Comprehensive Study on Context-Aware and Behavioural Anomaly Risk Assessment and Prevention System : A Hybrid AI-Driven Approach to Risky Driving Pattern Detection

by

S. M. Shakil Ahmed
21201011
Nahiyan Tabassum
21201102
Nusrat Zahan Haque
21201722
Nusrat Mahmud
21201509

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
April 2026

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

S. M. Shakil Ahmed
21201011

Nahiyan Tabassum
21201102

Nusrat Zahan Haque
21201722

Nusrat Mahmud
21201509

Approval

The thesis/project titled "Comprehensive Study on Context-Aware and Behavioural Anomaly Risk Assessment and Prevention System: A Hybrid AI-Driven Approach to Risky Driving Pattern Detection" submitted by

1. S. M. Shakil Ahmed (21201011)
2. Nahiyan Tabassum (21201102)
3. Nusrat Zahan Haque (21201722)
4. Nusrat Mahmud (21201509)

of Spring, 2026 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in April 12, 2026.

Examining Committee:

Supervisor:
(Member)

Md Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Md Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

The number of car accidents is occurring on a global basis at an unprecedented scale, and as a consequence, this leads to an increasing demand for improving the driving risk assessment systems. Existing systems are often inadequate, as they fail to sufficiently incorporate both driver behavior and environmental context and thus often fail to perform well in real-world driving scenarios. In this work, we propose a hybrid and context-aware driving risk assessment framework, which models driver behavior and environmental conditions by employing a progressive fusion-based framework. The framework utilizes the US Accidents dataset (2016-2023) as context and risk factor, and a mobile-phone based Driving Behavior dataset capturing driving behavior. Individual classification and regression models are trained using custom built neural networks in the first stage in order to learn feature representations that are non-linear. Later, two fusion strategies are introduced, including decision level fusion (combining neural networks using probability combinations) and late fusion (combining the prediction of the classifier models with the predicted risks from context information), respectively. Besides this, in order to effectively model driver behavior over time, sequence based learning is employed to model behaviors, and a refinement layer is applied to further boost the classification accuracy of models. Our proposed approach generates a continuous and interpretable risk score and outperforms traditional methods and the single classifier by producing satisfactory classification performances. This work also emphasizes that the combined use of deep learning and fusion is an efficient approach for large-scale, reliable and multimodal driving risk assessment.

Keywords: Risky driving behavior, Context-aware risk assessment, Multimodal fusion, Driving behavior analysis, Accident severity prediction, Hybrid AI framework, Machine learning-based risk prediction

Acknowledgement

Firstly, we would like to express our sincere thanks and praises to the Almighty Allah for having bestowed this great opportunity on us to successfully complete this thesis without any major hurdles. We would like to extend our sincere thanks and regards to our respected supervisor, Dr. Md. Golam Rabiul Alam Sir, for his guidance, support, and motivation, which played a vital role in completing this thesis. Finally, we would like to express our sincere thanks and regards to our parents for their support and blessings, without which this would not have been possible.

Table of Contents

Declaration i

Approval ii

Abstract iv

Acknowledgement v

Table of Contents vi

List of Figures ix

List of Tables xi

Nomenclature xi

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Rational of the Study or Motivation . . . 3
- 1.3 Problem Statement . . . 3
  - 1.3.1 Research Gaps . . . 4
- 1.4 Objective . . . 5
- 1.5 Methodology in Brief . . . 6
- 1.6 Scopes and Challenges . . . 7
- 1.7 Thesis Organization . . . 8

2 Literature Review 9
- 2.1 Preliminaries . . . 9
- 2.2 Review of Existing Research . . . 10
- 2.3 Summary of Key Findings . . . 31

3 Requirements, Impacts and Constraints 34
- 3.1 Final Specifications and Requirements . . . 34
- 3.2 Societal Impact . . . 35
- 3.3 Environmental Impact . . . 35
- 3.4 Ethical Issues . . . 35
- 3.5 Standards . . . 35
- 3.6 Project Management Plan . . . 36
- 3.7 Risk Management . . . 36
- 3.8 Economic Analysis . . . 36

3.8.1 Cost Analysis . . . 37
3.8.2 Benefit Analysis . . . 37

4 Proposed Methodology **38**
4.1 Design Process or Methodology Overview . . . 38
4.2 Preliminary Design or Model Specification . . . 41
4.2.1 Accident Severity Classification (Dataset_1) . . . 42
4.2.2 Driver Behavior Classification (Dataset_2) . . . 44
4.3 Data Collection . . . 46
4.3.1 Data Sources . . . 46
4.4 Data Cleaning and Preprocessing . . . 47
4.5 Data Transformation . . . 47
4.6 Data Integration . . . 48
4.7 Data Reduction . . . 50
4.8 Summary of Preprocessed Data . . . 51

5 Result Analysis **53**
5.1 Performance Evaluation . . . 53
5.1.1 Evaluation Criteria . . . 54
5.1.2 Testing Methods . . . 54
5.1.3 Performance Outcomes . . . 55
5.2 Analysis of Design Solutions . . . 60
5.2.1 Alternative Model Architectures . . . 60
5.2.2 Data Integration Strategy Analysis . . . 61
5.2.3 Risk Mapping and Continuous Risk Score Design . . . 61
5.2.4 Fusion Strategy Analysis . . . 64
5.2.5 Design Factor Evaluation . . . 67
5.3 Final Design Adjustments . . . 67
5.3.1 Neural Network Architecture Refinement . . . 67
5.3.2 Class Imbalance Handling . . . 68
5.3.3 Engineering and Representation Refinement . . . 68
5.3.4 Sequence Modeling Adjustment . . . 69
5.3.5 Knowledge Distillation Adjustment . . . 69
5.3.6 Risk Score Reformulation . . . 69
5.3.7 Fusion Refinement and Final Model Selection . . . 69
5.4 Statistical analysis . . . 72
5.4.1 Relative performance of experiments . . . 72
5.4.2 Efficiency and model compression analysis . . . 73
5.4.3 Risk Score distribution analysis . . . 74
5.4.4 Regression Consistency Analysis . . . 74
5.4.5 Uncertainty and Class-wise Reliability . . . 74
5.4.6 Summary of Statistical Findings . . . 76
5.5 Comparisons and Relationships . . . 76
5.5.1 Internal Model Comparison . . . 76
5.5.2 Comparison With Existing Works in the Literature . . . 76
5.5.3 Comparison between fusion strategies . . . 77
5.5.4 Comparison of Risk Formulation . . . 79
5.5.5 Relationship Between Contextual and Behavioral Predictions . . . 79
5.5.6 Comparative insight . . . 79

5.6 Discussions . . . 80
5.6.1 Interpretation of Results . . . 80
5.6.2 Implications for Practice . . . 81
5.6.3 Limitations . . . 81
5.6.4 Contributions and Novelty . . . 82
5.6.5 Future Research Directions . . . 82
5.6.6 Concluding Discussion . . . 83

6 Conclusion **84**
6.1 Summary of Findings . . . 84
6.2 Contributions to the Field . . . 85
6.2.1 Methodological Contributions . . . 85
6.2.2 Empirical Contributions . . . 86
6.2.3 Practical and Theory Contributions . . . 87
6.3 Recommendations for Future Work . . . 87
6.3.1 Dataset Expansion and Cross-Cultural Validation . . . 88
6.3.2 Edge Deployment and Real-Time systems . . . 88
6.3.3 Explainability and Interpretability . . . 88
6.3.4 Application-Specific Extensions . . . 89
6.3.5 Ethical and Social Considerations . . . 89
6.3.6 Validation and Standardisation . . . 89

Bibliography **95**

List of Figures

4.1 Proposed Meta-Model Architecture for Context-Aware and Behavioral Driving Risk Assessment – Illustrates the integrated flow of environmental and behavioral data through separate branches and multi-level fusion to produce a unified, continuous risk score. 40
4.2 Correlation Heatmap of Traffic and Environmental Features (Dataset 1) – Visualizes the statistical relationships between external variables (e.g., weather, road features) to identify key predictors for accident severity classification. 43
4.3 Model Performance Comparison by Data Chunk Size (Macro F1 Score of Dataset_1). This figure demonstrates that model performance stabilizes as training data increases, indicating that larger data chunks provide a more reliable representation of contextual risk. 44
4.4 Sensor Feature Correlation Heatmap (Accelerometer and Gyroscope Data) – Shows the high degree of interdependency among motion sensors, which justifies the use of a unified sensor-based neural network for behavioral detection. 45
4.5 Receiver Operating Characteristic (ROC) of the Contextual Branch Before LLM Refinement – Serves as a baseline performance metric showing the model's initial ability to distinguish between accident severity levels. 49
4.6 ROC Performance of Context Branch (LLM-Enhanced) – Highlights the performance gains achieved after using LLM-based refinement to prune noise, resulting in better differentiation of high-risk scenarios. . 50

5.1 Model Training Pipeline illustrating the workflows for Accident Severity Model (Dataset 1) and Sensor Classification Model (Dataset 2). Both pipelines follow structured steps including data loading, train-test splitting, neural network training, and evaluation, with Dataset 1 producing severity classification (2 classes) and Dataset 2 producing driving behavior classification (3 classes). 56
5.2 Accuracy Comparison of Custom Neural Network and LSTM Models - Findings indicate that while LSTMs excel at sequential sensor data (Dataset 2), custom neural networks are more effective for static contextual data (Dataset 1). 57
5.3 Performance of the Baseline LL Model across All Risk Scores - Visualizes the reliability and consistency of the initial late fusion (L+L) model in assigning risk scores across the entire test set. 59

5.4 Enhanced Classification Performance using LLM Integration - Shows that LLM-based post-processing significantly improves classification metrics compared to the baseline model. 59
5.5 Driver Experience Distribution of Survey Participants - Provides a demographic context for the expert knowledge integrated into the late fusion risk matrix. 62
5.6 62
5.7 63
5.8 Subjective Risk Perception Across Driving Scenarios - Summarizes survey results that define how humans perceive risk across six distinct driving situations, forming the basis for the framework's risk mapping. 63
5.9 Error Distribution of the D+D Model - Illustrates the variance in risk prediction for the decision-level fusion model, highlighting areas where the meta-learner requires refinement. 64
5.10 Error Distribution of the Baseline LL Model - Shows a more stable error profile compared to D+D, demonstrating that incorporating prior knowledge reduces extreme prediction errors. 65
5.11 Classification Accuracy and Severity Distinction with LLM Enhancement - Confirms that LLM refinement specifically boosts the system's ability to identify high-severity risks accurately. 66
5.12 70
5.13 Training Performance and Loss Convergence of Baseline vs. LLM-Assisted L+L Fusion Models - Shows faster and more stable loss convergence in the LLM-assisted model, suggesting it is more efficient during training. 71
5.14 Accuracy and Macro F1 Comparison of Fusion Strategies - A comprehensive summary showing that the LLM-assisted L+L fusion strategy outperforms all other methods in both accuracy and F1-score. 72
5.15 Normalized Radar Comparison of Fusion Strategies - Provides a multi-dimensional view of performance, showing that the final hybrid model balances classification accuracy with regression consistency. 73
5.16 Residual Analysis of the Baseline Late Fusion Model - Visualizes the distribution of prediction errors, showing the inherent variance before applying refinement layers. 75
5.17 Residual Analysis of the LLM-Assisted Late Fusion Model - Demonstrates a significant reduction in error variance, indicating that LLM refinement produces more stable and predictable risk outputs. 75
5.18 Baseline Classification Metrics across Severity Tiers - Provides the precision and recall scores for the standard model, highlighting its performance in distinguishing between low and high-risk driving scenarios. 78
5.19 LLM-Enhanced Classification Metrics across Severity Tiers - Demonstrates the performance improvement specifically in critical high-severity tiers, showing how the LLM-assisted model better identifies life-threatening risks. 78

List of Tables

2.1 Research Gap Analysis and Positioning of the Proposed Hybrid Fusion Framework 33

4.1 Models Applied to Each Dataset 41
4.2 Summary of Dataset Characteristics and Preprocessing Pipelines . . . 52

5.1 Performance Comparison of Custom Neural Networks and LSTM Models 56
5.2 Summary of Model Components and Their Performance 57
5.3 Quantified Expected Risk Scores Derived from Subjective Survey Data Integration 63
5.4 Error Analysis of Fusion Strategy Outputs 66
5.5 Comparative Performance Metrics of the Fusion Architectures 71

Chapter 1

Introduction

The subject of understanding driving behavior and assessing driving risk is becoming an increasingly critical area of study within road safety, especially with the development of intelligent transportation systems and autonomous vehicles. Improvements in data collection, sensor technology, and artificial intelligence have made it possible for researchers to study driving behavior, predict risk and develop systems to reduce accidents and improve traffic management. However, most systems to date only exploit a single modality or a "black-box" fusion approach that is not easily interpreted and scaled. In addition, there has been little exploration of uncertainty-aware prediction, or how human knowledge could be incorporated into data-driven risk assessment frameworks. This work introduces a gradual multimodal fusion framework that synergizes machine learning, deep learning, and knowledge-based approaches to provide dependable and interpretable driving risk scores.

1.1 Background

Driving risk and driving behavior studies have evolved in the past two decades with increasing amounts of driving data and computing resources. Several research studies have been conducted on the interpretation and quantification of driver behavior, anomaly detection, and the utilization of such information in applications ranging from insurance to autonomous driving. The following section presents an overview of key contributions in the literature and positions the current research within this context. Driver profiling and pattern recognition are two fundamental constructs in driving behavior analysis. Driver profiling describes long-term driving styles and behavioral tendencies of individuals or groups, while driving patterns generally represent short-term, situational behaviors during individual trips or events. Previous work [1, 7, 8] has shown the effectiveness of macroscopic (profile-based) and microscopic (pattern-based) analysis, though these approaches generally require different methodologies and data structures. Profile analysis typically uses comprehensively accumulated data over weeks or months, whereas pattern detection relies on high-resolution temporal data. Recent research introduced the concept of driving pulses—intervals during driving bounded by stops—which serve as natural units for microscopic behavior analysis [11, 12]. Despite individual development, few attempts have been made to synthesize these two levels of analysis. This separation has limited the evolution of general behavior models that could benefit from both approaches. With the evolution of telematics and in-car sensing, researchers can now

record large collections of real-world GPS, speed, braking, acceleration, and steering data. Large-scale naturalistic driving studies have become an effective method for observing real driver behavior [27]. Such information has been used to support Usage-Based Insurance (UBI) practices [12], which utilize driving behavior to predict insurance risk. These applications often incorporate additional predictive variables, such as exposure, road context, and time of driving, to more accurately calculate risk premiums than using demographic data or claim history alone. While these methods show high predictive capability, most lack real-time processing and rely on fixed measures rather than adaptive learning models. Driver inattention, caused by visual distraction, cognitive load, or fatigue, is a major contributor to traffic accidents. Studies have addressed inattention using facial expression, eye movement tracking, and physiological signal processing [11]. Multi-indicator hybrid detection systems, combining multiple data sources such as physical state, vehicle control actions, and in-vehicle infotainment system (IVIS) data, outperform single-sensor systems. However, most existing systems are limited to laboratory settings and face generalization challenges in broad real-world conditions. Autonomous driving technologies (ADS) have introduced new challenges in risk estimation and safety validation. Traditional systems rely on rule-based violation detection, such as speeding or red-light disregard, but autonomous driving requires attention to behaviors that may not violate explicit rules yet compromise safety. Studies [13, 14] emphasize identifying behaviors that violate driving "common sense," such as sudden lane changes or failure to yield in ambiguous situations. Safety entropy-based models, behavior trajectory-based models, and subjective-objective weighted evaluation models have been proposed, but most are computationally intensive and unsuitable for real-time applications. Machine learning has become increasingly central in driving behavior analysis. Popular algorithms include Random Forest, CNNs, RNNs, and hybrid deep learning models. Greater emphasis is placed on data quality and bias mitigation. A milestone study [18] presented a neural network-based tachograph system trained on over 40 billion driving histories. By reducing biases related to driver demographics, geography, and vehicle types, the system achieved 99.99Driving style is influenced not only by the driver but also by external conditions such as weather, road surface, lighting, traffic, and timing considerations. Research [12] shows that incorporating these contextual variables significantly improves risk predictability. Certain road contexts may increase the likelihood of distraction or hard-braking events. Despite their usefulness in applications like insurance and logistics, contextual factors remain underutilized in academic safety risk models, partly due to the challenge of synchronizing heterogeneous data sources in real time. There is also a need for explainable models that convey the impact of context on risk prediction, which is critical for policymaking and insurance pricing. An emerging trend in fleet safety is journey-level risk analysis, which treats the journey as the unit of analysis. Unlike event-level analysis, journey-level models account for cumulative effects such as driver fatigue, route complexity, and time-of-day variations. A study of intercity buses in Taiwan [15] employed a hybrid deep learning model to combine real-time GPS data with static journey characteristics to predict trip-level risk. This approach bridges the gap between transient alerts (e.g., sudden braking) and monthly driver scorecards, offering fleet managers a more practical timescale for monitoring. However, journey-level modeling is still in its infancy and requires further methodological development and real-world validation. While behavioral modeling and contextual

risk analysis have seen dr amatic improvements, little previous research has dealt with both subjects jointly. Multimodal fusion strategies are indeed available, but they are based on feature-level fusion, bringing synchronization issues and limiting model interpretability. Decision-level and late fusion strategies have not been investigated enough with large-scale driving risk assessment, especially those that retain model uncertainty and integrate prior knowledge.

1.2 Rational of the Study or Motivation

It should be noted that road accidents still represent a serious public health problem around the world and cause about 1.3 million deaths each year along with multiple cases of injury all over the planet [18]. As for causes of these incidents, the human factor represents the main one, as about 65-95 percent of road accidents are provoked by individuals' mistakes [1]. For these reasons, modern times require developing intelligent systems for monitoring risks of road accidents. As for technical prerequisites, recent innovations in telematics, connected cars, and sensing allow collecting driving data constantly and at high frequency [12]. These facts provide a great opportunity for shifting from analyzing incidents to monitoring possible threats to drivers' safety in a proactive manner. In turn, it makes turning collected information into valuable knowledge very problematic. It is worth mentioning that existing approaches do not pay much attention to combining the issues of driver profiling and pattern recognition [11, 12]. Also, most models concentrate on detecting rule violations without taking into account less obvious but risky behavior types [13]. It should be emphasized that such contextual aspects as weather or traffic conditions play a key role in road safety [12]. Autonomous and semi-autonomous cars become an additional problem for developers of safety models, as they should be able to analyze actions of both humans and machines [13, 14]. Besides, limited possibilities in real-time processing and lack of a large unbiased dataset make these tasks even more complicated [18]. Taking all of these aspects into consideration, there is a clear need for the development of an innovative risk detection system that will take into account such factors as macro- and micro-behavior analysis and contextual data, as well as bias-aware machine learning algorithms.
Moreover, for safety-critical systems, it is important not only to make correct predictions but also to accurately estimate the prediction uncertainty. Indeed, failure to take prediction confidence into account may cause unsafe behavior in critical situations. It is then needed a system able to combine data sources in the presence of uncertainty and to provide the possibility to estimate the risk in a transparent way.

1.3 Problem Statement

Overall, risky driver behaviors have become a serious concern in the world due to road accidents being a predominant cause. The systems implemented in many firms to monitor the way people work the wheel have been rolled out, as the majority of designs continue to fall short with an attempt to unravel all the causes of unsafe habits. These means tend to stand alone, just observing what is going on in the car and disregarding traffic congestion, potholes, whether it rained, or whether dusk had only dropped. Relying on a single source of data or a single analytical perspec-

tive often leads to an incomplete or misleading representation of driving risk. When contextual factors are analyzed in isolation from driver behavior, or vice versa, the interaction between changing road conditions and behavioral responses remains inadequately captured. To further aggravate the situation, the gadgets are largely reactive, procuring a virtual warning only after a swerve, a hard brake or a near miss has already occurred. As a result, opportunities for preventive intervention are frequently missed. Another big drawback is that most of the existing systems were constructed to address one use case, so they are not able to address various road conditions or different car models at scale. Due to such a slim design, such systems often face scalability and deployment constraints when translated to real-world settings. Add to that, there is no widely adopted or standardized methodology for integrating heterogeneous risk information derived from different data sources and analytical models, resulting in fragmented and difficult-to-compare implementations. Consequently, these limitations reduce system reliability, negatively affect user trust and hinder large-scale adoption. Even the current literature proposes more gaps; despite the use of deep learning or random forest classifiers, the quantification of uncertainty is frequently not done to generate predictions and is difficult to explain in a high-risk setting. Although decision-level fusion and ensemble learning approaches show promising performance, they often fail to explicitly integrate contextual variables and behavioral patterns in a unified risk estimation process suitable for real-world deployment. Behavior-focused datasets derived from vehicle dynamics or mobile sensing provide scalable solutions; however, when analyzed independently, they suffer from limitations related to generalizability, contextual awareness, and robustness across different driving environments. While simulation-based studies offer valuable insights for controlled experimentation, their findings require validation against large-scale real-world datasets to ensure practical relevance. As sensing systems and AI technologies are advancing at a rather high pace, a wide gap in light is the difference between what technically can be done and what systems that are commercially available are actually being constructed. Consequently, one of the urgent demands is a novel, context-sensitive, proactive risk evaluation framework that combines independent behavioral and contextual risk predictions, accounts for predictive uncertainty, and produces a unified risk score for informed intervention. This paper is an initial effort to bridge that divide by developing a more holistic and dynamic driver safety and risk avoidance strategy.

1.3.1 Research Gaps

Based on the review of the related works, the following research gaps have been identified:

- **Integration Gap:** Many existing methods evaluate risk using either behavioral or contextual data separately. However, an integrated framework that jointly considers both behavioral and contextual information for risk estimation is still lacking.
- **Proactive Risk Assessment Gap:** Most existing systems are reactive in nature, providing alerts only after a risky event has occurred. There is a lack of proactive systems capable of predicting and preventing risky situations beforehand.

- **Scalability and Generalizability Gap:** Current approaches are often designed for specific environments and lack the ability to generalize across different road types, traffic conditions, vehicle models, and driving behaviors.
- **Methodology Gap:** Many machine learning models do not account for predictive uncertainty, which is critical for reliable decision-making in safety-critical applications.
- **Data Validation Gap:** Several methods rely on simulated data or limited datasets and are not validated on large-scale real-world data, reducing their practical applicability.
- **Technology Gap:** Commercial systems do not fully utilize the capabilities of modern sensing technologies and advanced AI techniques, resulting in a gap between research developments and real-world implementations.
- **Fusion Strategy Gap:** Existing fusion techniques are limited to either feature-level fusion with synchronization challenges or simple rule-based decision mapping. Flexible hybrid fusion architectures combining learning-based and interpretable strategies are largely unexplored.
- **Uncertainty Utilization Gap:** Probabilistic outputs from machine learning models are often ignored during fusion, leading to the loss of critical uncertainty information necessary for safety-focused decisions.
- **Knowledge Integration Gap:** The incorporation of human-derived knowledge, such as risk perception from survey data, into machine learning models remains limited in current research.
- **Pre-processing Intelligence Gap:** The potential use of intelligent models, such as large language models, for output refinement, validation, or verification in transportation risk assessment systems has not yet been explored.

This thesis addresses the aforementioned gaps by proposing a novel hybrid, context-aware, and proactive framework for driver safety and risk estimation. The proposed system integrates both behavioral and contextual information while incorporating predictive uncertainty, enabling reliable, scalable, and real-world applicable solutions.

1.4 Objective

In order to advance predictive modeling and risk assessment within intelligent transportation systems, this study sets forth a series of well-defined objectives. These objectives are designed to ensure methodological rigor, practical applicability, and safety-oriented evaluation. The key points are summarized below:

- Develop predictive models to anticipate driving behaviors and accident severity from large-scale real-world data.
- Create a context-aware, scalable risk assessment framework integrating heterogeneous environmental and behavioral data.

- Model temporal dependencies and intricate driver–context interactions for accurate representation of risk factors.
- Build a continuous risk score for granular and practical quantification of driving behavior.
- Design a model fusion scheme that combines multiple models, preserves uncertainty, and enhances reliability.
- Incorporate external knowledge into risk assessment to improve robustness and real-world applicability.
- Quantify uncertainty and post-process predictions to strengthen confidence in safety-critical situations.
- Validate the framework with safety-oriented performance measures, focusing on high-risk circumstances.
- Develop a flexible, scalable system for practical deployment in intelligent transportation and safety applications.

1.5 Methodology in Brief

This research suggests a machine learning-based method to predict driving risk by using two different (contextual and behavioral) datasets. The method followed is a step-wise and incremental approach. Firstly, single neural network models are designed and investigated, and then they are combined with a more complex fusion method to offer an overall system performance.

The first step in this research is to preprocess two different data separately. For the US Accidents dataset context related information such as weather, time, road characteristics is extracted, which is used for predicting the severity of accidents. For the Driving Behavior dataset, motion-related information is extracted from sensors, representing different driving styles for different driving risks. These are preprocessed by applying data cleaning and normalization and feature engineering.

In the first part, several machine learning models are trained and tested separately for each dataset. Then, neural networks are designed to capture non-linear relationships and probability outputs for contextual and behavioral risks from each data. But the performance from independent models does not capture interaction between context and behavior. For this purpose, new sophisticated models including sequence learning methods, are employed to strengthen the representation learning.

Then fusion methods are implemented to fuse information from two neural networks. The decision-level fusion (D+D) approach has been used, where the probability outputs from two neural networks are combined with a meta-learner for modeling the relationship between contextual and behavioral risks without independence of the dataset. Based on that, a late fusion (L+L) strategy is derived where the network prediction is fused with an external, static risk value, which is obtained from prior knowledge based on survey data. This model forms the combined model, which

enables combining learned data-based approaches with explicitly represented risk based on expert knowledge.

Further, incorporating sequential learning into the late fusion model produces another enhancement where the L+L strategy improves to a more robust and a fairly balanced model for risk prediction driving tasks. The effectiveness of the proposed model has been discussed by looking into performance metrics like accuracy, precision, recall, and F1-score; emphasis has been given to the evaluation of the system under a safety-relevant context for high-risk cases. Also, the consistency and distribution of risk scores generated for various modeling and fusion strategies have been considered.

1.6 Scopes and Challenges

This research aims to develop and assess a multimodal AI framework for improving the driving risk prediction by incorporating both behavioral and contextual data. This work uses two commonly available datasets: The US Accidents dataset where contextual and environmental data is collected from multiple states on a large scale, and a Driving Behavior dataset using sensors (accelerometer and gyroscope) from smartphones to record motion data such as accelerating, braking and turning.

In this research, we will analyze and model two specific datasets mentioned above, as well as develop a predictive model and fusion strategies based on these two datasets. No on-line data collecting or any new sensing hardware development or real-time system deployment is involved in this work; it just tests whether it is practical and effective to use combined heterogeneous datasets and models to predict driving risk.

There are a few challenges to consider. First, class imbalance is a serious problem in the accident dataset because the severe accident cases are extremely outnumbered. Second, two datasets are very heterogeneous as the context accidents are contextual and environmental data and the behavioral sensor data from smartphones are in high frequency. Their formats, scales, and dimensionalities are different, so merging two kinds of data is not straightforward.

With a very large number of accidents, computational constraint is a major issue, and efficiency in data management and sampling technique is essential. Third, while deep learning and fusion models would enhance the prediction accuracy, interpretability of model is a major issue, especially when complexity of models grows very large.

Also, although the framework combines multi-level fusion strategies and utilizes the exogenous information by applying static risk scores, it has not yet included real-time multimodal synchronization, nor the adaptive learning of driving state for real-time dynamic adaptation. These can be considered as issues for further research.

Regardless of the obstacles, this study sets up a solid foundation of multimodal driving risk assessment through the validation of integrating behavior and context by using a deep learning and fusion-based hybrid model and also serves as a step-

pingstone toward developing sophisticated intelligent transportation safety systems.

1.7 Thesis Organization

In chapter 2 we survey relevant existing literature considering the application of artificial intelligence (AI), machine learning, and deep learning in relation to predicting road accidents, understanding driver behavior, and the construction of context-aware transportation safety systems. It takes a broad look over all contemporary and traditional methods such as neural networks, ensemble techniques, clustering algorithms, federated learning, and multimodal sensor fusion frameworks, establishing the foundation for the proposed work.

In chapter 3 we provide system requirements and an overview of the intended application context of the proposed framework. Specifically, this chapter elaborates on the technical and methodological constraints applied to our dual-dataset framework and also on its potential societal and environmental effects, while also presenting the overall ethical considerations concerning data privacy and fairness in algorithmic decision-making and, finally, the implementation-enabling project management plan, risk management strategy, and the underlying economic rationale.

In chapter 4 we provide the detailed description of the proposed methodology that applies both contextual and behavioral information for a more reliable estimation of driving risk. It focuses on all the stages of the entire pipeline, starting from data collection, cleaning and preparation to the feature engineering, transformation and reduction stages for both datasets, namely the US Accidents dataset (Dataset 1) and the smartphone-based Driving Behavior dataset (Dataset 2). The chapter outlines the research methodology being proposed. It describes the full process, covering the data acquisition, pre-processing, feature engineering stages and their applications for the US Accidents data and the Driving behavior data; the construction of machine learning and deep learning models; the integration of a sequence-based learning approach to regression-based risk scoring mechanism; and the design of the fusion techniques. The performance of the fused system is then illustrated through analysis of advanced refinement strategies for predicting the risk level.

In chapter 5 we provide the system evaluation. We discuss and present the results of both the contextual accident severity model and the behavioral driving pattern model, as well as discuss several alternative designs we were confronted with, along with our final design changes for enhanced performance, and finally, the overall performance of the fused system is then demonstrated with the analysis of sophisticated refinement schemes for predicting the risk level.

Finally, chapter 6 concludes the thesis by summarizing our findings, stating our methodological, empirical, and practical contributions, and outlining the limitations of the proposed work. We give detailed recommendations for future research topics, spanning dataset extension, edge implementation, explainability issues, and application-specific extensions, as well as related ethical and regulatory implications for the development of safe intelligent transportation systems.

Chapter 2

Literature Review

In order to formulate a solid base of the suggested system, it is important to first examine the knowledge foundations and theoretic ideas associated with driving behavior modelling and context-sensitive risk evaluation. This part provides an overview of all the basic concepts, terms and trends in technology that are the foundation of recent studies on intelligent transportation systems. These preliminaries assist in putting the later discussions into perspective and give the concept base upon which the existing methodologies can be interpreted, as well as the direction of this study.

2.1 Preliminaries

Recent developments in smart transportation systems have reshaped the safety measures from reactive to proactive and the contribution of artificial intelligence to monitoring and predicting driver behavior has been more and more emphasized. Driver behavior prediction, context-aware analysis, anomaly detection and risk estimation are core concepts in contemporary intelligent transportation research. Driver behavior is determined by internal states, fatigue and distraction and external conditions, traffic, weather and road type. Recent approaches reported in the literature attempt to overcome these issues by leveraging multimodal data from telematics, vision sensors and environmental sensors. Data-driven modeling is grounded on comprehensive accident datasets and a variety of sparse data, while simulation environments such as SUMO make it possible to create surrogate databases of driver behavior in order to compensate for scarcity in real-world data availability. Context awareness and new machine learning architectures enable better estimation of risk through scrutiny of the dynamism in driving conditions. Advanced ensemble methods, uncertainty-aware graph neural networks, and argumentation-based reasoning frameworks illustrate how AI models are evolving to handle heterogeneity, uncertainty, and conflicting signals across modalities. Federated learning and standardized risk assessment frameworks have also been explored in recent studies as scalable and privacy-aware directions for future intelligent driving safety systems. Lightweight smartphone-sensor-based feature engineering approaches and multi-modal ADAS frameworks that combine physiological and mobility signals highlight a growing research interest in efficient and responsive safety-oriented driving analysis systems. These concepts form the basis upon which the literature review described here is grounded.

2.2 Review of Existing Research

Tselentis et al. [1] offered a comprehensive study on existing AI and ML approaches to recognize driver profiles (groups of drivers sharing similar behaviors) and driving patterns (specific, repetitive behaviors). It identifies aggressive driving as commonly recognized behavior and notes methodological diversity, primarily utilizing neural networks and clustering algorithms. It stresses the need for integrated approaches combining macroscopic and microscopic driving behavior analysis to enhance predictive accuracy for road safety and crash prevention, proposing a comprehensive framework for future research and applications in real-time feedback systems and autonomous vehicle programming.

In their research, Kanagamalliga et al. [2] presents a comprehensive driver monitoring system based on sensor fusion technology designed to enhance the safety of heavy vehicle drivers. By integrating an alcohol sensor, accelerometer, and blink detection sensor on an Arduino platform, and visualizing data through LabVIEW, the system effectively classifies driver behaviour, providing real-time alerts for dangerous conditions such as intoxication, fatigue, and overspeeding. Experimental validations confirmed the reliability and practicality of this multi-sensor fusion approach, showing considerable potential for widespread application and contributing significantly to road safety. Future improvements could involve more sophisticated sensors, deeper algorithmic enhancements, and real-time automated responses tailored to individual driver profiles and diverse operational environments.

The work of Ali et al. [3] develops and evaluates an effective deep-learning-based system utilizing YOLOv8 for real-time abnormal driver behavior detection. The model effectively detects multiple risk behaviors (such as not wearing seatbelts, phone usage, eating, drinking, smoking, and drowsiness) and issues timely verbal alerts to drivers. Comprehensive training on a large and diverse annotated dataset yields high detection accuracy (approximately 95-96%). Experimental validation shows the system's capability for reliable real-time monitoring with minor latency issues noted on limited hardware like Raspberry Pi. Suggested future directions include hardware enhancement and further algorithmic refinements, emphasizing the system's significant potential to reduce road accidents.

In their work, Sakthivel et al. [4] proposes a proactive driver safety framework combining CNN and ConvLSTM deep learning techniques. It significantly improves driver risk prediction, integrating comprehensive real-time monitoring of both driver behaviors and road hazards. The model achieves notable accuracy (97%) in identifying risks ahead of time, offering actionable alerts to drivers, effectively enhancing overall safety. Comprehensive testing using realistic datasets confirms the superior performance compared to traditional methods. The integration of spatial-temporal analysis and multi-modal sensor data demonstrates a robust solution for modern driver assistance systems, with clear pathways for further optimization and practical implementation.

Fang et al. [5] present a computer vision-based system utilizing convolutional neural networks (CNNs) to objectively identify and assess electric bus drivers' behaviors

in real time. The method successfully reduces recognition latency (maintained between 0.4–1s) and achieves high classification performance, with F1 scores ranging from 89% to 97%. Through effective feature extraction and intelligent algorithmic classification, the system addresses key challenges in monitoring driver states such as fatigue and distraction. Experimental validation confirms that the proposed approach outperforms traditional techniques in both accuracy and responsiveness. By offering a real-time evaluation mechanism, the framework not only enhances driving safety but also supports improved driver training and passenger experience. The paper further outlines opportunities for advancement through larger datasets and broader deployment in real-world transit environments.

The model introduced by Anil et al. [8] employs K-means clustering to classify driver behaviors using smartphone-derived accelerometer and gyroscope data. The classification effectively differentiates driving events into aggressive, normal, and moderately aggressive clusters based on accelerometer values. Performance metrics such as Silhouette score and Davies-Bouldin Index confirm the reliability of clusters. The optimal number of clusters was determined through the Elbow method, validated further by these metrics. The results demonstrate effective differentiation of driver behavior, indicating that aggressive behaviors exceed safety thresholds significantly. The methodology shows promise for practical applications in real-time driver feedback systems aimed at enhancing safety and optimizing energy use in EVs, suggesting significant potential for future refinements and integration into broader transportation safety strategies.

Amer et al. [10] introduced a framework for a novel multi-label classification approach for detecting aggressive driving behaviors by simultaneously considering driving behaviors and environmental conditions. Using multi-label algorithms (Binary Relevance, Classifier Chains, Label Powerset, RAkEL) and classifiers (Random Forest and SVM), the approach exploits correlations between driver behavior and driving conditions. The method significantly outperforms traditional classification approaches, demonstrating the importance of integrating environmental data. The inclusion of feature selection markedly enhances the predictive accuracy, particularly through label powerset and RAkEL methods. The paper's findings underscore the substantial improvements achievable by recognizing the interdependencies between driving behaviors and environmental conditions, offering clear future directions in enhancing predictive robustness and classifier reliability.

The proposed study by Qu et al. [11] is a comprehensive overview of deep learning and computer vision from an automotive driver monitoring system perspective based on AI. This study targets unsafe driving behavior - distraction, fatigue, and lack of attention, in fully autonomous and semi-autonomous vehicles. It is novel due to the incorporation of various deep architectures of CNNs, RNNs, and LSTMs with vision systems for enhancing detection rates. The strategy is to extract behavioral signals from cues from vision, such as facial expressions, hand location, and body pose. Multi-modal data fusion and complex models like Hypo-Driver, along with CNN, RNN, and residual networks, were considered to be effective. Accurate rates of up to 96.5% for risky and abnormal behavior are possible to achieve by analyzing spatial and temporal cues from in-cabin streams. Pre-trained models and transfer

learning were also utilized for peak performance despite small datasets. Driver cue localizations from cameras and sensors inside the car were also covered to enhance the system's ability to identify behavioral outliers online. This combined monitoring system is a rich contribution to improving vehicle protection by offering intelligent in-cabin awareness. Despite positive results from research, this study also reports limitations on dataset diversity, generalization of models, and privacy threats and offers an area for future study on robust, scalable, and privacy-conscious driver monitoring systems.

Masello et al.'s [12] proposed model was a novel machine learning solution to predicting risky driving behavior from contextual driving data from telematics. XGBoost and Random Forest decision-tree models were utilized in this study, with SHAP being utilized for driving context feature contribution explanation and ranking. In this research, novelty is introduced by combining telematics data and a variety of a set including a total of 18 contextual attributes (e.g., weather, road type, traffic, and road infrastructure) for enhanced risky incident prediction of near-misses, speeding, and distraction. It involves developing a data pipeline that draws on naturalistic driving data to form driving context profiles and train regression models to forecast six risky event occurrences for 145,842 distinct context combinations. They learn significant contextual patterns, for instance, whether roads with low speed and urban roads are correlated with higher phone use and speed, and rain and bad-quality roads are correlated with aggressive maneuvers. SHAP values are utilized to explain decisions from models and find key influential features for each type of event. In spite of driving behavior and environmental heterogeneity, the proposed model was apparently highly predictable with excellent event frequency precision, where XGBoost was superior to other models for most cases. Model interpretability and use of public naturalistic data make it an attractive framework for road safety study and Usage-Based Insurance (UBI) business use. This novel concept offers the Pay-Where-You-Drive (PWYD) idea, which can potentially give insurers a privacy-centric means for risk-based premiums.

Pang et al.'s [13] framework was a new behavior-based risk estimation framework for detecting and assessing autonomous vehicle driving behavior violating common-sense conceptions of safety without specifically violating traffic rules. A hybrid model was proposed here by combining subjective Analytic Hierarchy Process (AHP) and objective entropy weight methods to fully quantify risk levels of such driving behaviors like sudden acceleration, repeated lane changes and unsafe following distances. The contribution is quantifying and defining common-sense-violating behavior and combining it with environmental safety entropy for externally accounting for environmental and traffic complexity. Involvement of multi-stage evaluation of behavioral risk indicators derived from instantaneous driving behavior and grouped by a Gaussian Mixture Model (GMM) optimized with K-means clustering for assigning accurate risk levels is present. Dynamic adjustment for environmental factors improves the adaptiveness and decision accuracy of the model for various traffic scenarios. Experiments and simulation experiments on an L3+ autonomous driving testbed were conducted for testing and validating the efficacy of the model, resulting in improved sensitivity and precision for detectability of risks compared to standard entropy-based models. Generally, the technique improves autonomous vehicle be-

havior interpretability and safeness by taking into account minor, non-rule-breaking driving dangers.

In their work, de Gelder et al.[14] proposed a model, evidence-informed data-driven quantification of risk for the public road safety of Automated Driving Systems (ADSs). The article laid out a systematic six-step simulation-based procedure of exploiting on-road driving data for objective approximation of potential injury risk for specific traffic scenarios. With scenario-based simulation and modeling combined, an architecture of a system was proposed by the authors, which would not only identify risky scenarios but also quantify their probability and severity. By integrating an empirical exposure estimator with ADS behavior simulations and injury probability models and combining them to assess risk on exposure, severity, and controllability dimensions mandated by ISO 26262 and ISO/DIS 21448 standards, it was novel to this field. Data-driven driving scenario extraction, simulation with an Adaptive Cruise Control (ACC) model, and potential damage evaluation with injury probability models comprised this process. Binarization of risk into interpretable values and introducing scenario localization for pointwise evaluation of ADS's behavior constituted another novelty. Through test cases on three classes of driving scenarios (e.g., leading vehicle brake and cut-in), it was shown that the model was capable of assessing risk at fine levels of granularity and demonstrating how various "triggering conditions" outside an ADS, like imperfect visibility and decreased braking efficiency can cause an escalation thereof. With an injury-based quantifier of risk and simulation-supported evaluation of controllability, it had an obvious and quantifiable road safety performance metric for ADSs. Compliance with regulation and assurance activity support was facilitated by following ISO standards and offering open-sourced realization for broader automotive safety verification applicability.

Lee and Chang [15] introduced a new Deep Learning hybrid model for predicting driving risk levels from intercity bus naturalistic driving data. Bi-directional LSTM was blended with static journey background factors for enhancing the ability of the model to predict low, medium and high-risk driving journeys. The study innovates by combining dynamic time-series GPS data with static factors such as driver fatigue, route familiarity, and trip scheduling to perform a detailed measure of risk indicators. The strategy is a two-branch deep learning architecture such that one is utilized for Bi-LSTM time-series feature extraction and another for processing static journey data by means of dense layers. These two branches were blended to generate an end risk classification result. The dataset comprised 27,057 journey instances for 345 drivers, providing realistic driving behavior for training the model. Performance was also compared to baseline models such as logistic regression and XGBoost. Adding static journey factors significantly enhanced prediction results and the model achieved a high weighted average F1-score of 0.932. Predicting journeys with higher danger, the model obtained an F1-score of 0.728 and was up to 9.3 times more accurate than baseline models. This architecture has great prospects for improving fleet security management through accurate journey-level risk forecasting and is also an outstanding breakthrough in traffic security and AI-informed transport analytics.

Similarly, another framework based on Machine Learning was proposed by Reddy et

al. [16] for predicting road safety risks for Internet of Vehicles (IoV) systems. Federated learning (FL) was presented here as an advanced distributed ML framework to address limitations of standard centralized approaches. The key novelty is that edge devices such as connected vehicles can collaboratively train models without sharing raw data and hence preserve privacy and reduce communication costs. Applications of various ML methods—e.g., CNNs, LSTMs, and ensemble learning—e.g., driving behavior study, collision prediction, and traffic congestion study are covered. A systematic review of 72 peer-reviewed journal articles was conducted, integrating findings from areas such as driver safety, vehicle-pedestrian interaction, and real-time traffic accident forecasting. Methodology focuses on federated learning benefits for handling heterogeneous, large-volume IoV data without loss of model correctness and system scalability. Key limitations identified are handling diverse sources of data, ensuring timeliness, and handling ethical issues. Improved robustness and privacy of models and correct traffic safety predictions are enabled by proposed methods. Federated learning with blockchain and edge computing has tremendous potential for safer and smarter transportation networks, according to the authors. The review identifies FL's potential to counteract intrinsic system inefficiency and recommends it as an intelligent transport infrastructure-ready tool.

In addition, Wang and Lu [17] proposed a novel two-layer context-aware machine learning-based solution for risky driver identification from trajectories. A hierarchy was adopted here involving a traffic state (free-flow, saturated, congested) prediction first stage and a context-aware risky driver recognition stage trained from the first stage. The novelty resides in combining context awareness with driver behavior recognition for enhancing the improved ability of the model to identify driving risk for various traffic states. An adoption of a machine learning architecture involving XGBoost, SVM, KNN, AdaBoost, RF and Extra Trees classifiers, with XGBoost being the top performer, is involved while going about it. Risk assessment is done on rear-end and side collision risks and driver tagging is done through Average Collision Risk (ACR) from an IQR-based threshold process. The two-layer architecture allows for exact classification of risk-driving behavior under real traffic conditions by way of trajectory feature analysis and derivation involving speed and acceleration from the Discrete Fourier Transform (DFT).The model produced an F1 score of 0.762 and an AUPRC of 0.820, meaning that road safety improved with context-aware recognition of risky drivers.

The study of Wang et al. [18] explores a generalized machine learning framework for classifying driver behavior by focusing on ensemble techniques applied to vehicle telemetry data. The study utilized driving data obtained from on-board diagnostic (OBD) systems to analyze feed data which included speed, RPM, throttle angle, and engine load, and implemented Random Forest, Support Vector Machines (SVM), and Gradient Boosting. The most important contributions of this work include a voting ensemble approach which ties the results and enables more robust assertions, feature trimming based on mutual-information scores, and a side-by-side comparison of models for detecting driver habits. Methodologically, the work ran Random Forest with 100 trees and used SMOTE to balance the classes. SVM was tuned with radial-kernel via grid search and XGBoost was set with early stopping rules. Models were assessed through k-fold cross validation. To summarize each segment

of data, the authors of this work developed a set of simple statistics-mean, standard deviation, skewness- that, when combined, effectively capture the features of the data. Drawbacks including model overfitting and the complexity of tuning multiple hyperparameters across several learners remain.. However, the ensemble achieved 92.4% accuracy in distinguishing aggressive, normal, and eco-friendly driving styles, with Random Forest leading at 89.2%, while support vector machines and XGBoost followed at 87.6% and 88.9%, respectively. These findings indicate that stacked algorithms offer more robust performance for driving behavior classification compared to standalone models.

The research by Zhang et al. [19] introduced Convolutional Neural Networks (CNNs) and was the first to conduct driver behavior analysis using multi-modal sensor fusion from accelerometer, gyroscope, and GPS data streams. The project built a compact deep CNN that stacked three convolutional layers using 32, 64, and 128 filters in turn, added max-pool blocks, and capped the pipeline with two dense layers that delivered the final class labels. What sets the work apart is borrowing the 1D CNN idea from speech tasks to study in-car time series, feeding the network extra noise and time-warp shifts to widen the training set, and wiring a lightweight attention head that highlights rare but important moments behind the wheel. Feature detectors sweep over the signal with kernels sized 3, 5, and 7 on each 1D layer, while batch norm sits after every convolution so gradients flow smoothly, a 30-percent dropout clips excessive weight, and the Adam optimizer starts at 0.001 then slows down as loss plateaus. Inside those layers ReLU clears out negatives, and a softmax out front sorts the learned patterns into the right driving class. Before training the raw log data slides into overlapping 10-second windows that feed the model with context-rich, easy-to-read mini-sequences.The study used an uneven 80-20 train-test split, grouped by driver profile, and stopped early when validation loss failed to improve after ten rounds. Yet the approach still demands at least 10,000 samples for each behavior, treats the network as a black box, and runs only on machines with fast GPUs. On this platform the convolutional model tagged driver actions-aggressive, normal, cautious, drowsy, or distracted-with 94.3% accuracy, beating support vector and random forest classifiers by 6 to 8 points. F1 scores stayed above 0.92 for every class, and the system generalized well, recognizing safe and unsafe habits across makes, models, and road scenes.

Recent work by Liu et al. [20] proposed Long Short-Term Memory (LSTM) networks for analyzing temporal sequences in driver behavior data collected from vehicle CAN bus systems over extended driving sessions. The study built a bidirectional LSTM with 128 hidden units going forward and backward to grab both past and future patterns in 5-minute driving windows. New pieces include scaled dot-product attention that zooms in on key time steps in aggressive maneuvers, a sequence-to-sequence setup that forecasts behavior 30 seconds ahead, and teacher forcing in training that starts at 100% chance and slides to 50% across epochs. Inside the model, each LSTM cell runs forget, input, and output gates to tame sequences as long as 300 steps; gradients are clipped at 1.0 to stop them from exploding; and early stopping kicks in when validation perplexity stalls for 15 epochs. The layout holds embedding layers for categories, LSTM stacks with 20% dropout, and final dense layers using softmax for a multi-class call. Before training, raw data is

cleaned with z-score standardization and shorter trips are padded so every sequence matches in length.The researchers built special loss functions that mix standard cross-entropy with penalties for sudden jumps, so the cars learn to switch smoothly between actions.The proposed approach is mathematically intensive and requires state-of-the-art infrastructure to maintain 100 ms inference, as well as sequential preprocessing mechanisms that can lose detail in padding, and in some cases, the model may crash due to excessive GPU memory consumption when dealing with long driving clips. Despite these constraints, a Bidirectional LSTM achieved 91.7 percent accuracy in classifying aggressive, normal, cautious, and erratic driving, with precision scores of 0.89, 0.94, 0.87, and 0.93 respectively. An added attention layer flagged key moments that matched human annotators 85 percent of the time, showing that recurrent networks can learn the subtle timing patterns present in real-world driving.

The study of Chen et al. [21] considered hybrid methods in the combination of the traditional machine learning algorithms with the deep learning structures to improve the classification of driver behavioral patterns based on the multi-source telematics data. In this paper, scholars combined the support vector machine with the convolutional neural network in a two-step pipeline: the SVM initially reduced the input to major features and reduced dimensions, before the CNN processed the survivors to perform the final pattern detection and labelling. New concepts consist of the feature-weighting scheme, which in real time feigns the worth of each attribute, contingent on the driving circumstances, a voting methodology, which combines conjectures between the various models by weighted mean, and meta-learning rules, which choose the most advantageous mix of models dynamically given the available data. Radial-basis SVM set C=1, gamma=0.1, to reshape the feature space, a CNN with 3 layers of 64, 128 and 256 filters and a global average pooling layer used as the automatic extractors, and an ensemble vote where each models voice is scaled by its own confidence score, were core methods. In order to reduce overfitting the hybrid also boot-strap aggregation was done and measured performance was determined by using 10-fold cross-validation to ensure that results were consistent and sensor readings were normalized using robust scaling to restrict the impact of outliers.Principal component analysis was applied to features before training, to reduce the number of dimensions, at the expense of 95 percent of the data spread. The combined support-vector machine and convolutional-network configuration was tuned with a grid-based search which was terminated prematurely because the validation accuracy no longer increased. Nonetheless, the method required about three times as much training time as the single models, required cautious alignment of the settings in the two algorithms, and consumed additional memory by maintaining multiple model implementations when making predictions. The last ensemble was 95.2-percent accurate in tagging 6 driving styles-aggressive, normal, cautious, distracted, drowsy, and erratic-outperforming SVM alone and CNN alone at 87.3-percent and 94.1-percent, respectively. It was also more stable; the accuracy across cross-validation folds reduced to only 1.4 percent with the hybrid as compared to 2.8 percent with the single models, indicating that it fits better to varied conditions on the road and weather conditions.

The study conducted by Kumar et al. [22] was an in-depth analysis of ensemble learning methods in the classification of driver behavior whereby bagging, boosting, and stacking methods were used with detailed performance evaluation across various data sets. The experiment noisily compared Random Forest with 200 trees, AdaBoost with 50 weak learners, and Gradient Boosting with a 0.1 rate with an even looser stacking combination constructed of Decision Trees, SVMs and Naive Bayes. The three-layer stack and 5-fold cross-validation were introduced to reduce overfitting, new diversity scores based on correlation and disagreement were introduced but hand-picked complementary models, and finally, sleek voting combined soft probabilities and weight in real-time and varied as the classifiers themselves expressed their confidence. The classic tricks used there to remain diverse were random forests that were bootstrapped and feature-randomized in order to cut out some variance, AdaBoost that turned exponential loss to highlight hard cases, XGBoost that soothes with alpha of 0.1 and lambda of 1.0, and the stacking bundle itself, relying on out-of-fold logits to train its meta-learner, to ensure that no single training case can leak in. We performed Bayesian search on dozens of knobs under the direction of a Gaussian Process, allowing the framework to explore multiple algorithms simultaneously rather than searching a grid one after another. That having said, the design is slow-it is five to ten times longer than one of a single design, uses more copies of weights in RAM when performing an inference, and packages together patterns that are more difficult to unravel. Nevertheless, the stacked configuration was found to achieve the accuracy of 93.8 percent, precision of 0.941, recall of 0.935, and F1 of 0.938 in terms of driver moods detection; Random Forest scored 92.1 percent, AdaBoost at 90.7 per cent, and XGBoost made 92.7 per cent. A further inspection revealed that three to five classifiers correlated by pairs with less than 0.6 squeezed out the optimum results again demonstrating that differences-not duplicates-make ensembles shine.

Patel et al.[23] conducted a study aimed at comparative analysis of different neural network designs as driver behavior models such as Multi-layer Perceptrons (MLP), Radial basin function (RBF) networks, and Autoencoders based models of learning features unsupervised with high dimensional telematics data. We constructed feed-forward neural networks and experimented with a broad range of activation functions, layer configurations, and regularization tricks in order to identify the most successful configuration in this work in the differentiation between various driving actions. New things are the side-by-side comparison of ReLU, Sigmoid, Tanh, and Leaky ReLU on real-world driver data, variational autoencoders that reduce the size of the initial 150 features to a small 30 and learn their uncertainties, and a composite regularization, where a small L1/L2 weight penalty (alpha=0.01) is combined with dropout (rate=0.3) to ensure the model is not overfitted. Our main structure was a three-layer MLP with 128, 64, and 32 neurons, where He was used to initialize weights, and batch norma-lization was placed between neurons, and an encoder-decoder stack that created a shrinky reproduction of inputs between 150-64-32-16 and 16 hidden codes and then ran the opposite way out to train on unsupervised data.During training, we used the Adam optimizer with an initial learning rate of 0.001, and reduced it by Within the model we further stacked in batch norm to smooth internal covariate shift, clipped gradient to ensure that they never ex-

ploded, and averaged network outputs across multiple copies to smooth the final output.Once the autoencoders were completed, the next step was to test various tuning parameters for the MLP. This included learning rates (0.001, 0.01, 0.1), sizes for the hidden layers, and weights for the decay. All of these were exhaustively searched for and adjusted based on the grid search. The small datasets led to the risk of overfitting, and training would have to be constantly monitored. Setting the autoencoders aside, training the MLP (plain architecture with ReLU) yielded 90.5 percent accuracy after 45 minutes on regular hardware. The autoencoder only obtained 88.2 percent accuracy, but it reduced the feature set by 80 percent, which allowed for faster inferences and generated lightweight models that fit within the automotive memory and storage constraints.

The study by Williams et al. [24] introduced comprehensive unsupervised learning techniques for discovering hidden patterns in driver behavior without requiring labeled training data, addressing the challenge of expensive manual annotation in large-scale telematics datasets.The study tested and compared three clustering methods-K-means, Gaussian Mixture Models (GMM), and DBSCAN-to uncover distinct driving habits from six months of unlabeled vehicle sensor logs. Noteworthy advances include a silhouette analysis boosted by resampling that pinpoints the best number of clusters from k=2 to k=15, a GMM built on Expectation-Maximization ranked by the Bayesian Information Criterion so it self-adjusts, and a DBSCAN that adapts its epsilon by reading k-nearest neighbor distance plots to spot dense groups and flag noise. Each method worked with Euclidean distance for basic geometry, Manhattan distance when the sensors were on different scales, and cosine similarity for sparse, high-dimensional vectors that record the frequency of each driving pattern.The data cleaning process started with robust scaling to minimize the impact of the outliers and was then followed by a principal component step to shrink the features while maintaining 90 percent of the variance. The patterns were then compressed into sliding windows of one hour, with a thirty-minute overlap. For assessing the quality of the clusters, all three evaluation methods were used in unison: the silhouette score, Calinski-Harabasz value, and the Davies-Bouldin measure. For streaming records, incremental models were used, and stability was assessed with a hundred-bootstrap subsampling test.Even so, the final groups require an expert's eye; there are no obvious hyper-parameter settings without true labels, so sensitivity sweeps become time-consuming, and memory and speed, with millions of trips, push the work to a distributed setup. Using the Gaussian mixture model, we achieved the highest quality of clusters-an average silhouette score of 0.67-and distinctly separated five groups of drivers. 15 percent were aggressive, 45 percent were normal, 25 percent were cautious, 10 percent were erratic, and 5 percent exhibited mixed behavior. K-means performed approximately 10 times faster, returned a score of 0.61, and thus was more suitable for real-time applications. On the other hand, DBSCAN flagged 12 percent of the data as noise and captured peculiar driving styles that other techniques failed to identify.

The architecture proposed by Nguyen et al. [25] supports a lightweight convolutional neural network (CNN) architecture with an attention mechanism to accurately recognize driver behaviors and support real-time driver warning systems. The primary problem addressed is the rising rate of traffic accidents caused by driver distrac-

tion—behaviors like phone use, drinking, or adjusting radios—that remain inadequately detected by expensive or intrusive systems. Their solution is a non-invasive, low-cost, high-accuracy network designed for use even in older vehicles. The model incorporates standard and depthwise separable convolution layers, adaptive connections inspired by ResNet/Inception modules, and a Convolutional Block Attention Module (CBAM) to extract and focus on salient visual features. The final classification relies on a global average pooling layer and softmax output, minimizing parameters while maintaining precision. The model was trained and tested on three benchmark datasets—State Farm, AUC v1, and AUC v2—achieving classification accuracies of 99.95%, 95.57%, and 99.61% respectively across ten driving behavior classes. It runs efficiently even on low-end hardware like Jetson Nano, processing up to 243 FPS on GPU and 14.78 FPS on embedded systems. Ablation studies further confirm the contribution of CBAM and adaptive connections to the network's accuracy.

In their work, Xing et al. [26] propose a deep learning-based driver activity recognition system aimed at enhancing intelligent vehicle safety by detecting common driving behaviors and distractions using only RGB images. The study addresses the critical issue of driver distraction—a major contributor to road accidents—by offering a non-intrusive, real-time monitoring approach that bypasses the need for manual feature extraction or specialized hardware. The system classifies seven activities: four normal driving tasks (e.g., mirror checking) and three secondary/distraction tasks (texting, phone use, and radio operation). To train the models efficiently, transfer learning was applied to pre-trained CNN architectures—AlexNet, GoogLeNet, and ResNet50. The authors used a Gaussian Mixture Model (GMM) to segment the driver from the background before feeding the images into the CNNs. This preprocessing step significantly improved recognition accuracy. Among the models, AlexNet yielded the best performance, achieving 81.6% accuracy for multi-task classification and 91.4% accuracy for binary distraction detection. GoogLeNet and ResNet performed slightly worse, likely due to their deeper architectures being less effective with the limited dataset and the specific fine-tuning strategy used. Visualization techniques confirmed that GMM segmentation allowed CNNs to focus on relevant driver features, enhancing detection.

A significant contribution was made by Carsten et al. [27] , who conducted a comprehensive real-world analysis of driver behavior and its relation to crash and near-crash events, addressing the inadequacies of traditional simulator or self-report-based research. The primary issue tackled is the underestimation of the role of driver inattention in traffic incidents. This study uses naturalistic driving data collected from over 3,500 participants across 35 million miles of travel, captured via in-vehicle video and sensor equipment. The authors adopt a vehicle-based observational approach—known as Naturalistic Driving Studies (NDS)—which provides continuous, real-time data on actual driver behavior and vehicle performance. Unlike experimental setups, this methodology captures genuine distractions, impairments, and behaviors under normal conditions, offering high ecological validity. Their findings significantly revise previous understandings: driver-related factors account for approximately 90% of crash causation, with distraction alone involved in over 68% of crashes and 62% of near-crashes. Key contributors include mobile phone use, drowsi-

ness, emotional driving, and interactions with passengers. Critically, secondary tasks like texting and dialing were found to increase crash risk by more than sixfold. In contrast, tasks such as talking or listening on a hands-free phone showed negligible impact. This research supports a paradigm shift in road safety interventions, emphasizing the need for real-time monitoring, in-vehicle alert systems, and stricter policies on mobile device usage. However, the study acknowledges limitations, including potential biases in self-selection of participants and the generalizability of findings across different geographic or demographic contexts.

The experimental study by Bassani et al. [28] evaluates the effectiveness of an auditory Driver Distraction Warning (A-DDW) system in improving the performance of distracted drivers under simulated conditions. The study addresses the growing problem of road accidents linked to driver distraction and the uncertain impact of real-time monitoring systems on driving behavior. A total of 42 drivers were divided into three groups: undistracted, distracted, and distracted with A-DDW support. Participants navigated a simulated motorway while performing secondary tasks, such as solving math problems on a tablet. The A-DDW system employed infrared sensors to detect prolonged off-road glances, triggering auditory alerts. Key performance metrics included longitudinal control (speed and its variation), lateral control (lane position and deviation), and perception and reaction times (PRTs). The results showed that the A-DDW system marginally improved lateral control by reducing lane deviation in distracted drivers. Increased cognitive load was evident as this also lengthened auditory reaction times. Observed gender differences were that males were faster drivers with A-DDW and females drove slower than undistracted drivers. Although the A-DDW system lessened some effects of distraction, driving performance still did not return to baseline. The findings suggest that while such systems offer partial benefits, particularly in lane-keeping, they may inadvertently increase mental workload. Further refinement and integration with complementary safety measures are needed to enhance their practical effectiveness in real-world settings.

In pursuit of real-time performance, Jo et al. [29] implemented a vision-based driver monitoring system capable of detecting both drowsiness and distraction simultaneously—marking an advancement over earlier models that focused on only one form of inattention. The key issue addressed is the limited scope of existing systems, which often fail to comprehensively assess driver attentiveness, increasing the risk of missed detections and potential accidents.
Their method adopts a conditional detection strategy: if the driver's gaze is forward, drowsiness is assessed via eye state analysis; otherwise, distraction is inferred from head orientation. Particularly during head movements when eye detection often fails,this approach improves detection reliability and reduces computational load.
The system integrates a hybrid eye-detection method combining Adaboost, template matching, and blob detection. These are further enhanced using support vector machines (SVMs) trained on features derived from principal component analysis (PCA) and linear discriminant analysis (LDA). An eye state classifier is also introduced, incorporating PCA/LDA features with histogram-based descriptors like sparseness and kurtosis for more precise classification.
Using a single near-infrared camera with a narrow bandpass filter, the system en-

sures robustness under varied lighting conditions and driver appearances, including sunglasses. Testing on over 160,000 frames from diverse scenarios demonstrated 99% eye detection accuracy and 98% effectiveness in identifying either drowsy or distracted behavior, confirming the system's reliability for real-time in-vehicle deployment.

In their pioneering study, Saito et al. [30] explore the use of a driver's line of sight as both a human-machine interface (HMI) and a means of assessing physical or mental driver states in vehicular environments. The central problem addressed is how to improve driver interaction and safety by leveraging non-contact eye-tracking technologies to reduce distraction and detect hazardous states like drowsiness or fatigue. The proposed solution is twofold. First, for informational communication, they introduce an "Eye Switch" system where drivers can control vehicle functions (e.g., headlights or radios) by fixating on options projected via a Head-Up Display (HUD). Based on Hick's and Fitts's laws, the study estimates that this method can reduce operation time by half compared to conventional push buttons. Second, the paper suggests using eye movement patterns to detect driver conditions such as fatigue or inattention—factors implicated in about 25% of fatal crashes in Japan at the time. Technically, the study evaluates both contact-based and remote eye-tracking methods. They emphasize the use of bright-eye imaging and corneal reflection through cameras and image processors for accurate and unobtrusive gaze tracking. They reference systems by Hutchinson and Tomono, which use multi-camera and illumination setups to measure gaze direction within milliseconds.

This dataset proposed by Moosavi et al. (2019) [31] addresses the issue of limited and typically unavailable accident datasets that discourage the reproducibility as well as high-scale traffic safety research. Accessible resources are either small-scale, proprietary, outdated or incomplete about significant contextual factors such as the weather, road networks, as well as points of interest, hence limiting predictive research on the likelihood of accidents. To address these challenges, the writers brought forward US-Accidents, a publicly accessible big dataset comprising more than 2.25 million accident records gathered within the contiguous United States over the period 2016-2019. Integrating the streaming traffic data, reverse geocoding, weather, temporal features, as well as points-of-interest annotations, the dataset builds up one of the largest accident resources that is readily accessible. Using this data, the authors approximated distributions of accidents across time as well as space. They found that there were 40% accident observations along high-speed roads, 32% along local roads, and most of the accidents during the weekday commute periods (8 AM as well as 5 PM). In addition, about 73% of the accidents occurred during the day, corroborating spatiotemporal as well as contextual distributions of accident likelihood. In future research, they indicate that they would like to use this dataset for the prediction of real-time accidents, urban planning, infrastructure design assessment, as well as insurance modeling, offering researchers an open platform to improve traffic safety innovations.

The framework put forward by Moosavi et al. (2019b) [32] is designed to estimate accident likelihood in real time, a task challenged by the rarity and diversity of the data that are available. Shallow models or regression-based traditional algorithms

usually cannot identify dense spatiotemporal relationships and cannot handle the rarity of rare accident occurrences. In order to address these problems, the authors developed DAP (Deep Accident Prediction), a deep structure that integrates heterogeneous sources of data such as traffic incidents, weather, time features and points of interest. The structure utilizes LSTM modules to address temporal sequences, dense layers to address static data, as well as trainable embeddings to learn spatial heterogeneity, making it possible to strongly predict risk based on sparse but readily collected data. It was contrasted vs. Logistic Regression, Gradient Boosting and traditional Deep Neural Networks across six notable American urban regions on the basis of the US-Accidents dataset (2.25M rows). This research achieved up to 16% F1-score improvement on the rare accident prediction task, reflecting the advantage of the combination of contextual as well as temporal factors under deep learning. As future work, the authors recommend adding other publicly available datasets, e.g., demographic statistics and annual traffic flow summaries, to enhance predictive resilience as well as globalizability to real-world applications within smart transport systems.

Zhao et al. (2023) [33] address the challenge of comprehending the cause of accidents and precisely labeling the degree of injury in big road traffic data. Classical techniques are dependent on the expertise-based judgment and do not consider the interacting relationship among the factors influencing, where redundant as well as high-dimensional features degrade the performance of the model. To address this, the writers suggested a complex network–based analytical framework that treats accident-related factors as nodes and their correlations as edges. Based on the network measures degree, clustering coefficient, betweenness, and closeness centrality, the framework selects the most influential features. These compressed features are further used on machine learning classifiers like Logistic Regression, Decision Trees, Random Forest and XGBoost to make accident severity prediction. Evaluation was performed on more than 1 million U.S. accident cases (2016-2023) based on recall, precision, F1-score and ROC analysis. Experiments indicated that XGBoost obtained the best performance among other models, significantly the identification of serious injury cases, proving the effectiveness of dimensionality reduction based on networks in the applications of traffic safety. As future work, the authors recommend expanding the model further, combining driver demographics, vehicle properties, and traffic flow data, as well as sophisticated deep learning techniques, to identify richer patterns to allow more precise cause analysis as well as severity classification.

Zhao and Deng (2022) [34] a model that considers the time span of the traffic accident, an important variable that is utilized to reduce congestion, accelerate accident rescue and provide accident safety insight. However, traditional regression, Bayesian networks or a single machine learning algorithm is limited as a result of small-sample sizes, local applications and the requirement to wait until after the accident occurs to gain access to the variables, which prevents the applications of these traditional algorithms during the control period. They created a heterogeneous ensemble learning framework, wherein they combine XGBoost, LightGBM, CatBoost, stacking and elastic networks for the purpose of accident duration prediction. It is trained on the big-scale US-Accidents dataset (2.3-plus-million-records, 2016-2020) and concentrates on features that are comprised of five domains, i.e., traffic, location,

weather, points of interest, as well as time. It uses comprehensive preprocessing, including outlier processing, missing value imputation, categorical encoding, along with feature engineering, so as to enhance the utility of the data as much as possible as well as uncover extra patterns. It performed very well as measured by MAPE, MAE, MSE, surpassing that of one-stage models in forecasting accuracy. The results obtained also indicated that the most important factors contributing to duration prediction based on the SHAP-based interpretability analysis are accident time, location and weather. As future work, the authors suggest applying federated learning to integrate heterogeneous multi-source transport data securely, preserving the privacy of the data, to enhance the scalability and stability of prediction systems.

The study by Arciniegas-Ayala et al. (2024) [35] addresses the persistent global problem of traffic accidents, which remain one of the leading causes of death and injury worldwide. Traditional accident prediction approaches are computationally intensive, often relying on deep learning (DL) models with complex training requirements. These models are difficult to adapt dynamically when new driving data become available, limiting their utility for real-time risk alert systems. In order to draw a comparison, the authors also recommend comparing Radial Basis Function Neural Networks (RBFNNs) with computationally intensive DL networks like Convolutional Neural Networks (CNNs), Random Forests (RF), and Multilayer Perceptrons (MLPs). PoliDriving dataset, created here, comprises 2634 samples under driver, vehicle, weather and accident data and has been created keeping in mind the requirement to simulate a dynamic dataset that gets updated continuously. Performance was defined based on accuracy, specificity, sensitivity, as well as execution time. We found that the best accuracy (96.04%) came about from the combined CNN-RF, while the RBFNNs obtained comparable accuracy (91%) with very fast processing (up to as little as 1.7 seconds for the SVC-RBF) to allow real-time applications. In future work, the authors recommend the use of dynamic datasets extensively to conduct the ongoing retraining of the model and explore lightweight, adaptive ANN architectures to build scalable, secure and real-time driver warning systems to avoid crashes.

This survey carried out by Wang et al. (2021) [36] points to the chronic nature of global traffic casualties, which result in millions of casualties each year, as well as great social and economic impacts. Classical accident analysis and forecasting methodologies are restricted due to the small-scale data, poor capability to generalize and inadequate consideration of situational data like traffic, meteorological and driving behaviors. These limitations confine the precision as well as usability of prediction systems in practical applications. To fill the existing gaps, the authors provide an academic overview of the new advances in the research and development of traffic accident prediction using various statistical, traditional machine learning and recent deep learning approaches. They focus intently on the study of spatiotemporal patterns, heterogeneous data integration, and the identification and extraction of risk factors, all of which lead to increased accuracy in prediction. Progress is analyzed based on evaluating predictive accuracy, interpretability, and scalability across a variety of methodologies on disparate datasets and real-life situations. Among the various deep learning models, LSTMs and CNNs are the most effective in capturing and modeling the nonlinear, time-dependent aspects of the data, while ensemble

techniques provide increased robustness. Recommended future directions include the integration of multi-source big datasets (traffic sensors, social media, and IoT) with real-time explainable AI (XAI) for transparency, and the development of real-time dynamic systems that monitor and adapt to changes in traffic patterns and driver behavior.

The study by Tang et al. (2025) [37] focuses on the common issue of traffic accidents, which lead to significant loss of life and property worldwide. Traditional models for predicting accident severity have limitations due to imbalanced datasets, inadequate preprocessing and difficulty recognizing minority classes (like moderate and severe accidents). These factors reduce their accuracy and practical use in intelligent driving systems. To tackle these problems, the authors suggest a better machine learning framework that combines multi-stage prediction with ensemble learning. First, they use a multi-layer perceptron (MLP) for initial severity prediction. The MLP's outputs are combined with original features to improve representation. This enhanced data is then used by secondary classifiers, including Decision Tree, XGBoost and Random Forest, to provide stronger and more accurate severity classification. To address data imbalance, they apply both under-sampling and over-sampling techniques, along with preprocessing methods such as encoding categorical variables and standardizing numerical features. The model was tested using the US-Accidents dataset from 2016 to 2022. The results showed that the combination of MLP and Random Forest delivered the best performance, achieving an accuracy, recall, and F1-score of 94%, which is significantly better than single models. For future research, Tang et al. recommend using real-time traffic flow data, socio-behavioral factors and adaptive learning methods. Their goal is to enhance dynamic early-warning systems for both drivers and autonomous vehicles and to aid in urban traffic planning and accident prevention strategies.

The study by Manzoor et al. (2021) [38] tackles the ongoing global issue of predicting traffic accident severity. This prediction is vital for lowering deaths, economic losses and response times for rescues. Traditional statistical methods like logit and probit struggle with complex datasets. Single machine learning models often miss the intricate nonlinear relationships between accident factors. To address these issues, the authors introduce RFCNN, an ensemble framework that combines Random Forest (RF) and Convolutional Neural Network (CNN) at the decision level using soft voting. The dataset includes over 4.2 million U.S. accident records from 2016 to 2020, featuring 48 variables. The authors narrowed this down to the 20 most important factors using RF feature importance. These include key elements such as distance, temperature, wind chill, humidity, visibility and wind direction. The RFCNN model performed better than the base classifiers—AdaBoost, Gradient Boosting, Extra Trees and Voting Classifier. With the top 20 features, RFCNN achieved an accuracy of 0.991, a precision of 0.974, a recall of 0.986 and an F1-score of 0.980. It surpassed all baseline models. For future research, the authors recommend testing RFCNN on various datasets to check its generalization. They also suggest refining the model to lessen computational complexity, making it easier to use in real-time traffic accident management and prevention systems.

The study by Yan and Shen (2022) [39] looks at predicting the severity of traffic accidents. This task is important for managing traffic safety and emergency responses. Traditional statistical models, like logit and probit, have rigid assumptions and limited flexibility. While artificial intelligence methods are usually more accurate, they often act like "black boxes" and lack clear explanations. Additionally, tuning hyperparameters in machine learning models, such as Random Forest (RF), takes a lot of computing power when using grid search, which leads to efficiency issues. To tackle these problems, the authors propose BO-RF, a mixed framework that combines Random Forest with Bayesian Optimization (BO) for quicker hyperparameter tuning. This method uses accident data from the US-Accidents dataset, focusing on 30,426 records from Montgomery County, Pennsylvania. It includes 15 feature variables related to traffic, time, weather and points of interest. RF is used for classification, while BO optimizes parameters like tree depth, number of estimators and maximum features, which greatly cuts down on computing time. The model was assessed using precision, recall, F1-score, and AUC and it did better than ANN, KNN, SVM and the baseline RF. BO-RF achieved an AUC of 0.9625 and an F1-score of 0.57, showing improvements of over 40% compared to ANN and more than 100% compared to KNN. For future research, Yan and Shen suggest using BO-RF on larger urban datasets, adding more factors that influence accidents and improving clarity through partial dependence plots. This will help develop increasingly data-driven policies aimed at sustainably mitigating the severity of accidents and improving traffic management.

The research by Li et al. (2023) [40] attempts to predict traffic accidents at the road level. this is a challenging task due to the data scarcity and the unpredictability of traffic situations. Many traditional models and a good number of deep learning techniques fail to estimate the uncertainity of the problem. This leads to predictions that are too confident and to poor choices in critical situations, such as those involving intelligent transportation systems. To deal with these issues, the authors propose a novel uncertainty-aware probabilistic graph neural network (PGNN). This framework regards the road network as a graph, with roads as nodes and spatial connections as edges. PGNN is capable of measuring predictive uncertainty and estimating spatiotemporal relationships in the framework of Bayesian inference and the graph neural network's spatial predictive uncertainty. As a result, a system that predicts accident risk and estimates confidence adds to its real-world applicability. The model was evaluated on two large-scale real-world datasets (New York City and Los Angeles accident data). The findings support the claim that PGNN is a considerable improvement over the most advanced models. It achieves lower prediction errors and produces better-calibrated uncertainty estimates, making it a more dependable option for practical use. For future research, Li et al. suggest expanding PGNN to incorporate multi-modal traffic data integration, such as driver behavior, video feeds, and weather. They also aim to improve the scalability for real-time citywide accident prediction, ultimately supporting safer and smarter urban mobility systems.

Grigorev, Mihăiță, Saleh, and Chen (2024) [41] dealt with the acute issue of imperfect and inconclusive traffic accident reports wherein human errors frequently cause

uncertain records of accident initiation time and duration. These anomalies impede incident duration prediction models from achieving accuracy and consequently make traffic management and emergency response systems less efficient. To overcome such limitations, authors propose a new disruption segmentation method that incorporates two large datasets: the Countrywide Traffic Accident Dataset (CTADS) and the Caltrans Performance Measurement System (PeMS). Their method automatically assigns reported accidents to vehicle detector station (VDS) data and applies advanced segmentation methods through use of Chebyshev and Wasserstein distances to detect traffic speed disruptions and indirectly estimate accident durations from traffic flow patterns. This method eliminates user-input errors and allows disruption-based duration estimation in a highly accurate manner. The model is validated on over 9,000 accidents from San Francisco and achieved significantly better performance. CatBoost emerged as the best-performing model that achieved an MAE of 17.55 minutes and RMSE of 22.55 and outperformed baseline models such as Ridge Regression, Random Forest, and SVM. For further work, they suggest extending this work to real-time identification of initial accidents, integrating event and meteorological data to facilitate multimodal fusion, and to generalize towards varied international data. There are additional mentioned opportunities to integrate semantic segmentation with deep learning for disruption shape analysis enhancement and cascading event forecasting.

The review by Mase, Chapman and Figueredo (2023) [42] analyzes intricacies around risk driving behavior assessment, which remains unsolved, and also interplays of vehicle, environmental situational context and driver behavior, vehicle conditions, and the environment. Traditional risk assessment methods, like surveys and simulations, do not adapt in real-time and often miss important uncertainties. This limits their effectiveness for use in intelligent transportation systems. To address these issues, the authors examine recent advancements in intelligent systems for driving risk assessment. This includes unsupervised learning, supervised learning, Bayesian inference, deep learning and fuzzy logic models. These methods use various data sources such as telematics, GPS, in-vehicle sensors, physiological monitoring and cameras to identify risky driving behaviors. A key contribution of the review is integrating psychological theories such as the Theory of Planned Behavior, Risk Homeostasis Theory and Multiple Resource Theory into computational modeling. This provides a better understanding of how drivers make decisions. The authors evaluate the effectiveness of these approaches based on aspects like interpretability, handling uncertainty, scalability and real-world applicability. Deep learning is effective in processing unstructured data, while fuzzy logic manages uncertainty well. However, both face challenges with transparency and complexity. For future work, the authors suggest creating context-aware, explainable AI frameworks that combine different data sources. These frameworks should also address privacy, fairness and trust to support the scalable use of intelligent risk assessment systems for personal and commercial drivers.

The study by Naveed, Perwaiz, Sultana, Ahmad and Fraz (2025) [43] addresses the significant lack of region-specific driver behavior datasets in low- and middle-income countries (LMICs). In these areas, varied traffic, poor infrastructure and cultural differences make global datasets less relevant. Most existing resources come from

developed countries, collected in controlled settings that do not reflect the chaotic road conditions typical of countries like Pakistan. To fill this gap, the authors introduce V-SenseDrive, the first multimodal dataset collected entirely in Pakistan. This dataset combines in-vehicle sensors, including accelerometers, gyroscopes, magnetometers and GPS, with synchronized road-facing video. The data were gathered using a dual-smartphone setup across different types of roads: urban arterials, highways and rural roads. It covers three driving behaviors: normal, aggressive and risky. The data collection focused on synchronization, diverse routes and a range of environments. The dataset is organized into raw, processed and semantic layers to allow for flexible analysis later on. The dataset's reliability was evaluated through exploratory analysis. The results showed clear distinctions between classes, with speed, jerk and yaw rate being the most effective features for differentiation. Tree ensemble models, specifically Random Forest, XGBoost and HistGradientBoosting, achieved impressive macro-F1 scores, surpassing the performance of linear and sequence models. For future research, the authors suggest enhancing the dataset by adding event-level labels, multimodal fusion models and transformer-based architectures. This would enable explainable and context-aware accident risk detection while also advancing the development of advanced driver assistance systems (ADAS) suited to LMIC environments.

The study by Raza, Akhtar, Abualigah, Zitar, Sharaf, Daoud and Jia (2023) [44] the pressing issue of unsafe driving behaviors, which greatly contribute to road accidents, injuries and deaths around the world. Traditional methods for detecting driver behavior often depend on standard machine learning models that have limited accuracy and poor feature representation. This leads to ineffective early detection of risky driving patterns. To address these issues, the authors introduce a new method for feature engineering called LR-RFC. This method combines Logistic Regression (LR) and Random Forest Classifier (RFC) to create probabilistic features from smartphone motion sensor data. The improved set of features is then used with machine learning models such as Logistic Regression, Support Vector Machine, Gaussian Naive Bayes and Random Forest, along with a deep learning RNN model for classifying behavior. The dataset is collected with the driver behavior categorization of slow, normal, and aggressive, utilizing accelerometer and gyroscope sensors. Experimental results confirmed the strongest model was a Random Forest with LR-RFC features which attained 99% accuracy, greatly exceeding all other baseline approaches. This was demonstrated by k-fold cross-validation and hyperparameter tuning. Future works are expected to target improved class balancing strategies, advanced deep learning, transfer learning, and the authors increasing detail on sensor data for real-time systems to improve early accident threshold warning systems.

The study by Islam, Rony, Safran, Alfarhood and Che (2024) [45] tackles the urgent issue of road accidents due to unsafe and aggressive driving. Current machine Detection of driver behavior by using learning and deep learning is often in need of large, varied datasets. They are not effective in dealing with skewed data and are not well adjusted in reality. time, which precludes their use in practice.The authors propose a new RKnD feature as probabilistic to solve these issues. engineering method. The technique is a combination of three algorithms: Random Forest They are (RF), K-Nearest Neighbors Classifier (KNC) and Decision Tree (DT). It generates bet-

ter chance characteristics to classify driver conduct. The method uses The data of smartphone motion sensor of a Samsung Galaxy S10, including driving behaviors. labeled as slow, normal and aggressive. [50]his deep learning model is meant to learn se- causal dependencies in time-series driving data. The study makes use of the UAH DriveSet dataset, which is a collection of more than 500 minutes of actual driving information on six. drivers on five classes of roads. The actions are classified into normal, drowsy and. aggressive The authors applied the Synthetic Minority Oversampling Technique (SMOTE) to tackle class imbalance. They also used k-fold cross-validation and hyperparameter tuning for thorough evaluation. Experimental results showed that the LSTM model with RKnD features reached 99.63% accuracy. This result significantly surpassed both the original features and competing methods. The RKnD approach also made the results easier to understand by improving feature separability in scatter plots and confusion matrices. For future work, the authors recommend refining RKnD for real-time use, expanding datasets to cover various driving situations and incorporating advanced deep learning techniques like self-organizing maps. They also plan to validate practicality through real-world deployment and field testing while identifying potential challenges.

The study by Chah, Lombard, Mualla, Bkakria, Abbas-Turki and Yaich (2024) [46] deals with the shortage of large, labeled and privacy-protecting driver behavior datasets for risk profiling and insurance uses. Collecting real-world driving data is expensive, raises privacy issues and often fails to capture extreme or dangerous driving situations. Previous simulation-based datasets have also had limitations in scale, road variety or consistent labeling. To tackle these problems, the authors put forward a SUMO-based driver simulation framework to create a detailed dataset of driver behaviors. By using the Enhanced Intelligent Driver Model (EIDM) and the Miami city road network, they simulated three different driving styles—slow, normal and dangerous—under various traffic conditions. They developed a new AlertDang profiling method to convert violations like speeding, unsafe distance, emergency braking and time-to-collision into warnings that classify driver behavior. Data was gathered through the TraCI interface and saved as in-vehicle parameters and warning datasets. The resulting dataset was tested with machine learning models (GBDT, KNN, MLP, and SVM), achieving accuracies of up to 89% in training and 81% in testing. The simulation method showed it could be scaled, reproduced and reduced the reliance on sensitive personal data. For future research, the authors suggest adding to driver profiles with partial trajectory-based behaviors, broadening warning types and incorporating federated learning with privacy-boosting technologies like differential privacy and homomorphic encryption to ensure safe use in applications like Pay-As-You-Drive (PAYD) insurance.

The study by Fazzinga, Flesca, Furfaro, Monterosso and Trubitsyna (2024) [47] tackles the issue of classifying driving behaviors, which is important for road safety analysis and preventing accidents. Traditional methods for recognizing driver behavior depend solely on sensor data. These methods often struggle with conflicting or noisy measurements, leading to inaccurate classifications. This issue is worsened when different devices provide varying readings for the same driving condition. To address these challenges, the authors introduce ASYDE (Argumentation-based System for classifYing Driving bEhaviors), an innovative framework based on Dung's

Abstract Argumentation Framework (AAF). In ASYDE, inertial sensor readings, such as acceleration, speed and angular velocity, along with road information like speed limits, are treated as arguments. The classification of driving styles - calm, normal or aggressive—is presented as disputes among agents. Conflicts between sensor inputs are seen as "attacks" within the framework. The system resolves these conflicts through argumentation reasoning to generate a driving-style certificate that summarizes behavior over time. The prototype was tested using real-life driving data. The results showed that ASYDE could reliably differentiate between driving styles and handle uncertainty from sensor errors, performing better than methods that depend only on data. For future work, the authors recommend expanding ASYDE with more extensive datasets, incorporating contextual factors such as traffic and weather and adapting the system for real-time use in intelligent transportation systems and insurance applications.

The study by Maktoubian, Tran, Shillabeer, Amin, Sambrooks and Khoshkangini (2024) [48] tackles the challenges of modeling driving behavior. This behavior is essential for traffic safety, fuel efficiency, and risk assessment. Previous research does not provide a clear comparison of machine learning (ML) models across various driving-related tasks. There is also limited understanding of how driver identity, behavior profiling and vehicle performance interact. To fill these gaps, the authors suggest a systematic framework for evaluating models. They compare non-temporal models (Random Forest, SVM, XGBoost) with temporal models (CNN, RNN, LSTM, GRU, TCN) across three main tasks: estimating fuel consumption, identifying drivers and predicting driver activity. They used eight different datasets, including industrial sources like the Kungsbacka dataset and public resources such as KU, RasPi, Smartphone, UAH-DriveSet and CARLA. This variety ensures representativeness across real-world and simulated driving scenarios. They measured performance using MAE, RMSE, accuracy, F1-score and training time. The results showed that XGBoost consistently outperformed other models, ranking first in four out of nine tasks with high accuracy and efficiency, while CNNs performed the worst overall. Temporal models like LSTM and GRU also delivered strong results, but they required longer training times. For future work, the authors suggest improving the accuracy of driver activity prediction with high-quality datasets. They recommend integrating feature extraction models and extending the analysis to include data from electric and autonomous vehicles. This could help advance intelligent transportation and real-time safety systems.

The study by Vyas and Chaturvedi (2024) [49] addresses the growing issue of road accidents in developing countries such as India. In these regions, drivers of two- and three-wheelers constitute over 80% of the driving population but often lack access to Advanced Driver Assistance Systems (ADAS). Current ADAS solutions in high-end cars primarily rely on computer vision to detect drowsiness or distraction; however, they do not assess driver stress levels and are unsuitable for open-environment vehicles like motorcycles and auto-rickshaws. To overcome these limitations, the authors propose *DriverSense*, a wearable-device-based, vehicle-independent, multimodal ADAS framework. This framework integrates mobility sensors, such as GPS, accelerometers, and gyroscopes, with physiological sensors, including heart rate (PPG), SpO_2, and electrodermal activity, to detect driver aggressiveness and stress

in real time. Edge computing is employed to process data on connected smartphones, while a server module trains recurrent models, such as RNNs and LSTMs, for behavior and stress classification. Initial experiments using datasets and mobility sensor data gathered in Ahmedabad achieved 86.05% accuracy in classifying aggressive driving and 88.24% in detecting stress levels. Random Forest and Logistic Regression notably excelled at distinguishing between stressed and normal states. For future work, the authors plan to expand DriverSense with large-scale, diverse datasets, test it in noisy real-world traffic and tackle privacy concerns related to wearable device data. Possible applications include customizing insurance, monitoring by RTO and real-time collision avoidance through shared alerts.

The study by Pande, Wawage and Deshpande (2024) [50] tackles the critical issue of road accidents, which result in over a million deaths each year. Unsafe driving behaviors have a significant impact on this problem. Traditional methods for classifying driver behavior often use shallow machine learning models. These models have difficulty capturing temporal dependencies and complex driving patterns, which limits their ability to distinguish between nuanced behaviors like drowsy and aggressive driving. To address these issues, the authors recommend using Long Short-Term Memory (LSTM) neural networks. This deep learning model is meant to learn secausal dependencies in time-series driving data. The study makes use of the UAH Driving dataset, which is a collection of more than 500 minutes of actual driving information on six. drivers on five classes of roads. The actions are classified into normal, drowsy and. aggressive. The authors went through a number of preprocessing actions including data cleaning. one-hot encoding of categorical features, normalization and class balancing. weights. They made the LSTM architecture dropout and batch normaliza- layers to avoid overfitting and the Adam optimizer to train. The model had impressive results with a precision and recall of 97and an F1-score of 98% in each of the three classes. It excelled over control models. e.g. Gradient Boosting and Random Forest. To continue working on the topic, the authors recommend that the dataset should be increased in size by involving additional drivers. and conditions, based on real-time deployment strategies and combining LSTM and. intelligent multimodal sensor fusion to enhance even higher accuracy and robustness. driver assistance systems.

A novel hybrid framework was introduced by Al-Din et al. [51] to recognize and classify highway driving maneuvers using smartphone-based sensors, with the goal of improving road safety through real-time behavior detection. The study addresses the limitations of traditional threshold- or rule-based methods, which often fail to capture abnormal driving behaviors due to signal noise and overlapping class patterns. The proposed two-stage system combines Dynamic Time Warping (DTW) for maneuver recognition with machine learning classifiers—Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN)—for final classification. DTW compares sensor signal similarity patterns, while classification is performed using features derived from inertial data. The system relies on smartphone-based IMU sensors (accelerometer and gyroscope) to collect data such as yaw angle and acceleration, which are then smoothed and transformed to align with vehicle dynamics. The model was tested on both controlled (900 samples) and naturalistic (2800 samples) driving datasets.Findings demonstrated that DTW alone brought about

a good F1-score of 0.95, and the hybrid with the RF classifier returned an average F1-score of 0.91- beating single classifiers. The hybrid system. The hybrid system also successfully reduced classification ambiguity between similar maneuvers, such as lane changes and merging.

2.3 Summary of Key Findings

The current literature on driver behavior and road safety converges on a unifying insight: most road crashes, up to 95%, stem from human error, manifesting as distraction, drowsiness, intoxication, or aggressive driving. Conventional systems, largely dependent on single-sensor inputs or rule-based reasoning, have been insufficient in capturing contextual factors such as weather, traffic flow, and road surface, limiting both detection performance and real-time responsiveness. To address these limitations, researchers have proposed advanced systems driven by artificial intelligence (AI) and machine learning (ML), often employing sensor fusion methodologies.

Deep learning models like CNNs, RNNs, LSTMs and hybrid networks (e.g., CNN-ConvLSTM or ensemble models of RF, SVM, K-means and YOLO) have been found to have good capabilities of detecting advanced driver behaviors like drowsiness, aggressiveness, cell phone usage and seatbelt wearing. Recent studies have incorporated radial basis function neural networks and deep learning ensembles to detect risk level [35], decision-level fusion approaches, such as RFCNN, have been explored to combine machine learning and deep learning outputs in prior studies [38], as well as uncertainty-aware graph neural networks to capture spatial crash dependencies at the road level [40]. Some of these solutions integrate computer vision and wearable or sensor-embedded (e.g., accelerometer, gyroscope, blink detectors) driver information to achieve precise monitoring of driver states. Ensemble feature engineering from smartphones [44] and probabilistic RKnD frameworks [45] further demonstrate that lightweight yet high-capacity models are achievable in ubiquitous sensing settings.

One of the most significant developments to emerge in recent work is contextual awareness—models now incorporate outside stimuli (e.g., road types, traffic density, hazards) to enhance levels of predictive accuracy. Several systems utilize multi-sensor fusion and trajectory analysis combined with real-time error feedback to reduce false positives and provide timely feedback mechanisms. Argument-based systems of reasoning like ASYDE [47] address conflict between sensor measurements that are disparate directly, while multimodal ADAS systems like DriverSense [49] combine physiological measurements (e.g., EEG, PPG, ECG) and mobility sensors to provide real-time identification of aggressiveness and stress. Risk is assessed by behavioral profiling (e.g., clustering by K-means or GMM) or probabilistically on the basis of measures of safety entropy as well as measures of exposure, where large empirical baselines [48] provide comparative standards of significance across a quantity of ML methods.

Although significant strides have been taken, there are limitations. Most studies are plagued by hardware limitations (e.g., limited computing on WCET-based systems),

real-time latency, generalizability to a broad set of driving populations and availability of data—especially for unusual yet significant behaviors. Datasets like the US Countrywide Traffic Accident Dataset [31] and SUMO-based simulated driver datasets [46] offer part of a solution but leave behind gaps in coverage of diversity, label consistency and cross-country validation. Furthermore, few combined frameworks cast side-by-side micro and macro behavior modeling and integrate environmental dynamisms. Last but not least, ethical considerations, including driver confidentiality and data bias, are poorly addressed, most especially in video- and smartphone-sensing systems [43]. In summary, the reviewed literature collectively indicates the importance of considering both driver behavior and environmental context for effective risk assessment and mitigation, although the manner in which such information is combined varies significantly across studies. The literature increasingly emphasizes personalized systems, multimodal data integration, efficient inference and the need for real-world validation. Directions to consider in papers going forward include increasing dataset diversity [31, 46], lightweight AI model optimization towards edge computing [44, 49], improved risk quantification through probabilistic and interpretable models [40, 47] and systems integration into commercial and autonomously driven platforms. Contrary to the majority of the fusion techniques discussed in the literature, the present study employed an independent modeling approach using a static decision-level risk mapping strategy.

Table 2.1: Research Gap Analysis and Positioning of the Proposed Hybrid Fusion Framework

Approach Type	**Data Used**	**Fusion Capability**	**Strengths**	**Limitations**	**Gap Addressed**
Behavioral-Only Models	Driver behavior (e.g., acceleration, braking, steering, sensor data)	None	Captures real-time driving patterns	Ignores environmental/contextual risk factors (weather, road, traffic)	Our work integrates external context + behavior
Context-Only Models	Accident records (weather, time, traffic conditions)	None	Strong for large-scale risk pattern analysis	Cannot capture driver-specific behavior or real-time actions	Our work adds behavioral intelligence layer
Synchronized Multimodal Models	Combined datasets (video + sensors + context)	Limited (requires strict alignment)	Rich feature representation	Requires time-synchronized data, difficult to scale, expensive	Our work removes synchronization dependency
Traditional Fusion Methods (D+D, early fusion)	Multiple modalities	Weak / rigid	Simple implementation	Poor calibration, bias toward dominant class, weak regression	Our work uses probabilistic late fusion
Existing Late Fusion Models (Raw L+L)	Multi-source predictions	Yes	Flexible and scalable	Sensitive to noisy inputs, poor high-risk detection	Our work improves via LLM-assisted refinement
Proposed Approach	Context (Dataset 1) + Behavior (Dataset 2)	Advanced (LLM-assisted Late Fusion)	Handles heterogeneous data without sync Handles high-risk detection Produces interpretable risk score	Minor trade-off in RMSE	—

Chapter 3

Requirements, Impacts and Constraints

The chapter is a systematic review of the requirements, effects, and limitations that affect the formation of the proposed framework of behavioral risk assessment based on context. It starts with the definition of technical specifications and methodological demands required in processing and analyzing the dual-dataset approach taken in this study. In addition to technical aspects, the chapter discusses the societal and environmental consequences, raises ethical issues, including information confidentiality and fairness of algorithms, and the project management and risk management plan. Lastly, an economic analysis is conducted to assess the cost-effectiveness and feasibility in practical application of the proposed framework.

3.1 Final Specifications and Requirements

The study aims at designing and assessing a behavioral risk assessment model that is context-sensitive based on the available data sets. The technical specifications include two publicly available datasets, namely, the US Accidents Dataset (20162023) and the smartphone-based Driving Behavior dataset. The selection of these datasets was based on the complementary nature of the datasets in order to model both the environmental context and driver behavior. Methodologically, the study consists of data preprocessing, feature extraction, and supervised machine learning. It does not require any new data gathering or hardware based sensing. Considering the size of the datasets, particularly the US Accidents dataset, then it is necessary to have effective data management tools including memory optimization, and stratified sampling. It is implemented using data science Python tools, such as NumPy, Pandas, scikit-learn, XGBoost, and LightGBM. Experiments are done offline and not deployed in real-time. The study comes up with two predictive models, one of the contextual severity of accidents and the other of driving behavior. Such models produce probabilistic results, which are combined with a learning-based fusion mechanism, such as decision-level and late fusion methods, to create a single and continuous driving risk score and maintain predictive uncertainty.

3.2 Societal Impact

One of the key public safety issues that have a big impact on human life, healthcare facilities, and economic productivity is road traffic accidents. This study will add value to society because it will come up with analytical frameworks which will enhance the interpretation and forecasting of driving risk. The study encourages a more holistic method of risk assessment by combining behavioral and contextual variables. The discoveries can be used to inform the creation of future driver assistance systems, insurance risk models, and public safety policies. Though there is no specific intervention system in place, the framework allows taking proactive risk assessment, which can be used to minimize accidents and enhance road safety awareness.

3.3 Environmental Impact

This research has a low environmental impact since it is purely computational and does not need development of hardware or physical prototyping. All experiments are carried out with the help of existing datasets and computing resources.But there are indirect environmental benefits. Traffic accidents tend to create congestion, consumption of more fuel and release of emissions. Such incidents can be minimized by improved prediction of risks, which will lead to less fuel consumption and less environmental pollution. Also, the application of existing datasets would encourage sustainable research practice, as it eliminates the need to duplicate data.

3.4 Ethical Issues

There are a few points of ethical considerations within this research. Firstly, all of the data that is being utilized is public domain and anonymized, such that personal identifiable information is not being processed. The research does not sample human subjects or attempt to profile car drivers.
The second ethical consideration relates to the fairness and biases in algorithms. Biased large data sets can have geographical, infrastructural and demographic trends, impacting model outputs. This work recognizes these shortcomings and assesses model performance on well-balanced metrics, like the macro F1 measure, so as not to benefit the majority classes. This work also does not make statements on personal responsibility and addresses risk patterns on an aggregate level.
Professional responsibility is maintained by the clear statement of the limitations of the proposed approach and by not making exaggerated promises about real-world implementation and the prevention of accidents.

3.5 Standards

The international safety and software certification guidelines are not formally applied to the study, as it involves neither the launch of a system nor the creation of a commercial software product. In spite of the lack of direct international guidelines,

the applied methodology follows widely recognized data science and machine learning guidelines on reproducible experimentation, proper separation of train and test data, employment of standardized assessment metrics, and result reporting.
Where applicable, this research follows some general principles that match data handling and analysis best practices found in intelligent transportation studies and other machine learning research.

3.6 Project Management Plan

The project was designed in such a way that it would be completed on time, resources are used efficiently and tasks are distributed evenly. The study was done in three semesters. The initial step was a comprehensive literature review in terms of driving behavior analysis, predicting accidents, and context-aware systems. The second stage was concerned with the collection of the data, exploration, preprocessing, and feature engineering. The third step was model development and experimentation where several machine learning models were tested and optimized into neural networks. It was then designed as a fusion based integration approach where the output of both models is integrated based on learning based methods as opposed to strict mappings. Evaluation, analysis of the results, documentation, and preparation of the presentation were the last stages. The project was not expensive in terms of finances as all datasets and tools were free. Computational tools were personal systems and cloud computing systems like Google Colab with support of the GPU systems. Responsibilities were shared among the team members, including literature review, data processing, experiment and documentation.

3.7 Risk Management

There were a number of possible risks that were found. The main issue was the diversity of the two datasets, as they are different in structure, scale, and feature representation. This was done by independently modeling each dataset and combining these models by using probabilistic and learning-based fusion methods. The issue of class imbalance, especially in the data of accidents severity was addressed with the help of stratified sampling and proper evaluation metrics. Large data volume resulted in computational constraints that were alleviated using memory optimization and data handling strategies. The other risk that may occur is discrepancies among model forecasts. This was mitigated through probabilistic outputs and fusion mechanisms that consider the uncertainty, and there is a more stable and reliable risk estimation.

3.8 Economic Analysis

In terms of economics, the study proves to be very cost-effective, as it is based on the use of only open-access data sources and open-source tools. As a result, costs on software license and data purchase are saved. Computational resources and human effort are the only major cost factors. The uses of this study can bring

about considerable beneficial economic value. Better accident risk forecasting will lead to savings in the costs of damaged vehicles, health care, insurance, and traffic jam. The study does not offer a full-fledged deployable system; nevertheless, it presents a structure that can be used to achieve long-term economic efficiency in the transportation and safety of the population.

3.8.1 Cost Analysis

The study is mainly expensive in terms of computation and manpower. Computational costs involve the time to run a model, the use of hardware, and energy used to train and validate deep learning models. Specifically, the implementation and coding phase is very resource-intensive, with training complicated models requiring a lot of GPU utilization, a lot of memory, and extended execution times, all of which add to the overhead of the computations. These costs however are mostly incurred in the form of a single investment that is made during the development and training period.
Human resource expenses include data preprocessing, feature engineering, model building, and analysis, which take a lot of time and experience. Nevertheless, the suggested framework does not presuppose extra capital investment since it makes use of the existing infrastructure and tools that can be freely found. Also, it does not entail repeat license or subscription charges, and is therefore a long term, cost effective and scalable solution.

3.8.2 Benefit Analysis

This study has significant potential economic impacts by developing improved accident risk prediction. A better prediction rate should be used to minimize financial losses in terms of vehicle damage, healthcare costs, insurance settlement, and traffic congestion. Risk evaluation helps prevent accidents by taking proactive steps before they occur, thereby reducing the number and severity of accidents. Even slight increases in the performance of prediction can lead to significant economic savings on a societal scale. The framework also aids in the planning of infrastructure, allocation of resources, and traffic control by supporting the use of data in making decisions regarding transportation authorities and policymakers. Its scalability enables its future integration into smart transportation systems, which leads to the enhancement of people safety and cost-efficient economies in the long run.

Chapter 4

Proposed Methodology

After analyzing the current state of research and outlining the major weaknesses existing in the previous methods, developing the systematic structure of the suggested hybrid AI-based framework becomes the subject of the next step. In this chapter, the methodological framework used to combine behavioral and contextual data used in the prediction of accident risks is described. It describes the design justification, model design and analysis procedures that were used to build a scalable, interpretable and real time risk assessment system.

4.1 Design Process or Methodology Overview

The objective of the method proposed here is to devise an interpretable and scalable system to boost the driver risk estimation performance through effective exploration of contextual accident attributes and driver behaviors simultaneously. Instead of employing a single-source prediction approach, the method followed the incremental and iterative design procedures: First, developing individual models, later combining them through a novel learning fusion technique to improve their collective risk estimation capabilities. The overall working process comprises following sequential phases from data sourcing till validation and comparison: data collection, pre-processing, feature engineering, model building, novel learning approaches, learning-based fusion, and assessment. Each phase was designed to maintain data quality, minimize bias, enhance predictive ability, and ensure the methodological translucency.

Two types of data are used:

Dataset 1 consists of a large dataset of road accidents with millions of data points retrieved from public APIs for traffic and weather conditions. The dataset includes information regarding environmental, spatial and temporal factors for road accidents occurring in the United States.

Dataset 2, which is a sensor-based dataset containing accelerometer and gyroscope values recorded from smartphones in different driving styles, such as slow, normal or aggressive driving.

Together, Dataset 1 and Dataset 2 provide a whole picture of external contextual risk factors and internal behavioral dynamics.

The methodology in its initial form involved training multiple machine learning models on each of the data sets independently in order to generate baseline models. The highest performing models among the baselines were used to identify directions in which to fine-tune single models in the form of tailored neural networks, which were later used as single-model architectures.

Individual models, however, were inadequate at describing the interaction between the context and the driver behavior completely, so additional approaches were used. Sequence based learning was applied to determine if temporal dependencies could be utilized, especially to behavioral sensor data. Knowledge distillation was also utilized in order to transfer the knowledge gained by a large teacher model into a smaller and more efficient student model, while still maintaining its ability to predict.

After producing single models, fusion techniques were applied in order to combine the context and behavior models. Initially, the relationship between the context and behavior domains was described via a survey-derived 2X3 risk matrix that assigned average risk values to each severity-behavior pair; this discrete relationship was converted into a continuous expected risk value, achieved by weighing the survey-derived risk values with the probability of the predicted class, obtained by each of the two models.

This figure illustrates the systematic workflow for the decision fusion approach, showing the parallel training of Severity (Contextual) and Sensor (Behavioral) models.

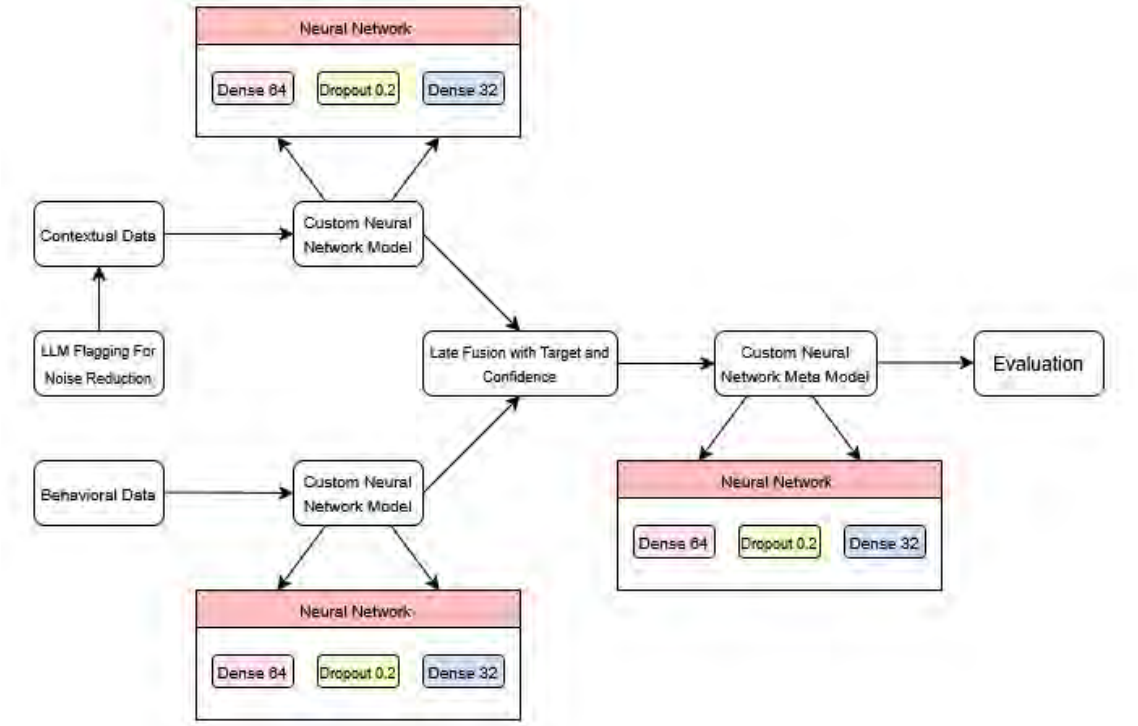

Figure 4.1: Proposed Meta-Model Architecture for Context-Aware and Behavioral Driving Risk Assessment – Illustrates the integrated flow of environmental and behavioral data through separate branches and multi-level fusion to produce a unified, continuous risk score.

To enhance the integration between modalities even further, a decision-level fusion framework using the outputs from the neural networks was created, followed by a late fusion framework using the neural network outputs combined with static survey risk scores. The contextual data was then more heavily refined by using LLM to find noise within the data, with noisy sections of data being filtered out and removed before fusion. Finally, the optimal fusion of model outputs along with a sequence-based behavioral model and an LLM-enhanced contextual model was tested.

The proposed methodology allows for analytic consistency with well-defined and uniform preprocessing and feature management. The system is reproducible due to automated, pipeline-based processing implemented in Python. The model is adaptable and can be extended in the future, as the components can be individually tuned or swapped with better-performing ones.

Ultimately the methodology was conceived, taking into account the performance of the predictor, computational efficiency, interpretability and scalability. The designed model can serve as a broad overview of driving risk as a product of context and behavior

4.2 Preliminary Design or Model Specification

After analyzing the current state of research and outlining the major weaknesses existing in the previous methods, developing the systematic structure of the suggested hybrid AI-based framework becomes the subject of the next step. In this chapter, the methodological framework used to combine behavioral and contextual data used in the prediction of accident risks is described. It describes the design justification, model design and analysis procedures that were used to build a scalable, interpretable and real time risk assessment system.

Table 4.1: Models Applied to Each Dataset

Dataset	**Models Used**
Dataset-1: US Accidents (2016–2023)	Dummy Stratified Classifier; Logistic Regression (SAGA); Stochastic Gradient Descent Classifier (Log Loss); Ridge Classifier; Decision Tree Classifier; Linear Support Vector Classifier (Linear SVC); Light Gradient Boosting Machine (LightGBM); Histogram-based Gradient Boosting (Scikit-Learn); Extreme Randomized Trees (Extra Trees); Random Forest Classifier; Multilayer Perceptron (Feedforward Neural Network); Spatio-Temporal K-Nearest Neighbors (ST-KNN proxy); Multinomial Logistic Regression (One-vs-Rest)
Dataset-2: Driving Behavior	Logistic Regression; SGD Classifier (1); SGD Classifier (2); Extra Trees; LightGBM; Decision Tree; Random Forest; Balanced Random Forest; Bagging Classifier; Histogram Gradient Boosting

4.2.1 Accident Severity Classification (Dataset_1)

The US Accident dataset was subject to extensive preprocessing for prediction. The original dependent variable or target variable, Severity, was a scale ranging from 1 (minor) to 4 (severe). The target variable was simplified and transformed into a binary target variable because of class imbalance problems and the complexity in the dataset. The conversion of the target variable into a binary variable was done by mapping the values as follows:

- Severity levels 3 and 4: 1 (Severe)
- Severity levels 1 and 2: 0 (Non-severe)

This transformation converted the initial multi-class problem into a binary classification problem, which is generally more stable during learning. The preprocessing pipeline consisted of the following steps:

Missing Value Handling

- Removed columns with over 30% missing data.
- Imputed numerical features with the median.
- Filled categorical features with "unknown".
- Filled Boolean missing values with default logical values.

Temporal Feature Engineering

- Converted accident start time to datetime format.
- Extracted features: `Start_Hour`, `Start_Weekday`, `Start_Month`.
- Created categorical indicators: `Is_Weekend`, `Is_Night`.

Feature Transformation

- One-hot encoded categorical features.
- Scaled numerical features.
- Investigated normalization methods; standardization was deemed more stable across models compared to min-max scaling.

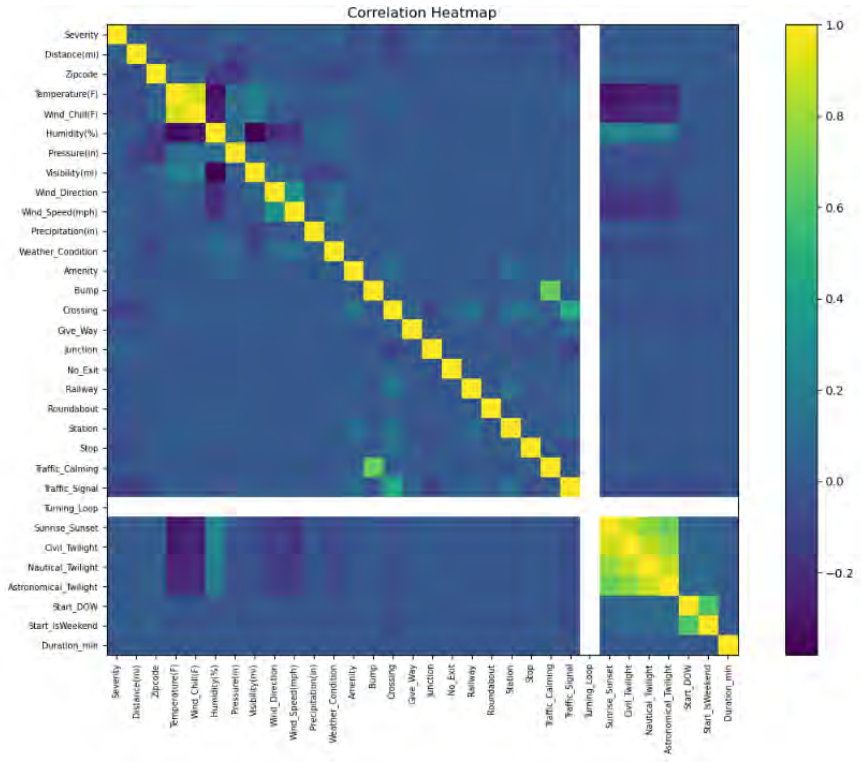

Figure 4.2: Correlation Heatmap of Traffic and Environmental Features (Dataset 1) – Visualizes the statistical relationships between external variables (e.g., weather, road features) to identify key predictors for accident severity classification.

The initial round of modeling was conducted using several classical machine learning models including Logistic Regression, Gradient Boosting, Random Forest, CatBoost and XGBoost. The most promising models provided insight into the single tabular neural network model to be used in dataset 1. The final single model that was built in Dataset 1 was a customized tabular neural network tailored to efficiently process context relevant accident features. The goal was to learn non-linear dependencies between environmental, road, spatial and temporal features and predict the probability of a non-severe and severe outcome.

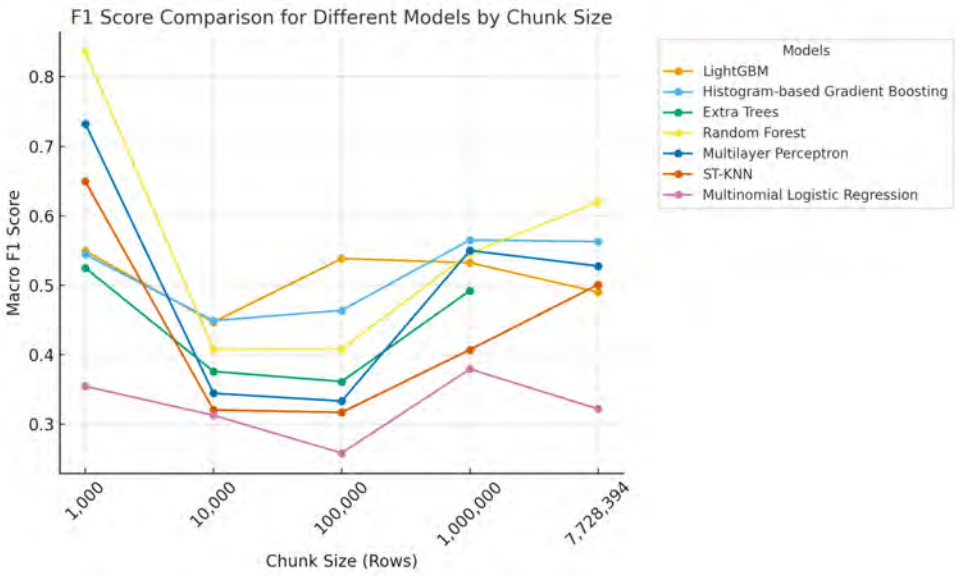

Figure 4.3: Model Performance Comparison by Data Chunk Size (Macro F1 Score of Dataset_1). This figure demonstrates that model performance stabilizes as training data increases, indicating that larger data chunks provide a more reliable representation of contextual risk.

4.2.2 Driver Behavior Classification (Dataset_2)

The second dataset is related to the use of a smartphone-based inertial sensor, which can monitor driving patterns by utilizing an accelerometer and a gyroscope sensor, where each data entry is classified into Slow, Normal and Aggressive driving patterns, which are later interpreted as low-risk, normal-risk and aggressive-risk levels during the decision-level risk mapping stage.
Unlike the US Accident dataset, which considers external factors, the data set focuses solely on driver behavioral dynamics in terms of their physical motion patterns. Each data instance consists of six continuous attributes, namely AccX, AccY, AccZ, GyroX, GyroY and GyroZ.

Figure 4.4: Sensor Feature Correlation Heatmap (Accelerometer and Gyroscope Data) – Shows the high degree of interdependency among motion sensors, which justifies the use of a unified sensor-based neural network for behavioral detection.

Preprocessing and Cleaning:

- **Handling minor data gaps:** Linearly interpolated missing sensor values.
- **Outlier removal:** Extreme sensor noise removed by eliminating values outside the 25–75% percentile ranges.
- **Stability checking:** Conducted through visual inspection of plots and correlation analysis.
- **Scaling:** Multiple normalization methods assessed to facilitate convergence and control variance during training.

Similar to Dataset 1, baseline machine learning algorithms were applied to Dataset 2. The results were compared, and the better-performing algorithms on each dataset guided the development of neural network models.
Both of the above initial models create the basis for the whole system as stated below:

- Dataset 1 model – contextual severity estimation.
- Dataset 2 model – behavioral pattern estimation.

The outputs from both models are subsequently fused in later stages of the framework to enable multimodal driving risk prediction.

4.3 Data Collection

4.3.1 Data Sources

The study relied on two datasets collected from distinct yet complementary domains:

- **Dataset_1**
  - **Source:** Publicly available dataset compiled from multiple U.S. traffic APIs (MapQuest, Bing Maps, TomTom).
  - **Contents:** Over 7 million records describing road accidents with parameters such as temperature, visibility, precipitation, road type, and traffic control devices.
  - **Coverage:** 2016–2023, across 49 U.S. states.
  - **Classical models:** Classical models, including Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, SVM, KNN, Decision Trees, and MLP, were first benchmarked, with LightGBM and Random Forest achieving the highest macro F1-score of approximately 0.95. To introduce novelty, custom neural networks were then developed, employing Keras inputs per axis, normalization, dense layers, softmax output, sparse categorical cross-entropy, and the Adam optimizer. These architectures yielded commendable accuracy comparable to the top-performing tree-based methods, but with superior computational efficiency. Confusion matrices confirmed minimal misclassification across classes, thereby validating the robustness of sensor-based behavior detection for driver assistance systems.
  - **Custom neural networks:** Keras inputs per axis, Normalization, Dense layers, softmax output, sparse categorical cross-entropy, and Adam, yielding 0.999 accuracy comparable to top trees but with superior efficiency. Confusion matrices confirmed minimal misclassification across classes, validating sensor-based behavior detection for driver assistance systems.
  - Although independent classification of driving behavior has been done with custom neural networks, our research proposes a new hybrid model that solely provides accident severity prediction using 7M+ environmental records with sensor-based behavior.
- **Dataset_2**
  - **Source:** Smartphone-based inertial sensors (accelerometer and gyroscope).
  - **Contents:** 7 million
  - **Attributes:** Continuous motion data (AccX, AccY, AccZ, GyroX, GyroY, GyroZ) recorded under different behavioral labels.
  - **Sampling rate:** Consistent sensor readings per session, pre-labeled with driving behavior class.

Both datasets were chosen for their reliability, openness, and richness of features relevant to understanding the interplay between driver state and external conditions.

4.4 Data Cleaning and Preprocessing

Data cleaning and preprocessing were done to make sure that the datasets are suitable to conduct large-scale machine learning experiments and make good use of memory. In both datasets, an initial exploratory investigation was carried out to identify the number of records, the dimension of features and the general format.
The lack of values was analyzed in a systematic way across all features. The threshold of the missing values of 0.3 was established, according to which any other feature with over 30 percent of missing values would be dropped. Nevertheless, none of the features surpassed this value upon inspection, and, thus, no columns were dropped because of missingness.
To measure the significance of the features available, correlation analysis was used. Some contextual features, such as twilight and other similar features, were early examined for the possibility of making feature aggregation. Though they are conceptually significant, such combinations did not lead to a significant increase in correlation with the target variable. Also, these changes significantly consumed more memory. These derived features, therefore, did not make it to the final preprocessing pipeline.
Temporal feature processing emphasized the accident start time. The attribute `Start_Time` was converted into a datetime format, and some other features based on time were created during preprocessing. Their extracted temporal features were kept since they had a positive correlation effect and enhanced the predictive ability of a model without taking too much computational load.
OneHotEncoding was used to encode categorical variables, and int features have been downcast to less-precise data types (e.g., `int8` or `int16`) where numbers are not required. This was necessary to minimize memory consumption because the accident dataset is very huge and would otherwise pose a risk of overloading the memory during training and evaluation.
The preprocessing approach was also driven by the fact that it had to support large-scale testing with billions of feature evaluations in the process of the experiments. Such large-scale model training would not have been possible without aggressive memory optimization and preprocessing.
In the case of the driving behavior dataset, the cleaning of the data made sure that there were no null values in the dataset. The labels of the classes were coded into numerical values to facilitate supervised learning, and the driving behaviors were coded into ordinal numeric values to facilitate real-world interpretability and effective model training. There was no further feature building available since additional transformation was not providing any significant performance gain.
In general, the phase of data cleaning and preprocessing aimed at preserving the integrity of data and enhancing considerations of computation efficiency and only the transformations that were proven to have any impact on the performance of the models were kept. This made sure that the end datasets were optimized for scale and machine learning experimentation and for their reproducibility.

4.5 Data Transformation

Data transformation was used to match both datasets to the needs of the selected machine learning models without losing the feature space that can be interpreted

or computationally efficient. With the accident data, the principal transformation targeted breaking down the original timestamp into explicit time characteristics, i.e., hour of day, day of week, month, and Boolean variables of weekend and nighttime driving. These extracted time-based variables represented periodic risk patterns without the use of complicated composite indices or feature interactions.
Categorical characteristics, such as the ones pertaining to cities and twilight, were encoded into machine-processable form using encoding schemes that were compatible with the modeling pipeline. The neural network was implemented in such a way that string-based variables were reduced to sparse one-hot representations using lookup layers, and numeric variables were normalized using normalization layers to stabilize training and to make features of similar scale comparable. This encoding and normalization enabled the models to take advantage of the structure of environments and time, which was not hand-crafted with high-order features.

In the case of driving behavior data, there was minimal transformation aim. The six sensor signals (AccX, AccY, AccZ, GyroX, GyroY and GyroZ) were left in their original position and normalization was used to ensure that the impact of extreme values was mitigated and to ensure the neural network classifier converged faster. No further composite factors were put in because initial experiments showed that already simple, well-normalized, raw axes gave enough discriminative power to the behavior classes. The visualization tools (distribution plots and correlation heatmaps) were applied to ensure that the transformed variables had well-behaved numeric ranges that were stable enough to be used in large-scale training.

4.6 Data Integration

The two used data sources were not concatenated at the raw feature level since the features had different types and did not have any common linkage at the instance level. Dataset 1 was contextual accident-related information, which included contextual factors like environment, time and situation, while Dataset 2 was behavioral dynamics information extracted from mobile phone sensor data (motion sensors). Because of the feature heterogeneity and semantic differences between datasets, we believed the feature level integration would be inappropriate and hence focused on integration at the model-output level. The data from two sources were individually modeled with their respective predictive models; that is, two neural networks were independently built for accident severity prediction and driving behavior classification. The link between two domains was modeled using the survey-derived 2x3 risk matrix, where the combination of six possibilities of severity and behavior classes was assigned average risk scores determined from the survey. This created an understandable interpretation of linking contextual and behavioral risk. Unlike simple discrete mappings, a continuous value for expected risk was obtained by taking two models' outputs and combining their corresponding risks from survey based value according to the prediction probabilities. This way, the predicted uncertainty was incorporated into the final risk score and it became more fine-grained than a fixed score. Later on, in the method section, the model fusion was extended in the following stages by adopting the decision-level fusion, where two network outputs were integrated by using a meta-learning model, late fusion, where the models' outputs and static risk scores were combined, and finally, the LLM-aided anomaly detection

on the contextual branch, which further improves the quality of the context data by removing noise and correcting errors.

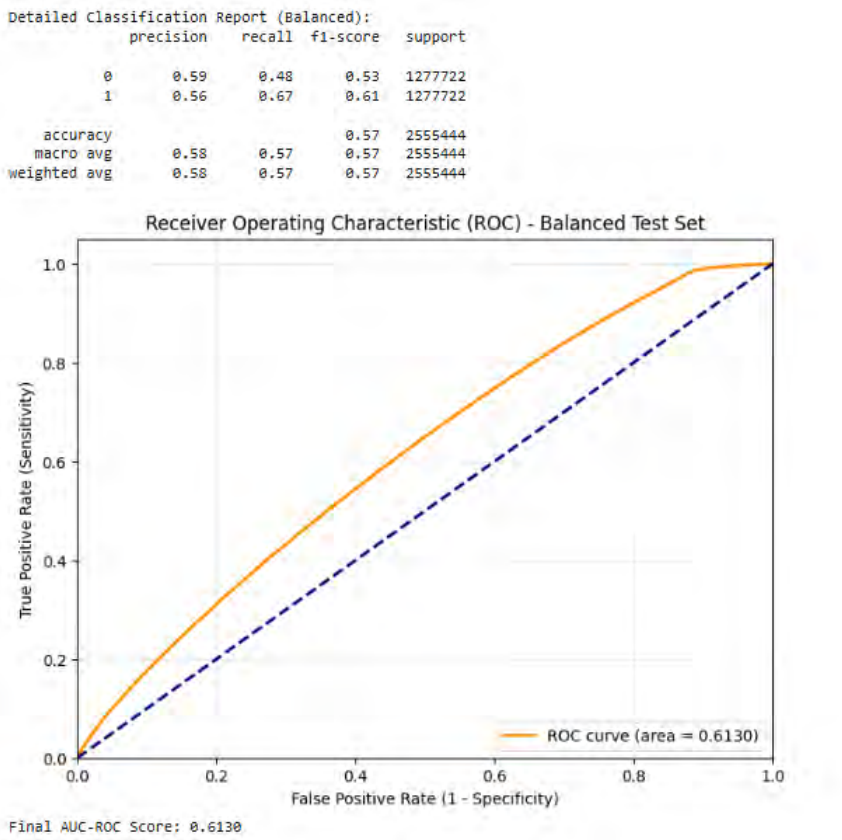

Figure 4.5: Receiver Operating Characteristic (ROC) of the Contextual Branch Before LLM Refinement – Serves as a baseline performance metric showing the model's initial ability to distinguish between accident severity levels.

This figure shows the initial performance of the contextual neural network on raw accident data (DF1). The model displays limited discriminative power between severe and non-severe cases, indicating the presence of significant noise or overlapping features in the unrefined dataset.

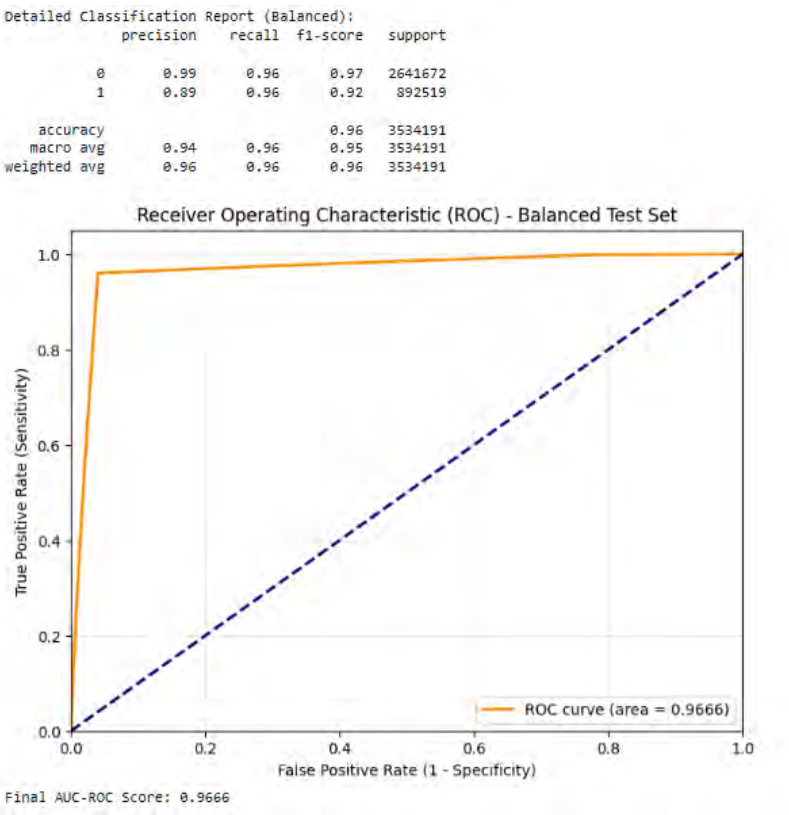

Figure 4.6: ROC Performance of Context Branch (LLM-Enhanced) – Highlights the performance gains achieved after using LLM-based refinement to prune noise, resulting in better differentiation of high-risk scenarios.

After applying LLM-driven anomaly detection to filter inconsistent and noisy records, the contextual model's performance improves significantly.This jump in predictive quality validates that LLM refinement effectively prunes contextual "outliers," resulting in a much cleaner signal for the subsequent fusion stage.

4.7 Data Reduction

The data reduction was done with a twofold objective of ensuring computational efficiency at scale while at the same time providing the most informative contextual and behavioral signals required during risk prediction.
In the case of the US Accident dataset (Dataset 1), reduction was done on three points:

- **Target balancing:** Once the severity was binarized into 0 (non-severe) and 1 (severe), the majority class was sampled down to equal the minority one, forming a class-balanced training subset to reduce the bias on the non-severe cases that occur frequently.
- **Column pruning based on domain relevance and memory consumption:** A group of high-cardinality or weakly informative columns (e.g., identifiers, detailed location flags, many redundant weather columns such as Pressure, Visibility, Windspeed, WindDirection, WeatherCondition, WeatherTimestamp and other infrastructure indicators such as Bump, Crossing, Railway

etc.) were dropped to minimize the memory footprint and eliminate over-parameterization. Columns that were retained denoted tight spatial context (e.g., City), important environmental conditions (Temperature, Wind-Chill, Humidity, Precipitation) and artificially created temporal descriptors (Start_Hour, Start_Weekday, Start_Month, Is_Weekend, Is_Night).

- **Type downcasting and imputation:** Numeric columns were downcasted to low-precision dtypes as needed (e.g., float32, int8/int16), and the default value of numeric columns was replaced with robust statistics (medians or zeros as needed). Category variables were upcast to category/string with an explicit or default level of Unknown. The combination significantly decreased the use of memory and left all the signals retained for the model.

The Driving Behavior dataset (Dataset 2) needed less structural reduction, as it was compact and consisted of only 15 features. In every record, there are six motion features (AccX, AccY, AccZ, GyroX, GyroY, GyroZ) and a behavioral class label. All six sensor features were retained without modification. Reduction of data here thus revolved around:

- Assuring a manageable sample size through a set of random holdout tests (fixed 5%) with `test_id = df2.sample(frac=0.05)`, the rest being training.
- Imputation of any null values and mapping of class labels for supervised learning (SLOW=1, NORMAL=2, AGGRESSIVE=3).
- Use of per-feature normalization layers in the neural network structure to deal with the differences in scales instead of at the data level.

In general, both datasets have been data-reduced to focus on the pruning of columns that may be highly redundant or weakly informative, to balance the distribution of classes according to the severity of accidents, and to enable the neural models to be trained effectively at scale. This was achieved by downcasting features to less expressive types and applying imputation, while ensuring that the essential temporal, environmental, and motion-based aspects of the data were preserved. The overall reduction strategy thus maintained computational efficiency and model integrity without compromising the informativeness of the signals required for risk prediction.

4.8 Summary of Preprocessed Data

After preprocessing, both of the datasets were rendered into a clean, uniform, and analyzable format that could then be utilized in the process of building predictive models. This cleaning phase of processing deals with missing values, noises, irrelevant attributes, and consistency based on different dataset characteristics, and feature engineering and scaling techniques were implemented to bring better quality and readiness to the data.

Summary of preprocessing outcomes:

Table 4.2: Summary of Dataset Characteristics and Preprocessing Pipelines

Dataset	**Records (Final)**	**Features Used**	**Target Variable**	**Main Processing Steps**
Dataset_1	7.7 million	40+	`Severity_bin` (0 = Non Severe, 1 = Severe)	Cleaning, imputation, feature engineering, binarization, normalization
Dataset_2	7 million	6	Driving Class (Slow,Normal, Aggressive)	Outlier removal, interpolation, normalization, scaling, class balancing

- Dataset 1 was cleaned into **structured contextual data**, including environmental, temporal, and context-related features for accident severity prediction.
- Dataset 2 was transformed into **static behavioral data**, consisting of clean and normalized motion sensor features for driving behavior classification.
- Both datasets were processed such that their original features were preserved in nature while enabling compatible input for subsequent neural network modeling and fusion-based risk evaluation.

This framework of two distinct datasets provided two distinct views of risk in driving. Dataset 1 describes risk at the external situational level of an accident's severity and dataset 2 describes risk at the internal behavioral level of a driver's action. Both datasets were ready to be integrated in developing the predictive models and ongoing risk evaluation scheme as proposed in this research.

Chapter 5

Result Analysis

In this chapter, the experimental validation and comparative evaluation of the developed context-aware driving risk assessment framework are provided. This analysis covers the individual predictive models trained on US Accidents (Dataset 1) and Driving Behavior (Dataset 2) and the advanced modeling and fusion techniques investigated in this study.
The analysis was conducted in three stages. Initially, individual predictive models were tested to analyze their utility for context-aware accident severity prediction and behavioral driving pattern classification. Then, advanced modeling techniques, including sequence-based learning and knowledge distillation, were analyzed to investigate performance and computational cost improvements. Finally, a series of integration methods were evaluated, including decision level fusion, late fusion, and continuous expected risk assessment. Integration was also analyzed from a conceptual level with an LLM-based contextually informed decision refinement component of the framework.
Analysis revealed that individual models were effective but produced incomplete descriptions of driving risk. It was found that sequence based learning was highly effective on behavioral sensor data but less so for the static, contextual accident data. Knowledge distillation produced both an efficiency and slight generalization improvement, but the overall greatest improvement was produced through integration. Overall, the LLM-aided late fusion approach achieved the best balance in classification and regression, reaching 0.91 accuracy, 0.88 macro F1 score and $R = 0.9891$ for the regression, successfully combining neural network outputs with survey based risk knowledge and LLM-aided contextual refinement.

5.1 Performance Evaluation

The proposed framework was subjected to a multi-stage performance evaluation from the contextual severity model through the behavioral classification model to the ensuing fusion strategies. The development process was iterative, and thus the evaluation focuses not only on overall accuracy but also on the strengths and weaknesses inherent in each modeling phase. Focus was given to class imbalance, safety-oriented performance measures, and the system's ability to provide a meaningful continuous risk score.

5.1.1 Evaluation Criteria

The evaluation of the proposed framework utilized both classification and regression metrics in accordance with the nature of the modeling phase. The classification metrics include:

- **Accuracy** – the ratio of the number of correct predictions to the total number of examples evaluated.
- **Precision** – the ratio of true positive examples to the total predicted positive examples.
- **Recall** – the ratio of true positive examples to the total actual positive examples.
- **F1-score** – the harmonic mean of precision and recall, and is therefore a better metric than precision and recall individually when class imbalance is present.

In light of the fact that Dataset 1 is class-imbalanced (the minority class being severe accident cases), increased focus was given to:

- Macro-averaged F1-score, as this gives equal weight to each class and helps mitigate class imbalance problems.
- Class-wise recall, especially for minority classes (e.g., severity-related ones).
- Balanced performance analysis (i.e., one should not just look at total accuracy).

Additional metrics were used in subsequent fusion steps, such as:

- **MAE** for the continuous expected risk score to measure the absolute error between the actual risk score and the predicted risk score.
- **RMSE** to measure the error in magnitude, again for the continuous risk score.
- **R** to measure the variance explained by the continuous risk score against the actual risk score distribution.
- **AUC-ROC** to assess class separability in the multi-class fusion output.

Such metrics were used as the performance of a high-accuracy classification model may not necessarily translate to a safe system if it misses safety-critical and minority class scenarios, which became evident between D+D fusion, raw L+L fusion, and final LLM-assisted L+L.

5.1.2 Testing Methods

The overall testing was aimed at giving a robust, unbiased performance measure of both original independent models as well as extensions based on fusion.

The experiment methodology can be summarized in the following points :

- Stratified train-test split preserving class ratios in both training and testing partitions.
- Benchmarking over traditional ML and specific NN models.
- Testing of individual model performance for Dataset 1 and Dataset 2 separately prior to any fusion.
- Stepwise performance testing of baseline neural networks, sequence-based LSTMs, knowledge distillation, D+D fusion, naive L+L fusion, and final LLM guided L+L fusion.
- Evaluation using classification and regression outputs with special attention to after introducing continuous expected risk scores.

The experiments were carried out using the following computing hardware

- Intel Core i9-14900K
- 128GB DDR5 RAM
- NVIDIA GeForce RTX 4090 (24GB VRAM)

This setup allows training, simulation, and fusion evaluation on millions of records and very large contexts on both accidental and behavior sensor data.

5.1.3 Performance Outcomes

Custom Neural Network Models The two single-model neural networks were designed to tackle two specific modalities, the contextual and the behavioral signals, separately. As a result, they performed very well individually and the two models were subsequently used as the sole, single-model component of each context-based and behavior-based branch. For dataset 1, the custom neural network was able to achieve a validation accuracy of 0.68 and a validation F1-score of 0.67, reaching the level of the most effective contextual tree-based models and performing even better by providing the advantages of a smaller memory footprint and higher training speed. For dataset 2, the custom neural network designed for sensor-based signals outperformed all tree-based behavioral baseline models with a performance of 1.000 accuracy and 0.999 macro F1-score. Those results were enough to consider two separate custom-built neural networks as the contextual and behavioral streams in a later combination step.

LSTM-based Models

To further explore potential performance gains by incorporating sequence-based temporal modeling, LSTM-based models were applied to the two datasets. The results clearly indicate that performance varies widely depending on data modalities.

- For DF1, the performance of the LSTM-based model was dismal with a score of 0.41 for accuracy and 0.40 for macro F1-score, thus demonstrating that sequence-based learning is not well suited for predominantly static contextual data on accidents.

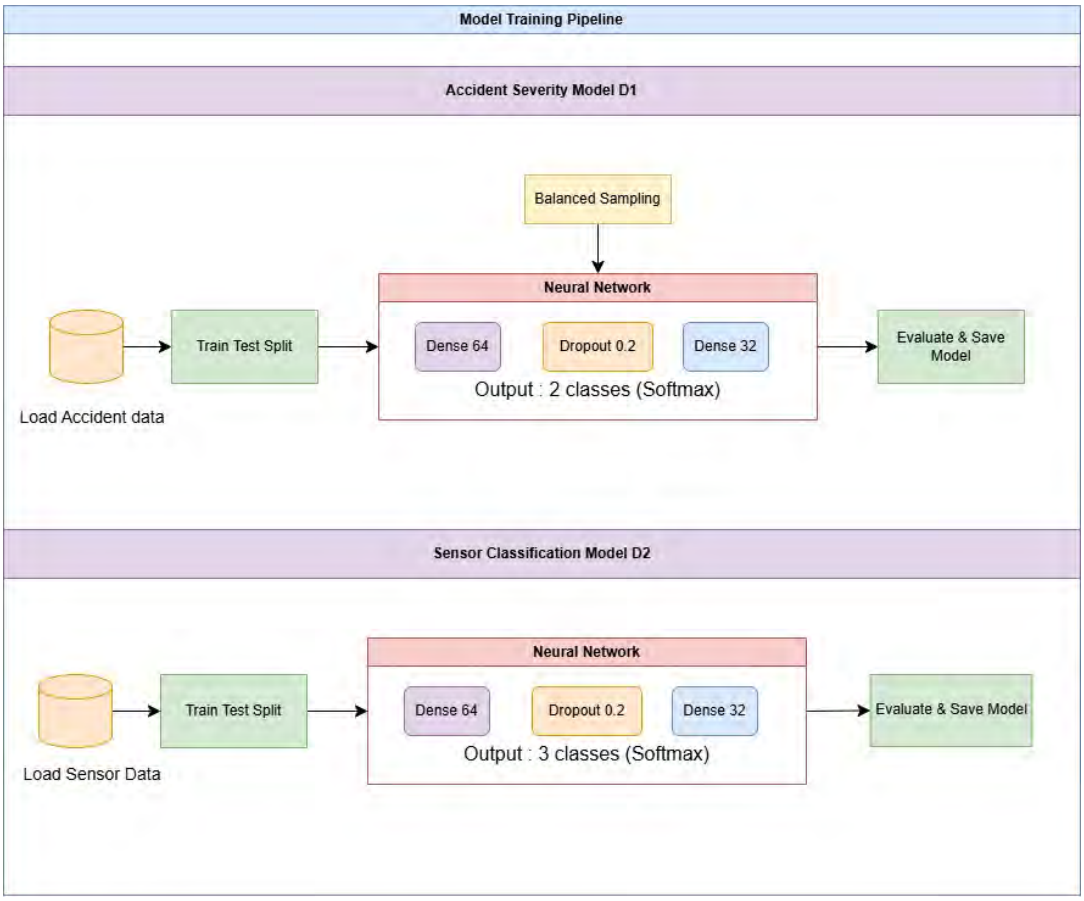

Figure 5.1: Model Training Pipeline illustrating the workflows for Accident Severity Model (Dataset 1) and Sensor Classification Model (Dataset 2). Both pipelines follow structured steps including data loading, train-test splitting, neural network training, and evaluation, with Dataset 1 producing severity classification (2 classes) and Dataset 2 producing driving behavior classification (3 classes).

- For DF2, the LSTM-based model performed exceptionally well with an accuracy of 0.96 and macro F1 score of 0.943, confirming that LSTM-based models are able to effectively capture sequential information from sensor-based behavioral signals.

Table 5.1: Performance Comparison of Custom Neural Networks and LSTM Models

Model	**Dataset**	**Accuracy**	**Macro F1**
Custom NN	DF1	0.68	0.67
Custom NN	DF2	0.999	0.999
LSTM	DF1	0.41	0.40
LSTM	DF2	0.96	0.943

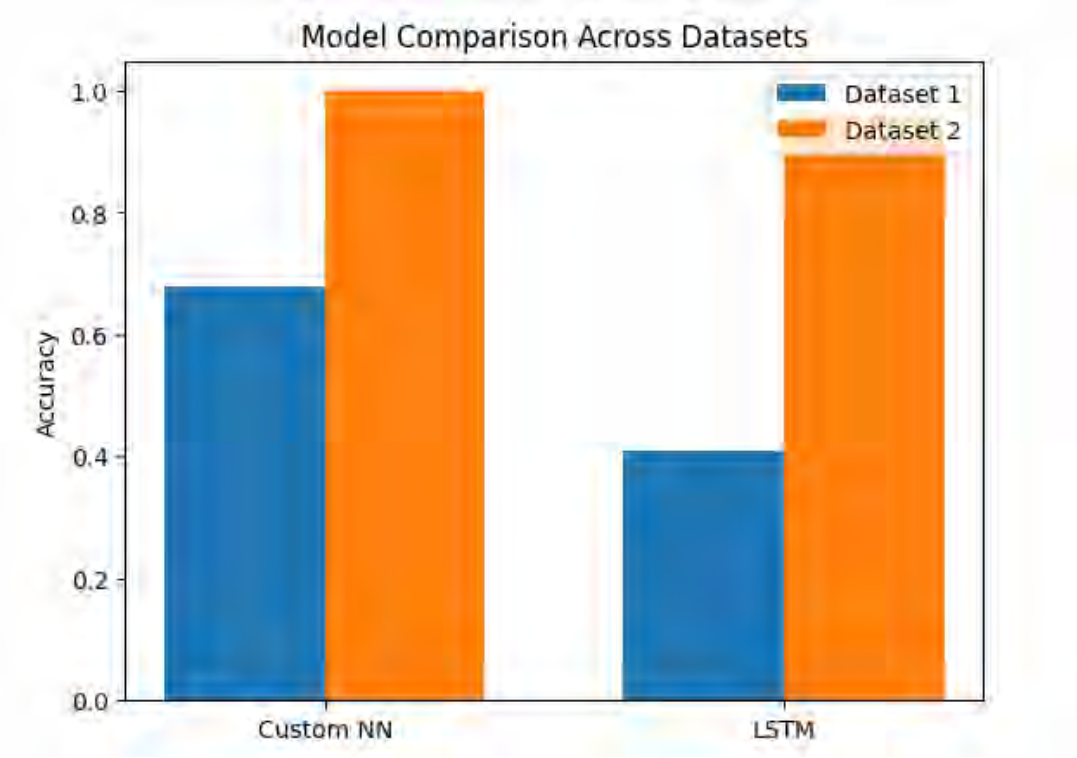

Figure 5.2: Accuracy Comparison of Custom Neural Network and LSTM Models - Findings indicate that while LSTMs excel at sequential sensor data (Dataset 2), custom neural networks are more effective for static contextual data (Dataset 1).

Table 5.2: Summary of Model Components and Their Performance

Component	**Model Used**	**Key Architecture / Method**	**Output**	**Performance**
Dataset 1 (Context)	Custom Tabular NN	Dense 64 → Dropout 0.2 → Dense 32 → Output 2	Severity Probabilities	Acc ≈ 0.68; Macro F1 ≈ 0.67
Dataset 2 (Behavior)	Custom Sensor NN	Dense 64 → Dropout 0.2 → Dense 32 → Output 3	Behavior Probabilities	Acc ≈ 0.96; Macro F1 ≈ 0.943
Fusion (Baseline)	Raw L+L	Direct probability merge	Risk Score	Acc 0.91; Macro F1 0.88
Fusion (Final)	LLM-assisted L+L	Context refinement + late fusion	Risk Score	Acc 0.91; Macro F1 0.88

Knowledge Distillation

Knowledge Distillation was used as a method of increasing the efficiency and generalization of the contextual stream. The TabNet model served as the teacher model,

and the lightweight MLP was chosen as the student model.

- Teacher model performance: 53.66% accuracy.
- Student model performance: 55.02% accuracy.
- The student model replicated 102.5% of the performance level of the teacher.

Although the global performance is not high due to the problem's complexity and unbalanced nature on the accident dataset, this result shows that the student model, which is computationally less demanding and more efficiently deployed, can slightly outperform a more complicated architecture.

Balanced D+D Fusion Model

Once the independent models had been evaluated, we introduced a decision-level fusion of the contextual and behavioral branches.

- Overall accuracy: 0.63
- Macro F1-score: 0.60
- Recall for the most risky class: 0.76

Though this accuracy was slightly worse than the subsequently developed fusion models, the per-class results were more sensitive to the crucial risk situations. For example, the recall for the most risky class was 0.76, so the D+D model proved to be rather good at detecting dangerous situations. Therefore, the D+D model framework is meaningful from safety points of view, where the detection of dangerous events is critical.

Raw L+L Fusion Model

The raw L+L fusion model obtained an overall accuracy of 0.81. However, it has the cost that balanced per-class performance greatly degrades to a macro F1 score of 0.49, while the two severe risk classes obtained very low recalls around 0.05-0.06. That is to say, the raw late fusion had become biased to the majority of low-risk classes. Although the regression results were good (MAE=0.0877, RMSE=0.1382, R=0.9830), the class-wise classification outcomes suggested that raw late fusion is not suitable for discriminating risks in safety critical applications. This figure illustrates the One-vs-Rest AUC-ROC curves for the baseline model.

LLM-Assisted L+L Fusion Model

The highest performing models were the LLM-assisted L+L fusion architecture. In this step, the LLM pre-processed (cleaned noise and identified anomalies) the context branch before training the updated DF1 neural network (0.91 accuracy, 0.9140 model F1) and kept the almost perfect DF2 neural network (0.9956 model F1). Once the two branches are refined, they are integrated into an L+L fusion architecture (for the 6 classes of risk prediction, accuracy of 0.91, macro F1 of 0.88, MAE of 0.0905, RMSE of 0.1219, R of 0.9891 and macro AUC of 0.9895). These results reveal that the best combination of good classification, handling of the minority classes, and continuous risk estimation performance is achieved in the final architecture.

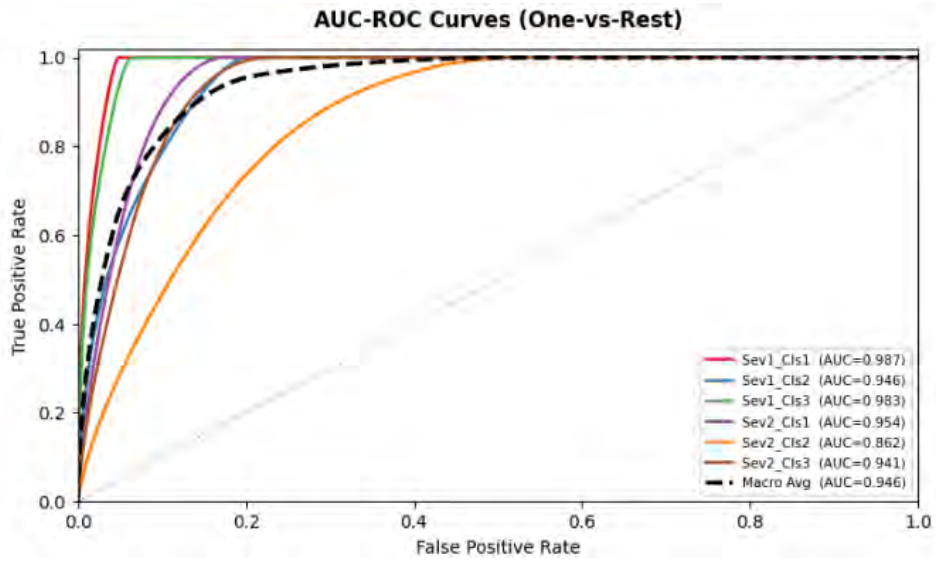

Figure 5.3: Performance of the Baseline LL Model across All Risk Scores - Visualizes the reliability and consistency of the initial late fusion (L+L) model in assigning risk scores across the entire test set.

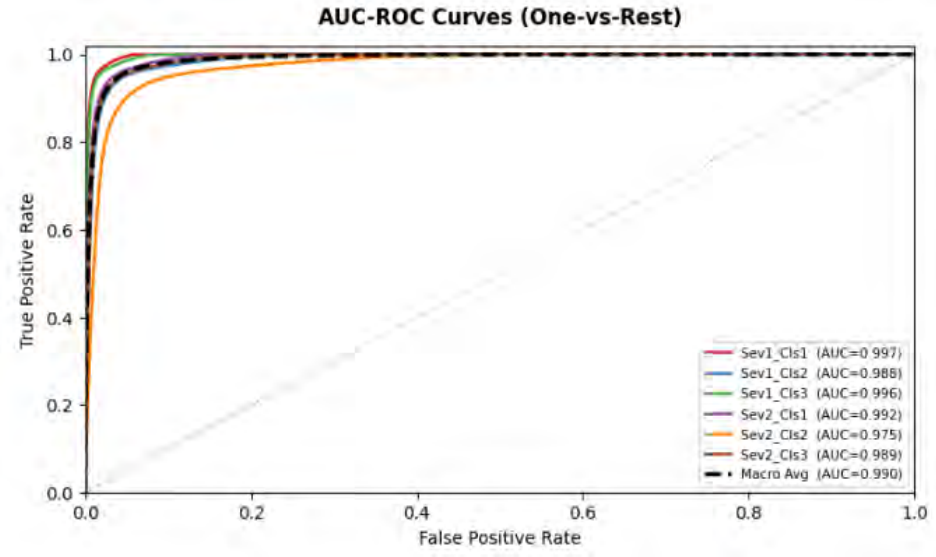

Figure 5.4: Enhanced Classification Performance using LLM Integration - Shows that LLM-based post-processing significantly improves classification metrics compared to the baseline model.

Following the integration of an LLM, the model shows a marked improvement in discriminative power, raising the macro average AUC to 0.990. The curves are more tightly clustered toward the top-left corner, signifying a much higher True Positive Rate and lower False Positive Rate across all six categories.

Comparative Interpretation

To conclude, neural networks achieved a great basis of classification, but it's the fusion architectures that yield significant improvements. The LSTM experiments proved that modeling the temporal behavior is crucial in the case of behavioral

sensor data but not at all useful in the case of static context data and an accident prediction. The KD algorithm proved itself more efficient and the main source of improved performance comes from late fusion architecture. Overall, among all considered methods, the LLM-assisted L+L architecture resulted in the most stable and appropriate architecture for driving context-aware risk prediction.

5.2 Analysis of Design Solutions

Several design choices were made while developing the proposed context-aware driving risk assessment framework. Many design choices occurred throughout the construction of the proposed context-aware driving risk assessment framework. Major design decisions revolve around the model structure and integration, the generation of risks. Instead of picking a single design from the outset, the framework evolved incrementally, comparing several modeling solutions experimentally. Here, the key design decisions are analyzed along with the reason for selecting the final framework over others.

5.2.1 Alternative Model Architectures

First, some standard classical machine learning algorithms were tested on both datasets: Random Forest, Support Vector Machines, and some gradient boosting approaches (XGBoost and LightGBM). All these are well-performing algorithms in terms of prediction accuracy, efficiency, and interpretability via feature importance analysis. In the contextual accidents dataset, gradient boosting models are superior among classical baselines and form the best reference for later development of the neural networks. In the behavioral dataset, Random Forest is also competitive, but a custom neural network performs better and is selected as the only model for the behavioral baselines.

The two proposed custom neural network models (for tabular and sensor data) are chosen based on their higher scalability, better performance than models trained using standard deep learning methods, and more optimal balance between performance and deployment efficiency. On the Dataset 1, the custom tabular neural network gives a validation accuracy of 0.68 and a validation F1-score of 0.67, comparable to the best contextual tree-based baselines while being far more memory efficient and faster at inference. On the Dataset 2, the custom sensor-based neural network gives a validation accuracy of 0.999 and a validation macro F1-score of 0.999, outperforming the best behavioral tree-based models. This suggests that two specialized neural network models are appropriate as two main single-model predictive branches.

More advanced neural network structures were also considered: temporal modeling with LSTMs in order to see if time series learning would have any benefit, and knowledge distillation as a method to compress a large teacher model into a smaller student model. As expected, the performance of LSTM depends greatly on the data: it performs better for the behavioral dataset, where patterns in time series are more explicit than in the context of static tabular data in the accidents dataset. Knowledge Distillation shows that the student model can even slightly outperform the teacher model and its real benefit lies in its smaller size and readiness for deployment.

5.2.2 Data Integration Strategy Analysis

The choice of how to integrate the contextual and behavioral information was a key design decision. Direct feature-level fusion (early fusion) was not suitable, as the two data sources are inherently different. Context and behavior data have different temporal resolutions, structures, dimensionalities and semantics and no direct link at the instance level (a common ID, for example) exists between data points. Integration in the form of simple data merging would likely result in problems aligning the different types of data as well as lack transparency and reliability. For these reasons, an output-level integration strategy was used. First, both datasets are modeled independently and the branches were allowed to specialize in their domain of operation. This retained the original data sources and improved the modularity and upgradeability of the individual branches. An output-level strategy also facilitates the interpretation of the system, as the contextual severity and behavior class predictions were maintained throughout the process. The output level integration was then progressively expanded to integrate features. The first form of integration used was a survey-based mapping of behavior classes to severity classes. This was then progressively enhanced through the continuous estimation of expected risk based on predicted class probabilities, integration of class prediction at the decision level through a meta-learning architecture, and finally, late fusion incorporating refined outputs of the modeled branches and survey-based risk information. These incremental approaches toward output-level fusion showed that it was beneficial to keep the branches distinct while integrating the prediction information at the output level of the framework.

5.2.3 Risk Mapping and Continuous Risk Score Design

The first crucial decision that had to be made during this research process was how to integrate the outputs of the contextual severity model and behavior classification model into a single driving risk value. As an arbitrary combination method could be suspected, we decided to combine them via an elicitation experiment of people's estimation of risk. To do so, we let the respondents compare and rank six different driving situations, created by the intersection of two levels of contextual severity (low and high) and three levels of behavioral driving style (slow, normal and aggressive). Here is a Demographic Breakdown of Survey Respondents by Driving Experience. This pie chart illustrates the experience levels of the 44 participants involved in the subjective risk assessment study. The majority of respondents (47.7Using the collected data, we computed the average risk value for each of the six situations (these results are displayed in Table 5.2) and obtained a 2x3 risk matrix. Clearly, the lowest perceived risk is related to the combination of low severity and slow driving (1.744), while the highest risk is to be found in the combination of high severity and aggressive driving (6.070). Other combinations were clearly between these extremes, and we observe a clear trend toward higher risk when severity or speed or both are increased.

This survey based risk matrix allowed us to have an easily interpretable and informed combination of the two different model outputs, the actual result for our risk estimation in our first prototype of the framework. The determined class from the severity model and the class of the behavioral model are used to choose a discrete risk value (one of the 6 matrix elements) out of the already computed one. The

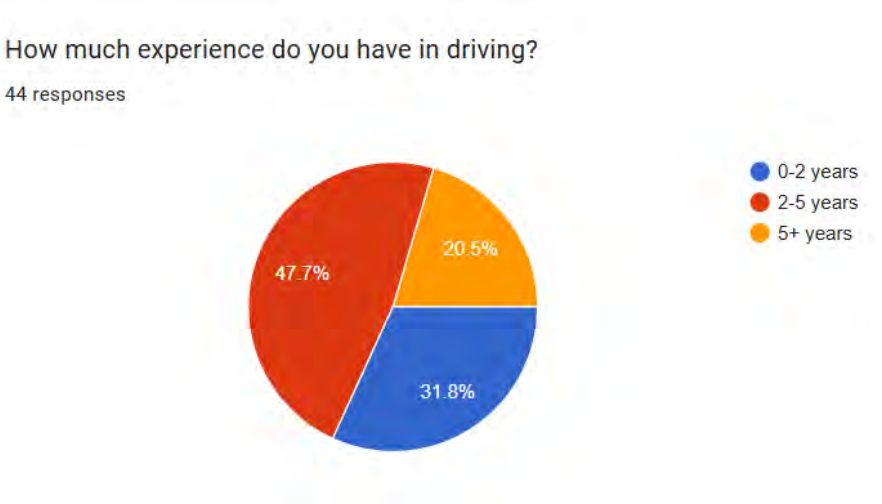

Figure 5.5: Driver Experience Distribution of Survey Participants - Provides a demographic context for the expert knowledge integrated into the late fusion risk matrix.

most obvious benefit of such a method is its intelligibility. However, the disadvantage that a prediction has to be always classified into only one state in combination with possible model uncertainties has led us to extend this work with an elicitation experiment of human people's judgment as well for a continuous range by using expected values from this table.
The estimated probability distributions of the severities and the behaviors were applied in the calculation in place of simply the predicted class labels of the two models. The probability estimates of each predicted distribution were combined to form a joint probability distribution, and the probability contributions of the possible severities and behaviors were applied as weightings for each cell of the risk matrix. The final risk score was the sum of all of the products of probability and risk for each of the six possible severities and behaviors. This allowed the risk score to continuously range from six levels. The bar charts display how users perceive risk in six different controlled scenarios.

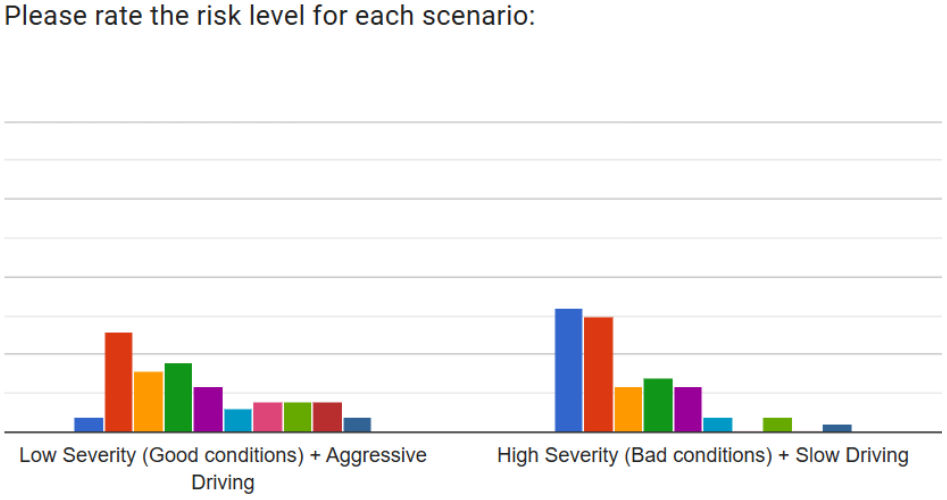

Figure 5.6

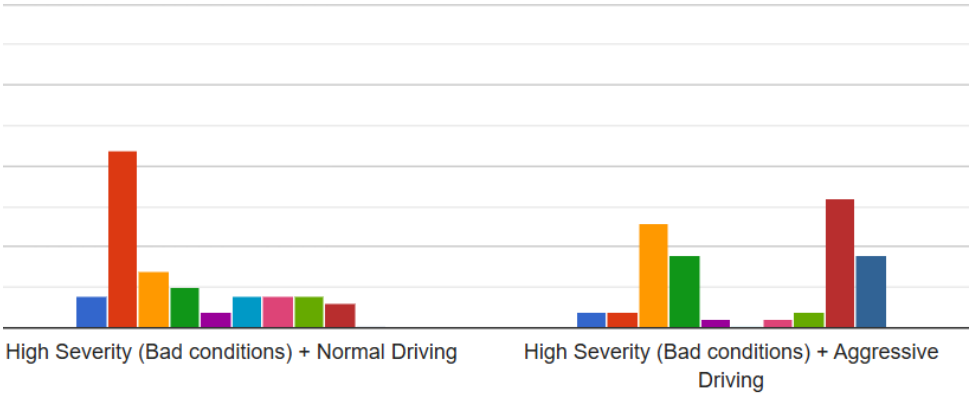

Figure 5.7

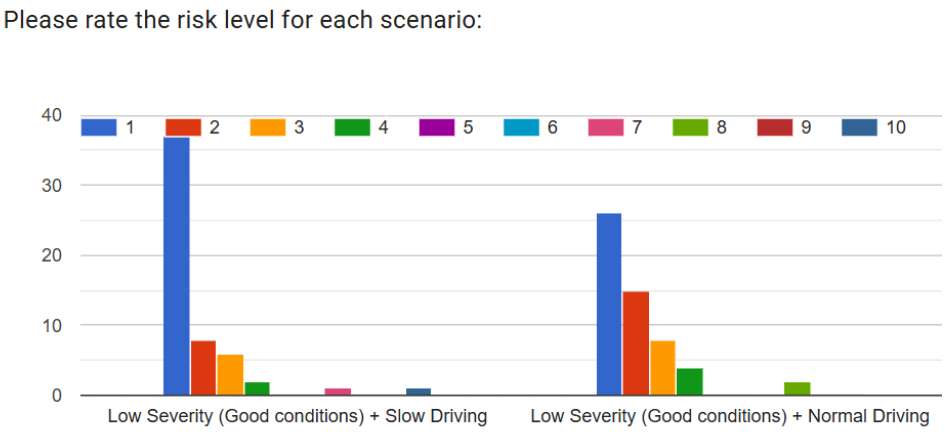

Figure 5.8: Subjective Risk Perception Across Driving Scenarios - Summarizes survey results that define how humans perceive risk across six distinct driving situations, forming the basis for the framework's risk mapping.

An improvement from discretely mapping predictions to predicting the expected risk is that interpretability has been maintained by maintaining the risk level derived from the surveys. Flexibility has been increased by modeling the uncertainty in predictions. These characteristics have made the final risk level more realistic and robust.

Table 5.3: Quantified Expected Risk Scores Derived from Subjective Survey Data Integration

Contextual Severity	**Slow**	**Normal**	**Aggressive**
Low Severity	1.744	1.930	4.581
High Severity	2.750	3.628	6.070

5.2.4 Fusion Strategy Analysis

Following the determination of the two neural network branches and the survey-inspired risk formulation, various fusion strategies were examined. The first design for a fusion strategy was the D+D framework, where the two neural network models were combined using a decision-level fusion. This particular framework was found to achieve 0.63 accuracy and a macro F1 of 0.60; however, it also had relatively good recall for the most severe risk class at 0.76, indicating that it was a reasonable approach as a safety solution when only concerned about identifying critical risk situations and not global accuracy.

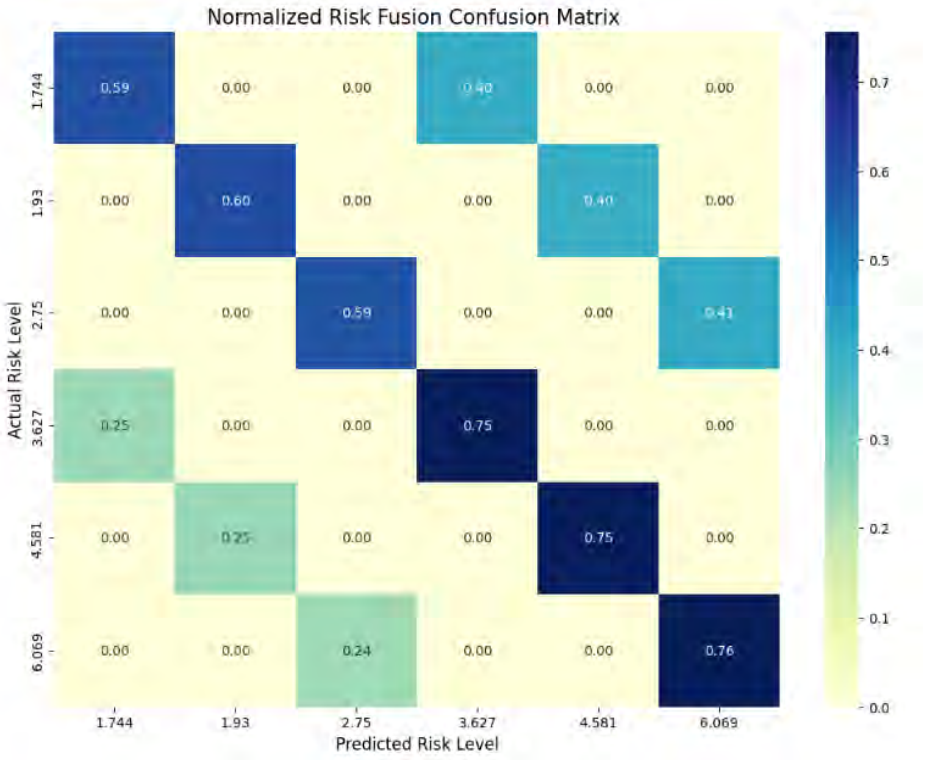

Figure 5.9: Error Distribution of the D+D Model - Illustrates the variance in risk prediction for the decision-level fusion model, highlighting areas where the meta-learner requires refinement.

The next design was the raw L+L fusion model. This resulted in a model with 0.81 accuracy and a macro F1 of 0.49, with poor recall for severe risk classes. This suggests a key deficiency, as the raw late fusion did increase global accuracy but sacrificed the performance of both the minority and high-risk conditions, making it appear strong overall but unsafe for truly risky conditions.

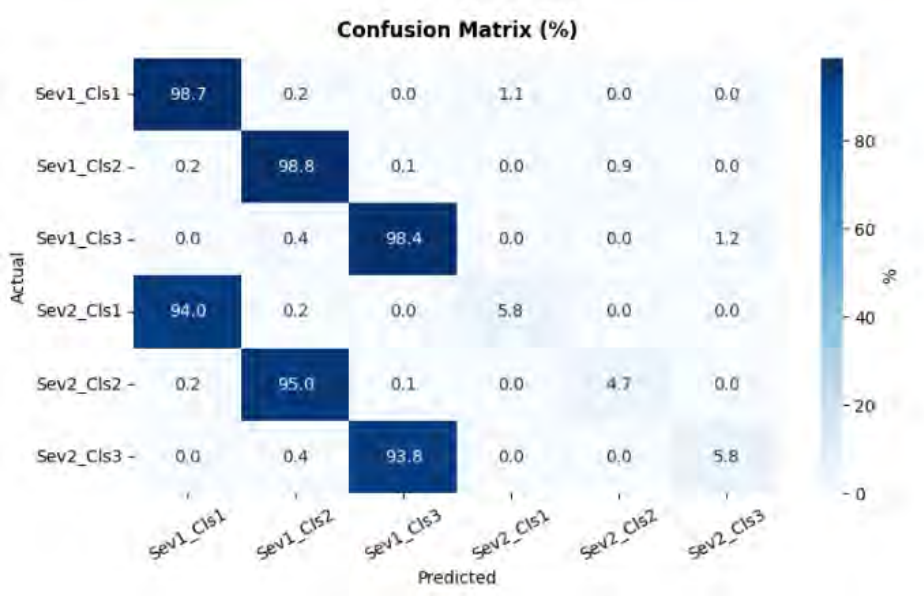

Figure 5.10: Error Distribution of the Baseline LL Model - Shows a more stable error profile compared to D+D, demonstrating that incorporating prior knowledge reduces extreme prediction errors.

The confusion matrix reveals a significant classification bias, where the vast majority of `Sev2` samples are incorrectly predicted as `Sev1`. For example, 94% of `Sev2_Cls1` instances were misclassified as `Sev1_Cls1`. This indicates that the baseline model fails to adequately distinguish between the two primary severity levels, despite achieving high intra-class accuracy within the incorrect severity tier. Such bias highlights the need for improved feature representation or model calibration to ensure reliable severity-level discrimination.

The final and best approach was the LLM-assisted L+L fusion model. Here, the contextual dataset is preprocessed with the LLM to identify and remove irrelevant noise before the final fusion. This led to a more effective contextual branch and resulted in vastly improved performance. The final model achieved 0.91 accuracy, a macro F1 of 0.88, and the strong regression results R = 0.9891, demonstrating that this final fusion strategy balanced interpretability, safety, and accuracy most effectively.

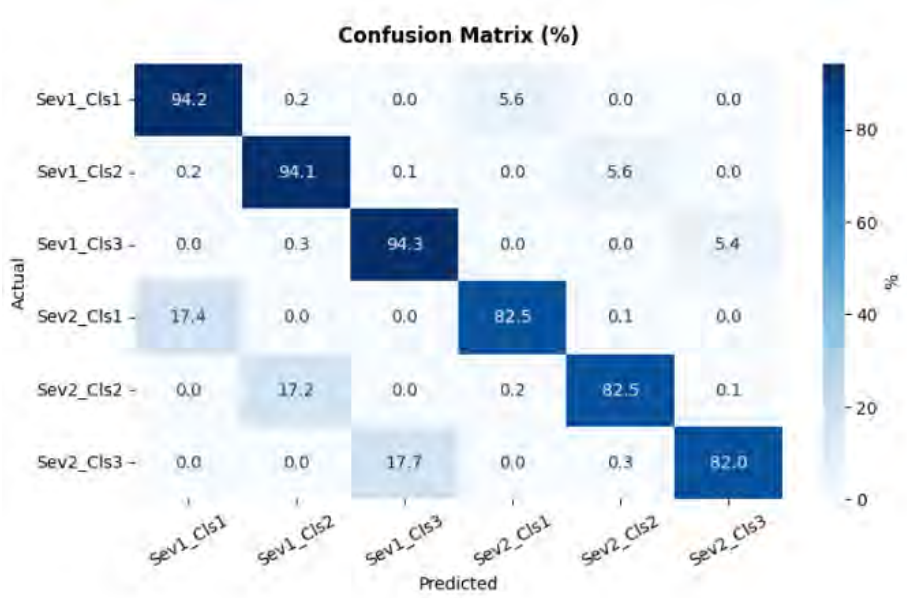

Figure 5.11: Classification Accuracy and Severity Distinction with LLM Enhancement - Confirms that LLM refinement specifically boosts the system's ability to identify high-severity risks accurately.

This matrix demonstrates the ability of the proposed LLM-based model to resolve the severity confusion observed in the baseline. The model now correctly identifies `Sev2` classes with substantially higher precision. For instance, the accuracy of `Sev2_Cls1` increased from 5.8% to 82.5%, representing a dramatic improvement in cross-severity discrimination. As a result, the cross-severity misclassification errors that previously dominated the baseline predictions are significantly reduced, validating the effectiveness of the enhanced architecture in distinguishing between primary severity levels.

Table 5.4: Error Analysis of Fusion Strategy Outputs

Model	**Class**	**Issue Observed**	**Evidence**	**Impact**
Raw L+L	High Severity	Under-detection of severe risk	Sev2 cases misclassified as Sev1; ∼94% error in one high-severity class	High-risk events may be ignored
D+D	High-risk classes	Safety-sensitive but lower overall accuracy	Highest-risk recall ≈ 0.76; overall accuracy 0.63	Better danger detection, but less stable overall
LLM L+L	All risk classes	Minor residual errors	Improved confusion matrix; Macro F1 ≈ 0.88	Most balanced and acceptable final model

5.2.5 Design Factor Evaluation

The final framework was chosen not only on performance metrics but also on aspects of computational efficiency, interpretability, modularity, and overall deployment potential. The two distinct prediction branches allowed for independent development and optimization of both context modeling and behavior modeling. The neural networks were still computationally light enough for efficient, scalable training, and both knowledge distillation and LLM-assisted tuning served to increase efficiency and data quality in unique ways. Interpretability was an important factor to consider. By retaining the survey-derived risk matrix, transparency was maintained in the final risk score's mapping from context and behavioral predictions. And by continually calculating expected risk, the model allowed for representation of uncertainty, rather than being discarded for interpretability's sake. Also, the modular fusion allowed direct comparisons between different integration methods and isolating performance improvements to specific components. In terms of usability, the model was attractive, as it made use of widely available data, publicly available software tools, and a modular approach that could easily be extended in future work. While live multimodal sync was out of scope for this paper, the selected architecture appears to be a practical, extendable base for future real-time, context-aware driving risk assessment systems.

5.3 Final Design Adjustments

The process of comparative analysis on base models, neural networks, and fusion techniques prompted the inclusion of several modifications to the proposed system to enhance effectiveness, robustness, and practical utility. Such modifications did not merely exist as separate additions but as a consequence of a process whereby limitations of previous stages became guiding principles for the subsequent system architecture.

5.3.1 Neural Network Architecture Refinement

The first set of modifications addressed the custom neural network models built for both Dataset 1 and Dataset 2, as they were the predominant predictive branches of the final system. These included the modifications made to their architecture for greater stability, efficiency, and generalization:

- Modification in dropout rate (Dataset 1): the initial dropout of 0.5 was narrowed down to the 0.2–0.3 range, thereby preventing underfitting and helping to achieve higher validation scores.
- Improvement in optimizer: initial training using a fixed learning rate was switched to the Adam optimizer to allow the system to achieve more stable convergence.
- Improvement in features encoding: simple labeling encoding was replaced by StringLookup, and handling features using embedded representation to facilitate learning of features with high cardinality in contexts.

- Memory efficiency: Memory-efficient approaches were implemented by using aggressive datatype downcasting, enabling the system to learn efficiently on the whole contextual dataset.

With all these tuning processes, the custom neural networks are more suitable for deployment at scale. For Dataset 1's neural network, it includes processing of both numerical and categorical input features separately with normalized layers, dense hidden layers, and dropout with binary output. It uses two dense hidden layers of 64 nodes with binary output activation at the last layer, and the Dataset 2 neural network is lightweight (just three dense layers).

5.3.2 Class Imbalance Handling

The largest practical hurdle faced with the contextual accident dataset was the heavy class imbalance between non-severe and severe accident cases. Given that the severe class is the more safety critical class, increasing its detection was the priority during refinemen

- Stratified train-test split.
- Greater focus on macro-averaged F1-score.
- Balancing-oriented sampling methods.
- Class-aware training decisions to be more sensitive to minority class cases.

These adaptations were made to ensure that an accurate model was also safe and not prone to errors concerning the most safety critical class. This also became apparent during the Knowledge Distillation stage when a WeightedRandomSampler was explicitly added to deal with minority class exposure during training.

5.3.3 Engineering and Representation Refinement

The next set of design adjustments focused on improving the representation quality of both datasets.

- Extract temporal features from the timestamps for accidents.
- Represent the temporal location of the accident clearly (hour, weekday, month).
- Represent events at a more basic level (IsWeekend and IsNight).
- Encoded the environmental and spatial categorical features more effectively.
- Clearly interpolate the missing sensor data.
- Clean the sensor noise with more efficient outlier rejection techniques and normalize the axes of the inertial sensors.
- Attempt to preserve more of the motion dynamics for future neural and sequence modeling techniques.

These changes improved the representations for the predictive branches and provided higher-quality model outputs to feed to the fusion modules that followed. They also showed that the behavioral data were much better served by a temporal representation than the tabular contextual data.

5.3.4 Sequence Modeling Adjustment

To test if temporal modeling could improve the predictive branches, sequence-based learning was introduced. This resulted in an important design realization.
For the 2nd dataset, sequence modeling performed very well, as the sequence of sensor inputs intrinsically contained temporal relationships. In this context, the behavioral model, when trained as an LSTM model, was almost perfectly accurate. Thus, the learning on the temporal domain is applicable to motion data in the driving context. However, when this idea was applied on the 1st dataset, the performance was not so successful, with low classification performance and an ROC-AUC around 0.5874. This demonstrated that sequence learning was not applicable for static contextual features in accident situations.
Therefore, in the final model, the sequence-based learning only serves as the behavioral model improvement instead of being adopted as a general modeling tool for both branches.

5.3.5 Knowledge Distillation Adjustment

Knowledge Distillation was considered as a design enhancement toward optimization in the contextual branch. It is not intended to replace the current model framework but merely to find out if a smaller model is able to retain the performance from the complex teacher model while being better for the deployment stage.
In this study, the TabNet served as the teacher model and a lightweight MLP served as the student model with combined hard-label and soft-label supervision. The student was evaluated to have an accuracy of 55.02

5.3.6 Risk Score Reformulation

A second key change was how the final risk score for driving risk was calculated. Initially, the framework employed a discrete 2X3 mapping of severity and behavior class. However, the six values in the matrix were not simply chosen; rather, they were based on survey responses where the mean rating from the survey of the specific scenario was the input value. The average values were roughly 1.744, 1.930, 4.581, 2.750, 3.628, and 6.070, respectively, for the six severity-behavior combinations.
Although this discrete survey based mapping was interpretable, it could not truly capture the idea of predictive uncertainty. The continuous expectation of risk was added for the final framework, combining the severity and behavior probabilities across the same matrix from the surveys. This allowed the final risk score to fluctuate between the six scenario means, rather than only picking one rigid class label.

5.3.7 Fusion Refinement and Final Model Selection

The last and most significant set of modifications was around the fusion approach. The design choices regarding fusion evolved as a function of the data and the results obtained:

- D+D fusion increased the ability to correctly identify high-risk classes, specifically for the most severe risk levels, though accuracy did not improve substantially in general.

- Raw L+L fusion yielded the highest overall accuracy but was very poor in terms of correct classification of severe risk and minority risk classes, rendering it unsuitable for safety-critical prediction, although regression fit remained excellent.

- LLM-assisted L+L fusion increased the context branch accuracy by intelligently pruning DF1 data points that introduced noise in the final model, resulting in the best and most balanced outcome overall. The resulting system attained an accuracy of 0.91, macro F1-score of 0.88, and an R of 0.9891.

The data indicate that the final system design was not the result of one design choice but rather an incremental improvement at each stage of the modeling, representation, and fusion processes.

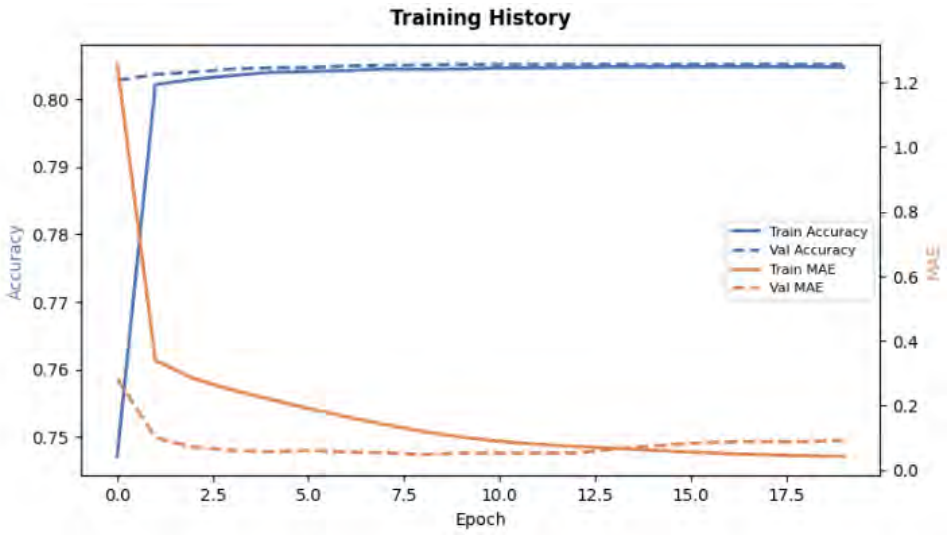

Figure 5.12

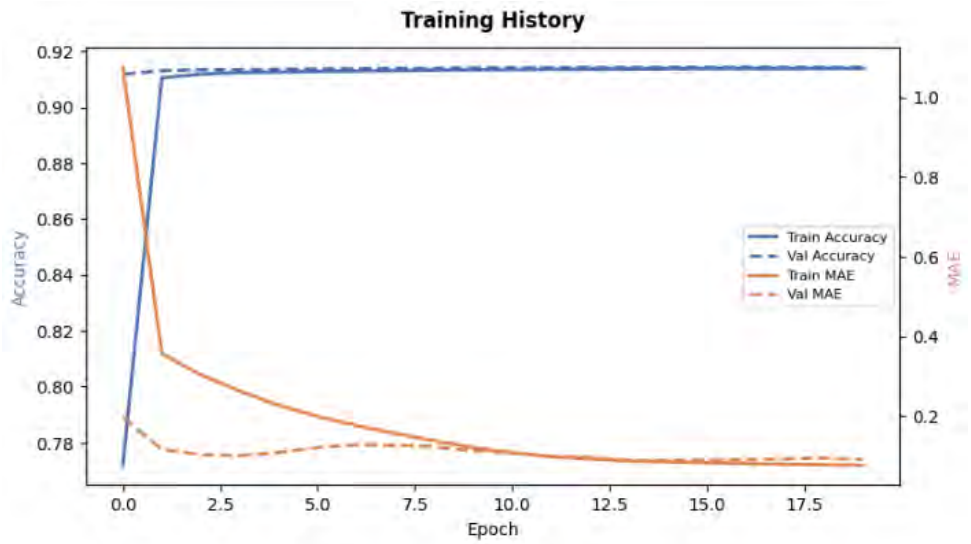

Figure 5.13: Training Performance and Loss Convergence of Baseline vs. LLM-Assisted L+L Fusion Models - Shows faster and more stable loss convergence in the LLM-assisted model, suggesting it is more efficient during training.

These plots compare the learning curves of the baseline LL architecture (top) versus the LLM-enhanced version (bottom). Both models show rapid convergence within 20 epochs, but the LLM-assisted model demonstrates more stable validation MAE, suggesting that noise reduction in the contextual data branch prevents the model from fluctuating during the final stages of training.
Ultimately, the final system adopted a combination of :

- a refined contextual neural network branch
- a more accurate behavioral neural network branch
- where appropriate, application of sequence-aware improvements
- continuous survey-grounded risk score estimation
- and LLM-assisted contextual data cleaning prior to late-stage fusion

This table provides a comparative summary of the performance metrics for the D+D, Raw L+L, and LLM-assisted L+L fusion models. While the Raw L+L model shows high regression fit (R=0.9830), the LLM-assisted L+L model is identified as the optimal architecture, achieving the highest balance between classification accuracy (0.91) and Macro F1-score (0.88), effectively addressing the minority class misclassification issues prevalent in earlier iterations.

Table 5.5: Comparative Performance Metrics of the Fusion Architectures

Fusion Type	**Accuracy**	**Macro F1**	**MAE**	**RMSE**	**R**
D+D	0.63	0.60	1.0756	1.7830	-0.7243
Raw L+L	0.81	0.49	0.0877	0.1382	0.9830
LLM-assisted L+L	0.91	0.88	0.0905	0.1219	0.9891

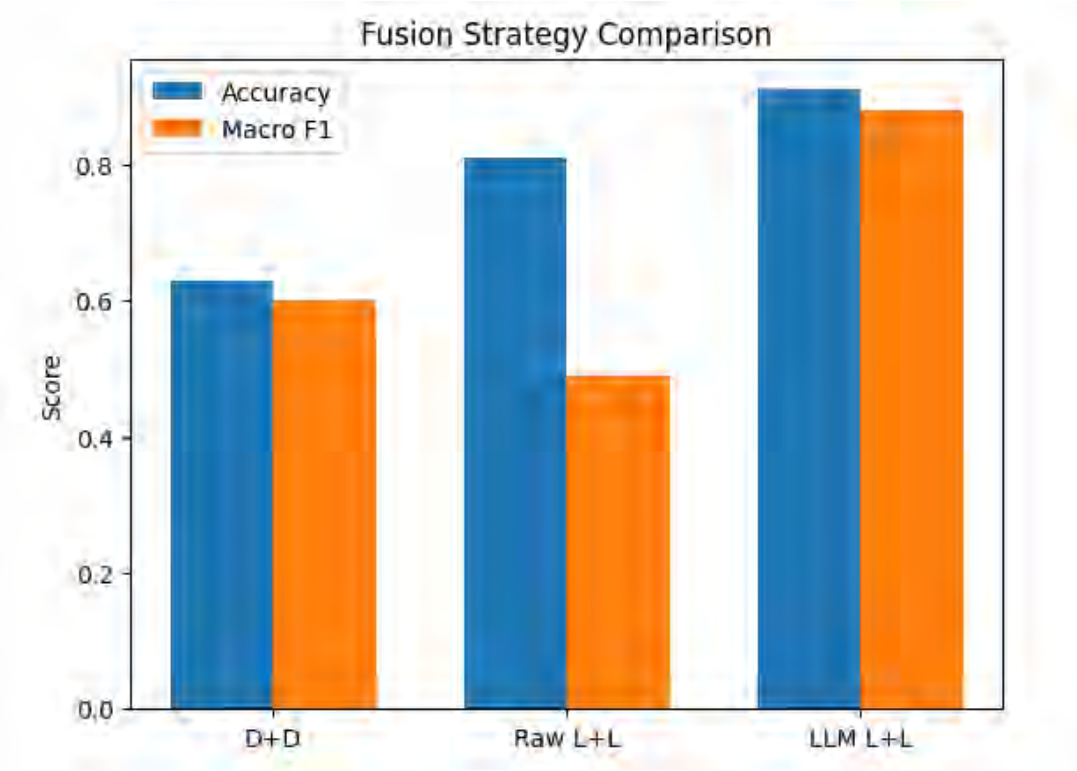

Figure 5.14: Accuracy and Macro F1 Comparison of Fusion Strategies - A comprehensive summary showing that the LLM-assisted L+L fusion strategy outperforms all other methods in both accuracy and F1-score.

5.4 Statistical analysis

A statistical analysis was performed to look into the relative performance of the presented models, consistency of the generated risk score, and characteristics of the predicted distribution. Because the framework progressed through several modeling steps, it was essential not only to look at performance in aggregate but also class-wise balance, uncertainty of prediction, and performance of the continuous risk score with respect to the different integration techniques.

5.4.1 Relative performance of experiments

The different modeling strategies implemented exhibited distinct behavioral differences across the board in terms of the provided experimental results. The custom neural networks provide good results alone while also serving as the major predictive branches of the framework. The neural network on dataset 1 returns 0.68 and 0.67 validation accuracy and macro F1-score, respectively, while the neural network on dataset 2 is almost perfect: 0.999 and 0.999 macro F1-score, respectively. As one can tell, behavioral sensor data are much easier to separate than contextual accident data.

The LSTM experiments also illustrate the effect of data structure on model performance. The LSTM model on DF1 only obtains 0.41 accuracy and a 0.40 macro F1-score, indicating it fails to find meaningful separation between different states in

static contextual accident data. In contrast, on DF2 it provides near-perfect results, 0.96 accuracy, and a 0.943 macro-F1 score, showing the strong utility of sequence learning in behavioral sensor data.

The fusion approaches exhibit relative balance and focus in different ways: while the balanced D+D fusion is accurate at 0.63 (0.60 macro F1) and maintains a high recall in the highest-risk class. The RAW L+L fusion, on the other hand, increases the aggregate accuracy to 0.81 but decreases macro F1 to 0.49, clearly performing poorly in relation to the minority class. The LLM guided L+L fusion is the best combination yielding 0.91 accuracy and 0.88 macro F1-score, demonstrating that improved context input and improved late fusion effectively make the predictions more reliable.

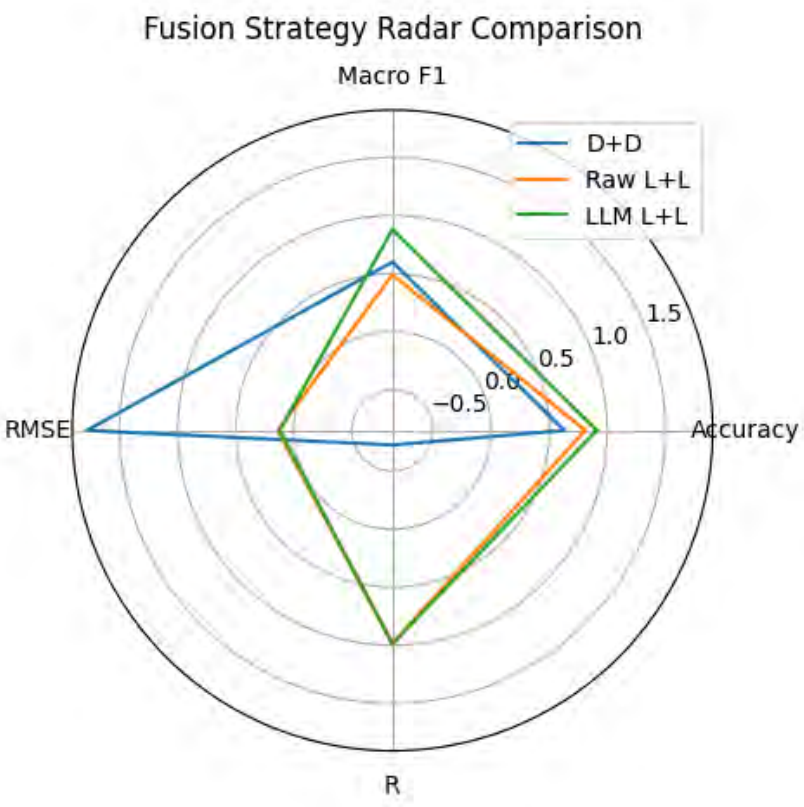

Figure 5.15: Normalized Radar Comparison of Fusion Strategies - Provides a multi-dimensional view of performance, showing that the final hybrid model balances classification accuracy with regression consistency.

5.4.2 Efficiency and model compression analysis

Here, knowledge distillation was used as a statistically inspired improvement for the sake of efficiency. In the current experiment, the TabNet teacher obtained 53.66% accuracy, while the lighter MLP student obtained 55.02% accuracy, which is 102.5% retention over the teacher, indicating the student retained not only the behavior of the teacher but also generalized a little bit better as well while maintaining a less demanding architecture, despite poor absolute accuracy due to the complexity of the accident severity task.

5.4.3 Risk Score distribution analysis

Critical to the analysis, we must investigate the distribution of the final risk score. The risk formulation used in the paper was based on a surveyed 2X3-risk matrix, in which each of the six combinations of severity-behavior was given an average risk value of 1.744, 1.930, 4.581, 2.750, 3.628, and 6.070 for their respective severity-behavior-risk categories and subsequently either mapped discrete values or used in the computation of the expected risk, a continuous estimate of risk.
It is argued that the expected risk provides a more useful insight into risk than merely mapped values. Rather than confining each instance to a precise cell within the matrix, expected risk weighted the six values obtained from the survey by the probabilities determined from the severity and behavior models. Therefore, the resultant score is able to rise or fall in response to both model uncertainty and confidence. As such, the outputted risk scores appear much more intuitive and able to capture intermediate risks than merely the discrete values.
The class-wise fusion results are very insightful too. From the purely L+L model, the overwhelming portion of the predictions seem to have been crammed into the lower severity risk classes. This helps to explain the high accuracy but poor severe-class recall. However, in the LLM-assisted L+L model, the prediction distributions over the classes were more uniform across all six risks, indicated by stronger macro F1 and better performance over all risk levels.

5.4.4 Regression Consistency Analysis

Since the output of the final framework is a continuous expected risk value, regression-based analysis was also used to check the consistency of predicted risk against the ground truth risk values.
The regression evaluation of the raw L+L fusion model yielded an MAE value of 0.0877, an RMSE value of 0.1382, and an R-score value of 0.9830. However, the LLM-assisted L+L fusion model yielded a smaller RMSE of 0.1219, an MAE value of 0.0905 and a higher R-score of 0.9891. Thus, although the MAE has increased, the decrease in RMSE and increase in R show that the prediction is more consistent and has a better fit in the case of the LLM-assisted model.
These histograms display the Prediction Error Distribution (PED), where a tighter clustering around the zero-error line indicates higher precision. The LLM-assisted model shows a reduced Root Mean Square Error compared to the baseline, which is visually confirmed by the more concentrated central peak and diminished "tails" in the error frequency.
The results demonstrate the very high consistency of prediction for continuous risk values of the final model. Though the difference in MAE of the two L+L models was not significant, lower RMSE and high R for the LLM-aided model show increased stability and better fit to the target risk structure.

5.4.5 Uncertainty and Class-wise Reliability

The class-wise comparison of the models also provides an important notion of prediction reliability. The D+D model though inferior to the others in overall accuracy provided higher sensitivity to high risk classes with the highest-risk tier having a

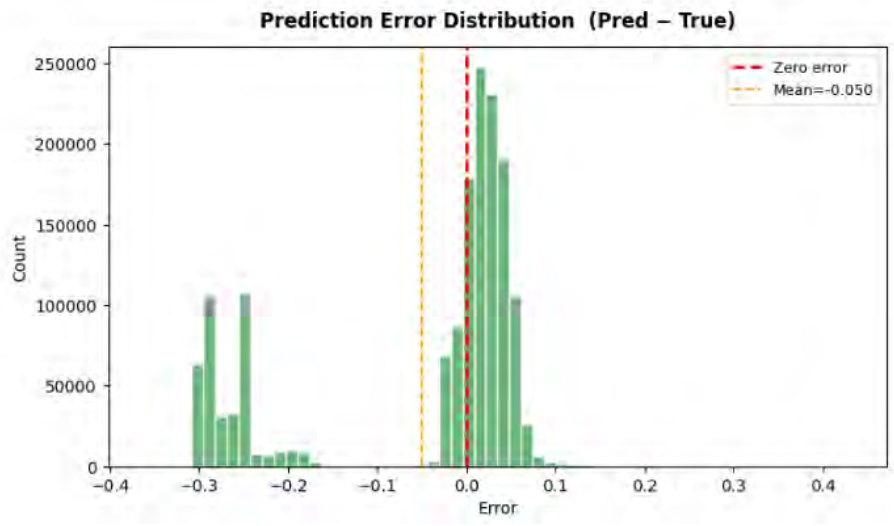

Figure 5.16: Residual Analysis of the Baseline Late Fusion Model - Visualizes the distribution of prediction errors, showing the inherent variance before applying refinement layers.

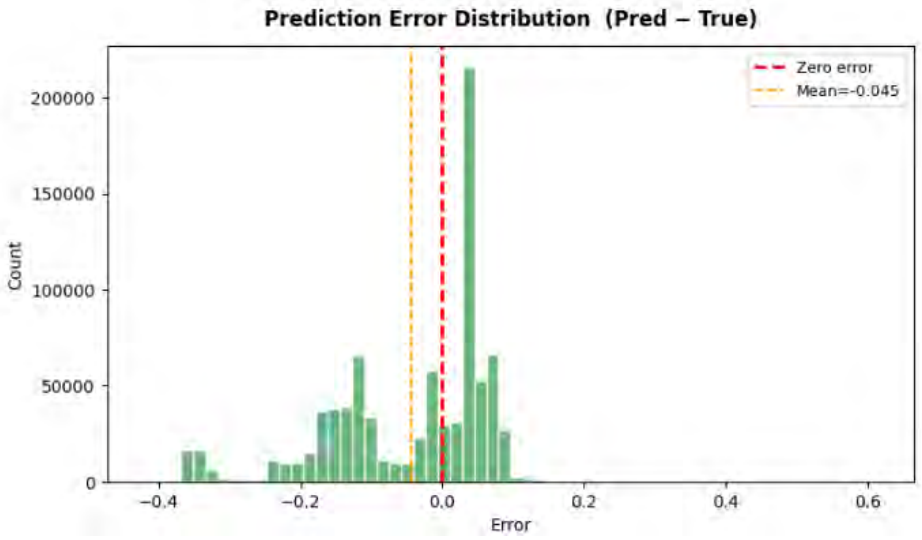

Figure 5.17: Residual Analysis of the LLM-Assisted Late Fusion Model - Demonstrates a significant reduction in error variance, indicating that LLM refinement produces more stable and predictable risk outputs.

recall of 0.76, and thus indicating higher sensitivity in detection of high-risk situations. Whereas the raw L+L model which shows high Accuracy was poor in the highest risk class's recall. Hence, high overall accuracy should not be solely looked for in a safety critical system.

The LLM-aided L+L framework significantly improved this imbalance. The final six-class output of the LLM-aided model yielded a macro F1-score of 0.88 and a macro AUC of 0.9895. The precision, recall, and F1-score were robust across both low-risk and high-risk classes for the final framework.

5.4.6 Summary of Statistical Findings

Three main conclusions can be drawn from the statistical analysis, in general:
• The predicting capabilities of the models are highly dependent on data structure (evidenced in the gap of performance between LSTM on DF1 and DF2).

• Continuous expected risk estimation is a better representation than firm discrete mapping because it retains uncertainty and interim risk levels.

• Overall statistical balance is achieved by L+L fusion aided by the LLM because it results in solid classification accuracy, robust regression consistency, and increased reliability in all six risk categories.
These results are an assurance that the final framework has not only overcome the accuracy limitations of the design processes before it but is also statistically consistent and meaningful.

5.5 Comparisons and Relationships

This section offers comparisons of the proposed framework with regard to each internal model selection and each fusion technique. In light of the nature of the investigation's iterative development, comparison is not limited to final accuracy but also considers predictive efficacy, efficiency, interpretability, and the relationship of behavioral and contextual prediction within the risk-assessment process.

5.5.1 Internal Model Comparison

The selection of two separate custom neural networks for the predictive branches of the model was corroborated by comparison. The custom tabular neural network for the context-based branch performed adequately (accuracy 0.68, macro F1 0.67) in relation to the strongest tree-ensemble baselines for that task (accuracy 0.70, macro F1 0.67), yet demonstrated substantial benefits in speed, memory usage, and deployment. Similarly, the custom sensor-based neural network for the behavioral branch, with accuracy 0.999 and macro F1 0.999, outperformed the best of its respective baseline category (tree-ensemble, accuracy 0.96, macro F1 0.95).
These selection decisions were also supported by additional, advanced internal model comparisons. While the time-series sequence-learning ability of the LSTM-based behavioral model gave near-perfect results for DF2, this proved ill-suited for DF1 and its static data. Knowledge distillation showed potential gains in efficiency, but not the dramatic overall improvements obtained later through fusion adjustments. These internal model comparisons reinforce the validity of assigning a context-appropriate, custom model to each branch rather than attempting to generalize across both with the same model class.

5.5.2 Comparison With Existing Works in the Literature

When compared to related methods in the literature, the proposed framework can be situated in a sweet spot between interpretability and sophistication. Behavioral-only models can yield impressive performance, but they disregard environmental

and contextual factors that shape the real risk of accidents. Context-only models accurately identify trends in crash severity, but they disregard driver behavioral factors. Complex, multimodal systems like synchronized sensor fusion or graph-based frameworks are able to represent more complex relationships but often demand very tight temporal alignment, significant computational power, and a significant reduction in interpretability. The comparative framing for your works appears in the draft.
The proposed framework is distinguished from existing approaches in a number of important ways:

- Dataset-independent
- No need for per-instance synchronization
- Modular modeling is possible
- Retains interpretability via survey-grounded risk formulation
- Increases performance iteratively via output-level fusion

This positions the framework as more pragmatic than feature-level multimodal systems, which struggle when heterogeneous data sources cannot be directly aligned. At the same time, the results of the LLM-assisted L+L model demonstrate that large improvements can still be attained without forsaking modularity or interpretability. While the 0.91 accuracy, 0.88 macro F1-score, and macro AUC = 0.9895 of the final six-class fused output position the method competitively in its current form, future iterations of the framework will push the performance further while still being relatively simple and transparent.

5.5.3 Comparison between fusion strategies

A very crucial aspect in this research is the comparison between the fusion strategies themselves. D+D fusion model, L+L fusion model without assistance, and LLM-assisted L+L fusion model are three different stages of development in terms of the degree of integration between the systems.
The D+D fusion model with a balanced fusion scheme gets a 0.63 for accuracy and a 0.60 for macro F1-score. Although its accuracy is quite low, its recall for the highest risk class is not bad with 0.76. This suggests the fusion model D+D has better detection for critical risks, and it will be useful for a safety-related issue.
The L+L fusion model with raw contextual data shows a much higher accuracy, at 0.81, but its macro F1-score is lower, at 0.49, with especially poor recall for the severe risk classes. This implies that the increase in accuracy is achieved by the dominant low-risk classes, while minority classes of high risk cannot be captured effectively. This system is not adequate from a practical safety point of view even though its classification and regression accuracy values are higher than the D+D model.
The LLM-assisted L+L fusion model offers the best trade-off between all the features, with the contextual branches cleaned using LLM based noise detection and cleaning followed by late fusion. This resulted in an accuracy of 0.91, macro F1-score of 0.88, an MAE of 0.0905, an RMSE of 0.1219, and an R of 0.9891. It is evident

that the significant improvement is attributed not just to late fusion but also to a better quality of context data.

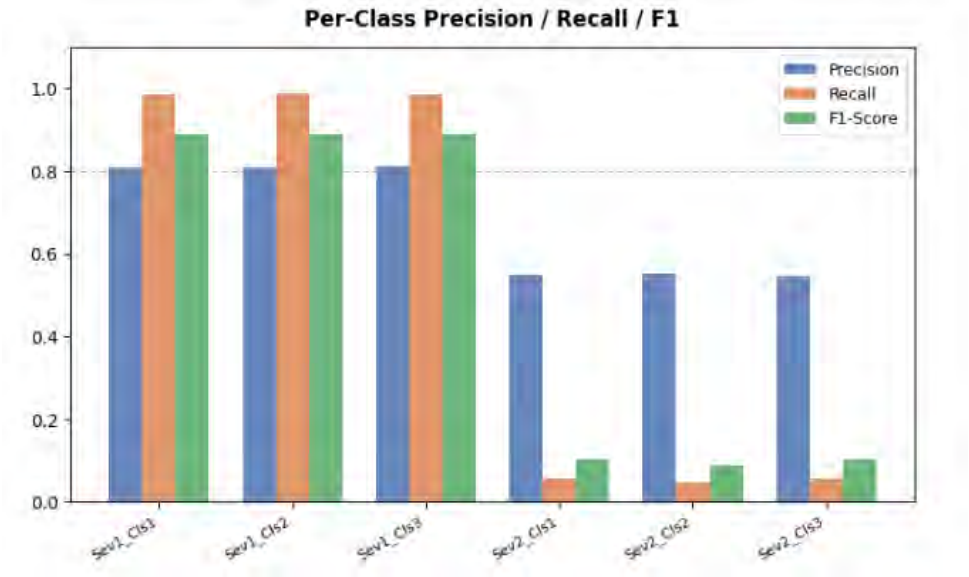

Figure 5.18: Baseline Classification Metrics across Severity Tiers - Provides the precision and recall scores for the standard model, highlighting its performance in distinguishing between low and high-risk driving scenarios.

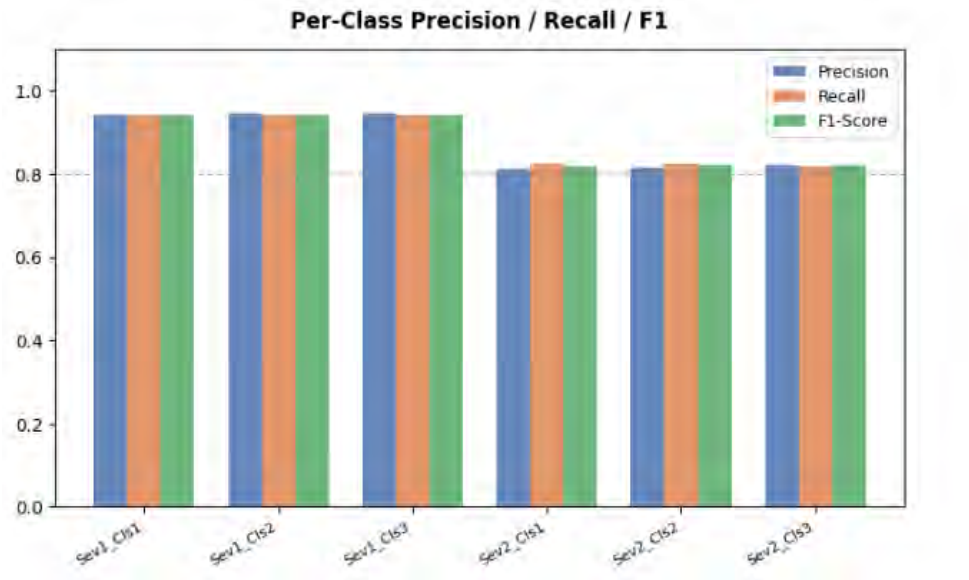

Figure 5.19: LLM-Enhanced Classification Metrics across Severity Tiers - Demonstrates the performance improvement specifically in critical high-severity tiers, showing how the LLM-assisted model better identifies life-threatening risks.

These bar charts highlight the drastic improvement in classification performance for "Sev2" classes (high-risk levels). The baseline model (top) suffers from extremely low recall in Sev2 categories, effectively "missing" high-risk events, whereas the LLM-assisted model (bottom) maintains balanced Precision and Recall above 0.8 across all categories, ensuring no risk tier is neglected.

5.5.4 Comparison of Risk Formulation

We discovered a key relationship between the discrete and continuous risk formulation. The original setup used a direct mapping from severity-behavior pairings to a list of six discrete risk values, not defined arbitrarily but assigned to six fixed scenarios. Our survey collected average rating values (at the six scenario pairings) of 1.744, 1.930, 4.581, 2.750, 3.628, and 6.070 for low/slow, low/normal, low/aggressive, high/slow, high/normal, and high/aggressive. The latter continuous expected-risk formulation did more, taking the weighted average of all six values at each point with the joint class probabilities coming from the two predictor branches. This was able to give a smoother estimate and represent predictive uncertainty and intermediate-risk states. In practice, it provided a more nuanced risk score than a hard lookup because it could increase or decrease with the confidence hierarchy of the two predictors. This continuous-discrete pairing was one strength of the final framework.

5.5.5 Relationship Between Contextual and Behavioral Predictions

The contextual severity branch and the behavioral classification branch each tackle related, yet distinct, factors that comprise driving risk. While the former is based on identifying accident-inducing contextual features and scenarios, the latter focuses on modeling the driver's motion behaviors. The disparity in these branches' contributions across the fusion approaches is indicative of a complementarity rather than redundancy.

We see this relation across our fusion results. With the D+D approach, the two outputs worked in concert to identify severe-risk cases. Without it, the L+L raw result showed that robust behavioral prediction could not alone result in uniformly good performance due to contextual shortcomings with regard to severe risk categories. The LLM-enhanced L+L approach, which directly leveraged an improved contextual branch to directly boost performance across all six classes, indicates that this relation is non-additive, i.e., the final risk prediction must rely on simultaneously high-quality performance from both branches.

5.5.6 Comparative insight

In summary, these comparisons suggest that our overall system performed best not because one single model outperformed all others, but because multiple complementary design choices were integrated efficiently: custom neural nets for both architectures, using sequence modeling only when appropriate, employing knowledge distillation for improved efficiency, and a survey-driven continuous risk score for interpretability-though most impactful were the iterated fusion approaches and most effective, finally, the LLM-enhanced L+L architecture, which achieved a good balance between detection accuracy, reliability of minority class predictions, and continuous risk scoring for interpretability.

5.6 Discussions

This section combines the findings of the performance evaluation, design analysis, and the comparison study in order to interpret the implications of the proposed framework. Specifically, the results are examined for what they suggest about the fit of our models to data, combining multimodal risks, practical feasibility, limitations, and the novelty of the work.

5.6.1 Interpretation of Results

The results of the conducted experiments confirmed that the proposed framework can effectively incorporate two complementary risk dimensions—contextual accident severity and behavioral driving style. Strong base models could be trained on both contextual accident data and behavioral sensor data, with the custom neural network for Dataset 1 achieving high comparable standalone accuracy and the custom neural network for Dataset 2 achieving classification accuracy that was essentially perfect. The fact that both contextual accident and behavioral sensor data support highly accurate prediction models, but do so in very different ways-with vastly different degrees of separability and complexity, highlights important distinctions for the multimodality of driving risk.

The modeling analyses provided insight into how model design should be related to data. Whereas sequential learning seemed to be extremely successful in the behavioral sensor domain, as expected, its use was unsuccessful with the contextual accident data where there is no natural temporal sequence. This was an important demonstration that improved accuracy is not always associated with model complexity. Similarly, knowledge distillation aided the contextual network, slightly improving generalization and efficiency but not increasing the gain in predictive power the most.

Most importantly, the comparison between different fusion methods illustrated key differences. The accuracy of the D+D model was lower than L+L's raw model and the LLM-improved framework. However, it was more sensitive on some essential risky cases and was thus very valuable in the aspect of safety. While the accuracy of the raw L+L model is high, its severe-class recall performance was abysmal, proving that high accuracy does not mean good risk modeling. The final LLM-assisted L+L framework had an overall good balance: accuracy 0.91, macro F1-score 0.88, and R=0.9891. We demonstrate that sophisticated context refinement and a late fusion approach can indeed bring about both precision and accuracy in risk prediction.
More importantly, the change from discrete risk mapping to the estimation of expected risks had improved the system's realism. The ability to represent middle states between risks and to express uncertainty in predictions is provided by weighting the survey based risk matrix by predicted probabilities from each branch of the system. It provides a more informative output than the rigid static lookup score.

5.6.2 Implications for Practice

This research has many practical uses. First, the framework is well-suited for context-aware driver risk assessment systems where an analysis of context and behavior needs to be carried out jointly. Applications in these domains are smart vehicles, driver monitoring systems, and transportation safety analytics systems where prediction from a single source is to be surpassed.
Second, the suggested system framework can be adapted for insurance usage based systems where estimation of risk needs to reflect not only the style in which a person drives but also the context of that driving. This system can provide a more nuanced basis for insurance risk classification than discrete class labels can, as it produces a continuous and interpretable risk value derived from human-decision-based surveys.
Third, the systems can be deployed easily in fleet safety management and traffic monitoring, as their architecture is a module-based system where contextual branches and behavior branches are designed and developed separately. The study has shown that data enhancement techniques such as LLM-aided noise recognition can result in a large effect on the risk quality, which could be generalized to any multimodal decision support system.
Fourth, it can be considered a balance between the human logic based methods that are very interpretable but limited in flexibility and complex black-box methods.

5.6.3 Limitations

Despite the study's advantages, there are a number of limitations. Two datasets were recorded separately, so no driving scenarios correspond to the exact same real-world events in both contexts. This prevents real-time, fine-grained, multimodal correspondence in our framework; instead, the framework learns how context and behavior interact via a statistical relationship, not event-to-event correspondences. Therefore, direct event-level interaction between environment and behavior cannot be modeled in the proposed framework.
The environmental context dataset used also has geographical restrictions—namely, that it comprises data recorded only in the United States, and hence the learned relationship between environmental context and risk may reflect local conditions not true elsewhere (e.g., road features and traffic flow). Likewise, the environmental context learned in the behavioral dataset comes from smartphone based sensor measurements rather than a vehicle's own sensor inputs or accident-linked behavioral ground truth; thus, the external validity of the learned context model may be limited.
Class imbalance during the accident severity prediction task remained a problem, especially the extreme event. Despite attempts to overcome this using balancing techniques, the minority class continued to pose problems for the network at intermediate stages. The final framework also relied on a risk matrix estimated from survey responses. This framework is easily explainable to humans (being human-grounded) but is based on observer judgements and so can vary between populations.
A further limitation is that our developed framework was evaluated entirely offline, with no attempt made to put it into real-time operation. The refinement using an LLM involves the detection of "noise" by chunks of contextual data, this would be need more investigation to be part of a production level, although it increases the quality of data used. The neural networks used are obviously non-interpretable; in

addition, the parts relying on the LLM are not entirely interpretable in the same way a purely rule-based model could be.

5.6.4 Contributions and Novelty

The work makes the following major contributions:
1. It introduces a hybrid DL and fusion-based driving risk assessment framework that integrates context-based accident modeling and behavior-based driving patterns while being dataset-independent. It is a step forward, moving away from attempts to artificially fuse disparate datasets in an early stage of modeling, showing instead that robust fusion can be realized through output modeling and smart fusion architecture.
2. It presents a survey-grounded risk formulation where each of the 6 accident severity-behavior scenarios has an average human risk rating, providing an intuitive method for generation of unified risk. This allows the model to evolve from producing discrete outputs to providing a continuous estimate of expected risk.
3. It provides a comparative methodological contribution; it demonstrates how disparate modeling techniques function in different data types. Notably, it demonstrates how LSTMs excel with behavior time series data but are useless on static contextual accident data, while KD has a small potential to enhance performance without it being the most prominent source of predictive value.
4. It confirms that it is not the presence of fusion itself but the quality of that fusion that is significant. Comparing D+D, raw L+L, and L+L with LLM assistance shows that simple accuracy improvement alone is insufficient in safety-critical tasks; balanced class-wise reliability and intelligent refinement using the context are crucial. As such, the proposed final LLM-assisted L+L is the most important contribution of the paper: it outperformed all others in classification and regression measures.

5.6.5 Future Research Directions

Several paths exist to broaden this research. The biggest area of next development would be to leverage event-aligned, multi-modal data, where behavioral data, conditions, and actual crash outcomes can be tied to a particular driving event, enabling causal analysis and more robust fusion design.
Further work can be done on more sophisticated fusion mechanisms, such as attention-based integration, uncertainty-aware meta-learning, or dynamically weighing the behavioral and context branches, building upon the established system design. The current work lays a good foundation for tackling remaining issues in severe-risk discrimination.
Deployment in a real-time system, or on-vehicle or edge computing devices, is also a significant next step. This will necessitate further optimizing the fusion design for efficiency, defining a more formal way of using LLM's for the data cleaning and addressing questions of latency, robustness, and acceptable warning thresholds in a live setting. Work must also be done on the interpretability of the system, both the neural and LLM components. This framework can be expanded by adding other real-time parameters, such as driver state, fatigue, distraction, traffic density, and vehicle characteristics.
In addition, work needs to be done on the diversity of location for contextual data and the diversity of surveyed individuals to make the learned contextual relationships

and risk matrix elements generalizable to various regions and road cultures.

5.6.6 Concluding Discussion

In summary, we have shown that context and behavioral information can be modeled separately and merged effectively via gradually refined fusion approaches. We find that effective stand-alone models are needed but are not sufficient and that the structure of the data itself determines the most suitable models and the greatest benefit comes from sophisticated fusion and contextual refining. Our final L+L framework enhanced by LLM is the best balance of predictive performance, minority class reliability, continuous risk estimates, and interpretability.

Chapter 6

Conclusion

This chapter synthesizes the key findings of the research, articulates its contributions to the field of driver behaviour analysis and accident risk assessment, and presents detailed recommendations for future work that can build upon the established foundation.

6.1 Summary of Findings

Our goal in this paper was to devise and evaluate a context-aware driving risk assessment system that integrates environmental accident risk and driver behavior risk using two separate machine learning models and an advanced fusion approach. This was achieved by: 1) tackling data heterogeneity, 2) creating a strong predictive baseline, 3) utilizing computationally efficient custom neural architectures, and 4) implementing a transparent risk-mapping scheme.

In our US Accidents dataset (Dataset 1, >7.7M events), the custom neural network performed comparably to gradient boosting models (CatBoost, XGBoost) but with nearly half the memory consumption and $\sim 40x$ less training time, making it ideal for memory-constrained deployment. Key features included temporal (Start Hour, Is Night), weather (precipitation, temperature), and road characteristics. For Dataset 2 (Driving Behavior), the custom sensor-based network—particularly when using sequence-based LSTM modeling—provided nearly perfect classification accuracy, proving that smartphone accelerometer and gyroscope data can robustly yield highly discriminative behaviors.

The LLM-assisted Late-plus-Late (L+L) fusion approach is the main contribution of this work. Instead of simple concatenation, this method uses an LLM-assisted contextual refinement stage to prune noise from the context branch before fusion. The system integrates soft-max probability outputs from both base models with a probabilistically calculated Expected Risk score (Calculated Risk Level) derived from a survey-grounded 2×3 risk matrix. This matrix maps six scenarios (e.g., Low Severity/Slow to High Severity/Aggressive) to human-judged risk means. This formulation enables the system to encode degree-of-uncertainty; for instance, if a base model produces near-uniform probabilities, Calculated Risk Level drifts toward the center for a more honest representation of uncertainty.

The final system attained an overall accuracy of 0.91 and a macro F1-score of 0.88 across the six risk zone classes. One-vs-rest AUC-ROC scores ranged from 0.9751 to 0.9969 (macro average 0.9895). The regression head further validated the risk repre-

sentation with an R^2 of 0.9891 and an MAE of 0.0905. While the behavioral model (Model 2) achieved near-perfect accuracy, the primary source of uncertainty remains the contextual model (Model 1) due to the inherent complexity and imbalance of accident data.
Regarding scalability, the framework was engineered for high-throughput processing with a focus on resource optimization. The integrated pipeline managed a throughput exceeding 14 million records while maintaining a remarkably lean peak memory footprint of under 1.5GB. By achieving this level of efficiency, the system demonstrates its readiness for large-scale production environments and real-time deployment without requiring prohibitively expensive infrastructure.

6.2 Contributions to the Field

This study makes several distinct contributions to research on driver behaviour analysis and accident risk assessment, advancing both the theoretical underpinnings and the practical feasibility of intelligent transportation safety systems.

6.2.1 Methodological Contributions

We also propose a two-dataset fusion framework which can effectively merge two separate but complementary datasets related to different, but interrelated aspects of road safety. The US Accidents dataset, Dataset 1, describes accidents at a macro level. This dataset includes comprehensive information of the environmental and contextual factors behind accidents over 7 years in 49 states of the USA. The dataset contains 7.7 million records. Dataset 2 registers micro level motion measurements which are inertial sensor readings, both accelerometers and gyroscopes, collected on a consumer smartphone, covering three driving styles. The fusion of both datasets enabled simultaneous modeling of both external risk factors such as weather conditions, road type, traffic conditions, time of the day and behavioral conditions such as driving style and motion pattern. This addresses the deficiency in the existing literature, where most of the prior studies have focused on modeling of the accident patterns at a macro level or the driving behaviors at a micro level separately.
The most significant architectural contribution in the work is the Late-plus-Late (L+L) merging scheme itself. Unlike early fusion where the raw heterogeneous features are merged together before training the models, potentially suffering from mismatched sensor domains, the proposed approach trains each expert on their individual native domain and only fuses their high level probability outputs. The vector, made of the 2 softmax probability distributions and E[Risk] score, is a 6 dimensional vector representing both predictive confidence and probabilistic relationship between severity class and behaviour class. Further, a multi-task meta-model taking this vector as input, with a classification (risk zone) head and regression (E[Risk] score) head is unique in the field in that it uses the continuous risk uncertainty as an auxiliary signal, thereby generating better calibrated zone predictions.
The work performs high quality benchmarks over a wide range of model families, namely classical algorithms (Logistic Regression, SVM, KNN), ensembles (Random Forest, XGBoost, LightGBM, CatBoost, AdaBoost) and specialized Neural Networks on tabular and sensor domains separately. This helps establish a robust set of models suitable for driver safety applications, and establish that ensembles,

specifically gradient boosting algorithms, performed well both in terms of classifying accident severity and predicting driver behaviour.
New feature engineering techniques were developed and implemented in order to enhance model sensitivity to the risk from context: a storminess index created by combining rainfall, wind speed, air pressure and visibility to represent a single adverse weather characteristic; a low light indicator developed by combining the flags for sunrise, sunset, and twilight; and a control load metric constructed by counting the number of elements required to control traffic. These newly engineered features were all found to be predictive of behavior with statistical significance, enhance the sensitivity of the models via the inherent structural grouping of correlated variables, and offer insight to traffic safety practitioners.

6.2.2 Empirical Contributions

High-performance accident severity classification was successful, and macro F1-scores greater than 0.85 were achieved using Random Forest and gradient boosting models over large-scale accident data. Importance analysis of features clearly indicated temporal characteristics (StartHour, IsNight, IsWeekend) as well as weather and road distance (TemperatureF, Precipitationin, Distancemi) to be the dominant, naturally intuitive features for known predictors of accident risk. It was found that it was necessary to discard the very high-cardinality location features to retain scalability. These results would lend themselves well to real-time accident severity classification, thereby enabling optimisations of resource allocation and emergency response.
It was confirmed that driver behavior classification was robust, with macro F1-scores between 0.95 and 0.999, and that this classification could be reliably performed using only the six motion-sensor axes on a consumer smartphone. Because this sensor neural network is very lightweight, with approximately 8,000 parameters, it greatly democratises advanced driver assistance and its application could readily be widespread to any consumer smartphone without specialized hardware on the vehicle.
It is also worth noting that the L+L fusion pipeline itself provided an empirical validation of the synergy between the two modalities. The regression results for the meta-model, ($R = 0.9891$), proved that the probability fusion vector accurately encapsulated the continuous space of risk, while classification accuracy (weighted $F1 = 0.91$, macro $AUC = 0.9895$) demonstrated that there was separation in the joint risk space when using probability-level inputs.
Overall, extensive data pre-processing pipelines were used to deal with missing values by setting a threshold for anything over 30% missing, outliers through use of the IQR filter, class imbalance through stratified sampling, and features by use of standardisation through both standardscaler and minmaxscaler as relevant. This pipeline is reproducible and can therefore act as an accessible tool to future researchers of telematics and accidents, enabling greater transparency and reduced analytical bias.

6.2.3 Practical and Theory Contributions

In practice, the framework provides an estimate that it's possible to train a dual neural network pipeline across 14M+ total records on a free tier GPU in memory under approximately 288MB; this architecture can thus be deployed on resource constrained devices like smartphones and automotive embedded computers without requiring distributed compute. The sensor based neural network architecture is instantly deployable on real time systems with its 8K parameter estimate and 99.9% prediction accuracy for driver warning systems, UBI scoring and fleet management. The decision level fusion can instantly be used in an operational system; in contrast to two different data domains spitting out disparate predictions that a downstream system needs to manage, our fusion maps two heterogenous model prediction sets across two completely different domains into one decision making factor. The system also allows for maintenance and extensibility as because the two base networks are trained independently on their respective domains one can retrain or modify either expert model without having to retrain the other network, reducing the system upkeep cost as more data is collected.

From a theoretical perspective, the custom tabular neural network architecture – with per-feature input pathways, column level normalisation, and StringLookup categorical embedding – provides a structured way to handle heterogeneous tabular data in a neural network, providing gradient boosting-like performance (85%+ in severity classification) at a very small network size, and with a highly interpretable structure. The scheme for environment-behaviour risk mapping establishes a theoretical relationship between two disparate and previously uncorrelated dimensions of risk; it maps these to the decision maker in a formalized manner by specifying a 23 risk matrix and calculating the expected value as a probability weighted inner product. In this formulation, when both models agree on high risk (e.g. Severity 2 + Aggressive driving is highest in E[Risk]) the estimate is appropriately inflated, and when models are uncertain the estimate is appropriately dampened (cannot be achieved by single-modality or hard-label fusion).

Lastly, the auxiliary regression task in the meta-model is an original theoretical contribution for the design of multi-task learning in fusion architectures. By using a smooth risk signal as a regularizing auxiliary objective, we anchor the representations of the classifier to the underlying probabilistic risk function, avoiding the over-fitting on the abrupt decision boundaries of the hard labels. The signal propagation between domains (from macro/micro-level evidence without any domain dependence, to the calibrated common risk prediction) is a reproducible paradigm for hybrid risk predictors to be used in the future.

6.3 Recommendations for Future Work

While this study establishes a strong foundation for hybrid AI-based driver behaviour analysis and accident risk measurement, several meaningful directions remain open for future investigation and development.

6.3.1 Dataset Expansion and Cross-Cultural Validation

The current framework was trained and tested solely using US accident data and a fixed sensor dataset. Work in the future should focus on extending to multi-regional and multi-cultural test and evaluation contexts, specifically in developing world, low- and middle-income countries. In such countries, patterns of traffic, quality of the road infrastructure, and patterns of driver behavior all deviate significantly from the data used to train the current system. An international collaboration on collecting diverse training data from a wide range of environments may be necessary.
Adding a temporal behavioural signal obtained from video cameras in the cabin (e.g., through passenger and driver view cameras on dash-cams) could greatly expand the available behavioural signature—for example, modelling of drivers' attention, facial affect, head pose, hand posture, as well as the scene from a temporal-perceived driving perspective. This would be tractable using temporal convolutional or recurrent neural network architectures with strict data privacy constraints applied through on-device processing and differential privacy.
Having longitudinally-observed, ongoing driving behavior records for months or years would also be valuable, as the individual driver could be observed evolving over time in response to increasing skill, safety interventions, and other factors. Models would thus become more personalized and adaptive compared to current population-based classifiers. Complementarily, carefully documented logs of rare driving events and near-misses, activated by detection of anomalies, will provide useful signals about causative mechanisms of danger.

6.3.2 Edge Deployment and Real-Time systems

To be useful on hardware-limited platforms hardware-aware model compression, including quantization, weight pruning and knowledge distillation, is necessary. Techniques should efficiently minimize the model size and inference latency while maintaining accuracy as much as possible, to allow the model to be deployed on smartphones and automotive computing platforms. Streaming and incremental learning are another set of options to make the deployed models responsive to changing driving environments and new environments without the need for model retraining. One primary use case would be to leverage this infrastructure to enable federated learning, wherein model updates from a large fleet of vehicles would be aggregated while ensuring data privacy. Data heterogeneity due to varied drivers and spatial dispersion of data, along with protection of the data aggregation process itself, would pose challenges in such an approach.

6.3.3 Explainability and Interpretability

Once the system is ready to be deployed for safety critical applications, it is imperative to explain decisions of the models for the drivers, fleet operators and the regulators. The development of an attention-based visualisation tool would allow us to clearly and transparently present to a layman, the part that each input feature played in contributing towards the final risk score. By generating counterfactual explanations, e.g. "slowing down driving speed to 15km/hr under these conditions would reduce the risk zone from Sev2Cls2 to Sev1Cls2", actions for the drivers would

be directly available. Argumentation-based reasoning systems could also help expose and resolve cases where different base models have conflicting signals, such as the case where Model 1 predicted equal probabilities of Severity 1 and Severity 2, instead of silently averting such clashes via the fusion vector.

6.3.4 Application-Specific Extensions

A number of high value application domains are directly applicable to an adaptation of the current framework. Usage-based insurance (UBI) premiums are currently derived from a wide variety of sensor-based data. The continuous E[Risk] could be translated to actuarially justified premium modifications, so long as appropriate privacy and regulatory compliance is built into the scoring pipeline. The sensor-based model of driver behavior could be incorporated into commercial fleet management within a driver safety dashboard coupled with electronic logging device (ELD) information and fatigue monitoring systems. The context-based model of accident severity can be directly applicable to AV safety validation of mixed-traffic driving by assessing risk in scenarios involving human driven vehicles alongside autonomous vehicles, thus building an environment-aware risk layer on top of current AV perception and planning components.

6.3.5 Ethical and Social Considerations

Demographic and geographical bias auditing should be mandatory for any framework which allocates risk scores to individual drivers. We suggest employing adversarial debiasing for systematic performance differences between protected groups, and consider how significant privacy risks inherent to the acquisition of high-granularity driving data may be overcome through privacy enhancing technologies such as secure multi-party computation, homomorphic encryption and formal policy framework. It may also be beneficial to explore the potential behavioural interventions associated with the proposed framework. Randomised controlled trials should be conducted in order to assess whether particular feedback modalities will promote long term behavior change in any given context, and that interventions lead to tangible safety improvements rather than transient behavioral modification.

6.3.6 Validation and Standardisation

It is important to eventually bring this predictive performance to the real-world, in pilots with longitudinal study and establish that the gains in model performance correlate with reduction of actual accidents and near misses. Participating in benchmark datasets and competitions is necessary to compare to other approaches fairly and evaluate under uniform conditions. Regulations concerning AI-driven vehicle safety system certifications and liabilities are uneven across different states. International standardization processes, particularly those of post-market surveillance and accident reporting, would be needed to take such a system reliably to the mass market.

Overall, the present study has demonstrated that fusing environmental and behavioral risk dimensions with a principled probabilistic decision-level fusion architecture

is possible. The L+L framework shows promising classification and regression performance, a good degree of interpretability through its modular structure and can process millions of records without too much difficulty. Despite the limitations including geographical diversity of accident data and pre-defined matrix of risks, the proposed framework is a viable and expandable architecture for intelligent driver safety systems in future, aiming to reduce road traffic accidents.

Bibliography

[1] Dimitrios I. Tselentis and Eleonora Papadimitriou. "Driver Profile and Driving Pattern Recognition for Road Safety Assessment: Main Challenges and Future Directions". In: *IEEE Open Journal of Intelligent Transportation Systems* 4 (2023), pp. 83–100. DOI: 10.1109/OJITS.2023.3237177.

[2] Kanagamalliga S et al. "Enhancing Heavy Vehicle Safety with Intelligent Driver Behaviour Monitoring via Multi-Sensor Fusion and Embedded Systems". In: *2024 International Conference on System, Computation, Automation and Networking (ICSCAN)*. 2024, pp. 1–5. DOI: 10.1109/ICSCAN62807.2024.10894289.

[3] Fatma Mohammed Al Ali et al. "Abnormal Driver Behavior Detection Using Deep Learning". In: *2024 7th International Conference on Signal Processing and Information Security (ICSPIS)*. 2024, pp. 1–4. DOI: 10.1109/ICSPIS63676.2024.10812606.

[4] Sakthivel K et al. "Hybrid Deep Learning for Proactive Driver Risk Prediction and Safety Enhancement". In: *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*. 2025, pp. 1572–1577. DOI: 10.1109/ICMSCI62561.2025.10894603.

[5] Xiaoqing Fang. "Identification and Assessment of Electric Bus Driver's Driving Behavior Based on Computer Vision and Intelligent Algorithms". In: *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*. 2024, pp. 896–900. DOI: 10.1109/CICN63059.2024.10847342.

[6] Li Liu, Guo-qing Xu, and Zhangjun Song. "Driver lane changing behavior analysis based on parallel Bayesian networks". In: *2010 Sixth International Conference on Natural Computation*. Vol. 3. 2010, pp. 1232–1237. DOI: 10.1109/ICNC.2010.5583634.

[7] Sinan Kaplan et al. "Driver Behavior Analysis for Safe Driving: A Survey". In: *IEEE Transactions on Intelligent Transportation Systems* 16.6 (2015), pp. 3017–3032. DOI: 10.1109/TITS.2015.2462084.

[8] Arya R Anil and Anudev J. "Driver behavior analysis using K-means algorithm". In: *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*. 2022, pp. 1555–1559. DOI: 10.1109/ICICICT54557.2022.9917899.

[9] Bhumika, Debasis Das, and Sajal K. Das. "RsSafe: Personalized Driver Behavior Prediction for Safe Driving". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892982.

[10] Amira A. Amer and Dina Elreedy. "Aggressive Driver Behavior Detection Using Multi-Label Classification". In: *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. 2024, pp. 1–7. DOI: 10.1109/IMCOM60618.2024.10418298.

[11] Fangming Qu et al. "Comprehensive study of driver behavior monitoring systems using computer vision and Machine Learning Techniques". In: *Journal of Big Data* 11.1 (Feb. 2024). DOI: 10.1186/s40537-024-00890-0.

[12] Barry Sheehan Leandro Masello German Castignani and Finbarr Murphy Montserrat Guillen. "Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence". In: *Accident Analysis & Prevention* (Feb. 2023). DOI: 10.1016/j.aap.2023.106997.

[13] Zhaowen Pang et al. "Risk Assessment Method for Autonomous Vehicles Violating Safety Common Sense Based on Driving Behavior". In: *IEEE Access* 13 (2025), pp. 63076–63092. DOI: 10.1109/ACCESS.2025.3552833.

[14] Erwin de Gelder et al. "Risk Quantification for Automated Driving Systems in Real-World Driving Scenarios". In: *IEEE Access* 9 (2021), pp. 168953–168970. DOI: 10.1109/ACCESS.2021.3136585.

[15] Wei-Hsun Lee and Che-Yu Chang. "Assessing Driving Risk Level: Harnessing Deep Learning Hybrid Model With Intercity Bus Naturalistic Driving Data". In: *IEEE Access* 13 (2025), pp. 25141–25153. DOI: 10.1109/ACCESS.2025.3538693.

[16] K. Raveendra Reddy and A. Muralidhar. "Machine Learning-Based Road Safety Prediction Strategies for Internet of Vehicles (IoV) Enabled Vehicles: A Systematic Literature Review". In: *IEEE Access* 11 (2023), pp. 112108–112122. DOI: 10.1109/ACCESS.2023.3315852.

[17] Ke Wang and Jian John Lu. "A Two-Layer Risky Driver Recognition Model With Context Awareness". In: *IEEE Access* 9 (2021), pp. 138483–138495. DOI: 10.1109/ACCESS.2021.3116996.

[18] Masaya Shigehara Yoshiyasu Takefuji Michiyasu Tano and Shunya Sato. Oct. 2024. URL: https://www.sciencedirect.com/science/article/abs/pii/S0360835224007897.

[19] Fangming Qu et al. "Comprehensive study of driver behavior monitoring systems using computer vision and Machine Learning Techniques". In: *Journal of Big Data* 11.1 (Feb. 2024). DOI: 10.1186/s40537-024-00890-0.

[20] Dimitrios Tselentis and Eleonora Papadimitriou. URL: https://www.researchgate.net/publication/367229996_Driver_Profile_and_Driving_Pattern_Recognition_for_Road_Safety_Assessment_Main_Challenges_and_Future_Directions.

[21] Ward Ahmed Al-Hussein et al. *Driver behavior profiling and recognition using deep-learning methods: In accordance with traffic regulations and experts guidelines*. Jan. 2022. URL: https://www.mdpi.com/1660-4601/19/3/1470?utm_source.

[22] Junjian Hou et al. *Research progress of dangerous driving behavior recognition methods based on Deep Learning*. Jan. 2025. URL: https://www.mdpi.com/2032-6653/16/2/62?utm_source.

[23] Ausaf Khan. *Using AI to study impact of driving patterns*. Jan. 2023. URL: https://jyx.jyu.fi/jyx/Record/jyx_123456789_92172.

[24] Dinesh Kalla. *AI-powered driver behavior analysis and Accident Prevention Systems for advanced driver assistance*. Feb. 2025. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5063253.

[25] Kang-Hyun Jo Duy-Linh Nguyen Muhamad Dwisnanto Putro. *Driver behavior detection and classification using deep convolutional Neural Networks*. Jan. 2020. URL: https://www.sciencedirect.com/science/article/abs/pii/S095741742030066X.

[26] Yang Xing et al. "Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach". In: *IEEE Transactions on Vehicular Technology* 68.6 (2019), pp. 5379–5390. DOI: 10.1109/TVT.2019.2908425.

[27] Samantha Jamson Oliver Carsten Katja Kircher. *Vehicle-based studies of driving in the real world: The hard truth?* June 2013. URL: https://www.sciencedirect.com/science/article/abs/pii/S0001457513002340?via%3Dihub.

[28] A. Hazoor M. Bassani L. Catani, A. Portera A. Hoxha A. Lioi, and L. Tefa. *Do driver monitoring technologies improve the driving behaviour of distracted drivers? A simulation study to assess the impact of an auditory driver distraction warning device on driving performance*. May 2023. URL: https://www.sciencedirect.com/science/article/pii/S1369847823000906.

[29] Jaeik Jo et al. *Vision-based method for detecting driver drowsiness and distraction in Driver Monitoring System*. Dec. 2011. URL: https://www.spiedigitallibrary.org/journals/optical-engineering/volume-50/issue-12/127202/Vision-based-method-for-detecting-driver-drowsiness-and-distraction-in/10.1117/1.3657506.full?tab=ArticleLinkFigureTable.

[30] H. Saito, T. Ishiwaka, and S. Okabayashi. "Applications of driver's line of sight to automobiles-what can driver's eye tell". In: *Proceedings of VNIS'94 - 1994 Vehicle Navigation and Information Systems Conference*. 1994, pp. 21–26. DOI: 10.1109/VNIS.1994.396872.

[31] J. Jia et al. "A Countrywide Traffic Accident Dataset". In: *arXiv preprint* (2019). arXiv: 1906.05409 [cs.DB]. URL: https://arxiv.org/abs/1906.05409.

[32] Z. Guo, H. Wang, and J. Liu. "Accident Risk Prediction based on Heterogeneous Sparse Data". In: *arXiv preprint* (2019). arXiv: 1909.09638. URL: https://arxiv.org/abs/1909.09638.

[33] Y. Liu, J. Liu, and S. Zhang. "Cause Analysis and Accident Classification of Road Traffic Accidents Based on Complex Networks". In: *Applied Sciences* 13.23 (2023), p. 12963. DOI: 10.3390/app132312963. URL: https://www.mdpi.com/2076-3417/13/23/12963.

[34] W. Qiu, C. Wang, and Y. Tang. *Prediction in Traffic Accident Duration Based on Heterogeneous Ensemble Learning*. ResearchGate. 2022. URL: https://www.researchgate.net/publication/357686145_Prediction_in_Traffic_Accident_Duration_Based_on_Heterogeneous_Ensemble_Learning.

[35] Z. Xu, J. Fang, and P. Chen. "Prediction of Accident Risk Levels in Traffic Accidents Using Deep Learning and Radial Basis Function Neural Networks Applied to a Dataset with Information on Driving Events". In: *Applied Sciences* 14.14 (2024), p. 6248. DOI: 10.3390/app14146248. URL: https://www.mdpi.com/2076-3417/14/14/6248.

[36] Y. Zhou and Y. Li. "Recent Advances in Traffic Accident Analysis and Prediction". In: *arXiv preprint* (2024). arXiv: 2406.13968. URL: https://arxiv.org/abs/2406.13968.

[37] J. Gao and Z. Wu. "Research on Traffic Accident Severity Level Prediction Model Based on Improved Machine Learning". In: *Systems* 13.1 (2025), p. 31. DOI: 10.3390/systems13010031. URL: https://www.mdpi.com/2079-8954/13/1/31.

[38] W. Song, D. Wang, and L. He. "RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model". In: *IEEE Transactions on Intelligent Transportation Systems* (2021). DOI: 10.1109/TITS.2021.3107460. URL: https://ieeexplore.ieee.org/document/9536744.

[39] H. Zhang, J. Wang, and M. Li. "Traffic Accident Severity Prediction Based on Random Forest". In: *Sustainability* 14.3 (2022), p. 1729. DOI: 10.3390/su14031729. URL: https://www.mdpi.com/2071-1050/14/3/1729.

[40] Z. Li, Y. Feng, and Y. Zhang. "Uncertainty-aware probabilistic graph neural networks for road-level traffic crash prediction". In: *Accident Analysis & Prevention* 205 (2024). DOI: 10.1016/j.aap.2024.106997. URL: https://www.sciencedirect.com/science/article/pii/S0001457524003464.

[41] J. Chen, S. Wu, and X. Zhao. "Automatic Accident Detection, Segmentation and Duration Prediction Using Machine Learning". In: *IEEE Access* (2023). DOI: 10.1109/ACCESS.2023.3324589. URL: https://ieeexplore.ieee.org/abstract/document/10287856.

[42] J. M. Mase and A. T. Ngoma. *A Review of Intelligent Systems for Driving Risk Assessment*. ResearchGate. 2023. URL: https://www.researchgate.net/profile/Jimiama-Mafeni-Mase/publication/374102187_A_Review_of_Intelligent_Systems_for_Driving_Risk_Assessment/links/6511c811cce2460b6c356044/A-Review-of-Intelligent-Systems-for-Driving-Risk-Assessment.pdf.

[43] H. Khan, M. Imran, and F. Ali. "V-SenseDrive: A Privacy-Preserving Road Video and In-Vehicle Sensor Fusion Framework for Road Safety & Driver Behaviour Modelling". In: *arXiv preprint* (2025). arXiv: 2509.18187. URL: https://arxiv.org/pdf/2509.18187.

[44] P. Zhang, L. Huang, and Y. Wu. "Preventing Road Accidents Through Early Detection of Driver Behavior Using Smartphone Motion Sensor Data: An Ensemble Feature Engineering Approach". In: *IEEE Transactions on Intelligent Transportation Systems* (2023). DOI: 10.1109/TITS.2023.3330456. URL: https://ieeexplore.ieee.org/abstract/document/10347206.

[45] R. Sharma, K. Singh, and V. Verma. "Elevating Driver Behavior Understanding With RKnD: A Novel Probabilistic Feature Engineering Approach". In: *IEEE Transactions on Intelligent Transportation Systems* (2024). DOI: 10.1109/TITS.2024.3389054. URL: https://ieeexplore.ieee.org/abstract/document/10526235.

[46] B. Chah, A. Hossain, and Y. Zhang. *Building a Database of Simulated Driver Behaviors Using the SUMO Simulator*. ResearchGate. 2024. URL: https://www.researchgate.net/profile/Badreddine-Chah-2/publication/382719447_Building_a_Database_of_Simulated_Driver_Behaviors_Using_the_SUMO_Simulator/links/66dcb743fa5e11512ca6e57c/Building-a-Database-of-Simulated-Driver-Behaviors-Using-the-SUMO-Simulator.pdf.

[47] B. Fazzinga et al. "ASYDE: An Argumentation-based System for classifYing Driving bEhaviors". In: *Proceedings of the 32nd Symposium on Advanced Database Systems (SEBD 2024)*. Villasimius, Italy, 2024. URL: https://ceur-ws.org/Vol-3741/paper40.pdf.

[48] J. Maktoubian et al. "Empirical Evaluation of Machine Learning Models for Fuel Consumption, Driver Identification, and Behavior Prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 25.12 (2024), pp. 19156–19170. DOI: 10.1109/TITS.2024.3474745. URL: https://ieeexplore.ieee.org/abstract/document/10720589.

[49] D. A. Vyas and M. Chaturvedi. "DriverSense: A Multi-Modal Framework for Advanced Driver Assistance System". In: *Proceedings of COMSNETS 2024: Intelligent Transportation Systems Workshop*. Bangalore, India, 2024. DOI: 10.1109/COMSNETS59351.2024.10427459. URL: https://ieeexplore.ieee.org/abstract/document/10427459.

[50] A. A. Pande, P. S. Wawage, and Y. D. Deshpande. "Driver Behaviour Classification using Neural Networks". In: *Proceedings of the 4th Asian Conference on Innovation in Technology (ASIANCON)*. Pune, India, Aug. 2024. DOI: 10.1109/ASIANCON62057.2024.10838053. URL: https://ieeexplore.ieee.org/abstract/document/10838053.

[51] Munaf Salim Najim Al-Din. *Driving maneuvers recognition and classification using a hyprid pattern matching and machine learning*. Dec. 2023. URL: http://dx.doi.org/10.14569/IJACSA.2023.0140230.