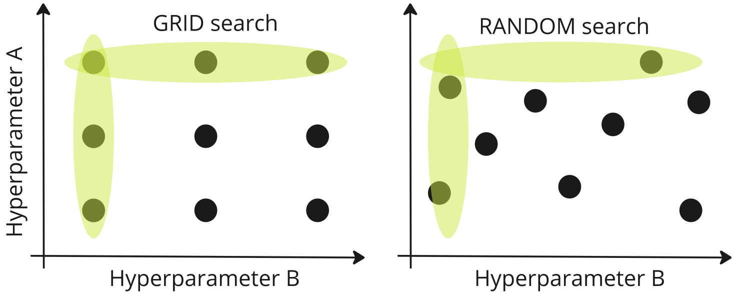


Figure 3.17: Hyperparameter Optimization Method: Grid Search and Random Search

To address this, strategies such as Grid Search and Random Search are utilized [16]. As shown in Fig. 3.17, Grid Search conducts a methodical exploration of specified hyperparameter combinations, ensuring comprehensive examination at the expense of high computational demand. Conversely, Random Search samples the hyperparameter space randomly, providing a more efficient yet potentially less thorough alternative. The choice between these methods hinges on the problem specific requirements, including the scope of hyperparameters and available computational resources. In this thesis, Grid Search was selected for (shown hyperparameter in Table. 3.7) optimization due to its systematic and exhaustive approach in identifying the most effective model parameters for predicting risk ratings. This method is aimed at achieving optimal performance in risk assessment tasks.

SVM is sensitive to the choice of the kernel and its parameters, which define the shape of the decision boundary. The *C* parameter controls the trade-off between the model's complexity and the degree to which deviations from a perfectly separating hyperplane are tolerated. Different *Kernels* allow the model to fit non-linear boundaries, and *gamma* determines the influence of individual training examples on the boundary, with larger values leading to more complex models.

ANNs are highly configurable models capable of capturing complex relationships in data. The *number of hidden units* and *layers* determines the model's capacity to learn non-linear functions but also its propensity to overfit. The *learning rate* is critical as it defines the size of the steps the

Table 3.7: Hyperparameters and their ranges for Grid Search optimization.

Model	Hyperparameter	Values to Compute
SVM	C	[0.1, 1, 10, 100]
	Kernel	[’linear’, ’poly’, ’rbf’, ’sigmoid’]
	Degree (for poly)	[2, 3, 4, 5]
	Gamma	[0.001, 0.01, 0.1, 1, ’scale’, ’auto’]
ANN	Number of hidden units	[10, 50, 100, 200]
	Batch size	[16, 32, 64, 128]
	Learning rate	[0.001, 0.01, 0.1, 0.2]
	Number of epochs	[10, 50, 100]
	Dropout rate	[0.0, 0.2, 0.5]
RF	n_estimators	[10, 50, 100, 200]
	max_depth	[10, 20, 30]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
	Max features	[’auto’, ’sqrt’]
	Bootstrap	[True, False]
GBDT	Number of trees	[50, 200, 300, 400, 500]
	Learning rate	[0.01, 0.05, 0.1, 0.5, 1]
	Sub sampling rate	[0.7, 0.8, 0.9, 1]

optimizer takes during training, with too high rate possibly overshooting minima and too low converging slowly. The *batch size* influences the stability of the learning process, and the *dropout rate* is a regularization technique to prevent overfitting.

RF models are ensembles of decision trees which tend to reduce overfitting by averaging the predictions. The *n estimators* parameter controls the number of trees in the forest, generally leading to better performance with more trees, albeit with diminishing returns. The *max depth* of the trees is a key parameter to control the complexity of the model, with deeper trees having a higher risk of overfitting. The *min samples split* and *min samples leaf* help provide constraints on tree growth and are forms of regularization.

GBDT is a sequential ensemble technique that builds one tree at a time, where each new tree helps to correct errors made by previously trained trees. The *number of learning cycles* determines the number of trees in the model, which can improve performance but also risk overfitting if too large. The *learning rate* controls the contribution of each tree to the final model and is typically set to a low value to allow for a more gradual and robust learning process. The *sub sampling rate* is part of the stochastic gradient boosting feature that helps in reducing overfitting by considering only a subset of data for each tree.

Each hyperparameter set is precisely aligned with the unique characteristics of the extracted features. This ensures both accuracy and computational efficiency in model predictions. Such meticulous calibration is integral to fully leveraging each algorithm's capabilities. Throughout the training process, the safety engineer observes and guides these adjustments, ensuring model efficacy. Safety engineer also maintains a log of changes in hyperparameter (as depicted in Fig. 3.8) for clarity on decision-making processes and justify specific model outcomes, enhancing transparency.

These change logs enable tracking of the model's performance over time. This allows teams to link specific alterations to shifts in the model's efficacy. If performance degradation or technical issues arise, a detailed log is invaluable for efficient debugging and troubleshooting. In collaborative and multi-departmental development environment, these logs are essential for knowledge sharing. They ensure that all contributors are aware of the modifications and understand their underlying reasons. This practice fosters a cohesive and informed approach to model development and maintenance for risk assessment.

3.6 Model Deployment

In the Assess phase, the best-performing model is deployed on ADS within a simulated environment. The primary goal is to use the model for predicting severity and controllability indicators. As shown in Fig. 3.18, this deployment allows for the recording of runtime risk assessment levels, capturing data on severity, controllability, and any anomalies, vital for safety engineer's analysis and ML model refinement. To maintain the integrity of the model, version control is strictly implemented, providing detailed documentation of each version of the risk assessment model. This enables clear change tracking, rollback capabilities, and audit processes.

Log Risk Assessment levels: The implementation of a log risk assessment levels is instrumental post-deployment of the ML model, particularly for early hazard identification. The log risk assessment module provides valuable evidence for ongoing runtime risk assessment. It captures runtime severity and controllability indicators and logs any anomalies or issues, enabling the safety engineer to promptly identify and address potential hazards. This monitoring includes incident reporting and tracking the model's performance, ensuring any deviations or unexpected behaviors are immediately flagged for review.

Furthermore, the module supports a structured feedback loop, where insights from the Assess phase are channeled back into the Plan and Do phases, fostering continuous improvement. Following the deployment of the ML model, the predictions it generates during the Assess phase play a crucial role in refining subsequent iterations of the Plan phase. Based on these predictions, the Plan phase can be adjusted to include new or altered

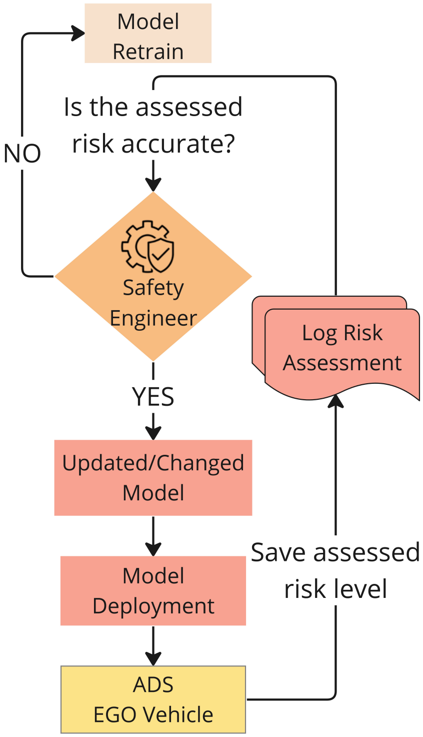


Figure 3.18: Post-Deployment Role of Safety Engineer

scenarios that target identified weaknesses or emerging risk patterns. For instance, if the model predicts a high risk of collision under certain traffic densities or environmental conditions, the Plan phase can be tailored to simulate and analyze these specific scenarios more extensively in the next cycle. This targeted adjustment allows for the development of more effective risk mitigation strategies and control measures, enhancing the overall safety of the ADS. Additionally, this adaptive approach enables the integration of real-time data and feedback from the operational environment, ensuring that the model remains relevant and effective against evolving driving conditions and risk landscapes. This continuous cycle of feedback and refinement fosters a proactive approach to risk management, dynamically enhancing the ADS's capabilities to predict and mitigate potential hazards.

Validation of the model takes place by applying it in diverse simulated driving conditions. The safety engineer assesses the model's performance across various scenarios to ensure its continued effectiveness and may make necessary adjustments to optimize its functionality based on these observations. During the validation phase, the model is rigorously tested across a spectrum of simulated ODDs that represent different traffic, environmental, and road conditions. Each ODD presents unique challenges and complexities, which helps to evaluate the model's performance under varied and sometimes extreme conditions.

Moreover, the complexity of the DDT is progressively escalated in these simulations. For instance, initial tests might focus on simple driving tasks such as maintaining lane-following under steady traffic conditions, whereas later tests might involve complex interactions like lane change maneuver, responding to erratic non-ADS driver behaviors, or emergency braking in slippery conditions. This stepwise increase in task complexity helps in assessing the model's decision-making capabilities and its ability to prioritize and react to dynamic risks accurately.

Monitoring the ML model's performance during runtime risk assessments introduces unique challenges, particularly when it encounters new, unlabeled data, and ground truths are unknown to the safety engineer. To navigate these challenges, safety engineers employ data visualization tools to scrutinize the model's predictions against new data, facilitating the discovery of emerging patterns, trends, or anomalies indicative of previously unrecognized risk indicators. Additionally, the application of model explainability and interpretability tools, like Shapley values, sheds light on the rationale behind the model's predictions in unfamiliar scenarios. This understanding aids safety engineers in pinpointing unseen risk indicators and grasping the model's decision-making logic.

Having outlined the LADRI framework for continuous risk assessment and early hazard detection, the focus now shifts to its practical implementation. The subsequent section demonstrates the application of LADRI framework´s theoretical constructs and components in simulations to assess risk in runtime, illustrating how theory translates into actionable insights in a simulated environment.

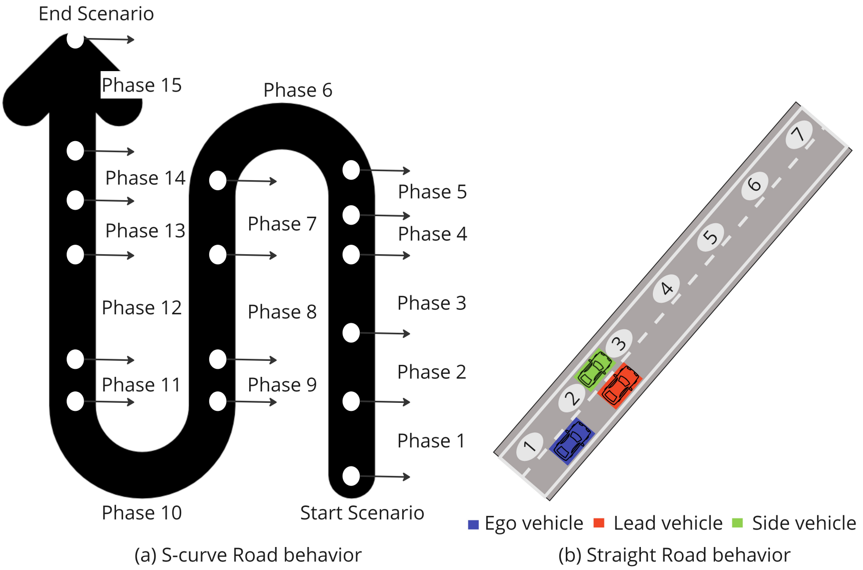
4 Evaluation

Building upon the methodology outlined in previous chapters, this evaluation chapter aims to forge a seamless connection, ensuring continuity with the LADRI framework's evaluation. In this section, the thesis delves into various aspects, including the contribution of scenarios and the verification of feature sets, where time-based, distance-based, and comprehensive features are tested to validate the model's efficacy. Subsequently, it explores the optimization of model performance through the application of the Adjust phase process, followed by the deployment of the model on ADS for validation across diverse driving scenarios. Additionally, this chapter provides insights into the model's explanation by the safety engineer, who evaluates the importance of risk features. Finally, the chapter concludes with a discussion on the application and limitations of the LADRI framework, offering a comprehensive overview of its impact and scope.

Chapter Content

4.1 Scenario Combination

To conduct a thorough evaluation of the LADRI framework, a strategic approach was adopted that encompassed model training, testing, and validation by deploying the model across a spectrum of simulated driving scenarios. This comprehensive strategy ensured that the LADRI framework covered various aspects of highway lane-following scenarios, thereby equipping it to predict severity and controllability indicators in a wide array of safety-critical situations.



(a) S-curve Road behavior

(b) Straight Road behavior

Figure 4.1: Scenario Map: Straight Road with Dynamic Driving Task

In an effort to generate a diverse set of scenarios, both S-curve and straight road configurations were utilized. These scenarios were crafted to include variations in speed, vehicle mass, behaviors of traffic participants, and levels of road friction to mimic environmental impacts, such as rain, and various failure conditions, as explained in Section 3.2. Specifically, Scenario S1 was chosen for the training and testing of ML models. All scenarios were devised based on predefined design and non-design parameters (as per Plan and Do phase), with only the speed, vehicle mass, road friction coefficient, and traffic participants' behavior being altered, as shown in a Table. 3.1.

For the S-curve road, the lead vehicle's speed changes over a certain distance in different phases (as illustrated in Fig. 4.1): starting with an acceleration to 30m/s, maintaining this speed for up to 500m, then slowing down to 5m/s for 50m. It accelerates again to 15m/s, holding this speed until 1800m, then speeds up to 30m/s. From this point, the speed fluctuates randomly between 25 and 35m/s until 2400m, followed by a slowdown to 12m/s. The speed then varies randomly between 10 and 15m/s up to 3600m, at which point it emergency brakes to 3m/s, holds for 20m, accelerates back to 20m/s, and maintains this velocity until 3800m. It concludes with another emergency brake to 3m/s, maintaining for another 20m. Such detailed simulation allowed for a nuanced exploration of vehicle dynamics under varied conditions.

Table 4.1: Summary of Simulation Scenarios for LADRI Evaluation

ID	Speed (m/s)	Vehicle Mass (kg)	Traffic Participant Behavior	Road Friction (μ)	Failure Mode
S1	Varied	1500	Normal	0.9	None
S2	12-36	2000	Normal	0.9	Unintended Acceleration (Ego)
S3	12-36	2000	Normal	0.3	Emergency Braking (Lead)
S4	12-36	2000	Cut-In	0.9	None
S5	12-36	2000	Cut-Out	0.9	None

On the straight road, scenarios S2, S3, S4, and S5 were simulated to validate the deployed model, as shown in Table. 4.1. In these scenarios, the lead vehicle performs a sequence of actions starting with an initial acceleration to a predetermined speed, followed by a phase of maintaining this speed. This sequence then introduces a deceleration phase, adding dynamic complexity, before maintaining a steady speed again. The cycle is completed by a re-acceleration to the initial speed and maintaining it, thus presenting adaptive challenges. The sequence concludes with the lead vehicle coming to a full stop to evaluate the ego vehicle's braking capabilities. These scenarios covered a broad range of speeds, varying from 12 m/s to 36 m/s in increments of 4 m/s to simulate different scenario combinations. Throughout these scenarios, unintended acceleration was introduced from the dashboard (shown in Fig. A.1) in the ego vehicle, and emergency braking maneuvers were executed in the lead vehicle. Cut-in and cut-out (side vehicle) maneuvers were also incorporated as planned behaviors at specific points on the straight road segment.

The rationale behind executing experiments in these phases, across both S-curve and straight road configurations, was driven by the goal to deeply understand the dynamic interactions between the lead and ego vehicles under a variety of conditions. By deliberately altering the lead vehicle's behavior, a rich dataset created from which risk features could be derived, as explained in Section 3.2.5. The effectiveness of the ML model, leveraging these diverse feature sets, is discussed in the following section. This structured approach not only ensures a robust evaluation of the LADRI framework but also enhances our understanding of vehicle dynamics and risk assessment in complex driving environments.

4.2 Verification of Feature Set

In this section, ML models are trained and tested (Train phase) on Scenario S1 using diverse risk feature sets to improve their predictive accuracy in assessing severity and controllability indicators. This thesis explores the performance enhancement of ML models through the selection of an optimal combination of risk features, as indicated in [74]. To validate the hypothesis that enhancing risk knowledge from diverse perspectives can improve model performance, the study systematically divides the feature set into three distinct categories: Time-based, Distance-based, and Comprehensive-based. The Time-based set focuses on temporal vehicle interactions, such as TTC, TTS, and TTE, crucial for immediate risk assessment. The Distance-based set emphasizes spatial considerations like MDAC, SFD, and STD, vital for strategic maneuvering. The Comprehensive-based feature set comprises temporal and spatial dimensions alongside additional metrics such as DRAC and KE, offering a rounded perspective on vehicle behavior and risk. Separate ML models are trained using each of the three feature sets. This process allows for a direct comparison of how the inclusion of different types of risk knowledge affects model performance. Conducting experiments and analyzing the results from these feature sets validate the effectiveness of the LADRI framework in predicting safety-critical scenarios.

4.2.1 Feature Set I - Time-based

The selection of this feature set is grounded in the hypothesis that the temporal distance to potential hazards, such as a front-end collision, provides a direct measure of the urgency and severity of a risk scenario. These features offer insights into the available reaction window to initiate corrective actions, thus directly influencing the controllability aspect as well. From a risk assessment perspective, utilizing time-based features allows for evaluating how swiftly a situation might escalate into a hazard. The time dimension reflects not just the vehicle's current state but its dynamic interaction with the environment, including the behavior of other road users and changing road conditions.

Utilizing only time-based features, ML models demonstrated higher accuracy in predicting severity indicators compared to controllability as shown in Fig. 4.2. This can be attributed to the direct correlation between time margins and the severity of potential impacts, which is more straightforward to predict. Models like ANN and SVM tend to perform well with these features due to their ability to model complex nonlinear relationships in time-constrained scenarios. However, the predictability of controllability is less pronounced, likely because controllability involves additional spatial and dynamic considerations not captured by time-alone metrics.

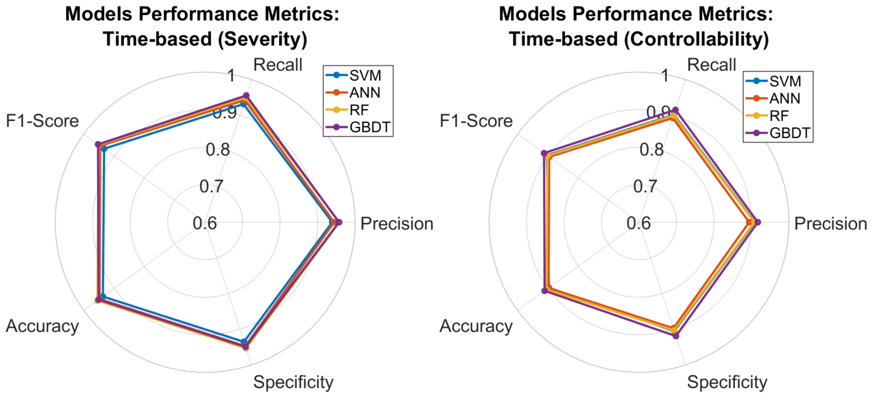


Figure 4.2: Model Performance Metrics: Time-based Feature

4.2.2 Feature Set II - Distance-based

The ego vehicle's braking efficiency, acceleration capabilities, and the current velocity directly influence distance-based features. These features are crucial for understanding the physical space needed to prevent collisions under various conditions. A vehicle with higher braking capabilities may require a shorter stopping distance, positively affecting the controllability indicators. However, in high-speed scenarios or conditions with reduced traction, even vehicles with advanced braking systems may face increased stopping distances, elevating the severity indicators.

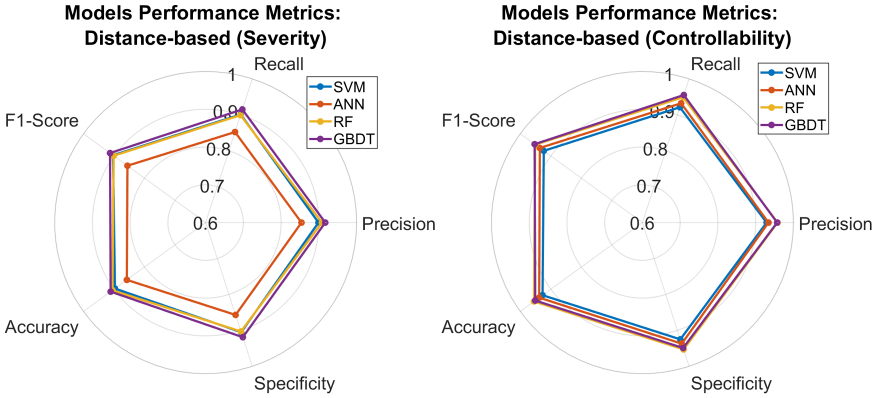


Figure 4.3: Model Performance Metrics: Distance-based Feature

Utilizing only distance-based features, the ML models showed an enhanced ability to predict controllability over severity as shown in Fig. 4.3. This improvement in controllability predictions arises because these features directly relate to the vehicle's capacity to maintain or regain safe positioning relative to other vehicles and obstacles. The RF and GBDT models, known for their strength in handling feature interactions and dependencies, excel in utilizing distance-based features, reflecting their effectiveness in spatial aspect of the vehicle's interaction with its surroundings. The higher performance in controllability suggests that spatial dynamics are more indicative of the vehicle's ability to control or to avoid hazards.

4.2.3 Feature Set III - Comprehensive

The comprehensive-based feature set integrates both time-based and distance-based features along with additional features such as DRAC and KE, offering a holistic view of the risk landscape. This integrative approach is predicated on the understanding that a multifaceted assessment incorporating both the temporal urgency and spatial requirements of risk scenarios, along with the vehicle's physical response capabilities, yields the most accurate risk predictions. This feature set allows for a nuanced risk assessment that considers immediate threats (time-based), spatial safety margins (distance-based), and the vehicle's operational dynamics (deceleration rate and kinetic energy).

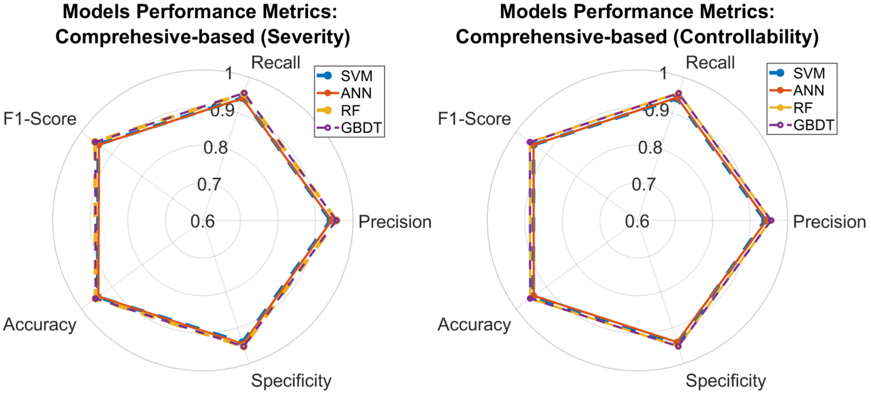


Figure 4.4: Model Performance Metrics: Comprehensive-based Feature

Incorporating a comprehensive feature set resulted in a balanced accuracy for both severity and controllability predictions across all ML models as shown in Fig. 4.4. The observed performance patterns, higher accuracy in predicting severity with time-based features and better controllability predictions with distance-based features, underline the importance of feature selection based on the risk dimension being assessed. This suggests that a blend of temporal, spatial, and impact dynamics provides a robust foundation for assessing risk.

However, performance optimization, potentially through techniques such as Grid Search, is identified as necessary to further enhance the model's performance. This need for optimization suggests that while a comprehensive set of features enriches the model's input, identifying the most effective feature interactions and model parameters is crucial for maximizing predictive performance. This aspect will be discussed in the following section.

4.3 Optimizing Model Performance

In this thesis, the focus was on optimizing four ML models: SVM, RF, GBDT, and ANN, by meticulously selecting and tuning a variety of hyperparameters to achieve the best performance outcomes and Adjust the model.

For the SVM model, optimization focused on the kernel function and the box constraint value (C), using a grid search to evaluate performance across different kernel functions (linear, radial basis function (rbf), and polynomial) and box constraint values ([0.1, 1, 10]). This method allowed the identification of the optimal combination that maximized test set accuracy.

Regarding the RF model, the number of trees hyperparameter was the primary focus, crucial for the model's generalization capabilities without overfitting. After thorough testing, setting the number of trees to [10, 50, 100, 200] was determined to provide a balanced blend of accuracy and computational efficiency.

Table 4.2: Model Performance Metrics: Severity

Model	Accuracy	Precision	Recall	F1-Score	Specificity
GBDT	0.9992	0.9982	0.9993	0.9988	0.9998
ANN	0.9102	0.8679	0.9241	0.8951	0.9718
RF	0.9992	0.9984	0.9993	0.9989	0.9997
SVM	0.9937	0.9905	0.9938	0.9922	0.9976

The ANN model's optimization involved using the Adam optimization algorithm, adjusting the learning rate schedule ([0.001, 0.01, 0.1, 0.2]), and varying the number of epochs ([10, 50, 100]), all of which significantly improved the model's performance.

For the GBDT model, attention was focused on adjusting the learning rate hyperparameter (set at 0.01, 0.05, 0.1, 0.5, 1) to balance capturing the nuances of the training data and preventing overfitting.

Table 4.3: Model Performance Metrics: Controllability

Model	Accuracy	Precision	Recall	F1-Score	Specificity
GBDT	0.9989	0.9982	0.9993	0.9988	0.9996
ANN	0.8690	0.8162	0.8979	0.8551	0.9549
RF	0.9989	0.9981	0.9993	0.9987	0.9997
SVM	0.9907	0.9834	0.9940	0.9887	0.9969

The optimization process for each model was guided by an extensive exploration of various hyperparameters, as detailed in Table 3.7. Due to computational cost considerations, not all potential hyperparameters were computed; instead, a strategic selection was made based on their impact on model performance. This careful selection was crucial in achieving the high accuracy levels documented in our results for severity in Table 4.2 and controllability in Table 4.3 for predictions.

4.4 Validation in Diverse Driving Scenarios

Based on the optimization process and its results, this thesis proceeds to practically validate ML models across various driving scenarios. This deployment phase aimed to Assess how well the models generalize to situations they have not encountered during training, thereby validating their predictions against ground truth. Among all the models, SVM, RF, and GBDT demonstrated good performance. However, for the sake of brevity, the GBDT model was selected for detailed analysis due to its superior performance over the others, as evidenced by the polar plots for both severity and controllability in Fig. 4.5.

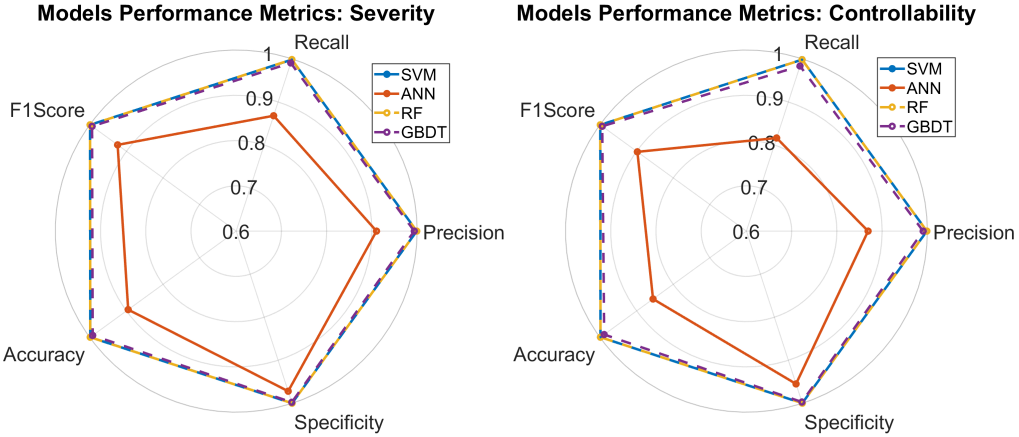


Figure 4.5: Model Performance: Validation of Diverse Driving Scenarios

To illustrate the practical application and efficacy of the GBDT model in dynamic driving conditions, the subsequent analysis focuses on its deployment on an ADS-equipped ego vehicle. This examination specifically assesses the model's precision in forecasting the severity and controllability indicators during complex vehicular maneuvers, including acceleration, deceleration, and speed adaptation in relation to dynamic traffic participants.

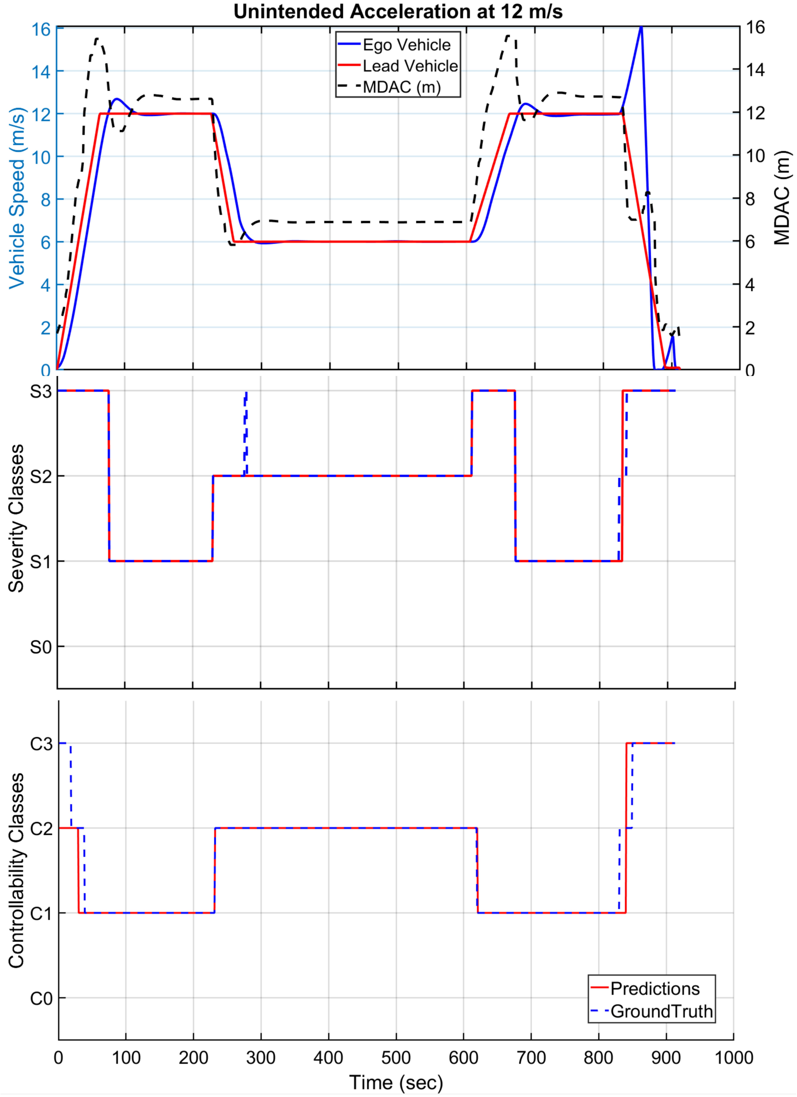


Figure 4.6: Model Validation: Unintended Acceleration at 12 m/s

4.4.1 Ego Vehicle Unintended Acceleration

As depicted in Fig. 4.6, the top plot presents the velocity of both the ego and lead vehicles, alongside the MDAC feature over time. MDAC is a key feature for classifying both severity and controllability. Initially, up to 800 seconds, the ego vehicle maintains a controlled distance from the lead vehicle. However, post-800 seconds, an unintended acceleration occurs, significantly reducing the MDAC and potentially escalating the risk of collision.

The middle and bottom plots extends the further evaluation, showcasing the severity (S0 to S3) and controllability (C0 to C3) indicators, respectively. The severity plot shows varying severity levels, mainly between S1 and S2, indicating that while the situation is generally under control, it becomes critical when unintended acceleration occurs.

Moreover, the controllability classes plot reveals that the model's predictions (red solid line) closely mirror the ground truth (blue dashed line), showcasing the model's adeptness at runtime risk assessment. Before the unintended acceleration event, the controllability ranges between C1 and C2, indicating moderate control. However, after 800 seconds, controllability deteriorates to C3, marking a high-risk scenario with diminished control. Also, the model's predictions are in strong agreement with the ground truth, particularly in critical events like unintended acceleration, highlighting its reliability in runtime risk evaluation.

In an escalated scenario within this thesis, the ego vehicle encounters unintended acceleration at a velocity of 27 m/s (as shown in Fig. 4.7). A deviation in the velocity profile from the previous 12 m/s case is evident, with the spike in speed signaling a severe unintended acceleration, thus presenting a significant challenge for front-end collision avoidance. The resultant sharp decrease in MDAC suggests diminished reaction time, escalating the risk of collision in high-speed conditions.

The model's performance, depicted in in Fig. 4.7, the severity plot, reveals an acute rise to the S3 level following the acceleration, accentuating the event's seriousness. The sudden shift to the highest severity level indicates the possibility of serious consequences due to high-speed irregularities. Concurrently, the controllability classes plot displays intervals with a C3 controllability, indicating severely hampered vehicle handling.

Despite these challenges, the GBDT model's predictions include a noteworthy aspect: the presence of false positives and false negatives is minimal. The model adeptly identifies critical situations, reflecting its robustness and the scalability of its predictive accuracy. This is commendable, considering the dynamic and sometimes unpredictable nature of high-speed unintended acceleration incidents.

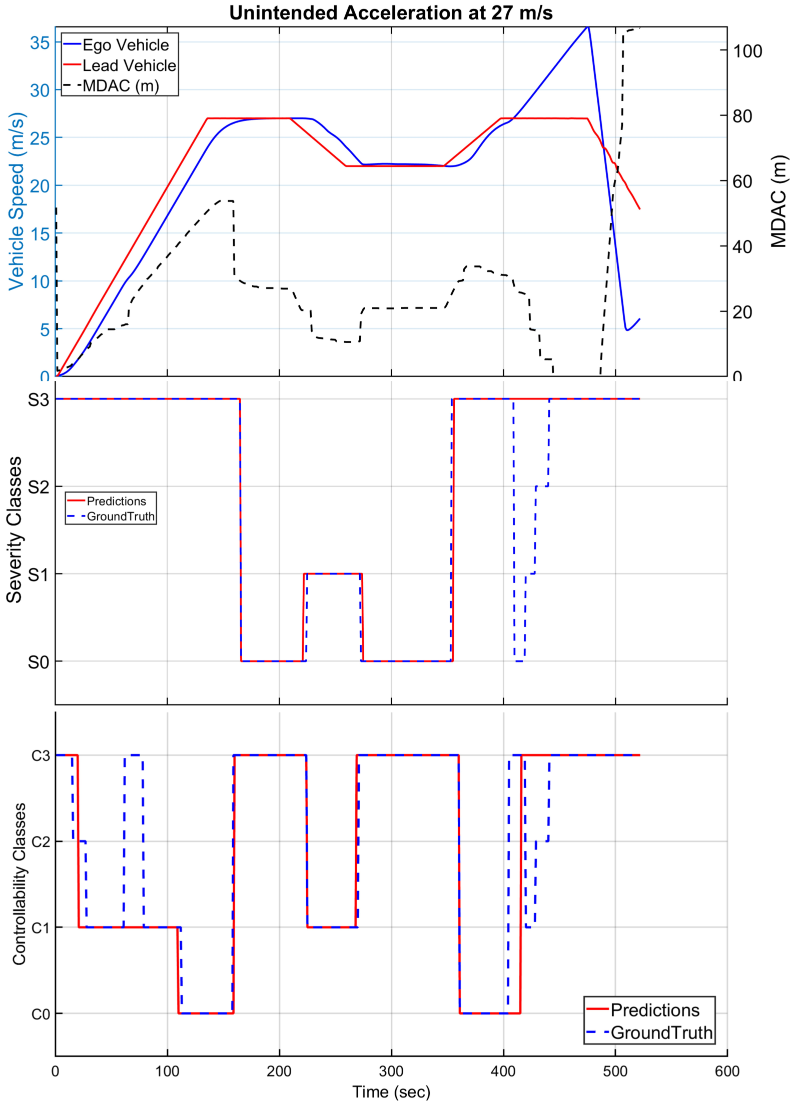


Figure 4.7: Model Validation: Unintended Acceleration at 27 m/s

4.4.2 Lead Vehicle Emergency Braking

In this scenario, the performance of the GBDT model is meticulously evaluated in a lead vehicle emergency braking scenario under heavy rain conditions. This situation introduces additional complexities related to vehicle dynamics and sensor fidelity. Heavy rain not only diminishes visibility but also impacts the road’s friction coefficient, a pivotal element in vehicle dynamics that affects braking behavior and stopping distances.

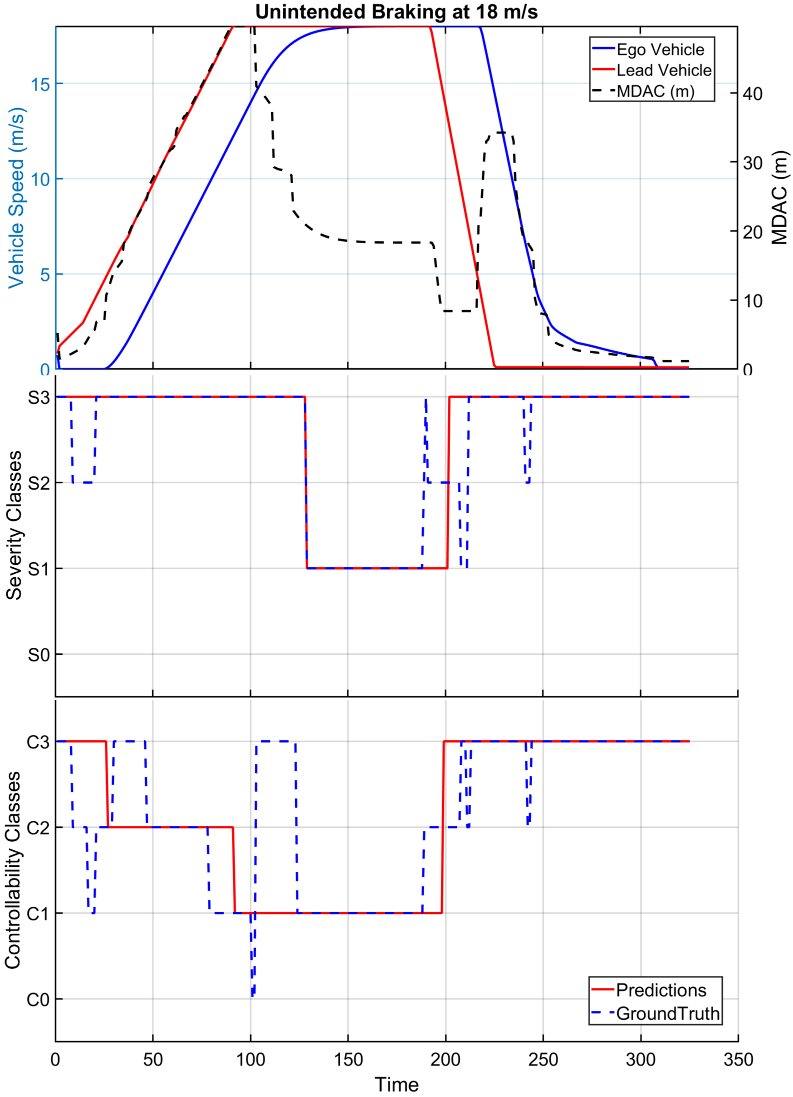


Figure 4.8: Model Validation: Emergency Braking at 18 m/s

The velocity plot distinctly captures the ego vehicle's response to the lead vehicle's emergency braking at 18 m/s. Despite the challenging weather conditions, the ego vehicle successfully increases the MDAC, indicating an effective emergency response. This enhancement suggests that the model takes into account factors of vehicle dynamics, recognizing the extended distances required for coming to a safe halt on wet surfaces.

Furthermore, the severity plot illustrates the model's prompt identification of the escalated risk to S3 during the emergency braking peak. This swift adjustment indicates the model's capacity to integrate runtime data on environmental conditions, modifying risk levels in response to the diminished traction and elongated stopping distances caused by heavy rainfall.

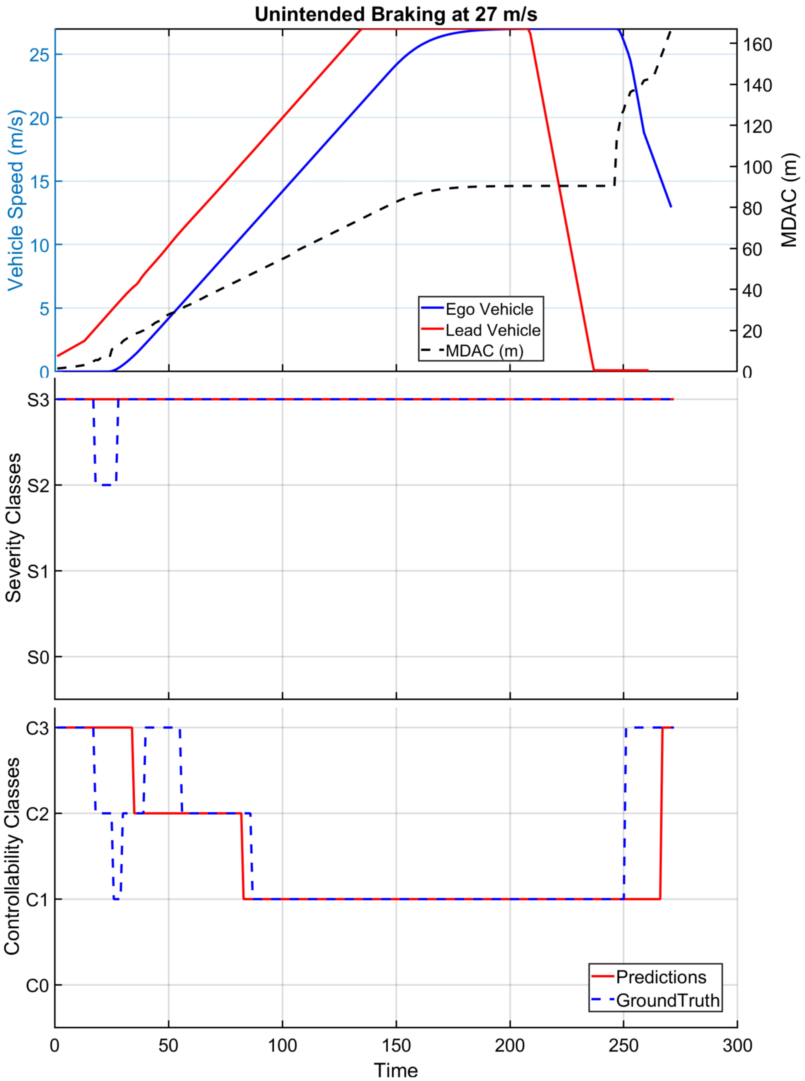


Figure 4.9: Model Validation: Emergency Braking at 27 m/s

In the controllability plot, transient discrepancies between the model's predictions and the ground truth during the emergency braking episode underscore the challenges imposed by dynamic driving conditions. However, the model's swift adjustment to precise controllability evaluations after the incident demonstrates its capability to adapt to changing dynamics caused by wet road conditions.

Similarly, another scenario evaluates the GBDT model's capability in the context of lead vehicle emergency braking at a higher velocity of 27 m/s (as shown in Fig. 4.9). The velocity graph depicts the ego vehicle reaching a higher speed before a drastic deceleration, contrasting with the lead vehicle's more gradual speed reduction. The MDAC increases sharply during the ego vehicle's deceleration, indicating a rapid response to the unexpected event. The severity plot remains at S3 for the duration of the scenario, suggesting that the model consistently recognizes the high-risk potential of sudden deceleration at such speeds. Meanwhile, the controllability plot shows the model's predictions fluctuating before stabilizing, reflecting the challenge of immediate risk classification during abrupt changes in vehicle dynamics.

Both these scenarios underscores the GBDT model's proficiency in managing complex driving situations influenced by adverse weather and dynamic vehicle behaviors, affirming its utility in enhancing vehicular safety under varied environmental conditions. The consistent S3 prediction highlights the model's conservative approach to safety, emphasizing an immediate response to potential high-impact events. However, the model's delayed C3 prediction during rapid high-velocity events highlights the difficulty of effectively managing such critical situations.

4.4.3 Traffic Participants Behavior: Cut-In

The challenge of representing the dynamic nature of a cut-in scenario in 2D plots is difficult, however, the shown Fig. 4.10, presented valuable insights. The green line depicts a side vehicle executing a cut-in maneuver at approximately 400 seconds, compelling the ego vehicle to decelerate abruptly to maintain a safe following distance. The velocity profile prior to the cut-in event is relatively stable, with the ego and lead vehicles maintaining consistent speeds. However, the introduction of the side vehicle at 400 seconds is a critical moment, sharply impacting the velocity of the ego vehicle as it decelerates to avoid a front-end collision. The absence of a trailing vehicle is fortunate, as the scenario could have escalated to include the additional risk of a rear-end collision.

The MDAC peaks following the cut-in, showing the ego vehicle's response to increasing the distance from the side vehicle (which is now lead vehicle) to avoid a crash. The severity and controllability plots demonstrate a fluctuating assessment of risk, with severity reaching S3 and controllability oscillating, as the model reacts to the sudden change in traffic dynamics. The fluctuating controllability levels highlight the model's sensitivity to the rapidly changing situation and the inherent challenges of accurately predicting vehicle behavior in complex scenarios. This evaluation illustrates the intricacies of ADS when addressing real-world driving behaviors. The ability of the GBDT model to adapt to a sudden cut-in, emphasizes its potential to contribute to the development of an advanced risk assessment model that can handle unpredictable traffic events.

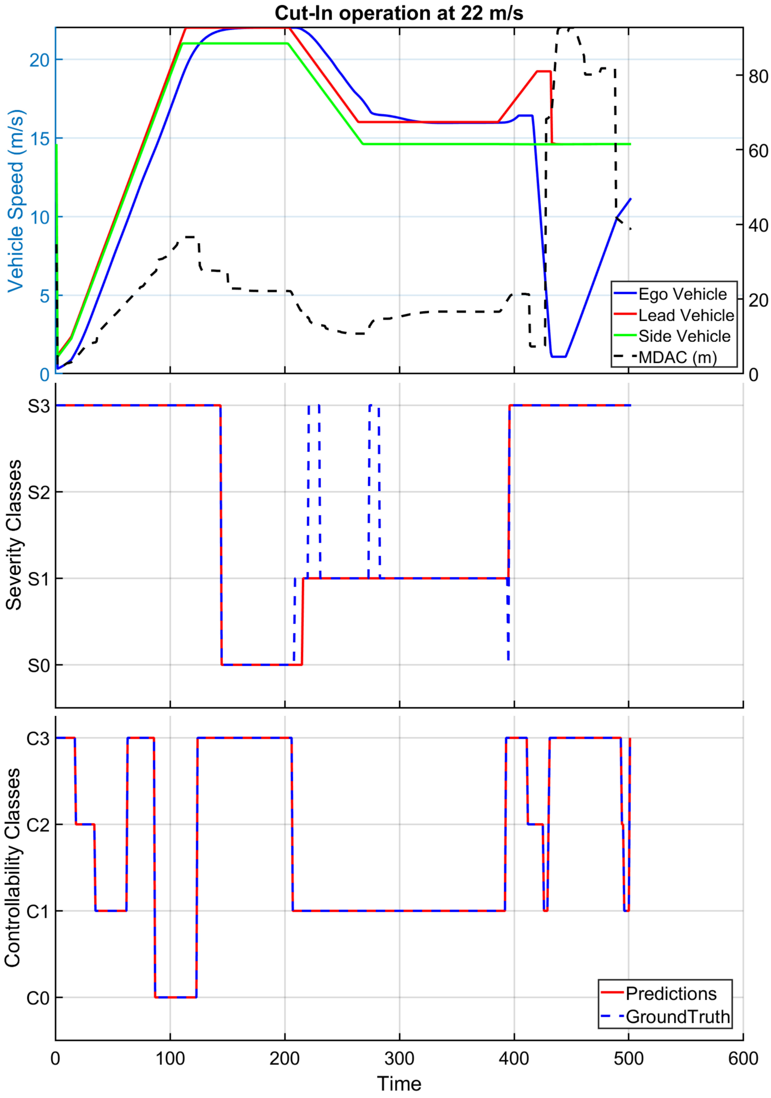


Figure 4.10: Model Validation: Analyzing Traffic Participant Behavior at 22 m/s (Cut-In)

4.4.4 Traffic Participants Behavior: Cut-Out

As shown in the Fig. 4.11, a cut-out scenario at a speed of 22 m/s, a driving event where a leading vehicle moves out of the lane, resulting in an increased following distance (MDAC) for the ego vehicle. The vehicle speed plot shows the ego vehicle decelerating shortly after the lead vehicle's cut-out, which increases the MDAC significantly. This action likely reflects an automated response to suddenly improved road visibility and available space ahead, which in real-world dynamics would allow for a variety of driver responses based on the new driving context. The sever-

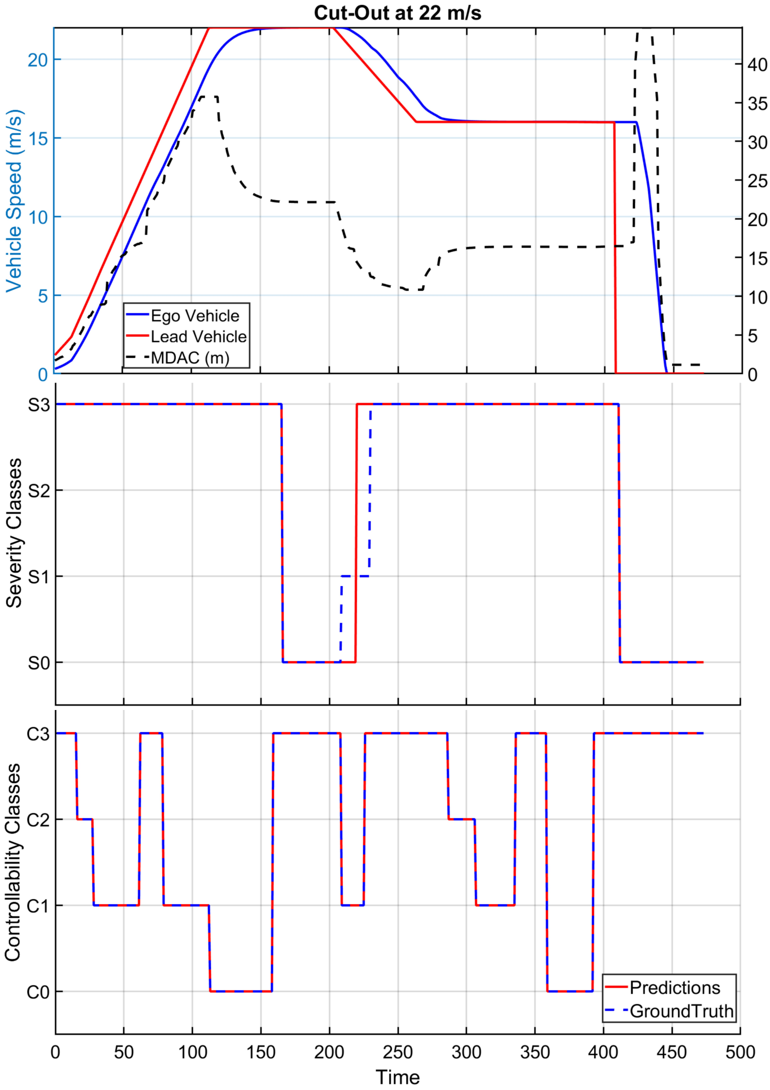


Figure 4.11: Model Validation: Analyzing Traffic Participant Behavior at 22 m/s (Cut-Out)

ity plot shows a decrease to S0 post cut-out, which might initially seem counterintuitive given the increased MDAC. However, in the context of vehicle dynamics, the ego vehicle's risk of front-end collision has decreased substantially due to the lead vehicle's absence, thus reducing the severity assessment.

Contrastingly, the controllability plot shifts to C3 during this period. This could be attributed to the ego vehicle's need to recalibrate its driving strategy after the lead vehicle's maneuver. It might also indicate a response to other objects or vehicles in the environment, suggesting a momentary loss of control or increased difficulty in maintaining the desired trajectory.

This analysis underscores the model's sensitivity to changing driving contexts and its ability to dynamically adjust risk assessments. The shift from a state of higher severity and controllability to S0 and C3 respectively, following a cut-out, reflects a complex interplay between ADS perception and decision-making processes, influenced by the vehicle's dynamics and the immediate driving environment. Understanding the reasons behind these transitions is critical for refining safety strategies for ADS.

4.4.5 Assessment of Model Validation Against Key Quality Criteria

The validation of GBDT model in predicting risk ratings can be summarized by examining several quality aspects:

- Completeness: The model captures various driving scenarios such as unintended acceleration, emergency braking, cut-in, and cut-out maneuvers, indicating a comprehensive approach to scenario coverage. Severity and controllability plots indicate that the model can predict a range of outcomes, suggesting a level of completeness in risk assessment.
- Correctness: The close alignment of predictions with ground truth in several plots suggests high correctness in standard driving conditions. However, discrepancies observed in the controllability during complex maneuvers indicate that correctness may vary with the driving context's complexity.
- Adaptability: The fluctuations in controllability ratings in response to dynamic driving events show the model's adaptability to sudden changes in the driving environment. The model adapts its severity assessment appropriately when MDAC increases due to a cut-out scenario, highlighting its contextual adaptability.
- Reliability: The consistent performance of the model across different scenarios suggests a degree of reliability in its predictive capabilities. Occasional misalignment between predictions and ground truth under certain conditions may affect perceived reliability and warrant further investigation.
- Robustness: The model's robustness is evidenced by its ability to maintain performance despite the introduction of high-speed anomalies and environmental challenges. The model's ability to manage sudden events such as unintended acceleration and emergency braking without critical errors indicates a robust design.
- Variability: The model demonstrates variability in its predictions, as seen by the shifts in severity and controllability classes during different maneuvers. This variability reflects the model's responsiveness to changes in the vehicle's dynamics and environmental conditions, indicating a nuanced understanding of driving behaviors.

- Complexity and Scalability: The model's performance in varied and high-speed scenarios indicates its capability to handle complexity and scalability. The ability to adjust predictions in runtime to events like cut-in and cut-out maneuvers demonstrates the model's scalability to different traffic situations.
- Correlation vs Causality. The model appears to correlate driving behaviors with risk levels effectively; however, it is unclear if it truly understands causality, such as recognizing if increased MDAC is a result of the ego vehicle's actions or other external factors.

4.5 Explaining Risk Feature Importance

In the quest to elucidate the intricacies of model behavior, the Shapley summary plots stand out as an useful tool for model explanation, particularly following the extensive validation of ML models across diverse scenarios. The Shapley value analysis methodically quantifies the contribution of each risk feature towards the predictive outcomes of the model, adhering to principles that ensure the fidelity and fairness of the attribution.

The Shapley values fulfill two critical properties that bolster their utility in model explanation. First, local accuracy guarantees that the simplified explanation model's output aligns with the original model at all feature vectors. Second, consistency demands that if the contribution of a feature to the prediction increases or remains constant, its attributed importance should proportionally reflect this change. As demonstrated in [90], Shapley values uniquely satisfy both local accuracy and consistency. The summation of individual feature Shapley values equates to the overall deviation of the model's prediction for a given sample from the average model prediction, providing a comprehensive understanding of feature impact.

It allows us to identify the specific impact of each feature, whether temporal or spatial, on the model's evaluation of severity and controllability indicators. By leveraging the clarity provided by Shapley values, safety engineer can optimize the LADRI framework to enhance its predictive accuracy, ensuring that it serves as a reliable and transparent tool for evaluating safety-critical situations.

Shapley values can be computed on both the dataset used for training the model and on completely new, unseen datasets. Typically, Shapley values are used to interpret the model's predictions on individual instances, which means they can be applied to explain the contribution of each risk feature to a particular prediction, regardless of whether that data point was part of the training set or not. When applied to training data (scenario S1), Shapley values can provide insights into how the model learned to make predictions based on the training features. In this thesis, Shapley values are calculated for the new unseen data (scenarios S2, S3, S4, S5) to understand how the model generalizes its learned patterns to make predictions on previously unencountered data. This flexibility makes Shapley values a powerful tool for model interpretation and explanation across different scenarios.

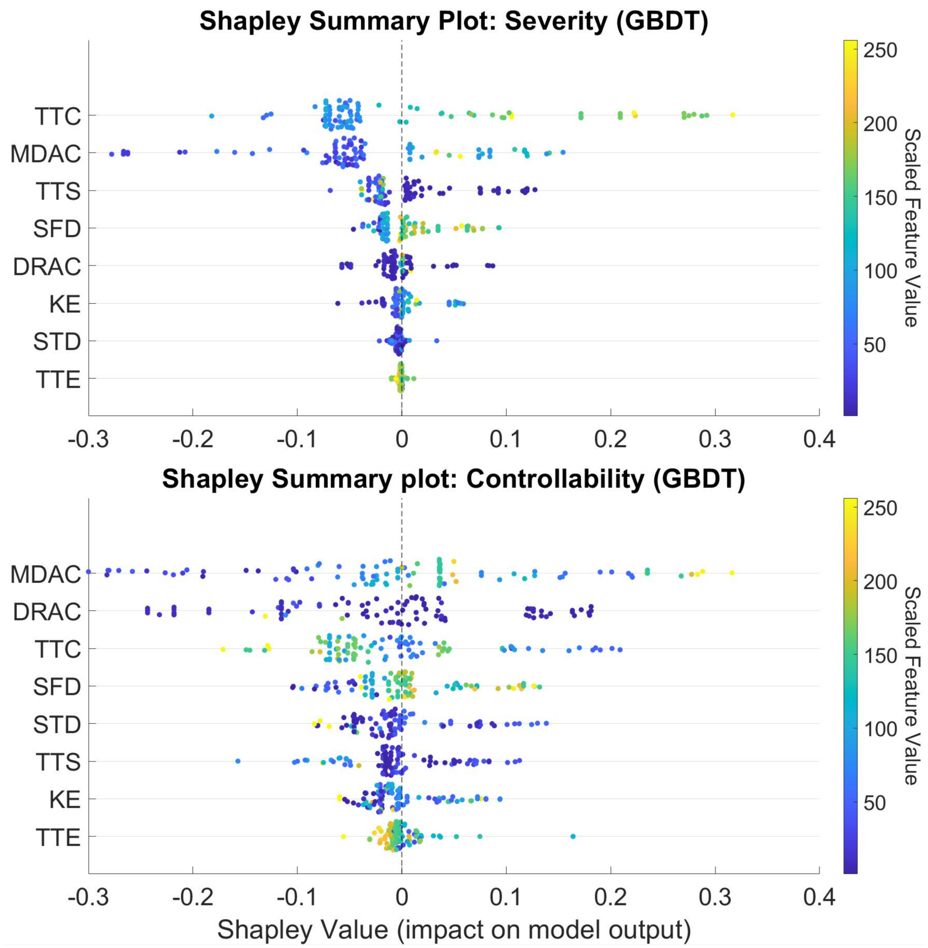


Figure 4.12: Shapley Summary Plot: Severity and Controllability (GBDT Model)

As shown in Fig. 4.12, a Shapley value summary plot visually represents the impact of each feature on a model's prediction. The horizontal axis displays the Shapley values, where positive numbers indicate a feature generally increases the model's prediction, and negative ones suggest a decrease. Vertically, the plot lists each feature, with color-coded points reflecting the feature's value in a specific instance, ranging from low (blue) to high (yellow). The spread of points across the horizontal axis shows the variability of each feature's impact, with a wider spread indicating greater variability. A vertical dashed line at zero serves as a reference, distinguishing features that positively contribute from those that negatively contribute to the model's prediction. Jittering of points aids in discerning the density of overlapping data.

The evaluation of severity using GBDT model highlights the significant impact of MDAC and TTC on severity predictions, as demonstrated by their shapley values being notably distinct from zero. This finding supports the hypothesis that both the proximity of a potential collision and the available space to avoid it are critical factors in assessing the severity of incidents. Furthermore, the variation in shapley values for MDAC and TTC points to a scenario-dependent influence on severity, with a range of driving situations from normal traffic to critical events like unintended acceleration being adeptly handled by the model. This variability, mirrored in the color gradation of the plot, underscores the model's ability to adapt and accurately gauge severity across diverse driving conditions. It showcases model's capability to integrate real-world driving dynamics, where temporal and spatial factors are crucial in determining the severity of incidents.

In the evaluation of controllability using the GBDT model, MDAC emerges as a crucial factor, with its Shapley values indicating a significant impact on controllability assessments, albeit variable. In contrast, the DRAC shows less variability but maintains a consistent influence across scenarios, highlighting its steady contribution to controllability predictions. The GBDT model demonstrates a stable impact from most features on controllability, with MDAC and DRAC as exceptions, suggesting a focused understanding of what determines controllability. The model appears to have learned that controllability is primarily dictated by the ability to decelerate or control acceleration and DRAC, which might relate to vehicle dynamics in response to events like cut-in or cut-out operations.

It is important to note that while MDAC, TTC, and DRAC are crucial, they do not work alone. The GBDT model's performance also hints at the influence of additional, perhaps subtler, risk features that were not directly employed in labeling. This indicates that factors influencing severity and controllability go beyond our primary risk features, highlighting the complexity of driving scenarios and the model's ability to incorporate various inputs to improve predictions.

4.6 Iterative Risk Assessment Process

The LADRI framework's iterative risk assessment process exemplifies a methodical approach to enhancing the safety and reliability of ADS across various ODDs. Each cycle of this process builds on insights and risk indicators from previous evaluations, systematically enhancing safety measures and refining the system's response to potential hazards. As depicted in the described Fig. 4.13, the progression from one ODD to the next, alongside the escalation in driving task complexity, underscores the framework's capacity to adapt and respond to an expanding array of driving scenarios and environmental conditions.

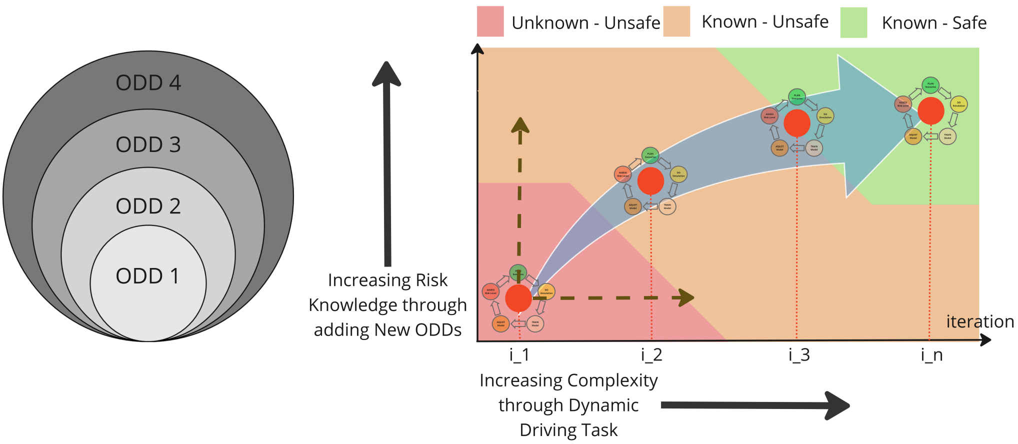

Unknown - Unsafe
Known - Unsafe
Known - Safe
ODD 4
ODD 3
ODD 2
ODD 1
Increasing Risk Knowledge through adding New ODDs
iteration
i_1
i_2
i_3
i_n
Increasing Complexity through Dynamic Driving Task

Figure 4.13: Advancement of Risk Knowledge Through PDTAA process

For example, if highway lane following on a straight road is considered as ODD 1, subsequent ODDs might include curved roads as ODD 2, uphill scenarios as ODD 3, and downhill scenarios as ODD 4, respectively. The iterations on the X-axis correspond to increasing complexity in DDT; for example, the first iteration might cover basic acceleration/deceleration and braking operations, the second iteration could introduce steering maneuvers, the third might involve overtaking maneuvers, and the fourth could include exiting and entering the highway.

The phased exploration of increasing complexity in DDT, depicted along the X-axis, reflects the LADRI framework's strategic approach to gradually introducing and mastering each aspect of the driving environment. This systematic escalation not only assesses the ADS's capabilities in a controlled manner but also ensures that safety engineers and the system can gradually adapt to and learn from each new set of challenges presented.

This iterative nature of the LADRI framework ensures a comprehensive exploration of the ADS's behavior in diverse settings, systematically addressing and mitigating previously unidentified or untested unsafe conditions. By sequentially incorporating new scenarios, ranging from straightforward to complex driving tasks such as steering adjustments, overtaking maneuvers, and navigating exit/entering highway, the framework continually enhances its predictive models. This progression not only broadens the scope of risk assessment but also deepens the system's understanding of dynamic driving challenges.

Moreover, the cyclic repetition of assessing risk indicators (as shown in Fig. 4.14) following safety modifications fosters a continuous improvement loop, which is discussed in detail in the next section. This loop continuously refines ADS functionalities by understanding the unique challenges and specific failures of each ODD. The runtime risk indicators acquired through risk knowledge, can assist safety engineers in augmenting ADS

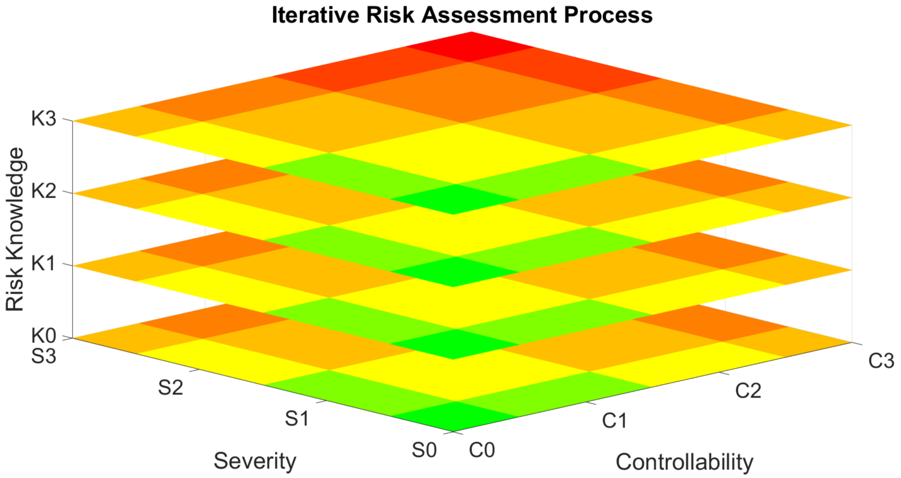


Figure 4.14: Iterative Risk Assessment Process: Severity, Controllability, and Risk Knowledge

design to increase redundancy or incorporate fault-tolerant mechanisms, relying on empirical evidence and objective indicators rather than subjective and biased assessments that do not fully consider system capabilities and environmental factors.

4.7 Application and Limitation of the LADRI Framework

The LADRI framework represents a significant departure from traditional risk assessment methods, standing at the crossroads of safety engineering, vehicle engineering, and ML engineering. It aims to enhance the operational safety of ADS in dynamic environments, particularly focusing on highway lane-following scenarios.

In this thesis, severity and controllability risk ratings are crucial for transitioning from static safety goals to tailored, context-specific safety goals. As shown in Fig. 4.15, these actions are deeply rooted in risk-informed decision-making, utilizing runtime risk assessments to enhance decision-making processes. By leveraging these risk ratings, safety experts can identify underlying causes of risks, bolster safety mechanisms, and improve cross-domain communication. The outputs of the model are utilized to establish specific safety goals that merge quantitative data with expert insights, enabling a thorough review of the ADS to identify potential improvements.

This continuous cycle of evaluation and refinement ensures the ADS not only meets technological standards but also adheres to rigorous safety norms. Additionally, the ongoing optimization of both the ML model and safety protocols, driven by operational feedback, ensures that the framework remains both data-driven and aligned with evolving safety standards.

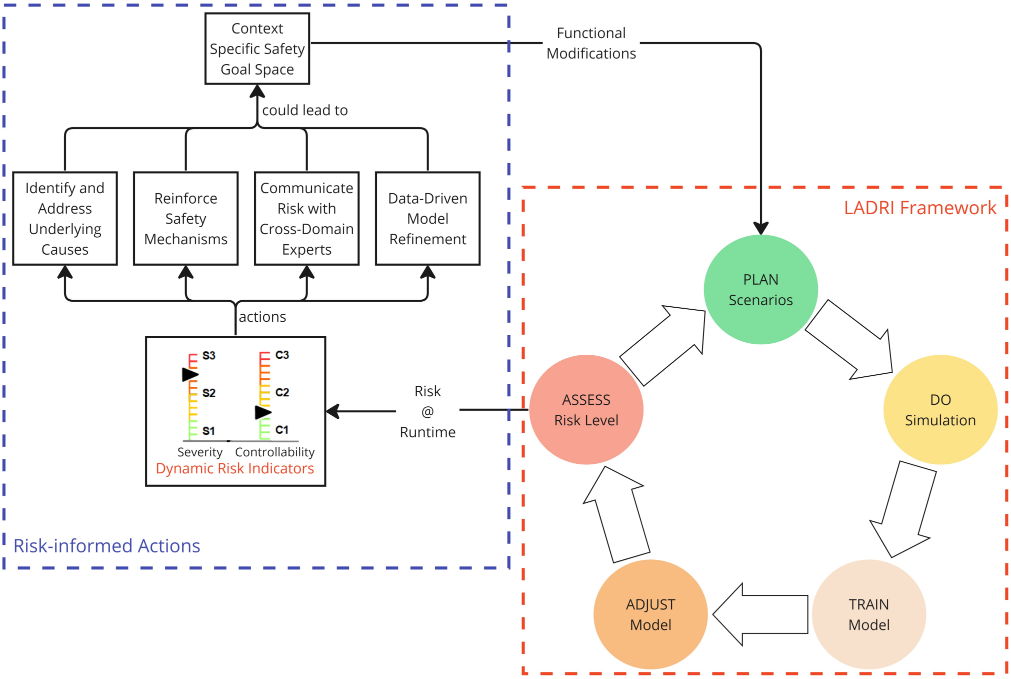


Figure 4.15: Overview of Risk-Informed Actions Enabled by LADRI Framework for Pre-Deployment Safety Enhancement of ADS (adapted from [114])

This dynamic approach ensures a comprehensive integration of empirical evidence and safety values crucial for the effectiveness of ADS. The following four key actions are instrumental in facilitating this transition [114]:

Identifying and Analyzing Underlying Causes: LADRI excels in DRA by adapting to runtime data and changing conditions, such as in highway lane-following scenarios. It dynamically adjusts risk indicators based on the relative movements of the ego, lead, and side vehicles, whether they are accelerating, decelerating, or braking, thus enabling ADS to make informed decisions swiftly. Through detailed analysis of predicted risk ratings, it becomes possible to pinpoint the underlying causes of risks. For example, a high severity rating during highway driving could suggest problems with the vehicle's ACC in specific situations. Such insights guide focused investigations into particular system behaviors or environmental factors that may increase risk levels. However, the efficacy of LADRI is deeply reliant on the quality and immediacy of sensor inputs. In rapidly changing highway environments, any lag in data processing or sensor failure could lead to inaccuracies in risk assessment.

Continuous Improvement and Reinforced Safety Mechanisms: The insights derived from analyzing risk ratings can be used to enhance the safety mechanisms in ADS. When specific scenarios consistently result in high-risk ratings, it indicates a need to strengthen the related safety protocols or system responses. For instance, scenarios that exhibit low controllability could require improvements to the vehicle's emergency braking system or updates to the algorithms that control evasive maneuvers. The framework is designed for iterative updates, enabling the integration of new insights and data. This ensures LADRI remains current with the latest traffic patterns, construction zones, and environmental conditions on highways, facilitating up-to-date risk assessments. However, this continuous improvement cycle requires ongoing data collection and model retraining, which can be resource-intensive and face practical constraints like computational limits, especially in onboard systems during operation.

Communicate Risk and Enhanced Decision-Making: Quantified risk ratings are crucial for effective communication among cross-domain experts. Clear, data-driven risk assessments facilitate targeted and efficient discussions with safety engineers, software developers, and other key stakeholders. Such collaborative efforts are vital for developing an all-encompassing safety strategy, which ensures thorough examination and optimization of every element of the ADS, from software algorithms to hardware reliability. Also, refining severity and controllability indicators to a more granular 5-level scale enhances decision-making by allowing ADS to respond more appropriately to detected hazards. It also improves communication with human operators or other systems by conveying risk levels more precisely. However, this finer granularity introduces challenges such as the need for larger training datasets, more complex calibration of thresholds, potential ambiguity in ratings, and increased demands on sensing and processing capabilities.

Increase Learning Capability through Model Refinement: LADRI leverages ML to enhance risk prediction capabilities over time. Continuous data feeds from highway scenarios allow the system to recognize subtle patterns preceding incidents, thereby improving its predictive accuracy for assessing severity and controllability. Nonetheless, ML models demand extensive and varied data for effective learning. LADRI's learning effectiveness might diminish in unique or rare scenarios where data is scarce, potentially increasing the margin of error in risk predictions. Designed to generalize across various driving scenarios, LADRI is particularly beneficial for highway driving, where conditions can significantly differ. It can apply risk assessments learned from one highway segment to another by recognizing common risk indicators, despite changing external conditions. Achieving true generalization is challenging. Highways feature various environments and behaviors, and models effective in one scenario may not perform well in others without significant adjustments. Additionally, over-generalization can overlook specific risks unique to certain highway segments.

The continuous risk assessment through the LADRI framework, encounters several challenges due to the ADS's complexity and evolving nature. These challenges include the complex interconnectivity of ADS components, from advanced algorithms to runtime data processing modules, which may complicate accurate modeling and prediction of outcomes. Additionally, fully comprehending both functional and non-functional requirements in dynamic environments proves challenging, particularly in scenarios that are unique or rare. The integration of diverse sensors and subsystems essential for precise risk modeling further complicates this task, given the potential for data nuances and inaccuracies. Lastly, the extensive computational demands required to model a wide range of driving conditions and interactions can hinder the process's efficiency, potentially causing delays in risk assessment and decision-making.

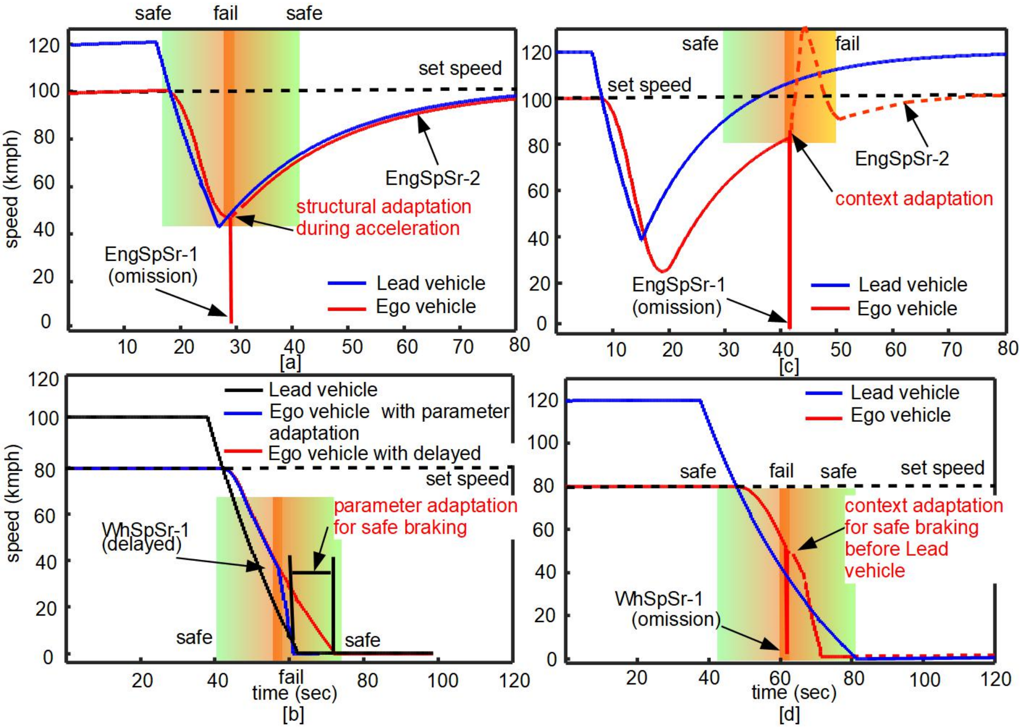


Figure 4.16: Application of LADRI in Adaptation: [a] Structural adaptation during acceleration [b] Parameter adaptation during braking [c] Context adaptation during acceleration [d] Context adaptation during braking [112]

To effectively manage some of these challenges, implementing adaptive techniques as part of the reconfiguration process could help mitigate risk during ADS operation. As discussed in Section 1.2.2, "Why systems need to perform adaptation?" outlines how runtime risk assessment initiate the adaptation process. Similarly, predicted dynamic risk indicators can activate appropriate safety measures, enhancing the vehicle's ability to dynamically respond to predicted risks, as shown in Fig. 4.16. Adaptation in the reconfiguration process involves three techniques: parameter-based reconfiguration to adjust or tune specific parameters, structural reconfiguration to replace a failed component with a redundant one, and context reconfiguration that merges both methods [1].

Utilizing dynamic risk indicators for adaptation not only enhances immediate responses to detected risks but also contributes to the long-term evolution of vehicle safety systems. By continuously learning from the adaptations made and the outcomes achieved, the ADS can refine its algorithms, improve its predictive accuracy, and enhance its overall adaptability. The ability of dynamic risk indicators to provide detailed risk reasoning is vital for the execution of adaptive responses. This reasoning involves analyzing the data behind each indicator to understand the cause of risk and its potential impact. This allows the ADS to prioritize the most necessary and urgent adaptations. This ensures that the system's responses are not only timely but also appropriate to the level of risk encountered, optimizing safety outcomes.

Implementing adaptive reconfiguration techniques in ADS enhances vehicle safety but introduces several challenges. These include the increased complexity of integrating multiple systems, a high dependency on accurate and timely data, and the substantial computational resources required for real-time operations. Additionally, there's a risk of overfitting to specific scenarios, which can compromise the system's effectiveness in new conditions. The testing and validation of these systems are complex due to the unpredictability of driving scenarios, and regulatory standards may struggle to keep pace with rapid technological advancements. Moreover, continuous learning and updates needed to maintain system effectiveness can increase operational costs and logistical complexities. This highlights the importance of managing these adaptive systems carefully to ensure they enhance safety without introducing new risks.

Despite these challenges, the LADRI framework's dynamic, data-driven approach, combined with its learning capabilities, potential for continuous improvement, and ability to generalize across scenarios, stands out. However, it faces obstacles related to data dependency, extensive training needs, the resource requirements for continuous updates, and the complexity of generalization across varied highway scenarios. Balancing these strengths and limitations is crucial for the successful deployment and operation of LADRI framework for ADS development.

Using predicted risk ratings as quantitative outputs lays a strong foundation for developing ADS, but incorporating qualitative aspects improves the system's robustness and relevance before deployment. By quantifying risk indicators such as severity and controllability, developers gain precise metrics to refine and validate the ADS's performance. However, incorporating qualitative assessments allows for a deeper understanding of the contextual nuances that might influence these risk ratings. This dual

[1] The detailed explanation of the behaviors associated with these adaptation techniques is provided in [112]. To maintain the flow and focus of this thesis, these explanations have not been reiterated here.

approach ensures that the ADS not only meets specific numeric risk acceptance thresholds but also aligns with broader safety expectations and scenarios that may not be fully captured through quantitative data alone. Such integration of qualitative insights with quantitative data helps in constructing a comprehensive safety case, making the ADS development process more thorough and aligned with real-world operational needs. This ensures that the system is well-prepared to handle diverse driving conditions and can adapt to unexpected situations, ultimately enhancing safety and reliability before the ADS is deployed.

5 Summary

This thesis summarizing the scientific advancements contributed by recalling research objectives, discussing the challenges encountered during the research and implementation phases of the framework, and outlining potential directions for future research that could further develop or expand upon the contributions of this work.

Ensuring the safety of ADS necessitates an integrated approach that encompasses multiple technical domains, including safety engineering, computing hardware, software algorithms, and human-machine interaction. This interdisciplinary approach is essential for addressing the challenges of validating ADS against unexpected failures and ensuring their reliability for widespread deployment. As automation levels increase, the responsibility for safety transitions from the driver to the system developers. This shift demands a reevaluation of safety assumptions and necessitates the creation of a risk assessment framework to assess system failures across an ADS. Moreover, the complexity of ADS, marked by their reliance on inductive reasoning and the variability of real-world data, poses significant challenges in ensuring system safety. It requires a comprehensive validation strategy that goes beyond traditional testing methods to quantify and address the risks associated with these systems.

5.1 Conclusion

The LADRI framework introduced in this thesis marks a step forward in advancing risk assessment within ADS. By integrating the domains of safety engineering, vehicle engineering, and ML engineering, LADRI emerges as a comprehensive tool designed to navigate the dynamic complexities of operational safety in ADS. Its foundation lies in the PDTAA cyclic process, a methodical approach that ensures continuous refinement and adaptability of risk indicators to ever-evolving operational conditions.

Central to the LADRI framework, the PDTAA cycle encompasses planning simulations, executing these simulations to gather risk-specific context information, training ML models with this data, adjusting the models based on assessment outcomes, and applying these models for ongoing operational risk assessment. This process ensures a perpetual loop of learning and improvement, allowing for the constant evolution of the risk assessment framework in line with the ADS's exposure to new data and operational shifts. By detailing each phase of the PDTAA cycle, Plan, Do, Train, Adjust, and Assess, the framework outlines a clear strategy for integrating a wide array of parameters and potential failure conditions, elucidating the operational dynamics of the LADRI framework and highlighting how each phase feeds into the next, fostering continuous enhancement of risk knowledge.

1. Research Objective:

Conduct a thorough investigation into dynamic factors affecting ADS risks, such as environmental variability, traffic participants and ADS behavior.

This cyclic process systematically integrates both controllable design parameters and uncontrollable non-design parameters alongside potential failure conditions. By considering design parameters, such as sensor configurations and vehicle control strategies, alongside non-design parameters, like environmental conditions and traffic behavior, LADRI offers a comprehensive understanding of the myriad operational risks and challenges ADS may encounter. This holistic view goes beyond the scope of traditional risk assessment methods, providing a more nuanced and thorough perspective on the risk that ADS may face, and facilitates a deeper understanding for a more accurate prediction of risks.

The strategic use of advanced simulation tools facilitates the generation and exploration of a broad spectrum of hypothetical scenarios, enabling proactive identification and assessment of potential risks. By simulating real-world driving conditions of varying complexity and potential failures, the LADRI framework validates itself in a controlled environment and evaluates the effectiveness of different mitigation strategies. This proactive stance towards risk assessment is instrumental in identifying and addressing potential risks before their manifestation in real-world operations, thus contributing to the enhancement of ADS deployment.

2. Research Objective:

Develop and test ML algorithms for assessing risks dynamically, with the goal of making the process automatic, more accurate, and better at adjusting to new situations compared to the current HARA method.

Extensive testing across scenarios from S-curves to straight roads under various conditions has validated the framework's robust capability to predict severity and controllability indicators accurately. The analysis indicates the crucial roles of both time-based and distance-based features in assessing risk, with a holistic integration of these features proving essential for a comprehensive risk assessment approach. The fine-tuning of ML models, particularly the standout performance of the GBDT and RF models, highlights the importance of optimization in achieving high accuracy and reliability in predictions. Practical validation in diverse driving conditions further demonstrates the GBDT model's exceptional adaptability and reliability, reinforcing the LADRI framework's utility in complex environments. These key findings, supported by detailed performance metrics and scenario-based evaluations, establish the LADRI framework as an advancement in the field of risk assessment.

In the Adjust phase, the LADRI framework empowers safety engineers to optimize model performance by updating risk feature thresholds for classification or refining the classification rule set through the addition of new risk features. This phase is crucial for maintaining the framework's relevance and effectiveness in changing conditions and emerging risk scenarios. Moreover, safety engineers have the option to engage in hyperparameter optimization, a process that further increases model performance by fine-tuning the ML algorithms based on the assessment outcomes.

The LADRI framework emphasizes the critical risk indicators of severity and controllability for hazardous events, highlighting key aspects of the risk landscape for ADS. This focus allows the framework to comprehensively address both the impact of hazardous events and the ADS's ability to manage such events. By prioritizing severity and controllability, which are system-dependent attributes, the LADRI framework enhances the operational understanding of specific risks, enabling more targeted safety measures.

The challenge of monitoring ML model performance during runtime risk assessments for ADS, particularly when encountering new, unlabeled data, is addressed through the use of data visualization and model explainability tools, such as Shapley values. This thesis demonstrates the effectiveness of Shapley values in elucidating the impact of critical risk features on severity and controllability assessments. This method not only underscores the significance of spatial and temporal factors in risk assessment but also showcases the model's adaptability to a variety of driving conditions. The identification of influential risk features beyond MDAC, TTC, and DRAC highlights the complexity of driving scenarios and the model's capacity to process a broad spectrum of inputs. This refined understanding of ADS risk assessment, nurtured by the LADRI framework, represents a significant advancement in the field of explainable risk assessment.

3. Research Objective:

Design and validate a comprehensive and iterative framework that incorporates dynamic factors into the risk assessment process for ADS, utilizing runtime data and predictive analysis.

The LADRI framework, through its implementation of an iterative cycle, fosters the continuous enhancement of risk knowledge. This approach ensures that the risk assessment framework evolves in alignment with new insights and the changing operational conditions of ADS. A key feature of this iterative process is its emphasis on continuous learning and adaptation, which allows the framework to remain both relevant and effective over time. This capability is crucial in facilitating ongoing improvements in safety measures and deepening the understanding of ODDs and the complexities of DDT. Each cycle within the PDTAA process is meticulously designed to explore and address previously unknown or unsafe situations across various ODDs, thereby progressively enhancing the safety strategies for ADS. This structured exploration methodically expands the risk knowledge base with each iteration, addressing increasingly complex driving scenarios and potential failure conditions.

5.2 Future Work

Future research on the LADRI framework should focus on deepening the understanding of causal links between risk features and risk indicators, clearly differentiating between correlation and causality using causal inference models. This will refine the framework's capability to identify and mitigate risks more effectively. Expanding the exploration of ODD to encompass complex urban environments, varied weather, and traffic conditions will broaden the framework's applicability. The integration of comprehensive risk features related to vehicle dynamics and environmental characteristics will facilitate nuanced risk assessments across a diverse range of scenarios.

Addressing the challenges of transparency and interpretability associated with "black box" ML models is essential. The development of explainable AI (XAI) methods within the LADRI framework will enhance the transparency of risk assessments, rendering the process more understandable and trustworthy for stakeholders. Optimizing the LADRI framework for computational efficiency, possibly through edge computing solutions and the development of less computationally intensive algorithms, is also critical. A key focus will be on mitigating sensor data uncertainty, requiring advanced filtering techniques to ensure the reliability of risk assessments under various conditions. Enhancing data fusion algorithms will help address discrepancies and inaccuracies in sensor data, supporting reliable risk assessment even in scenarios with incomplete information.

References

[1] S.A. Adedigba, F. Khan, and M. Yang. Dynamic failure analysis of process systems using neural networks. *Process Safety and Environmental Protection*, 111:529–543, 2017.

[2] National Highway Traffic Safety Administration et al. Summary Report: Standing General Order on Crash Reporting for Automated Driving Systems, 2022.

[3] B.W. Al-Mistarehi, A.H. Alomari, R. Imam, and M. Mashaqba. Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS. *Frontiers in Built Environment*, 8:860805, 2022.

[4] B. Ale. Risk analysis and big data. In *Safety and reliability*, volume 36, pages 153–165. Taylor & Francis, 2016.

[5] A. Aleti et al. Identifying Safety-critical Scenarios for Autonomous Vehicles via Key Features. *arXiv preprint arXiv:2212.07566*, 2022.

[6] R. Alexander, D. Kazakov, and T. Kelly. System of systems hazard analysis using simulation and machine learning. In *International Conference on Computer Safety, Reliability, and Security*, pages 1–14. Springer, 2006.

[7] W.K.M. Alhajyaseen. The integration of conflict probability and severity for the safety assessment of intersections. *Arabian Journal for Science and Engineering*, 40: 421–430, 2015.

[8] L. Anthony (Tony) Cox Jr. What's wrong with risk matrices? *Risk Analysis: An International Journal*, 28(2):497–512, 2008.

[9] J. Archer. *Indicators for traffic safety assessment and prediction and their application in micro-simulation modelling: A study of urban and suburban intersections*. PhD thesis, KTH, 2005.

[10] T. Aven. On the meaning of a black swan in a risk context. *Safety science*, 57: 44–51, 2013.

[11] T. Aven and B.S. Krohn. A new perspective on how to understand, assess and manage risk and the unforeseen. *Reliability Engineering & System Safety*, 121: 1–10, 2014.

[12] A. Avizienis, J.-C. Laprie, and B. Randell. Fundamental concepts of dependability. *Department of Computing Science Technical Report Series*, 2001.

[13] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE transactions on dependable and secure computing*, 1(1):11–33, 2004.

[14] V.E. Balas, R. Kumar, R. Srivastava, et al. *Recent trends and advances in artificial intelligence and internet of things*. Springer, 2020.

[15] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22, 2014.

[16] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[17] M.S. Blumenthal, L. Fraade-Blanar, R. Best, and J.L. Irwin. Safe enough. Technical report, RAND Corporation, 2020.

[18] T. Bokaba, W. Doorsamy, and B.S. Paul. Comparative study of machine learning classifiers for modelling road traffic accidents. *Applied Sciences*, 12(2):828, 2022.

[19] Y. Castro and Y.J. Kim. Data mining on road safety: factor assessment on vehicle accidents using classification models. *International journal of crashworthiness*, 21 (2):104–111, 2016.

[20] C. Chen, G. Zhang, Z. Qian, R.A. Tarefder, and Z. Tian. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, 90:128–139, 2016.

[21] W.M.D. Chia, S.L. Keoh, C. Goh, and C. Johnson. Risk assessment methodologies for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(10):16923–16939, 2022.

[22] A. Chinea and M. Parent. Risk assessment algorithms based on recursive neural networks. In *2007 International Joint Conference on Neural Networks*, pages 1434–1440. IEEE, 2007.

[23] F. Cunto. *Assessing safety performance of transportation systems using microscopic simulation*. PhD thesis, University of Waterloo, 2008.

[24] E. de Gelder, A.K. Saberi, and H. Elrofai. A method for scenario risk quantification for automated driving systems. In *26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*. Mira Smart, 2019.

[25] E. De Gelder, H. Elrofai, A.K. Saberi, J.-P. Paardekooper, O.O. Den Camp, and B. De Schutter. Risk quantification for automated driving systems in real-world driving scenarios. *Ieee Access*, 9:168953–168970, 2021.

[26] A. Dixit, R.K. Chidambaram, and Z. Allam. Safety and risk analysis of autonomous vehicles using computer vision and neural networks. *Vehicles*, 3(3):595–617, 2021.

[27] J. Dobaj, C. Schmittner, M. Krisper, and G. Macher. Towards integrated quantitative security and safety risk assessment. In *Computer Safety, Reliability, and Security: SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Turku, Finland, September 10, 2019, Proceedings 38*, pages 102–116. Springer, 2019.

[28] N. Dogru and A. Subasi. Traffic accident detection using random forest classifier. In *2018 15th Learning and Technology Conference (L&T)*, pages 40–45. IEEE, 2018.

[29] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[30] S. Duan and J. Zhao. A model based on hierarchical safety distance algorithm for ACC control mode switching strategy. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 904–908. IEEE, 2017.

[31] A. Duracz, A. Aljarbouh, F.A. Bartha, J. Masood, R. Philippsen, H. Eriksson, J. Duracz, F. Xu, Y. Zeng, and C. Grante. Advanced hazard analysis and risk assessment in the ISO 26262 functional safety standard using rigorous simulation. In *Cyber Physical Systems. Model-Based Design: 9th International Workshop, CyPhy 2019, and 15th International Workshop, WESE 2019, New York City, NY, USA, October 17-18, 2019, Revised Selected Papers 9*, pages 108–126. Springer, 2020.

[32] S. Elyassami, Y. Hamid, and T. Habuza. Road crashes analysis and prediction using gradient boosted and random forest trees. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pages 520–525. IEEE, 2021.

[33] H.M. Fahmy, M.A. Abd El Ghany, and G. Baumann. Vehicle risk assessment and control for lane-keeping and collision avoidance at low-speed and high-speed scenarios. *IEEE Transactions on Vehicular Technology*, 67(6):4806–4818, 2018.

[34] S. Feng, Y. Feng, H. Sun, Y. Zhang, and H.X. Liu. Testing scenario library generation for connected and automated vehicles: An adaptive framework. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):1213–1222, 2020.

[35] P. Feth, R. Adler, T. Fukuda, T. Ishigooka, S. Otsuka, D. Schneider, D. Uecker, and K. Yoshimura. Multi-aspect safety engineering for highly automated driving: Looking beyond functional safety and established standards and methodologies. In *Computer Safety, Reliability, and Security: 37th International Conference, SAFECOMP 2018, Västerås, Sweden, September 19-21, 2018, Proceedings 37*, pages 59–72. Springer, 2018.

[36] P. Feth, M.N. Akram, R. Schuster, and O. Wasenmüller. Dynamic risk assessment for vehicles of higher automation levels by deep learning. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, pages 535–547. Springer, 2018.

[37] International Organization for Standardization (ISO). IEC 61025: Fault Tree Analysis (FTA). Technical report, Technical Report, 2006.

[38] International Organization for Standardization (ISO). ISO 11270: 2014 — Intelligent transport systems — Lane keeping assistance systems (LKAS) — performance requirements and test procedures. *Geneva, Switzerland*, 2014.

[39] International Organization for Standardization (ISO). ISO 9001: 2015 Quality Management Systems-Requirements. *Geneva, Switzerland*, 2015.

[40] International Organization for Standardization (ISO). ISO 15622: 2018 — Intelligent transport systems — adaptive cruise control systems — performance requirements and test procedures. *Geneva, Switzerland*, 2018.

[41] International Organization for Standardization (ISO). ISO 26262: 2018 - Road Vehicles — Functional Safety. *Geneva, Switzerland*, 2018.

[42] International Organization for Standardization (ISO). IEC 31010: 2019 Risk Management — Risk Assessment Techniques. *Geneva, Switzerland*, 2019.

[43] International Organization for Standardization (ISO). ISO 21448: 2019 -Road Vehicles - Safety of the Intended Functionality. *Geneva, Switzerland*, 2019.

[44] L. Fraade-Blanar, M.S. Blumenthal, J.M. Anderson, and N. Kalra. Measuring automated vehicle safety: Forging a framework. Technical report, RAND Corporation, 2018.

[45] T. Goehler. The Application of Explainable AI for Robust and Reliable Dynamic Risk Assessment. Master's thesis, RPTU Kaiserslautern-Landau, 2024.

[46] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. ISBN 9780262035613. URL `http://www.deeplearningbook.org`.

[47] T. Griffon, J.L. Sauvaget, S. Geronimi, and R. Brouwer. Description and taxonomy of automated driving functions. *Deliverable D4: L3Pilot consortium*, 1, 2019.

[48] Z. Halim, M. Sulaiman, M. Waqas, and D. Aydın. Deep neural network-based identification of driving risk utilizing driver dependent vehicle driving features: A scheme for critical infrastructure protection. *Journal of Ambient Intelligence and Humanized Computing*, 14(9):11747–11765, 2023.

[49] S. Hallerbach, Y. Xia, U. Eberle, and F. Koester. Simulation-based identification of critical scenarios for cooperative and automated vehicles. *SAE International Journal of Connected and Automated Vehicles*, 1(2018-01-1066):93–106, 2018.

[50] S. Hallerbach, U. Eberle, and F. Koester. Simulation-Enabled Methods for Development, Testing, and Validation of Cooperative and Automated Vehicles. *Autonomes Fahren Ein Treiber zukünftiger Mobilität*, page 30, 2022.

[51] A. Hanzlik and E. Kristen. Towards a framework for simulation based design, validation and performance analysis of electronic control systems. In *Computer Safety, Reliability, and Security: SAFECOMP 2012 Workshops: Sassur, ASCoMS, DESEC4LCCI, ERCIM/EWICS, IWDE, Magdeburg, Germany, September 25-28, 2012. Proceedings 31*, pages 373–381. Springer, 2012.

[52] J. Hegde and B. Rokseth. Applications of machine learning methods for engineering risk assessment–A review. *Safety Science*, 122:104492, 2020.

[53] J. Hegde, I.B. Utne, and I. Schjølberg. Development of collision risk indicators for autonomous subsea inspection maintenance and repair. *Journal of Loss Prevention in the Process Industries*, 44:440–452, 2016.

[54] M. Hilgers. From Advanced Driver Assistance Systems to Automated Driving. In *Electrical Systems and Mechatronics*, pages 55–71. Springer, 2023.

[55] J. Hillenbrand, K. Kroschel, and V. Schmid. Situation assessment algorithm for a collision prevention assistant. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 459–465. IEEE, 2005.

[56] J. Hillenbrand, A.M. Spieker, and K. Kroschel. A multilevel collision mitigation approach—Its situation assessment, decision making, and performance tradeoffs. *IEEE Transactions on intelligent transportation systems*, 7(4):528–540, 2006.

[57] Q.V.E. Hommes. Review and assessment of the iso 26262 draft road vehicle-functional safety. Technical report, SAE Technical Paper, 2012.

[58] S. Hu and G. Zheng. Driver drowsiness detection with eyelid related parameters by support vector machine. *Expert Systems with Applications*, 36(4):7651–7658, 2009.

[59] D. Hubbard and D. Evans. Problems with scoring methods and ordinal scales in risk assessment. *IBM Journal of Research and Development*, 54(3):2–1, 2010.

[60] Sae International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE Int.*, 4970(724):1–5, 2018.

[61] S.U. Jan, Y.-D. Lee, J. Shin, and I. Koo. Sensor fault classification based on support vector machine and statistical time-domain features. *IEEE Access*, 5:8682–8690, 2017.

[62] J. Jansson. *Collision Avoidance Theory: With application to automotive collision mitigation*. PhD thesis, Linköping University Electronic Press, 2005.

[63] S. Jesenski, N. Tiemann, J.E. Stellet, and J.M. Zöllner. Scalable generation of statistical evidence for the safety of automated vehicles by the use of importance sampling. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2020.

[64] R. Johansson and J. Nilsson. The need for an environment perception block to address all asil levels simultaneously. In *2016 IEEE intelligent vehicles symposium (IV)*, pages 1–4. IEEE, 2016.

[65] R. Johansson, S. Alissa, S. Bengtsson, C. Bergenhem, O. Bridal, A. Cassel, D.-J. Chen, M. Gassilewski, J. Nilsson, A. Sandberg, et al. A strategy for assessing safe use of sensors in autonomous road vehicles. In *Computer Safety, Reliability, and Security: 36th International Conference, SAFECOMP 2017, Trento, Italy, September 13-15, 2017, Proceedings 36*, pages 149–161. Springer, 2017.

[66] C. Johnsson, A. Laureshyn, and T. De Ceunynck. In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators. *Transport reviews*, 38(6):765–785, 2018.

[67] G. Juez, E. Amparan, R. Lattarulo, A. Ruíz, J. Pérez, and H. Espinoza. Early safety assessment of automotive systems using sabotage simulation-based fault injection framework. In *Computer Safety, Reliability, and Security: 36th International*

Conference, SAFECOMP 2017, Trento, Italy, September 13-15, 2017, Proceedings 36, pages 255–269. Springer, 2017.

[68] P.M. Junietz. *Microscopic and macroscopic risk metrics for the safety validation of automated driving*. PhD thesis, Technische Universität Darmstadt, 2019.

[69] C. Katrakazas, M. Quddus, and W.-H. Chen. A new integrated collision risk assessment methodology for autonomous vehicles. *Accident Analysis & Prevention*, 127:61–79, 2019.

[70] S. Kemmann. *SAHARA-a structured approach for hazard analysis and risk assessments*. PhD thesis, Technische Universität Kaiserslautern, 2015.

[71] S. Khastgir, S. Birrell, G. Dhadyalla, H. Sivencrona, and P. Jennings. Towards increased reliability by objectification of Hazard Analysis and Risk Assessment (HARA) of automated automotive systems. *Safety Science*, 99:166–177, 2017.

[72] S. Khastgir, S. Brewerton, J. Thomas, and P. Jennings. Systems approach to creating test scenarios for automated driving systems. *Reliability engineering & system safety*, 215:107610, 2021.

[73] M. Khatun, R. Jung, and M. Glaß. Scenario-based collision detection using machine learning for highly automated driving systems. *Systems Science & Control Engineering*, 11(1):2169384, 2023.

[74] I.-H. Kim, J.-H. Bong, J. Park, and S. Park. Prediction of driver's intention of lane change by augmenting sensor information using machine learning techniques. *Sensors*, 17(6):1350, 2017.

[75] P. Koopman. *How safe is safe enough?: Measuring and predicting Autonomous Vehicle Safety*. Carnegie Mellon University, 2022.

[76] P. Koopman and M. Wagner. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 4(1):15–24, 2016.

[77] P. Koopman and M. Wagner. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1):90–96, 2017.

[78] P. Koopman, B. Osyk, and J. Weast. Autonomous vehicles meet the physical world: RSS, variability, uncertainty, and proving safety. In *Computer Safety, Reliability, and Security: 38th International Conference, SAFECOMP 2019, Turku, Finland, September 11–13, 2019, Proceedings*, volume 38, pages 245–253. Springer International Publishing, 2019.

[79] B. Kramer, C. Neurohr, M. Büker, E. Böde, M. Fränzle, and W. Damm. Identification and quantification of hazardous scenarios for automated driving. In *International symposium on model-based safety and assessment*, pages 163–178. Springer, 2020.

[80] O. Kumtepe, G.B. Akar, and E. Yuncu. Driver aggressiveness detection via multisensory data fusion. *EURASIP Journal on Image and Video Processing*, 2016: 1–16, 2016.

[81] G. Landucci, N. Paltrinieri, et al. Dynamic evaluation of risk: From safety indicators to proactive techniques. *Chemical engineering transactions*, 53:169–174, 2016.

[82] A. Laureshyn, T. De Ceunynck, C. Karlsson, Å. Svensson, and S. Daniels. In search of the severity dimension of traffic events: Extended Delta-V as a traffic conflict indicator. *Accident Analysis & Prevention*, 98:46–56, 2017.

[83] J. Leroy, D. Gruyer, O. Orfila, and N.-E. El Faouzi. Adapted risk indicator for autonomous driving system with uncertainties and multi-dimensional configurations modeling. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2034–2041. IEEE, 2021.

[84] N.G. Levenson. Engineering a safer world. *Systems Thinking Applied to Safety, The MIT Press, Cambridge, MA, USA*, 2011.

[85] G. Li, Y. Yang, T. Zhang, X. Qu, D. Cao, B. Cheng, and K. Li. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. *Transportation research part C: emerging technologies*, 122:102820, 2021.

[86] Z. Li, P. Liu, W. Wang, and C. Xu. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*, 45:478–486, 2012.

[87] Z. Li, I. Kolmanovsky, E. Atkins, J. Lu, D.P. Filev, and J. Michelini. Road risk modeling and cloud-aided safety-based route planning. *IEEE Transactions on Cybernetics*, 46 (11):2473–2483, 2015.

[88] Y. Liang, M.L. Reyes, and J.D. Lee. Real-time detection of driver cognitive distraction using support vector machines. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):340–350, 2007.

[89] P. Liu, L. Wang, and C. Vincent. Self-driving vehicles against human drivers: Equal safety is far from enough. *Journal of Experimental Psychology: Applied*, 26(4):692, 2020.

[90] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[91] Y. Ma, M. Chowdhury, A. Sadek, and M. Jeihani. Real-time highway traffic condition assessment framework using vehicle–infrastructure integration (VII) with artificial intelligence (AI). *IEEE Transactions on Intelligent Transportation Systems*, 10 (4):615–627, 2009.

[92] M. Machin, F. Dufossé, J. Guiochet, D. Powell, M. Roy, and H. Waeselynck. Model-checking and game theory for synthesis of safety rules. In *2015 IEEE 16th International Symposium on High Assurance Systems Engineering*, pages 36–43. IEEE, 2015.

[93] K. Madala, H. Do, and C. Avalos-Gonzalez. A Dependency-based Combinatorial Approach for Reducing Effort for Scenario-based Safety Analysis of Autonomous Vehicles. In *VEHITS*, pages 235–246, 2021.

[94] S.M.S. Mahmud, L. Ferreira, M.S. Hoque, and A. Tavassoli. Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. *IATSS research*, 41(4):153–163, 2017.

[95] H. Martin, K. Tschabuschnig, O. Bridal, and D. Watzenig. Functional safety of automated driving systems: Does ISO 26262 meet the challenges? In *Automated Driving: Safer and More Efficient Future Driving*, pages 387–416. Springer, 2016.

[96] M. Mauritz, F. Howar, and A. Rausch. Assuring the safety of advanced driver assistance systems through a combination of simulation and runtime monitoring. In *Leveraging Applications of Formal Methods, Verification and Validation: Discussion, Dissemination, Applications: 7th International Symposium, ISoLA 2016, Imperial, Corfu, Greece, October 10-14, 2016, Proceedings, Part II 7*, pages 672–687. Springer, 2016.

[97] J.A. McDermid, M. Nicholson, D.J. Pumfrey, and P. Fenelon. Experience with the application of HAZOP to computer-based systems. In *COMPASS'95 Proceedings of the Tenth Annual Conference on Computer Assurance Systems Integrity, Software Safety and Process Security'*, pages 37–48. IEEE, 1995.

[98] M. Megnidio-Tchoukouegno and J.A. Adedeji. Machine learning for road traffic accident improvement and environmental resource management in the transportation sector. *Sustainability*, 15(3):2014, 2023.

[99] N. Merat, B. Seppelt, T. Louw, J. Engström, J.D. Lee, E. Johansson, C.A. Green, S. Katazaki, C. Monk, M. Itoh, et al. The "Out-of-the-Loop" concept in automated driving: proposed definition, measures and implications. *Cognition, Technology & Work*, 21:87–98, 2019.

[100] H. Monkhouse, I. Habli, and J. Mcdermid. The Notion of Controllability in an autonmous vehicle context. In *CARS 2015-Critical Automotive applications: Robustness & Safety*, 2015.

[101] G.E. Mullins, P.G. Stankiewicz, and S.K. Gupta. Automated generation of diverse and challenging scenarios for test and evaluation of autonomous vehicles. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1443–1450. IEEE, 2017.

[102] P. Mundt, I. Kumara, W.-J. Van Den Heuvel, D.A. Tamburri, and A.S Andreou. Knowgo: An adaptive learning-based multi-model framework for dynamic automotive risk assessment. In *International Symposium on Business Modeling and Software Design*, pages 268–278. Springer, 2022.

[103] R. Nahata, D. Omeiza, R. Howard, and L. Kunze. Assessing and explaining collision risk in dynamic environments for autonomous driving safety. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 223–230. IEEE, 2021.

[104] D. Nalic, T. Mihalj, F. Orucevic, M. Schabauer, C. Lex, W. Sinz, and A. Eichberger. Criticality assessment method for automated driving systems by introducing fictive vehicles and variable criticality thresholds. *Sensors*, 22(22):8780, 2022.

[105] C. Neurohr, L. Westhofen, T. Henning, T. De Graaff, E. Möhlmann, and E. Böde. Fundamental considerations around scenario-based testing for automated driving. In *2020 IEEE intelligent vehicles symposium (IV)*, pages 121–127. IEEE, 2020.

[106] K. Øien. Risk indicators as a tool for risk control. *Reliability Engineering & System Safety*, 74(2):129–145, 2001.

[107] K. Ozbay, H. Yang, B. Bartin, and S. Mudigonda. Derivation and validation of new simulation-based surrogate safety measure. *Transportation research record*, 2083 (1):105–113, 2008.

[108] N. Paltrinieri, L. Comfort, and G. Reniers. Learning about risk: Machine learning for risk assessment. *Safety science*, 118:475–486, 2019.

[109] A.R. Patel and P. Liggesmeyer. Machine learning based dynamic risk assessment for autonomous vehicles. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pages 73–77. IEEE, 2021.

[110] A.R. Patel and P. Liggesmeyer. LADRI: LeArning-based Dynamic Risk Indicator in Automated Driving System. *arXiv preprint arXiv:2401.02199*, 2024.

[111] A.R. Patel, N.B. Haupt, and P. Liggesmeyer. Exploiting Adaptation Behavior of an Autonomous Vehicle to Achieve Fail-Safe Reconfiguration. In *Commercial Vehicle Technology 2020/2021: Proceedings of the 6th Commercial Vehicle Technology Symposium*, pages 389–402. Springer Fachmedien Wiesbaden, 2021.

[112] A.R. Patel, N.B. Haupt, and P. Liggesmeyer. A Conceptual Framework of Dynamic Risk Management for Autonomous Vehicles. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 475–486. IOS Press, 2022.

[113] A.R. Patel, S. Gorasiya, and P. Liggesmeyer. Dynamic Risk Assessment for Automated Driving System using Artificial Neural Network. In *Commercial Vehicle Technology 2024: Proceedings of the 8th Commercial Vehicle Technology Symposium*. Springer Fachmedien Wiesbaden, 2024.

[114] A.R. Patel, K. Thummar, and P. Liggesmeyer. Continuous Risk Assessment for Automated Driving Systems using Random Forest. In *CARS 2024*, 2024.

[115] Đ. Petrović, R. Mijailović, and D. Pešić. Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. *Transportation research procedia*, 45:161–168, 2020.

[116] A.B.J. Rafael and Z. Bachir. SAHARA: Simulation aided hazard analysis and risk assessment methodology. *Risk Analysis XII*, 129:41, 2020.

[117] J. Reich and M. Trapp. SINADRA: towards a framework for assurable situation-aware dynamic risk assessment of autonomous vehicles. In *2020 16th European dependable computing conference (EDCC)*, pages 47–50. IEEE, 2020.

[118] J. Reich, M. Wellstein, I. Sorokos, F. Oboril, and K.-U. Scholl. Towards a software component to perform situation-aware dynamic risk assessment for autonomous vehicles. In *Dependable Computing-EDCC 2021 Workshops: DREAMS, DSOGRI,*

SERENE 2021, Munich, Germany, September 13, 2021, Proceedings 17, volume 17, pages 3–11. Springer International Publishing, 2021.

[119] SAE International. Taxonomy & Definitions for Operational Design Domain (ODD) for Driving Automation Systems J3259, 2024. URL `https://www.sae.org/standards/content/j3259/`.

[120] R. Salay and K. Czarnecki. Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262. *arXiv preprint arXiv:1808.01614*, 2018.

[121] K. Saleh, M. Hossny, and S. Nahavandi. Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2017.

[122] M. Salehie and L. Tahvildari. Self-adaptive software: Landscape and research challenges. *ACM transactions on autonomous and adaptive systems (TAAS)*, 4(2): 1–42, 2009.

[123] A. Schoeters, N. De Vos, and F. Slootmans. European Road Safety Observatory National Road Safety Profile-Cyprus. Report, European Road Safety Observatory, 2021.

[124] X. Shi, Y.D. Wong, M.Z.-F. Li, C. Palanisamy, and C. Chai. A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention*, 129:170–179, 2019.

[125] S. Singh. Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. Technical report, Traffic Safety Facts Crash Stats, Report No. DOT HS 812 115, Washington, DC, 2015.

[126] J. Stilgoe. How can we know a self-driving car is safe? *Ethics and Information Technology*, 23(4):635–647, 2021.

[127] M. Taamneh, S. Taamneh, and S. Alkheder. Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks. *International journal of injury control and safety promotion*, 24(3):388–395, 2017.

[128] A. Tamke, T. Dang, and G. Breuel. A flexible method for criticality assessment in driver assistance systems. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 697–702. IEEE, 2011.

[129] The MathWorks Inc. MATLAB version: 9.13.0 (R2022b). Natick, Massachusetts: The MathWorks Inc., 2022. URL `https://www.mathworks.com`.

[130] S. Thomas and K.M. Groth. Toward a hybrid causal framework for autonomous vehicle safety analysis. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 237(2):367–388, 2023.

[131] M. Trapp and G. Weiß. Towards dynamic safety management for autonomous systems. In *Safety-Critical Systems Symposium*, 2019.

[132] M. Trapp, D. Schneider, and G. Weiss. Towards safety-awareness and dynamic safety management. In *2018 14th European Dependable Computing Conference*, pages 107–111. IEEE, 2018.

[133] Q. Van Eikema Hommes et al. Assessment of safety standards for automotive electronic control systems. Technical report, United States. Department of Transportation. National Highway Traffic Safety . . . , 2016.

[134] S. Wagner, K. Groh, T. Kuhbeck, M. Dorfel, and A. Knoll. Using time-to-react based on naturalistic traffic object behavior for scenario-based risk assessment of automated driving. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1521–1528. IEEE, 2018.

[135] H. Wang, M. Gu, S. Wu, and C. Wang. A driver's car-following behavior prediction model based on multi-sensors data. *EURASIP Journal on Wireless Communications and Networking*, 2020:1–12, 2020.

[136] Y. Wang and J. Kato. Collision risk rating of traffic scene from dashboard cameras. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2017.

[137] F. Warg, M. Gassilewski, J. Tryggvesson, V. Izosimov, A. Werneman, and R. Johansson. Defining autonomous functions using iterative hazard analysis and requirements refinement. In *Computer Safety, Reliability, and Security: SAFECOMP 2016 Workshops, ASSURE, DECSoS, SASSUR, and TIPS, Trondheim, Norway, September 20, 2016, Proceedings 35*, pages 286–297. Springer, 2016.

[138] F. Warg, M. Skoglund, A. Thorsén, R. Johansson, M. Brännström, M. Gyllenhammar, and M. Sanfridson. The quantitative risk norm-a proposed tailoring of HARA for ADS. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pages 86–93. IEEE, 2020.

[139] F. Warg, A. Thorsén, V. Vu, and C. Bergenhem. A unified taxonomy for automated vehicles: Individual, cooperative, collaborative, on-road, and off-road. *arXiv preprint arXiv:2304.02705*, 2023.

[140] H. Weber, J. Bock, J. Klimke, C. Roesener, J. Hiller, R. Krajewski, A. Zlocki, and L. Eckstein. A framework for definition of logical scenarios for safety assurance of automated driving. *Traffic injury prevention*, 20(sup1):S65–S70, 2019.

[141] L. Westhofen, C. Neurohr, T. Koopmann, M. Butz, B. Schütt, F. Utesch, B. Neurohr, C. Gutenkunst, and E. Böde. Criticality metrics for automated driving: A review and suitability analysis of the state of the art. *Archives of Computational Methods in Engineering*, 30(1):1–35, 2023.

[142] H. Winner, S. Hakuli, F. Lotz, and C. Singer, editors. *Handbook of Driver Assistance Systems*. Springer International Publishing, Amsterdam, The Netherlands, 2014.

[143] H. Winner, K. Lemmer, T. Form, and J. Mazzega. PEGASUS—First steps for the safe introduction of automated driving. In *Road Vehicle Automation 5*, pages 185–195. Springer, 2019.

[144] H. Woo, H. Madokoro, K. Sato, Y. Tamura, A. Yamashita, and H. Asama. Advanced adaptive cruise control based on operation characteristic estimation and trajectory prediction. *Applied Sciences*, 9(22):4875, 2019.

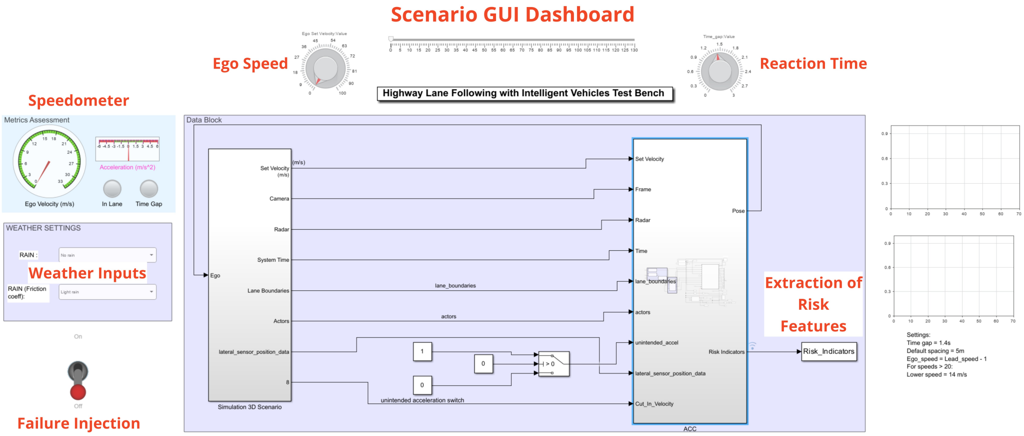
[145] X. Yan, J. He, C. Zhang, Z. Liu, B. Qiao, and H. Zhang. Single-vehicle crash severity outcome prediction and determinant extraction using tree-based and other non-parametric models. *Accident Analysis & Prevention*, 153:106034, 2021.

[146] X. Zhang, J. Tao, K. Tan, M. Törngren, J.M.G. Sanchez, M.R. Ramli, X. Tao, M. Gyllenhammar, F. Wotawa, N. Mohan, et al. Finding critical scenarios for automated driving systems: A systematic mapping study. *IEEE Transactions on Software Engineering*, 49(3):991–1026, 2022.

[147] E. Zio. The future of risk assessment. *Reliability Engineering & System Safety*, 177: 176–190, 2018.

A Appendix

A.1 Scenario Creation Dashboard



Settings:
Time gap = 1.4s
Default spacing = 5m
Ego_speed = Lead_speed - 1
For speeds > 20:
Lower speed = 14 m/s

Figure A.1: Interactive Dashboard for Scenario Modification and Failure Injection

A.2 Ego Vehicle Dynamics Model

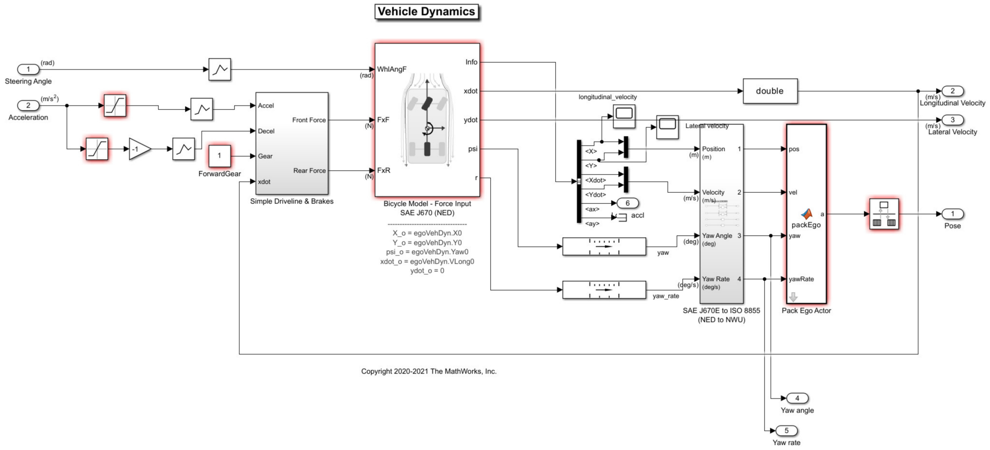


Figure A.2: Modified Vehicle Dynamics Model [129]

A.3 Simulation 3D Scenario

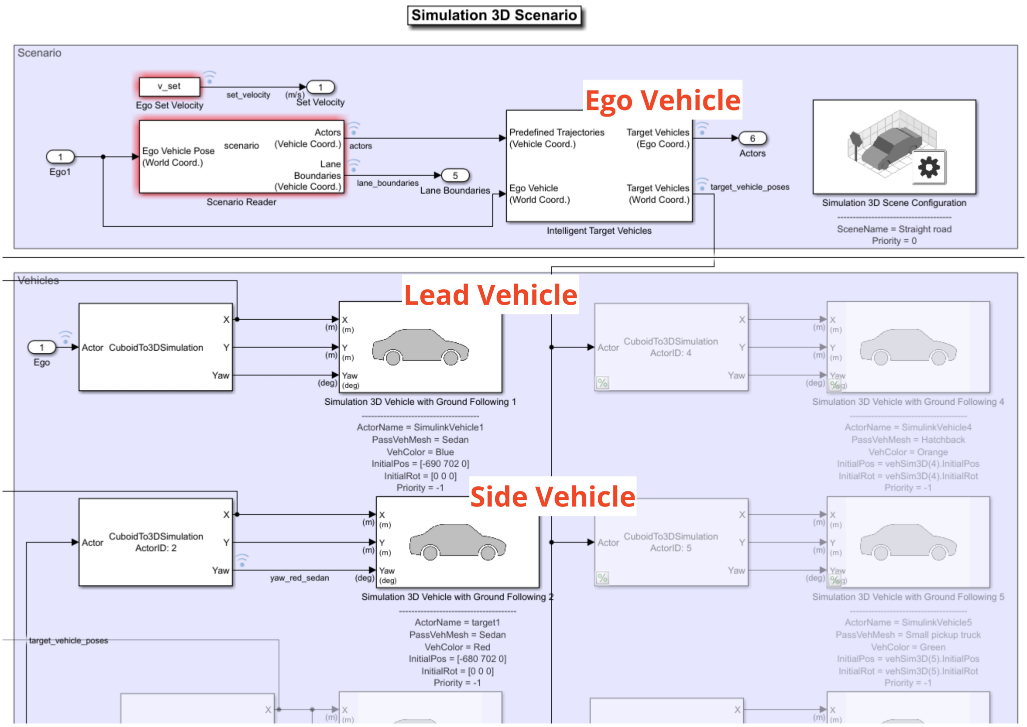


Figure A.3: Simulink Model of 3D Scenarios

A.4 Risk Feature Extraction

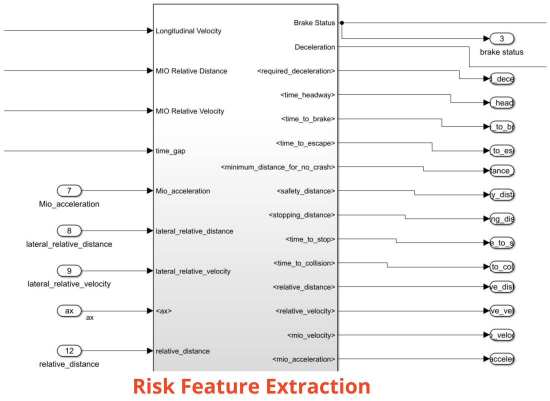


Figure A.4: Risk Feature Extraction block in Simulink

A.5 Matlab Function for Supervised ML Classification

```
svmModel = fitcecoc(X_train, y_train, 'Learners', template, 'Coding', 'onevsall', 'ClassNames', unique(y_train));
```

Listing A.1: MATLAB function for SVM

```
net = trainNetwork(X_train, y_train, layers, options);
```

Listing A.2: MATLAB function for ANN

```
rfModel = TreeBagger(numTrees, X_train, y_train, 'Method', 'classification');
```

Listing A.3: MATLAB function for RF

```
gbdtModel = fitcensemble(X_train, y_train, 'Method', 'Bag', 'NumLearningCycles', 30);
```

Listing A.4: MATLAB function for GBDT

List of Publication

1. Patel, A. R., Göhler, T., Liggesmeyer, P. (2024). Integrating Explainable AI to Enhance Dynamic Risk Assessment in Automated Driving Systems. In 18th IEEE International Conference on Vehicular Electronics and Safety, IEEE.
2. Patel, A. R., Dhingra, A., Liggesmeyer, P. (2024). Leveraging the Brier Score to Enhance Predictive Accuracy in Learning-Based Risk Assessment. In 8th International Conference on System Reliability and Safety, IEEE.
3. Patel, A. R., Shah, T. M., Liggesmeyer, P. (2024). Adaptive Risk Feature Thresholds in Automated Driving Systems: A Deep Q-Learning Approach. In 27th IEEE International Conference on Intelligent Transportation Systems, IEEE.
4. Patel, A. R., Liggesmeyer, P. (2024). Enhancing Continuous Risk Assessment: The Role of Safety Engineers in Early Hazard Identification. In 2nd International Workshop on Verification and Validation of Dependable Cyber-Physical Systems, in 54th Annual Conference on Dependable Systems and Networks (DSN-W), IEEE.
5. Patel, A. R., Thummar, K., Liggesmeyer, P. (2024). Dynamic Risk Assessment: Leveraging Ensemble Learning for Context-Specific Risk Features. In 35th IEEE Intelligent Vehicle Symposium, IEEE.
6. Patel, A. R., Thummar, K., Liggesmeyer, P. (2024). Continuous Risk Assessment for Automated Driving Systems using Random Forest. In 8th Workshop on Critical Automotive Applications: Robustness and Safety (CARS), in 19th European Dependable Computing Conference, HAL science.
7. Patel, A. R., Gorasiya, S., Liggesmeyer, P. (2024). Dynamic Risk Assessment for Automated Driving System using Artificial Neural Network. In 8th International Commercial Vehicle Symposium Kaiserslautern, Springer.
8. Patel, A. R., Liggesmeyer, P. (2024). Adaptive Safety Measures: A Concept to Optimize Safety in Automated Driving Systems. In 32nd Safety-Critical Systems Symposium (Poster)
9. Patel, A. R., Liggesmeyer, P. (2023). LADRI: LeArning-based Dynamic Risk Indicator in Automated Driving System. In 8th IEEE Automotive Reliability, Test and Safety (ARTS) Workshop in International Test Conference, IEEE.
10. Patel, A. R., Haupt, N. B., Adler. R., Elberzhager. F., Liggesmeyer, P. (2023). Exploring Safety Challenges in Dynamic Systems-of-Systems for Flood Management. In 18th Annual System of Systems Engineering Conference, IEEE.
11. Patel, A. R., Haupt, N. B., Liggesmeyer, P. (2022). A Conceptual Framework of Dynamic Risk Management for Autonomous Vehicles. In New Trends in Intelligent Software Methodologies, Tools and Techniques (pp. 475-486). IOS Press.
12. Patel, A. R., Haupt, N. B., Liggesmeyer, P. (2021). Prediction of Dynamic Adaptation Technique for Autonomous Vehicles using Decision Trees. In 29th Safety-Critical Systems Symposium (Poster).

13. Patel, A. R., Liggesmeyer, P. (2021). Machine Learning based Dynamic Risk Assessment for Autonomous Vehicles. In 2021 International Symposium on Computer Science and Intelligent Controls, IEEE.

14. Patel, A. R., Haupt, N. B., Liggesmeyer, P. (2020) Exploiting Adaptation Behavior of an Autonomous Vehicle to Achieve Fail-Safe Reconfiguration. In 6th International Commercial Vehicle Technology Symposium Kaiserslautern, Springer.

Lebenslauf / Curriculum Vitae

Name	Anil Ranjitbhai Patel
Schulbildung	N.G.Zaveri Jain High School, Surat
1995 – 2005	Abitur
Studium	Maschinenbau
2005 – 2009	Gujarat Technological University
	Diplom
Studium	Maschinenbau
2009 – 2012	Gujarat Technological University
	B.Sc
Studium	Commercial Vehicle Technology
2015 – 2019	Technische Universität Kaiserslautern
	M.Sc
Berufstätigkeit	Wissenschaftlicher Mitarbeiter
2019 – 2024	Lehrstuhl Software Engineering
	Fachbereich Informatik der RPTU Kaiserslautern-Landau