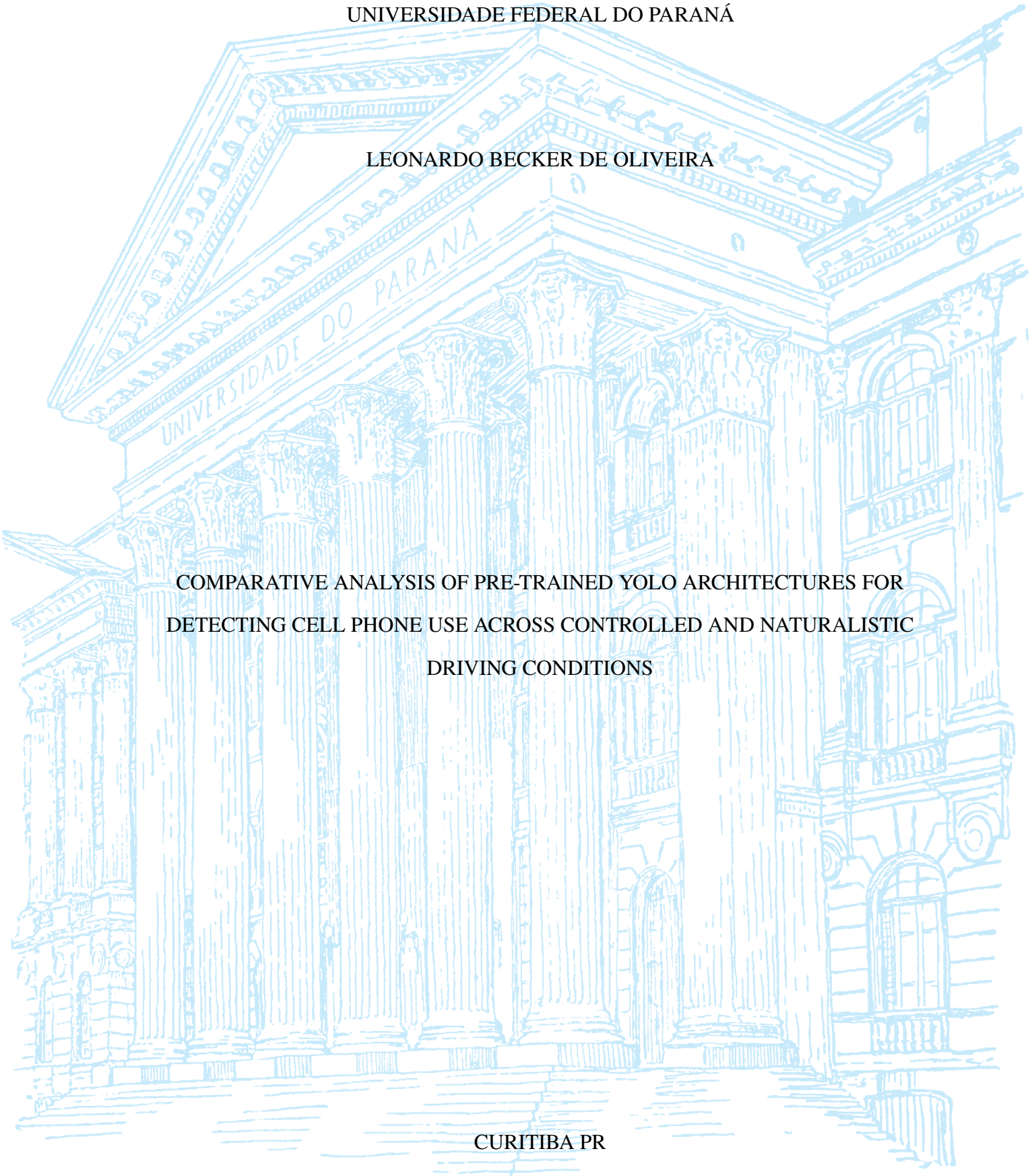
UNIVERSIDADE FEDERAL DO PARANÁ

LEONARDO BECKER DE OLIVEIRA



COMPARATIVE ANALYSIS OF PRE-TRAINED YOLO ARCHITECTURES FOR DETECTING CELL PHONE USE ACROSS CONTROLLED AND NATURALISTIC DRIVING CONDITIONS

CURITIBA PR

2025

LEONARDO BECKER DE OLIVEIRA

COMPARATIVE ANALYSIS OF PRE-TRAINED YOLO ARCHITECTURES FOR DETECTING CELL PHONE USE ACROSS CONTROLLED AND NATURALISTIC DRIVING CONDITIONS

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: Computação.

Orientador: David Menotti Gomes.

CURITIBA PR

2025

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho: ***COMPARATIVE ANALYSIS OF PRE-TRAINED YOLO ARCHITECTURES FOR DETECTING CELL PHONE USE ACROSS CONTROLLED AND NATURALISTIC DRIVING CONDITIONS***

Autor(es):

GRR 20211779 None: LEONARDO BECKER DE OLIVEIRA

GRR _____ Nome: _____

GRR _____ Nome: _____

Apresentação: Data: 26 / 11 / 2025 Hora: 18:00 Local: Lab4 – Departamento de Informática

Orientador:	DAVID MENOTTI GOMES	
Membro 1:	LUCAS MATHEUS LEITE WOJCIK	
Membro 2:	GABRIEL EDUARDO LIMA	
	(nome)	(assinatura)

AVALIAÇÃO – Produto escrito		ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo	(00-40)				25
Referência Bibliográfica	(00-10)				10
Formato	(00-05)				04
AVALIAÇÃO – Apresentação Oral					
Domínio do Assunto	(00-15)				15
Desenvolvimento do Assunto	(00-05)				05
Técnica de Apresentação	(00-03)				03
Uso do Tempo	(00-02)				02
AVALIAÇÃO – Desenvolvimento					
Nota do Orientador	(00-20)		**********	**********	16
NOTA FINAL		**********	**********	**********	**80**

Pesos indicados são sugestões.
Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramita o processo para CCOMP (Coordenação Ciência da Computação).

ACKNOWLEDGEMENTS

Agradeço primeiramente à minha mãe Maria de Lourdes Pivovar, por seu amor incondicional, dedicação, e apoio contínuo em cada etapa desta jornada. À minha irmã Luiza Becker de Oliveira, pelo incentivo e presença constante, que sempre trouxeram força e equilíbrio aos momentos desafiadores.

Aos meus amigos, que contribuíram com palavras de motivação, compreensão e companheirismo, tornando o processo mais leve e significativo.

Por fim, deixo meu sincero agradecimento ao meu professor orientador David Menotti, cuja orientação, paciência e conhecimento foram fundamentais para o desenvolvimento deste trabalho e para meu amadurecimento acadêmico e profissional.

RESUMO

Esta monografia apresenta uma avaliação comparativa abrangente das arquiteturas YOLO mais recentes (versões 8, 10, 11 e 12) para a detecção de uso de telefone celular durante a condução, em cenários controlados e naturalísticos. Motivado pelo crescente impacto da distração do motorista nos acidentes de trânsito e pelas limitações operacionais da fiscalização manual, este estudo investiga como os avanços arquiteturais da família YOLO influenciam o desempenho em sistemas de monitoramento veicular em tempo real. A análise é conduzida utilizando dois conjuntos de dados complementares: o State Farm Distracted Driver Detection (SFDDD), que oferece imagens controladas e de alta qualidade, e o Naturalistic Driving Study - Brazil (NDS-BR), o primeiro estudo naturalístico de direção em larga escala no Brasil, caracterizado por longas sequências de vídeo, em diferentes ambientes e comportamentos realistas. Os experimentos foram divididos em duas etapas. Na primeira, através das imagens estáticas foi feita uma comparação independente do dataset entre todas as versões e granularidades dos modelos YOLO. Na segunda, é aplicada inferência quadro a quadro nos vídeos brutos do NDS-BR, seguida de uma agregação temporal, a fim de avaliar os benefícios e limitações da redundância temporal. Todos os modelos são avaliados com precisão, revocação, especificidade, acurácia e tempo de inferência, com ênfase especial no equilíbrio entre ganhos de revocação e o aumento de falsos positivos em cenários naturalísticos. Os resultados mostram que o YOLOv12 atinge o melhor desempenho geral, especialmente em ambientes naturalísticos, beneficiando-se dos mecanismos de atenção. O YOLOv8 permanece altamente competitivo em cenários controlados, enquanto YOLOv10 e YOLOv11 apresentam desempenho intermediário com boas relações entre eficiência e acurácia. A agregação temporal melhora consistentemente a revocação para eventos breves ou intermitentes de uso de celular, mas também amplifica falsos positivos, destacando implicações éticas e operacionais para implantação real. As conclusões oferecem diretrizes práticas para seleção e uso de arquiteturas YOLO em sistemas de monitoramento do motorista, ao mesmo tempo em que reforçam a importância de avaliar frameworks modernos de detecção em datasets naturalísticos e comportamentais como o NDS-BR.

Palavras-chave: Detecção de Distração do Motorista. Detecção de Uso de Celular. Arquiteturas YOLO.

ABSTRACT

This dissertation presents a comprehensive comparative evaluation of recent YOLO architectures—versions 8, 10, 11, and 12—for detecting cell phone use while driving under controlled and naturalistic conditions. Motivated by the growing role of driver distraction in traffic accidents and the operational limitations of manual enforcement, this study investigates how architectural advances of YOLO family influence performance in real-time driver monitoring systems. The analysis is conducted using two complementary datasets: the State Farm Distracted Driver Detection (SFDDD) dataset, which offers high-quality, controlled images, and the Naturalistic Driving Study - Brazil (NDS-BR) dataset, Brazil's first large-scale naturalistic driving study, characterized by long video sequences, environmental variability, and realistic behavior patterns. A two-stage experimental design is implemented. First, static image experiments provide a dataset-agnostic comparison across all YOLO versions and model granularities. Second, frame-wise inference is applied to raw NDS-BR videos, followed by temporal aggregation to assess the benefits and drawbacks of temporal redundancy. All models are evaluated using precision, recall, specificity, accuracy, and inference time, with special attention to the trade-off between increased recall and the rise in false positives under naturalistic noise. Results show that YOLOv12 achieves the best overall performance, particularly in naturalistic settings, benefiting from attention mechanisms. YOLOv8 remains competitive in controlled environments, while YOLOv10 and YOLOv11 exhibit intermediate performance with improved efficiency–accuracy trade-offs. Temporal aggregation consistently increases recall for short or intermittent phone-use events but also amplifies false positives, highlighting ethical and operational considerations for real-world deployment. The findings provide empirical guidance on selecting appropriate YOLO architectures and model sizes for practical driver-monitoring systems and underscore the importance of evaluating modern detection frameworks within naturalistic, behaviorally rich datasets such as NDS-BR.

Keywords: Driver Distraction Detection. Cell Phone Use Detection. YOLO Architectures.

LIST OF FIGURES

1.1 Controlled Environment (SFDDD): Consistent lighting, fixed camera angle, and staged actors posing specific distractions. 15

1.2 Naturalistic Environment (NDS-BR): Highly variable lighting (day/night), occlusion, dynamic backgrounds, and spontaneous driver behavior.. 16

2.1 Visual representation of general architecture of the YOLO family. Source: Kateb et al. (2021).. 19

2.2 Comparison between the C3 and C2f modules. Source: Liu et al. (2024). 20

2.3 Consistent dual assignments for NMS-free training. Source: Wang et al. (2024).. 21

2.4 Architectural diagrams of the modules used in YOLOv11 (a) C2PSA and (b) C3K2. Source: Meng et al. (2025).. 22

2.5 Illustrative components referenced across YOLO architectures. Source: Tian et al. (2025).. 23

4.1 Naturalistic Environment (NDS-BR): Realism and visual complexity centered around dynamic driving scenarios. 32

4.2 State Farm Distracted Driver Detection classes. Source: Montoya et al. (2016). . 33

4.3 Confusion matrix. This matrix allows for the calculation of key model metrics such as recall, specificity, precision, and accuracy. Source: (Cabot and Ross, 2023). 35

5.1 Visual representation of the Accuracy by granularity for different YOLO models and datasets. (a) SFDDD. (b) NDS-BR. 37

5.2 Visual representation of the confusion matrix for different YOLOv8 model sizes. (a) YOLOv8 nano. (b) YOLOv8 extra-large. 38

5.3 Relationship between accuracy and mean inference time for different YOLOv8 model sizes. 39

5.4 Visual representation of the confusion matrix for different YOLOv12 model sizes. (a) YOLOv12 nano. (b) YOLOv12 extra-large. 39

5.5 Relationship between accuracy and mean inference time for different YOLOv12 model sizes. 40

5.6 Visual representation of the confusion matrix for the Medium granularity. (a) YOLOv8 Medium. (b) YOLOv12 Medium. 41

5.7 Example of an image labeled as safe driving in the NDS-BR dataset. 43

5.8 Visual representation of the confusion matrix for different granularities of YOLOv8 on NDS-BR.. 44

5.9 Relationship between accuracy and mean inference time for different YOLOv8 model sizes. 45

5.10 Visual representation of the confusion matrix for different granularities of YOLOv12 on NDS-BR. 45
5.11 Relationship between accuracy and mean inference time for different YOLOv12 model sizes. 46
5.12 Confusion Matrix comparing YOLOv8 and YOLOv12 in the frame-by-frame analysis. 48
5.13 Accuracy comparing YOLOv8 and YOLOv12 in the threshold aggregation analysis. 49
5.14 Confusion Matrix comparing YOLOv8 and YOLOv12 in the aggregation analysis (threshold=1). 50

LIST OF TABLES

2.1 Comparison of innovations and mAP on the COCO val2017 dataset (Lin et al., 2015). mAP values compiled from the official documentation and publications of each model. Source: Jocher (2024). 24

3.1 Summary of related work on detecting distraction while driving. 29

4.1 Distribution size of Samples (Images/Seconds) Across Experimental Datasets. . . 34

5.1 Comparison of Metrics by Model Granularity (v8 vs. v12) considering Nano, Medium, and Extra-Large in the SFDDD dataset. 42

5.2 Comparison of Metrics by Model Granularity (v8 vs. v12) considering Nano, Medium, Large, and Extra-Large in the NDS-BR dataset. 48

5.3 Summary of the best-performing models across all experimental conditions. . . . 51

LIST OF ACRONYMS

DInf	Departamento de Informática
PPGInf	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná
NDS-BR	Naturalistic Driving Study - Brazil
SFDDD	State Farm Driver Distraction Detection
WHO	World Health Organization
NHTSA	National Highway Traffic Safety Administration
ADAS	Advanced Driver Assistance Systems
ABS	Anti-lock Braking System
CDC	Centers for Disease Control and Prevention
CNN	Convolutional Neural Network
YOLO	You Only Look Once
RCNN	Region-based Convolutional Neural Network
FPN	Feature Pyramid Network
PANet	Path Aggregation Network
C2f	Contextualized Feature Fusion
NMS	Non-Maximum Suppression
A2	Area Attention
R-ELAN	Residual Efficient Layer Aggregation Networks
CSP	Cross Stage Partial
SPPF	Spatial Pyramid Pooling - Fast
mAP	mean Average Precision
FPS	Frames Per Second
ITS	Intelligent Transportation Systems

LIST OF SYMBOLS

T	Threshold

CONTENTS

1 INTRODUCTION 13
1.1 MOTIVATION 13
1.2 OBJECTIVES 14
1.3 CONTRIBUTION 16
1.4 DOCUMENT ORGANIZATION 17
2 BACKGROUND 18
2.1 THE YOLO ARCHITECTURE: PARADIGM AND FUNDAMENTAL COMPONENTS 18
2.1.1 YOLOv8 19
2.1.2 YOLOv10 20
2.1.3 YOLOv11 21
2.1.4 YOLOv12 22
2.1.5 Size Variants of the YOLO Family 23
2.1.6 Comparative Table of Innovations and Performance 23
2.1.7 General Comparison Across Versions 24
2.2 BENCHMARK EVALUATION METRICS 24
2.3 CONCLUDING BACKGROUND 25
3 RELATED WORKS 26
3.1 CNN-BASED APPROACHES 26
3.2 SVM-BASED APPROACHES 28
3.3 TRANSFORMER-BASED APPROACHES 28
3.4 MULTIMODAL APPROACHES WITH CLIP 28
3.5 FINAL CONSIDERATIONS 29
4 MATERIALS AND METHODS 31
4.1 DATASETS EMPLOYED 31
4.1.1 NDS-BR Dataset 31
4.1.2 SFDDD Dataset 32
4.1.3 General Comparison Between the Datasets 33
4.2 EXPERIMENTAL STRUCTURE 33
4.3 PROCESSING PIPELINE DESCRIPTION 34
4.4 TEMPORAL AGGREGATION AND DETECTION HANDLING 35
4.5 COMPUTATIONAL ENVIRONMENT AND EXPERIMENT EXECUTION . . 35
4.6 EVALUATION METRICS AND RESULTS ANALYSIS 35

5 RESULTS 37
5.1 GENERAL COMPARISON 37
5.2 SFDDD ANALYSIS 38
5.2.1 YOLOv8 38
5.2.2 YOLOv12 39
5.2.3 Analysis of the Medium Granularity 40
5.2.4 Comparative Considerations 41
5.3 NDS-BR ANALYSIS 42
5.3.1 YOLOv8 43
5.3.2 YOLOv12 45
5.3.3 General Architectural Behavior 46
5.3.4 Comparative Considerations 47
5.4 TEMPORAL AGGREGATION ANALYSIS ON NDS-BR 48
5.4.1 Baseline Frame-Level Analysis 48
5.4.2 Threshold Recall and Optimization 49
5.4.3 Aggregation Performance with Threshold $T = 1$ 50
5.5 FINAL COMPARATIVE SYNTHESIS 51
6 CONCLUSION 52
REFERENCES 53

1 INTRODUCTION

Road safety remains one of the greatest global public health challenges, with direct impacts on mortality and urban quality of life. Despite significant advancements in vehicle engineering, assistive technologies, and road infrastructure, international statistics continue to reveal an alarming scenario. According to the World Health Organization (WHO), traffic crashes are among the leading causes of death worldwide (WHO, 2023). In the United States alone, the National Highway Traffic Safety Administration (NHTSA) has reported an average of 40,000 fatalities per year over the past two years, exceeding the average for the previous decade (NHTSA, 2025).

This stagnation in the reduction of fatalities, despite the widespread adoption of advanced driver assistance systems (ADAS), such as Anti-lock Braking Systems (ABS), lane-keeping assistance, and traction control, reveals a structural paradox: technological progress in vehicles has been systematically neutralized by human behavior. Indeed, human error remains the predominant factor in traffic accidents. A widely cited study (Dingus et al., 2016) attributes 90% of traffic crashes involving injuries or property damage to human failures such as distraction, inattention, fatigue, or poor decision-making.

Among these risk factors, driver distraction has become one of the most prevalent and dangerous behaviors. According to the Centers for Disease Control and Prevention (CDC), driver distraction is defined as any activity that diverts attention from the primary task of safe driving (CDC, 2025). This phenomenon encompasses three interrelated dimensions: visual (eyes off the road), manual (hands off the wheel), and cognitive (mind off the task). Mobile phone use is particularly critical as it activates all three dimensions simultaneously. In this context, detecting the physical presence of a mobile phone serves as a crucial proxy for identifying cognitive distraction and potential visual impairment. The WHO estimates that drivers who use mobile phones while driving are approximately four times more likely to be involved in a crash (WHO, 2023). Furthermore, a single textual interaction can distract the driver for up to five seconds, enough time to travel the length of a football field at 90 km/h without visual attention to the road (GTSC, 2025).

In Brazil, the Federal Highway Police (PRF, 2025) reported 73,156 crashes on federal highways in 2024 alone, resulting in 6,160 deaths and over 84,000 injuries. Contributing factors such as “driver mobile phone use,” “lack of reaction,” and “delayed reaction” consistently rank among the most frequent presumed causes of these accidents. Brazilian states such as Minas Gerais, Paraná, and Santa Catarina are particularly affected, illustrating the national scale of the problem.

1.1 MOTIVATION

Despite legislation that classifies mobile phone use while driving as a serious offense, traditional enforcement remains limited in effectiveness. Visual identification of infractions by human agents is operationally challenging, subjective, and difficult to scale. Consequently, automated video-based driver monitoring systems have emerged as an objective solution for detecting risky behaviors. Among these, computer vision techniques have demonstrated significant promise.

Current literature heavily explores deep convolutional neural networks (CNNs) trained on specific datasets to classify distraction, achieving high accuracy with architectures like VGG16 or ResNet50 (Hossain et al., 2022). Nonetheless, real-time detection requires architectures that

balance accuracy with speed. In this context, the You Only Look Once (YOLO) family (Redmon et al., 2016) has established itself as a leading architecture for real-time object detection. Its ability to balance high detection accuracy with low computational latency makes it particularly suitable for embedded driver monitoring systems, where rapid response is critical.

However, a significant gap exists in the literature regarding the generalization capability of these models. Most studies validate solutions on "controlled" datasets—where lighting, camera angles, and driver poses are staged—leaving an open question about how state-of-the-art models perform in "naturalistic" conditions, which are characterized by chaotic lighting, occlusion, and spontaneous behavior.

This work specifically addresses the evolution of YOLO architectures (v8, v10, v11, and v12). These versions were selected due to their recent architectural advancements, such as anchor-free detection (Jocher, 2025) and attention-based mechanisms (Tian et al., 2025). Methodologically, this study adopts a zero-shot evaluation strategy. Instead of fine-tuning the models on small, specific driver datasets, we assess the performance of pre-trained models (on the generic COCO dataset) to detect mobile phones in driver monitoring scenarios. This approach allows us to measure the inherent robustness of the architectures and their applicability to the safety domain without the need for extensive retraining.

1.2 OBJECTIVES

The central research question driving this study is: How do state-of-the-art computer vision models (YOLO family) perform when transitioning from controlled experimental environments to the noisy, unpredictable reality of naturalistic driving without task-specific fine-tuning? To answer this, the general objective of this work is to validate and benchmark the behavior of YOLO architectures (v8 through v12) in detecting mobile phone usage, explicitly quantifying the performance gap between controlled datasets and naturalistic environments.

Specifically, this research evaluates the architectural evolution of the YOLO family, comparing versions 8, 10, 11, and 12 to understand how recent innovations, such as attention mechanisms and NMS-free training, impact the detection of small objects like mobile phones. Concurrently, the study investigates the trade-offs between accuracy and inference speed across different model granularities (Nano to Extra-Large) to determine the viability of lightweight models for embedded monitoring.

A central focus of this work is to quantify the "Reality Gap" by assessing performance degradation when models transition between distinct data domains. This comparison is visually illustrated in figure bellow, which contrasts the two scenarios evaluated. As shown in Figure 1.1, the State Farm Distracted Driver Detection (SFDDD) dataset represents a controlled environment with consistent lighting and staged actors (Montoya et al., 2016). In sharp contrast, Figure 1.2 depicts the Naturalistic Driving Study - Brazil (NDS-BR) dataset, which captures the complexity of the naturalistic environment, characterized by variable lighting, motion blur, and spontaneous behavior (Bastos et al., 2020). Finally, the study assesses operational feasibility by analyzing the trade-off between recall and false positives, evaluating the impact of temporal aggregation strategies to mitigate noise in continuous video monitoring.

Figure 1.1: Controlled Environment (SFDDD): Consistent lighting, fixed camera angle, and staged actors posing specific distractions.

Figure 1.2: Naturalistic Environment (NDS-BR): Highly variable lighting (day/night), occlusion, dynamic backgrounds, and spontaneous driver behavior.

1.3 CONTRIBUTION

This work provides a systematic and empirically grounded investigation into the evolution of the most recent YOLO architectures—versions 8, 10, 11, and 12—and their effectiveness in detecting distracted driving behaviors, with particular emphasis on mobile phone detection. Unlike prior studies that either rely exclusively on controlled datasets or evaluate isolated models without cross-version comparison, this research advances the field by offering the first comprehensive assessment of these architectures across both controlled and naturalistic scenarios. By jointly analyzing the SFDDD dataset and NDS-BR dataset, the study examines how architectural innovations, dataset realism, and model granularity interact to shape detection performance.

A further contribution lies in the dual experimental design, which evaluates the models under two complementary paradigms. The first treats images as independent samples, enabling a controlled comparison across datasets through homogeneous preprocessing. The second incorporates frame-wise inference on continuous video sequences from the NDS-BR, followed by temporal aggregation strategies designed to mitigate noise and characterize the trade-offs inherent to fine-grained temporal analysis. This two-level approach allows the study to quantify the gains in recall derived from temporal redundancy, while simultaneously assessing the increase in false positives that may emerge from high-frequency predictions under naturalistic noise.

Additionally, this work provides detailed insights into the operational implications of deploying YOLO models for real-time driver monitoring. By analyzing precision, recall, specificity, and inference time across versions and granularities, the research identifies the

conditions under which higher-capacity models meaningfully improve detection reliability, as well as the scenarios in which lightweight architectures may be preferable due to computational constraints. This evaluation extends beyond accuracy, incorporating ethical and practical considerations related to false positives, false negatives, and the broader applicability of such systems in safety-critical settings. Altogether, the study contributes to establishing an empirical foundation for selecting and deploying state-of-the-art detection architectures in real-world driver monitoring systems, particularly within the context of emerging naturalistic datasets in Brazil.

1.4 DOCUMENT ORGANIZATION

This document provides a concise overview of the theoretical, methodological, and empirical components developed throughout this study. It first introduces the architectural principles, innovations of the YOLO family (v8, v10, v11, v12) and benchmark evaluation metrics, as detailed in Chapter 2. Relevant literature on driver-distraction detection—spanning convolutional, SVM-based, Transformer, and multimodal CLIP-based approaches—is reviewed in Chapter 3, situating the present work within the broader research landscape and emphasizing the scarcity of studies relying on naturalistic data.

Chapter 4 describes the datasets used and their preparation, experimental setup, processing pipeline, temporal aggregation strategies, and evaluation metrics adopted, establishing the methodological basis for the analyses. The performance of the YOLO architectures on both controlled (SFDDD) and naturalistic (NDS-BR) environments is then presented in Chapter 5, highlighting differences across granularities, contrasts between YOLOv8 and YOLOv12, and the recall–specificity dynamics observed in static and temporal inference. Together, these chapters provide an integrated synthesis of findings and their implications for real-time driver monitoring.

Building upon these findings, Chapter 6 synthesizes the main conclusions derived from the experimental results, discussing their implications for real-time driver monitoring and model deployment in naturalistic settings.

2 BACKGROUND

This chapter is structured to progressively and coherently present the theoretical and empirical elements that support this study. Section 2.1 begins by discussing the architectural evolution of the You Only Look Once (YOLO) family, highlighting its principles, components, and innovations across the selected versions (YOLOv8, YOLOv10, YOLOv11, and YOLOv12). This part establishes the conceptual foundation required to understand how each model operates and which developments motivate their use in real-time detection scenarios. Subsequently, Section 2.2 introduces the benchmark evaluation metrics used throughout this study, providing the necessary theoretical basis for interpreting and comparing model performance.

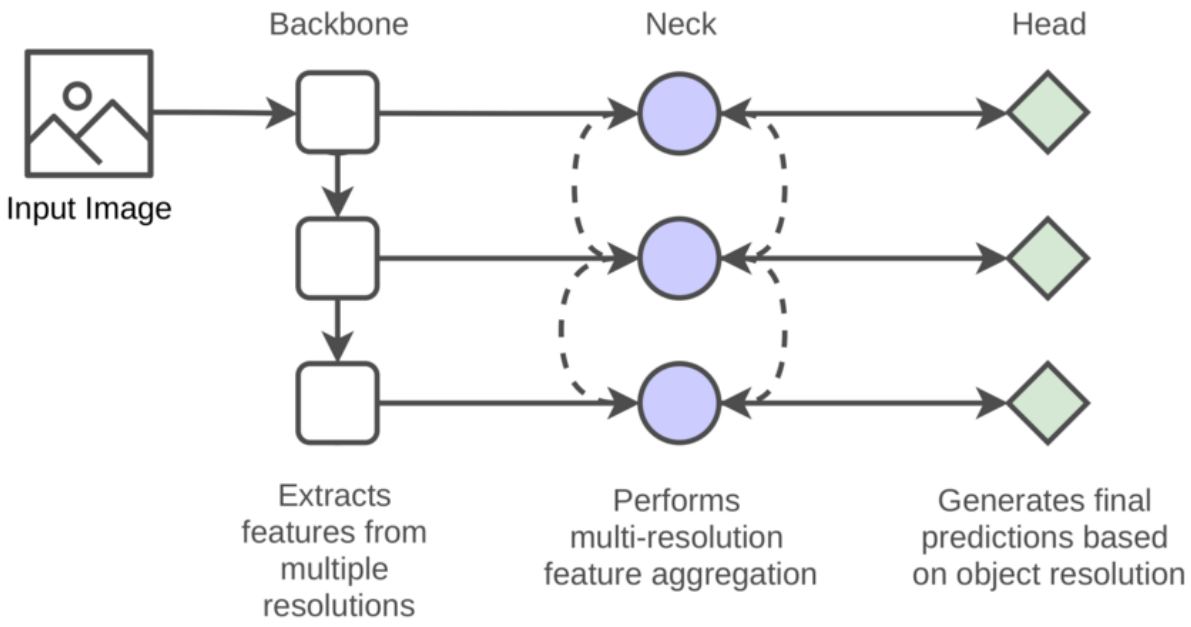
2.1 THE YOLO ARCHITECTURE: PARADIGM AND FUNDAMENTAL COMPONENTS

The YOLO approach, introduced by Redmon et al. (2016), represented a major shift from the prevailing object detection methods. Two-stage detectors, such as the Region-based Convolutional Neural Network (R-CNN) family, first generate regions of interest and subsequently classify them. In contrast, YOLO unifies these steps into a single neural network.

This reformulation turns detection into a direct regression problem. The network processes the entire image in a single pass, simultaneously predicting bounding boxes and class probabilities. This global approach allows the model to reason holistically about image context, reducing false positives and achieving inference speeds compatible with real-time applications.

The longevity and adaptability of the YOLO family stem from its modular architecture, generally divided into three components: Backbone, Neck, and Head (Figure 2.1). The Backbone extracts features from the input image, transforming pixels into rich representations. The Neck serves as a linking structure that performs multiscale feature fusion through modules such as Feature Pyramid Network (FPN) and Path Aggregation Network (PANet). Finally, the Head is responsible for producing the final predictions. The evolution of these three components has been the main driver of innovation across YOLO versions.

To understand the specific capabilities evaluated in this study, the following subsections detail the distinct characteristics and architectural innovations of the selected versions. Each of these iterations introduces specific improvements in feature aggregation, attention mechanisms, and processing efficiency that are critical to the performance analysis conducted in Chapter 5.



(a)

Figure 2.1: Visual representation of general architecture of the YOLO family. Source: Kateb et al. (2021).

2.1.1 YOLOv8

Released by Ultralytics in 2023, YOLOv8 consolidated advancements that redefined the balance between performance and computational cost. The network introduced optimizations resulting in more stable, accurate, and efficient detection.

According to Jocher (2025), the main innovations introduced in YOLOv8 include:

- **New C2f Module:** Replacing the C3 (Figure 2.2) module from YOLOv5, the C2f (Cross Stage Partial with 2 fusions) enhances gradient flow by concatenating intermediate outputs from internal blocks, resulting in more robust feature extraction, especially for small objects.
- **Anchor-Free Detection Head:** The model eliminates dependence on predefined anchor boxes. Instead, it directly predicts object properties (center point, width, and height), simplifying training and increasing generalization.
- **Decoupled Head:** Localization (bounding box regression) and classification tasks are separated into independent branches, reducing conflicts between spatial and semantic optimizations and enabling each branch to improve independently.

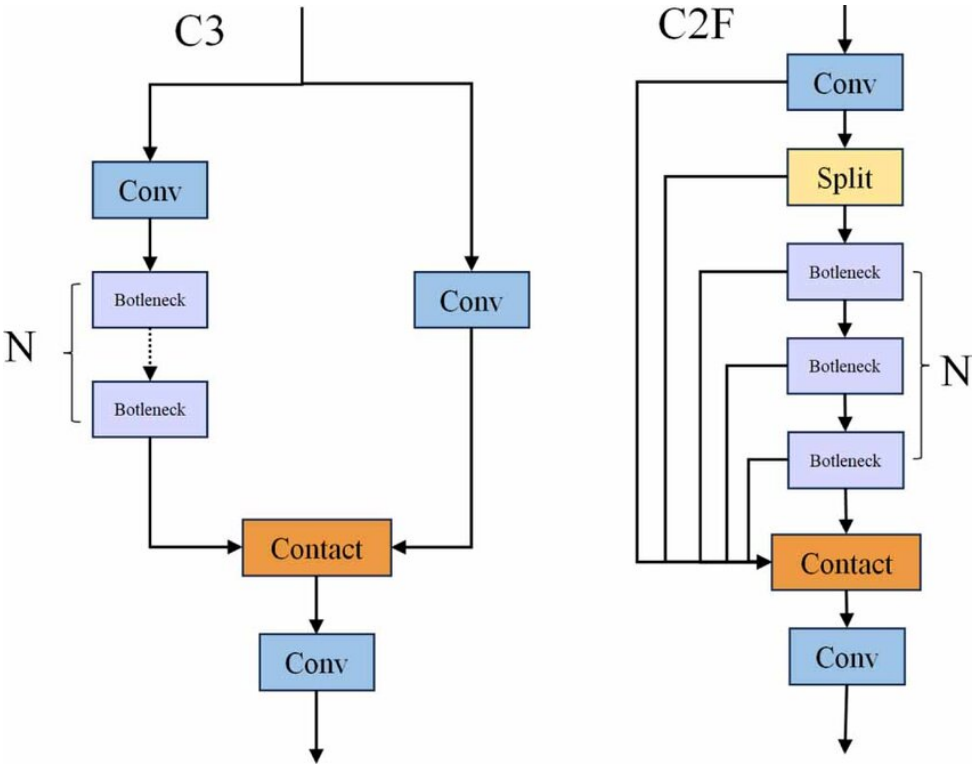


Figure 2.2: Comparison between the C3 and C2f modules. Source: Liu et al. (2024).

2.1.2 YOLOv10

Developed by Tsinghua University in collaboration with Ultralytics, YOLOv10 (Wang et al., 2024) introduced a major breakthrough that is the removal of Non-Maximum Suppression (NMS) during inference. Traditionally, NMS is used to eliminate redundant predictions, but it introduces a significant latency bottleneck.

The core innovations that make YOLOv10 truly end-to-end are:

- **End-to-End Training (NMS-Free):** The model internally learns to eliminate duplicate predictions during training, enabling faster and more direct inference without post-processing.
- **Consistent Dual Assignments:** The architecture employs two prediction heads during training (one-to-many and one-to-one) (Figure 2.3). The former provides dense and rich supervision, while the latter enforces a unique association between object and prediction.
- **Consistent Matching Metric:** Ensures that both heads optimize in an aligned manner, allowing the one-to-one head to govern inference.
- **Optimized Architecture:** The model adopts a Cross Stage Partial Network (CSPNet)-based backbone and a modified Path Aggregation Network (PAN) neck, incorporating large-kernel convolutions and partial attention to expand the receptive field with minimal computational cost.

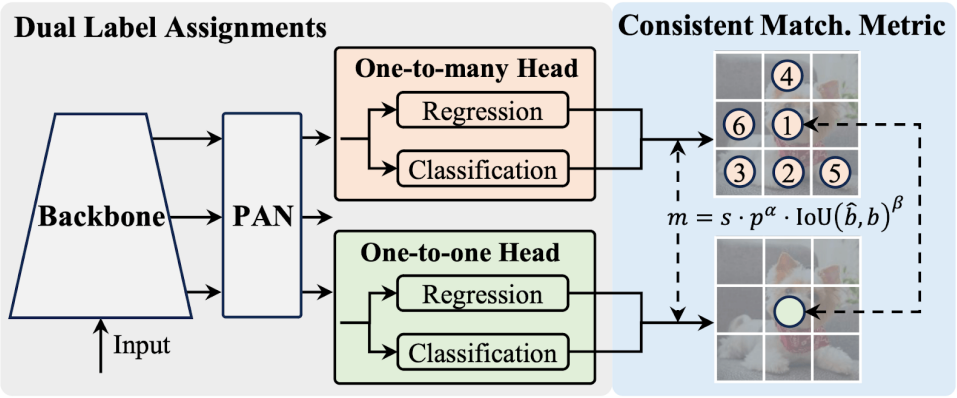


Figure 2.3: Consistent dual assignments for NMS-free training. Source: Wang et al. (2024).

This combination (Figure 2.3) yields a highly efficient model with competitive accuracy and significantly reduced inference time.

2.1.3 YOLOv11

YOLOv11 (Khanam and Hussain, 2024) introduces targeted architectural refinements aimed at improving computational efficiency and feature representation while preserving the convolution-centric design characteristic of previous Ultralytics models. Rather than redefining the detection paradigm, YOLOv11 focuses on simplifying blocks and enhancing spatial awareness within the backbone and neck.

The main architectural components include:

- **C2PSA (Cross-Stage Partial + Spatial Attention):** A streamlined spatial-attention module placed after the Spatial Pyramid Pooling - Fast (SPPF) block, enabling the model to highlight relevant local regions without adding significant overhead.
- **C3k2 Block:** A lighter variant of the Cross Stage Partial (CSP) bottleneck that replaces C2f, using smaller kernels to reduce computation while maintaining feature extraction capacity.

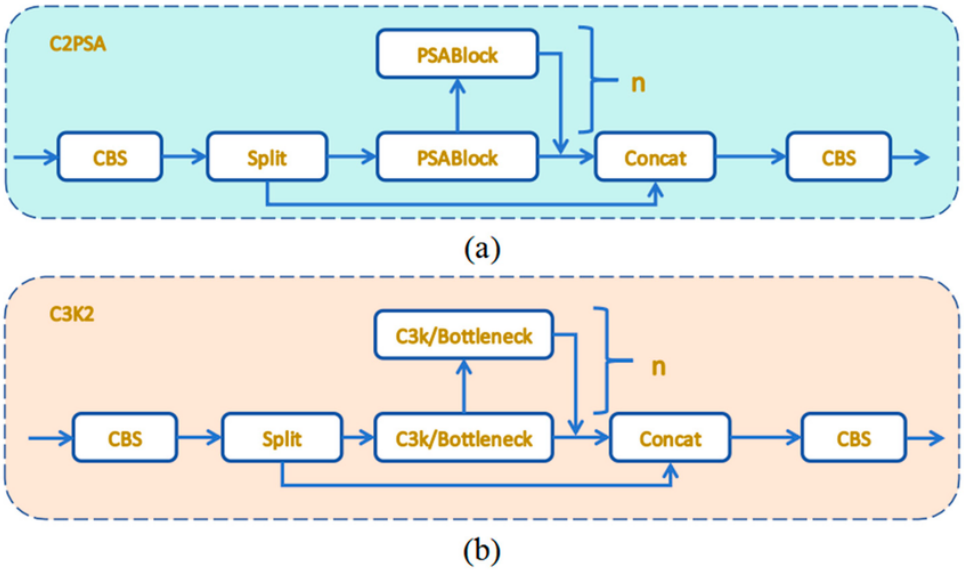


Figure 2.4: Architectural diagrams of the modules used in YOLOv11 (a) C2PSA and (b) C3K2. Source: Meng et al. (2025).

These modifications (Figure 2.4) lead to improved efficiency–accuracy trade-offs across model scales (Nano to Extra-Large), while maintaining compatibility with multiple vision tasks such as detection, segmentation, pose estimation, and oriented bounding boxes.

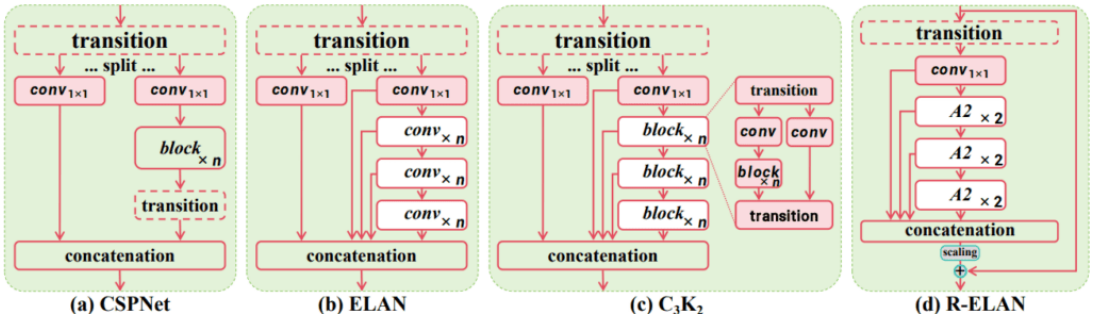
2.1.4 YOLOv12

YOLOv12 (Tian et al., 2025) marks a shift toward an attention-centric architecture, addressing historical limitations of self-attention in real-time detection—specifically, computational complexity and inefficient memory operations.

Its principal components are:

- **Area Attention (A2):** A simple region-based partitioning strategy that reduces attention cost by limiting token interactions, providing a larger receptive field without windowing overhead (Figure 2.5(a)).
- **R-ELAN:** A residual version of the ELAN aggregation block that stabilizes training in deeper attention-based architectures while lowering FLOPs and parameters (Figure 2.5(b)).
- **FlashAttention:** An optimized attention implementation that minimizes memory transfers, enabling high-speed inference even under attention-heavy workloads.

(a) Illustration of spatial attention mechanisms in YOLO architectures.



(b) Comparison between traditional CSP/ELAN modules and residual aggregation approaches.

Figure 2.5: Illustrative components referenced across YOLO architectures. Source: Tian et al. (2025).

With these elements, YOLOv12 achieves higher accuracy than YOLOv10 and YOLOv11 across all model scales, while maintaining real-time latency and reducing computational inefficiencies commonly associated with transformer-based modules.

2.1.5 Size Variants of the YOLO Family

YOLO architectures are available in multiple scales to optimize the trade-off between speed, accuracy, and computational requirements. The variants are nano (n), small (s), medium (m), large (l), and extra-large (x). This flexibility is achieved through two primary scaling factors: `depth_multiple`, which adjusts the number of layers of the network, and `width_multiple`, which adjusts its number of channels per layer. Smaller models, such as YOLOv8n, are optimized for edge devices and low latency, while larger variants, such as YOLOv12x, increase representation capacity, yielding better performance in complex scenarios at the cost of higher computational demands.

2.1.6 Comparative Table of Innovations and Performance

The table below (Table 2.1) summarizes the main architectural innovations and reported performance (mAP on the COCO dataset) for all versions of each analyzed model.

Version	Main Architectural Innovation	mAPval 50-95				
		Nano	Small	Medium	Large	X-Large
YOLOv8	C2f Module; Anchor-Free Head; Decoupled Head.	37.3%	44.9%	50.2%	52.9%	53.9%
YOLOv10	NMS-Free (End-to-End); Consistent Dual Assignments (one-to-one and one-to-many).	38.5%	46.3%	51.1%	53.2%	54.4%
YOLOv11	C2PSA (Cross-Stage Partial + Spatial Attention); C3k2 Block (Lighter CSP variant).	39.5%	47.0%	51.5%	53.4%	54.7%
YOLOv12	Area Attention (A2); R-ELAN (Residual ELAN); FlashAttention Integration.	40.6%	48.0%	52.5%	53.7%	55.2%

Table 2.1: Comparison of innovations and mAP on the COCO val2017 dataset (Lin et al., 2015). mAP values compiled from the official documentation and publications of each model. Source: Jocher (2024).

2.1.7 General Comparison Across Versions

In summary, YOLO versions maintain the principle of real-time one-stage detection but continually evolve toward greater efficiency and accuracy. YOLOv8 consolidated the use of decoupled and anchor-free heads. YOLOv10 advanced to fully end-to-end detection by removing NMS and reducing latency.

YOLOv11 focused on computational efficiency, introducing simplified blocks like C3k2 and a lightweight spatial attention module, C2PSA, to refine feature extraction without overhead. Finally, YOLOv12 marked a shift toward an attention-centric architecture, using Area Attention (A2) and the residual aggregation R-ELAN for greater efficiency, while also integrating FlashAttention to optimize memory transfers and accelerate inference.

Given the rapid release cycle and varying performance nomenclature across versions, critical comparison is necessary. Versions 8, 10, 11, and 12 were selected for this study because they report the highest mAP levels in their official publications, enabling evaluation under controlled conditions. The comparative analysis also focuses on the internal scaling of each version (n, s, m, l, x). Earlier or parallel versions were excluded because many use different scaling strategies or evaluation metrics, which would hinder fair comparison across model sizes.

2.2 BENCHMARK EVALUATION METRICS

To evaluate the effectiveness and reliability of object detection models, this study adopts a set of threshold-based evaluation metrics traditionally derived from the confusion matrix. These metrics—accuracy, recall, precision, specificity, and negative predictive value (NPV)—quantify different aspects of predictive performance and offer complementary perspectives regarding correct classification, error characterization, and class-related trade-offs. Their definitions and interpretations are presented in (Hossin and Sulaiman, 2015).

Accuracy measures the overall proportion of correctly classified instances with respect to all evaluated samples. Although widely used for both binary and multiclass tasks, accuracy alone may be non-discriminative, insensitive to minority classes, and unable to reveal trade-offs between types of errors. Precision quantifies the fraction of correctly predicted positive instances among all positive predictions, emphasizing the reliability of the model when it assigns a positive label. Recall measures the proportion of true positive instances that are correctly identified, reflecting the model's ability to detect relevant patterns. Specificity, in contrast, evaluates the proportion of correctly classified negative instances, ensuring that the model is not excessively

prone to false alarms. Finally, the Negative Predictive Value (NPV) denotes the fraction of true negatives among all predicted negatives, complementing precision by focusing on the correctness of negative predictions.

As discussed by Hossin and Sulaiman (2015), these metrics arise directly from the four confusion-matrix components—true positives, false positives, true negatives, and false negatives—and collectively provide a more informative characterization of classifier behavior compared to accuracy alone. Their combined use helps mitigate common issues related to imbalance, lack of discriminability, and insufficient representation of minority classes, all of which are critical considerations when comparing detection architectures such as the YOLO family.

2.3 CONCLUDING BACKGROUND

With the theoretical foundation now established—covering the architectural evolution of the selected YOLO versions and the benchmark evaluation metrics employed to assess their performance—the document proceeds to review related works employing convolutional networks, YOLO-based detectors, Transformers, and multimodal approaches.

3 RELATED WORKS

Driver distraction detection has become a significant focus of research, with a growing number of approaches employing deep learning and machine learning-based methods. In this work, the literature is categorized into four major strands: approaches based on Convolutional Neural Networks (CNNs) (Section 3.1), Support Vector Machine (SVM) classifiers (Section 3.2), Transformer architectures (Section 3.3), and more recently, multimodal models using CLIP (Section 3.4).

3.1 CNN-BASED APPROACHES

Several studies have explored the application of Convolutional Neural Networks (CNNs) for classifying images of drivers in different attention states. Hossain et al. (2022) proposed a system that leverages pre-trained CNNs with transfer learning, employing architectures such as VGG-16, ResNet50, and MobileNetV2 to detect behaviors like mobile phone use, drowsiness, and interaction with passengers. Using the public SFDDD dataset, the MobileNetV2-based model showed the best performance, achieving 98.12% accuracy.

Xing et al. (2019) developed a system for driver activity classification with an emphasis on distractions. The methodology included an initial unsupervised segmentation step using Gaussian Mixture Models (GMM) to isolate the driver's body in the images. Transfer learning was subsequently applied using models such as AlexNet, GoogLeNet, and ResNet50. In its best result, with the proposed segmentation step, AlexNet achieved 81.60% accuracy, compared to 69.20% without it. The dataset used was private, consisting of 34,000 images from 10 drivers captured with a Kinect camera.

Baheti et al. (2018) proposed a distraction detection system based on deep CNNs. The VGG-16 architecture was used with structural modifications and regularization techniques. The model achieved 96.31% accuracy. An optimized version further reduced the number of parameters by approximately 90.00%, maintaining an accuracy of 95.54%. Although the dataset used was public, containing 17,000 images, no source code was released.

Similarly, Masood et al. (2020) proposed a system not only aimed at detecting distraction but also identifying its specific cause. The study explored VGG-16 and VGG-19 architectures on the SFDDD dataset, comparing the performance of models with randomly initialized weights against those pre-trained on ImageNet. The VGG-16 pre-trained approach proved to be the most effective, achieving an average accuracy of 99.57%, compared to 99.39% from the VGG-19 model without pretraining. Although a public dataset was used, favoring comparison, the authors did not release the source code or final weights.

In a hybrid model approach, Huang et al. (2020) introduced the Hybrid CNN Framework (HCF), a system that combines pre-trained models ResNet50, Inception V3, and Xception to extract visual features at different scales in parallel. The extracted features are concatenated and passed through classification layers designed to reduce overfitting. Evaluated on the SFDDD dataset, the HCF model achieved 96.74% accuracy, outperforming the individual base models. Despite the promising results, the study does not provide source code or trained weights.

Li et al. (2020) proposed a two-module pipeline for detecting manual driver distraction. The first module employs a YOLOv3 model to detect the bounding boxes of the driver's right hand and right ear, while the second module uses a multi-layer perceptron (MLP) to classify the distraction type based on the spatial information from these detections. The system was trained

and evaluated on a private dataset collected from 20 participants in a driving simulator. The framework achieved an F1-score of 59.00% for overall multi-class distraction detection. However, the use of a private simulator dataset and the lack of publicly available source code or trained models were noted.

Sajid et al. (2021) proposed an object detection-based framework for identifying driver distractions, leveraging the EfficientNet/EfficientDet architectures (variants D0–D4). The system was designed to detect not only the distraction class but also to localize distracting objects and relevant body-part Regions of Interest (ROIs). The models were evaluated on a subset of the SFDDD. In a systematic comparison of its variants, the EfficientDet-D3 model achieved the best performance, with a mean Average Precision (mAP) of 99.16%, outperforming other architectures like Faster R-CNN and YOLOv3. However, the study relied on a single dataset and the absence of public source code or pre-trained weights.

Poon et al. (2022) explored the use of lightweight YOLO variants for real-time, in-cabin monitoring. The study evaluated different architectures, including YOLOv3-tiny and YOLO-fastest. The models were trained and tested on a private dataset. The YOLO-fastest-three-scales variant demonstrated the best performance, achieving a mAP of 97.16%. However, the study used a private dataset with limited behavioral categories, and the lack of publicly available trained models or specific implementation code was an observation.

Still within CNN-based architectures, YOLO has been applied to the detection and classification of driver distraction. Liu et al. (2022) proposed CEAM-YOLOv7, a variation of YOLOv7 that integrates a Global Attention Mechanism (GAM) and a Channel Expansion (CE) algorithm to enhance detection in infrared (IR) images. The method achieved a mean Average Precision (mAP) of 73.60% on a private dataset, though no code was made available.

Ma et al. (2024b) adopted the YOLOv8, developing an improved variant of this architecture. Enhancements included the addition of a Multi-Head Self-Attention (MHSA) module in the network backbone and a convolutional module for emotion classification. The model achieved a mAP of 81.40% for distraction detection and 73.30% for emotion classification. The dataset used was private, collected from only 20 participants, and no source code was released.

Neamah et al. (2024) introduced a real-time driver distraction detection system designed for deployment on embedded hardware. The approach utilizes a YOLOv8 classification model running on an NVIDIA Jetson Nano, integrated with a camera and a GSM/GPS module for data transmission. Evaluated on the public SFDDD, the system achieved a final accuracy of 99.75%. Despite the focus on a practical implementation, the authors did not release the source code or pre-trained models.

Khalil et al. (2025) developed a low-cost driver monitoring system featuring a hybrid processing approach on a Raspberry Pi Zero 2 W. The system runs a modified YOLOv8n object detection model on the CPU, while simultaneously executing a head-pose estimation algorithm on the device's GPU via OpenCL to enhance prediction accuracy. The models were trained on a diverse, publicly available dataset created by the authors, which combines five existing datasets with their own captures. The system achieved an overall accuracy of over 90.00%; however, validation performance on the celphone detection class was significantly lower at 75.00% mAP50.

Fu et al. (2024) proposed GD-YOLO, an optimized version of the YOLOv7 architecture developed for detecting smoking and phone use by drivers. The model incorporates a module with deformable convolutions (D-LAN) to improve small-object feature extraction and another with GhostConv layers (G-LAN) to reduce the backbone's complexity. The system was evaluated on a private dataset. The solution achieved a mAP50 of 86.00%. Although the results are promising, the use of a private dataset was noted.

Ge et al. (2025) introduced YOLO-AFR, a modified YOLOv12 architecture for the real-time detection of dangerous driving behaviors. The model incorporates components such as a feedforward network and attention mechanisms. Evaluated on the public YawDD-E and SFDDD, the model achieved a mAP50 of 97.60% and 98.90%, respectively. Despite its strong performance on public datasets, the study does not provide a public repository for the source code.

3.2 SVM-BASED APPROACHES

The use of Support Vector Machines (SVM) has also been explored, particularly in scenarios with limited labeled data. Liu et al. (2016) investigated the use of Laplacian SVM (LapSVM) and Semi-Supervised Extreme Learning Machine (SS-ELM) to detect cognitive distraction based on eye and head movements. The method achieved a G-mean of up to 95.70% in binary classification, outperforming traditional SVMs. However, the dataset was private, and no code was shared.

Tango and Botta (2013) proposed a system based exclusively on vehicle dynamics data, employing an SVM with RBF kernel. The model achieved up to 97.00% accuracy in binary classification in an experiment involving 20 drivers. Neither datasets nor code repositories were made available.

3.3 TRANSFORMER-BASED APPROACHES

With the advancement of Transformer architectures, some studies have explored this methodology. Mohammed et al. (2024) proposed a hybrid CNN-Transformer model trained in a semi-supervised fashion using pseudo-labels generated by a mean-teacher architecture. Evaluated on the public SFDDD and 3MDAD datasets, the model achieved 95.43% and 70.39% accuracy, respectively. Despite the promising results, the source code was not released.

Ma et al. (2024a) presented ViT-DD, a purely Vision Transformer-based approach with multitask learning for distraction detection and emotion recognition. Using the public SFDDD and AUCD2 datasets, the model achieved 92.51% accuracy in driver-wise splits and up to 99.63% in image-wise classification. The study provides code, trained models, and replication scripts in a public repository, which is essential for research validation.

3.4 MULTIMODAL APPROACHES WITH CLIP

Hasan et al. (2024) proposed DriveCLIP, an adaptation of OpenAI's CLIP (Contrastive Language–Image Pretraining) architecture for driver distraction detection (Radford et al., 2021). This approach employs contrastive learning between images and textual descriptions, enabling inference through the alignment of both information sources. The model was evaluated on the public SynDD and SFDDD datasets, achieving 81.85% and 83.15% accuracy, respectively. The source code, trained weights, and usage instructions are publicly available, ensuring result reproducibility.

Author (Year)	Architecture	Model(s)	Dataset	Performance
Hossain et al. (2022)	CNN	VGG-16, ResNet50, MobileNetV2	SFDDD	Accuracy of 98.12%
Xing et al. (2019)	CNN	AlexNet, GoogLeNet, ResNet50	Private (Kinect, 10 drivers)	Accuracy of 81.6% (best case)
Baheti et al. (2018)	CNN	Modified VGG-16	Private (Abouelnaga et al., 2018)	Accuracy of 96.31%, lightweight version: 95.54%
Masood et al. (2020)	CNN	VGG-16, VGG-19	SFDDD	Accuracy of 99.57% (VGG-16), 99.39% (VGG-19)
Huang et al. (2020)	CNN	HCF (ResNet50, Inception V3, Xception)	SFDDD	Accuracy of 96.74%
Sajid et al. (2021)	CNN	EfficientDet	SFDDD	mAP 99.16%
Li et al. (2020)	CNN	YOLOv3	Private	F1-score of 59.00%
Poon et al. (2022)	CNN	YOLOv3	Private	mAP 97.16%
Liu et al. (2022)	CNN	CEAM-YOLOv7	Private	mAP 73.60%
Ma et al. (2024)	CNN	YOLOv8 with MHSA	Private (20 participants)	mAP 81.40% (distraction), 73.30% (emotion)
Neamah et al. (2024)	CNN	YOLOv8	SFDDD	Accuracy of 99.75%
Khalli et al. (2025)	CNN	YOLOv8	Mix of public Datasets	Accuracy of over 90.00%
Fu et al. (2024)	CNN	YOLOv7	Private	mAP 86.00%
Ge et al. (2025)	CNN	YOLOv12	Public	mAP 98.90%
Liu et al. (2016)	SVM	LapSVM, SS-ELM	Private	G-mean up to 95.70%
Tango and Botta (2013)	SVM	SVM, FFNN, ANFIS	Private	Accuracy of 97.00% (best case)
Mohammed et al. (2024)	Transformer	Hybrid (MobileViT-like)	SFDDD, 3MDAD	Accuracy of 95.43% (SFDDD), 70.39% (3MDAD)
Ma and Wang (2022)	Transformer	ViT-DD	SFDDD, AUCD2	Accuracy of 99.63% (SFDDD), 93.59% (AUCD2)
Zahid et al. (2023)	CLIP	DriveCLIP	SFDDD, SynDD1	Accuracy of 83.15% (SFDDD), 81.85% (SynDD1)

Table 3.1: Summary of related work on detecting distraction while driving.

3.5 FINAL CONSIDERATIONS

The literature review highlights consistent progress in driver distraction detection through convolutional networks (especially YOLO-based models), Transformers, and multimodal approaches such as CLIP. Nevertheless, the review also identified several critical limitations. Most existing studies rely on data collected in controlled environments or international traffic contexts, which often presents low representativeness regarding the specific behavioral patterns and scenarios of Brazilian traffic (Table 3.1). Furthermore, concerning the Brazilian Naturalistic Driving Study (NDS-BR), an initiative led by (Bastos et al., 2020), there is a lack of systematic application and evaluation of state-of-the-art deep learning architectures.

Given this context, the present work aims to bridge these gaps by undertaking a rigorous evaluation of the YOLO family of architectures for cell phone detection, utilizing both static image datasets (SFDDD and NDS-BR) and frame-by-frame temporal analysis on NDS-BR

videos. This methodology is designed to quantify performance gains from continuous analysis and provide fully reproducible results.

4 MATERIALS AND METHODS

This section details the materials and methods employed in this research, which aimed to compare the performance of different versions and granularities of the YOLO family in the task of detecting cell phone usage while driving. The experimental design is structured into two complementary stages. The first involves a static, image-wise comparison between the controlled State Farm Distracted Driver Detection (SFDDD) dataset and the naturalistic NDS-BR (Naturalistic Driving Study – Brazil). This configuration allows for a controlled evaluation of domain generalization, preserving the structure of original annotations without the influence of temporal factors. The second stage expands this analysis to the temporal domain using raw NDS-BR videos, applying frame-by-frame inference followed by temporal aggregation. This strategy aims to verify whether continuous detection yields sensitivity gains over discrete evaluation and to what extent this additional granularity impacts the trade-off between recall and false positives.

The remainder of this chapter is organized as follows: Section 4.1 characterizes the SFDDD and NDS-BR datasets used in the experiments. Section 4.2 outlines the structural design of the static and temporal experiments. The processing pipeline implementation is described in Section 4.3, followed by the specific strategies for temporal aggregation and detection handling in Section 4.4. Finally, the computational environment is specified in Section 4.5, and the evaluation metrics are defined in Section 4.6.

4.1 DATASETS EMPLOYED

Characterizing the datasets used in this study is essential for understanding the empirical conditions in which the YOLO architectures will be evaluated. The diversity, structure, and capture context of these datasets determine the levels of visual variability, environmental conditions, and behavioral patterns present in the samples, directly influencing the challenges each model will face during inference. For this research, two widely datasets in the field of computer vision applied to driver behavior monitoring were selected: Natutalistic Driving Study Brazil (NDS-BR), representing Brazilian naturalistic contexts, and State Farm Distracted Driver Detection (SFDDD), derived from controlled scenarios and widely used in international benchmarks.

4.1.1 NDS-BR Dataset

The NDS-BR is the first large-scale multimodal dataset dedicated to studying driver behavior in the Brazilian context. As described in (Bastos et al., 2020), NDS-BR was built from a continuous data collection protocol conducted over extended driving periods using a set of sensors installed in the participants vehicles. This methodological design follows the paradigm of naturalistic driving studies, which aim to record driver activity without experimental interference and in full interaction with the real traffic environment.

Vehicles were equipped with front, lateral, and interior cameras, as well as additional sensors, enabling simultaneous capture of multiple perspectives of driver behavior. Inside the vehicle, the camera positioned toward the driver's torso and face is the primary source of information relevant to this study, as it records hand gestures, head movements, and object manipulation—key elements for behavioral assessment.

Across all monitored participants, the naturalistic protocol resulted in 201 valid trips, corresponding to approximately 58.38 hours of effective driving, generating a large volume of

continuous video sequences from the internal camera. These recordings constitute the core visual material used for the analyses developed in this work.

Figure 4.1: Naturalistic Environment (NDS-BR): Realism and visual complexity centered around dynamic driving scenarios.

NDS-BR captures all these variations continuously, producing long video sequences subject to fluctuations in lighting, vehicle vibration, sudden environmental changes, occlusions caused by driver movements, and external interferences such as passing vehicles or direct sunlight (Figure 4.1).

These characteristics give NDS-BR a high degree of realism and visual complexity, distinguishing it significantly from datasets collected in controlled environments. Its naturalistic variability constitutes a structural element of the dataset, enabling analysis of model performance under unpredictable and highly dynamic conditions. Furthermore, NDS-BR includes multiple drivers, different vehicle types, urban and highway environments, and both daytime and nighttime recordings, adding considerable diversity. Thus, NDS-BR provides a broad, heterogeneous, and realistic panorama of driver behavior in Brazil.

4.1.2 SFDDD Dataset

The SFDDD is one of the most widely used datasets in the literature focused on detecting distracted driving behaviors. Originally released for a Kaggle competition in 2016 (Montoya et al., 2016), SFDDD was developed to provide a standardized and labeled set of in-vehicle images, enabling direct comparison between different computer vision approaches.

The dataset consists of static images recorded from a camera mounted on the vehicle dashboard, with a fixed framing directed at the driver. Unlike naturalistic datasets, SFDDD was constructed in a semi-controlled environment, where drivers were instructed to perform a predefined set of behaviors. These include safe driving (no manual interaction) and nine typical distraction behaviors, such as phone use with one or both hands, talking on the phone, adjusting the radio, drinking, performing personal actions, and interacting with passengers (Figure 4.2).

Figure 4.2: State Farm Distracted Driver Detection classes. Source: Montoya et al. (2016).

SFDDD images exhibit high sharpness and visual clarity, with moderate variation among drivers, head poses, clothing, and slight body movements. Although it lacks naturalistic unpredictability, the dataset stands out for its organization and its clear, well-defined class structure—ideal for comparative studies.

Containing over 22,000 training images from 26 distinct drivers, it offers reasonable internal diversity, though limited to the simulated environment of a vehicle cabin. Its widespread adoption in the scientific community has solidified SFDDD as a reference benchmark for classification and detection models in traffic-safety-related tasks.

4.1.3 General Comparison Between the Datasets

The complementarity between NDS-BR and SFDDD becomes evident when analyzing their capture nature, data structure, and levels of visual variability. While NDS-BR represents driver behavior in real-world conditions, with long video sequences and substantial environmental variation (Figure 4.1), SFDDD provides a controlled scenario with clearly defined classes and independent images (Figure 4.2). This distinction offers two different perspectives on behavioral patterns: one rooted in the continuous dynamics of the real world and the other in the semantic clarity of simulated environments.

4.2 EXPERIMENTAL STRUCTURE

To make the experiments comparable, the NDS-BR was initially processed to generate a dataset of static images in the same format as the SFDDD (Table 4.1). Each frame was converted to PNG and treated as an independent instance, ensuring that both sets could be evaluated under equivalent conditions. This step allowed the performance of the YOLO versions to be analyzed without the influence of temporal factors, ensuring a direct image-wise comparison between the two databases.

Subsequently, leveraging the availability of the original videos, a second experiment was conducted on the NDS-BR itself, this time preserving the temporal sequence of the frames. In this configuration, each video was read in its entirety and processed frame-by-frame, with results aggregated into one-second windows. The objective was to investigate whether continuous video

flow analysis, in contrast to the discrete approach, offers any effective gain in terms of detection, particularly for short or intermittent cell phone usage events.

For the purpose of the comparative experiments, the original ten classes of the SFDDD dataset were binarized into two labels: Cellphone Use (Yes) and No Cellphone Use (No). Classes C1, C2, C3, and C4 (texting/talking on the phone, right/left hand) were mapped to Cellphone Use (Yes). This mapping resulted in a dataset split of approximately 41% positive samples and 59% negative samples, resulting in a total of 22,424 images, used for the static image comparison.

To match the scale of the SFDDD, an initial sample of approximately 20,000 images was extracted from the NDS-BR. Crucially, this selection was performed by randomly sampling frames from the entire video, rather than extracting continuous sequences. This approach ensures that the static dataset consists of statistically independent instances, avoiding the temporal redundancy inherent in video data. For both the NDS-BR static image experiment and the NDS-BR video experiment, the dataset was balanced to approximate a 50%/50% distribution between Cellphone Use (Yes) and No Cellphone Use (No) to ensure a balanced evaluation of model performance. The final distribution of samples used across all three experimental setups is detailed in Table 4.1.

Dataset / Experiment	**Cellphone Use (No)**	**Cellphone Use (Yes)**	**Total Samples**	**Pos. (%) / Neg. (%)**
SFDDD Static Images	13,168	9,256	22,424	41.28% / 58.72%
NDS-BR Static Images	12,884	12,662	25,546	49.57% / 50.43%
NDS-BR Video Temporal	11,278	10,672	21,950	48.62% / 51.38%

Table 4.1: Distribution size of Samples (Images/Seconds) Across Experimental Datasets.

4.3 PROCESSING PIPELINE DESCRIPTION

The processing pipeline was divided into two complementary fronts, both based on the Ultralytics YOLO library. In the first, dedicated to static images, the script iterates through all instances of the dataset, executing inference using the weights corresponding to each model and architecture size (v8, v10, v11, and v12; in variations nano, small, medium, large, and extra-large). For each image, the original metadata (class, subject, and identifier) are loaded, and inference is performed considering only the class of interest—cell phone, index 67 in the Common Objects in Context dataset (COCO) (Lin et al., 2015). When the detection confidence is equal to or greater than 0.2, the image is registered as positive and saved along with the respective label file containing the normalized bounding box coordinates. All outputs are stored in a CSV file containing the main variables of interest: filename, detection, class, model used, confidence, and inference time.

The video-oriented pipeline follows the same general logic but incorporates additional mechanisms for temporal extraction and association with the ground truth. Each video is processed frame-by-frame; the driver is identified from the file path, and, whenever possible, frames are synchronized with reference annotations by reading the timestamps embedded in the image. This correspondence is performed by an optical character recognition (OCR) function that attempts to locate and validate the date and time fields in every frame, mapping them to the base table. When the reading temporarily fails, the system retains the last valid value to preserve temporal coherence.

During inference, each frame is submitted to the loaded model (YOLOv8 or YOLOv12), and only the cellphone class is evaluated. The minimum confidence threshold is maintained at 0.2, and in the event of a detection, the confidence and box coordinates are recorded. The entire process is timed, allowing for the subsequent calculation of the average inference time per frame. Results are saved in a CSV file containing, for each frame, information regarding the video, frame

number, detection, annotated action, driver, date, time, model, and processing time. This uniform log structure enables aggregated analyses and comparisons between models and granularities.

4.4 TEMPORAL AGGREGATION AND DETECTION HANDLING

Frame-by-frame analysis, while offering greater detail, can be susceptible to instantaneous fluctuations in predictions, especially in scenarios with visual noise, reflections, or brief device occlusions. To mitigate this effect, temporal aggregation was applied to the video results. The logic consists of grouping all frames corresponding to the same second of recording and determining the final label for that interval in two alternative ways: (i) considering the second as positive if at least one frame presents a cell phone detection, or (ii) requiring that a minimum number of frames (threshold) within that second satisfy the confidence criterion. This second approach is more conservative and aims to reduce isolated false positives.

4.5 COMPUTATIONAL ENVIRONMENT AND EXPERIMENT EXECUTION

All experiments were conducted in a controlled environment configured to ensure the reproducibility of results. Simulations were executed on a machine featuring an AMD Ryzen 7 4800H processor with Radeon Graphics, 8 cores (16 threads), a maximum frequency of 2.9 GHz, and 32 GB of RAM, running on an x86_64 architecture with a 64-bit Linux operating system. GPU acceleration was utilized for inferences, leveraging CUDA extensions compatible with the Ultralytics YOLO library. The execution environment was configured with Python 3.10, and all necessary dependencies were pinned to fixed versions to ensure consistency across runs.

4.6 EVALUATION METRICS AND RESULTS ANALYSIS

To evaluate the models, classic binary classification metrics were used, calculated both on a frame scale and a temporal scale aggregated per second. These include: precision, recall, specificity, negative predictive value (NPV), accuracy, and average inference time per frame. The corresponding confusion matrices (Figure 4.3) were generated for each experimental scenario, enabling a direct comparison between architecture versions and granularities.

		Predicted		
		Positive	**Negative**	
Ground Truth	**Positive**	True Positive (TP)	False Negative (FN) [Type II Error]	**Sensitivity (Recall)** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) [Type I Error]	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision (PPV) $\frac{TP}{(TP + FP)}$	**NPV** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 4.3: Confusion matrix. This matrix allows for the calculation of key model metrics such as recall, specificity, precision, and accuracy. Source: (Cabot and Ross, 2023).

In the case of the video experiment, complementary analyses were conducted comparing the frame-by-frame inference results with those obtained from the image dataset derived from the NDS-BR itself. This comparison allowed for the quantification of the effective sensitivity gain resulting from continuous analysis, as well as the expected increase in false positives associated with this finer granularity. The interpretation of results takes into account the presence of noise in the human annotations of the NDS-BR and ambiguous scenarios, which imposes an upper bound on accuracy metrics and must be weighed in the discussions. Thus, the combination of static and temporal experiments, associated with a detailed analysis of the trade-offs between computational cost and performance, composes the methodological basis that supports the conclusions presented in the following section.

5 RESULTS

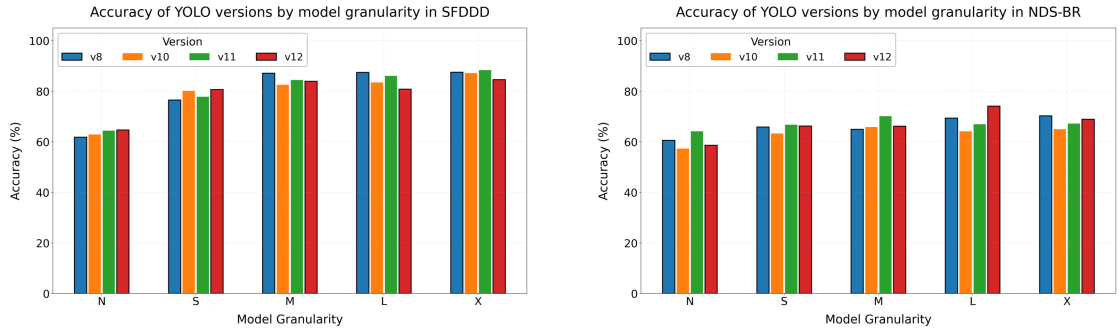
This chapter presents the experimental results obtained from the evaluation of the YOLO architectures. The analysis is structured to provide a progressive understanding of the models' behaviors, moving from a macro-level comparison to specific case studies in controlled and naturalistic environments.

Section 5.1 offers a general comparison of the aggregated metrics across all tested versions and granularities. Section 5.2 details the performance on the SFDDD dataset, focusing on the trade-off between detection capability (recall) and reliability (precision) in a controlled setting. Finally, Section 5.3 explores the results on the NDS-BR dataset, discussing the challenges imposed by real-world constraints and the practical implications for embedded driver monitoring systems.

5.1 GENERAL COMPARISON

The results presented here were obtained through the execution of an experimental pipeline (described in Section 4), in which models from the YOLOv8, v10, v11, and v12 architectures—across their five granularities (Nano, Small, Medium, Large, and X-Large)—were subjected to inference on the State Farm Distracted Driver Detection (SFDDD) and Naturalistic Driving Study Brazil (NDS-BR) dataset.

A macro-level analysis of the aggregated results reveals an expected positive correlation between model complexity and accuracy, as illustrated in Figure 5.1. This performance scaling directly reflects the enhanced representational capacity and feature learning capabilities afforded by an increased parameter count. A similar trend is partially observed in the NDS dataset, see Figure 5.1, although with more pronounced variability across granularities and versions, likely due to the higher intra-class diversity and temporal complexity of the video-based samples.



(a) Accuracy by granularity and version on the SFDDD dataset.

(b) Accuracy by granularity and version on the NDS dataset.

Figure 5.1: Visual representation of the Accuracy by granularity for different YOLO models and datasets. (a) SFDDD. (b) NDS-BR.

Although all versions demonstrate broadly comparable overall behavior, the subsequent analyses focus specifically on YOLOv8 and YOLOv12. The inclusion of YOLOv8 is justified by its consolidated status as a state-of-the-art benchmark, widely adopted in the literature and consistently achieving top-tier accuracy—either leading or within a few percentage points of the best results—across both datasets and all granularities. Conversely, YOLOv12 was selected to evaluate whether the architectural and algorithmic advances introduced in the most recent

iteration of the YOLO family effectively translate into measurable improvements. While its initial accuracy remains aligned with earlier versions, further exploration of complementary performance metrics, primarily precision and recall, reveals qualitative gains in detection reliability and class discrimination.

5.2 SFDDD ANALYSIS

The analysis of the results on the SFDDD dataset, without fine-tuning, aims to elucidate the trade-off between detection capability (recall) and reliability (precision), alongside computational cost (inference time) within the YOLOv8 and YOLOv12 architectures (Table 5.1).

The central objective is to identify which combination of version and granularity yields the optimal operating point—balancing the need to detect distractions without overwhelming the user with false alarms—considering practical applicability in real-time vehicular monitoring scenarios.

5.2.1 YOLOv8

The eighth version of the YOLO architecture exhibits a well-defined behavior concerning the trade-off between recall and computational cost, which is evidenced by the expressive differences across its granularities.

Matrix Confusion of Dataset SFDDD for YOLO v8 Nano.
Sample size: 22424

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 886	FN 8370	Recall 9.57%
	N	FP 194	TN 12974	Specificity 98.53%
		Precision 82.04%	NPV 60.79%	Accuracy 61.81%

(a) Nano

Matrix Confusion of Dataset SFDDD for YOLO v8 Extra-Large.
Sample size: 22424

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 7729	FN 1527	Recall 83.50%
	N	FP 1272	TN 11896	Specificity 90.34%
		Precision 85.87%	NPV 88.62%	Accuracy 87.52%

(b) Extra-Large

Figure 5.2: Visual representation of the confusion matrix for different YOLOv8 model sizes. (a) YOLOv8 nano. (b) YOLOv8 extra-large.

The YOLOv8-Nano model (Figure 5.2) adopts a highly conservative stance. Its recall of 9.57% is extremely low, indicating that it fails to identify the vast majority of distraction events, essentially, most moments when the driver uses a cell phone go unnoticed. However, this conservative approach results in fewer false positives. In practical terms, this signifies that the model is well-suited for automated systems operating without human supervision—such as auditing or event-logging platforms—where reliability is paramount. However, its critically low recall compromises its utility in preventive contexts, such as real-time alerts, where identifying the maximum possible number of cell phone usage incidents is essential.

At the other extreme, the YOLOv8-Extra-Large (Figure 5.2(b)) inverts this strategy and prioritizes detection coverage. Its recall rises to 83.50%, allowing it to identify the majority of distracted drivers—approximately 8 out of 10 real cases of cell phone use are recognized. This improvement is accompanied by a natural trade-off in precision (85.87%), indicating a proportional increase in false positives compared to smaller models. In practice, this behavior makes it more suitable for embedded alert scenarios, where it is preferable to detect more cases (even if some are incorrect) than to let real events go unnoticed. Despite this aggressiveness,

the precision remains high enough to keep the model reliable for active monitoring applications, such as corporate fleets or assisted autonomous vehicles.

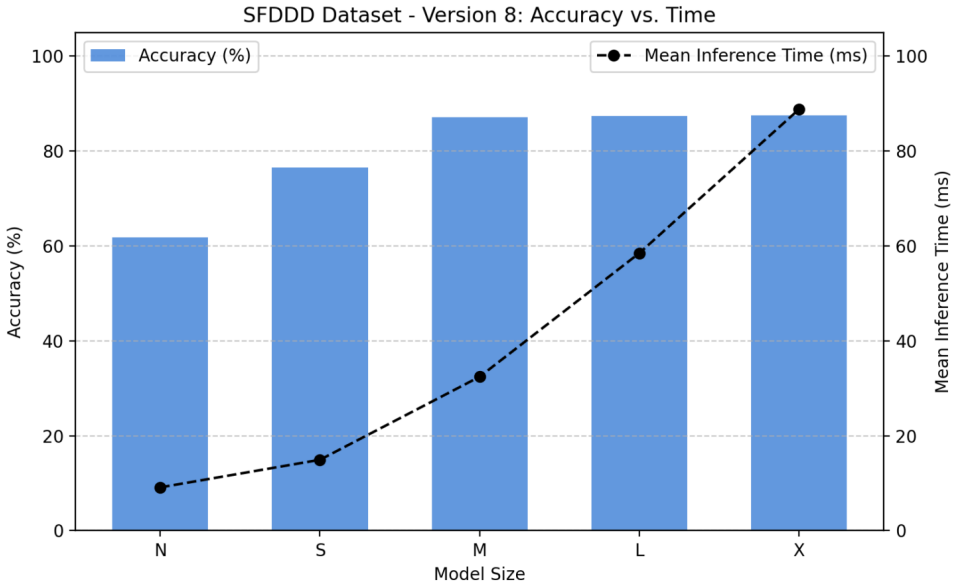


Figure 5.3: Relationship between accuracy and mean inference time for different YOLOv8 model sizes.

The Accuracy vs. Time graph (Figure 5.3) is crucial for contextualizing these metrics. A clear pattern of stabilization is observed. The leap in accuracy from the Nano model (62%) to the Medium (87%) is expressive. However, the gain from Medium to Extra-Large is marginal (1 p.p.), whereas the inference time nearly triples (from 33ms to 89ms). In a vehicular scenario, this difference translates to the ability to process frames nearly three times faster—a determining factor in embedded systems that must respond in real-time to distraction events. Thus, YOLOv8-Medium emerges as the most efficient model for continuous driver monitoring, reconciling precision and responsiveness.

5.2.2 YOLOv12

The twelfth version of the architecture maintains the standard trade-offs but introduces structural adjustments that affect its performance progression and the balance between granularities.

Matrix Confusion of Dataset SFDDD for YOLO v12 Nano.
Sample size: 22424

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 1404	FN 7852	Recall 15.17%
	N	FP 71	TN 13097	Specificity 99.46%
		Precision 95.19%	NPV 62.52%	Accuracy 64.67%

(a) Nano

Matrix Confusion of Dataset SFDDD for YOLO v12 Extra-Large.
Sample size: 22424

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 7218	FN 2038	Recall 77.98%
	N	FP 1422	TN 11746	Specificity 89.20%
		Precision 83.54%	NPV 85.21%	Accuracy 84.57%

(b) Extra-Large

Figure 5.4: Visual representation of the confusion matrix for different YOLOv12 model sizes. (a) YOLOv12 nano. (b) YOLOv12 extra-large.

The YOLOv12 Nano (Figure 5.4) reinforces an even more conservative stance than its predecessor. Its high precision of 95.19% makes it the most reliable model among all when it

issues a positive detection. In a traffic context, this means that each alert issued would have a very high probability of corresponding to a real distraction. However, its recall of only 15.17% indicates that the majority of distracted drivers would go unnoticed. Such a profile is more appropriate for auditing systems or generating post-processed reports where false alarms are unacceptable, rather than real-time safety applications.

Conversely, the YOLOv12 Extra-Large expands the recall to 77.98%, allowing it to identify a significant portion of drivers using cell phones, but at the cost of reducing precision to 83.54%. Similar to the YOLOv8 Extra-Large, the model becomes more "reactive" but also more prone to signaling false alarms—which can be problematic in unsupervised automated systems. Its inference time (118 ms) reinforces this practical limitation, making it more suitable for processing on servers rather than directly in vehicles.

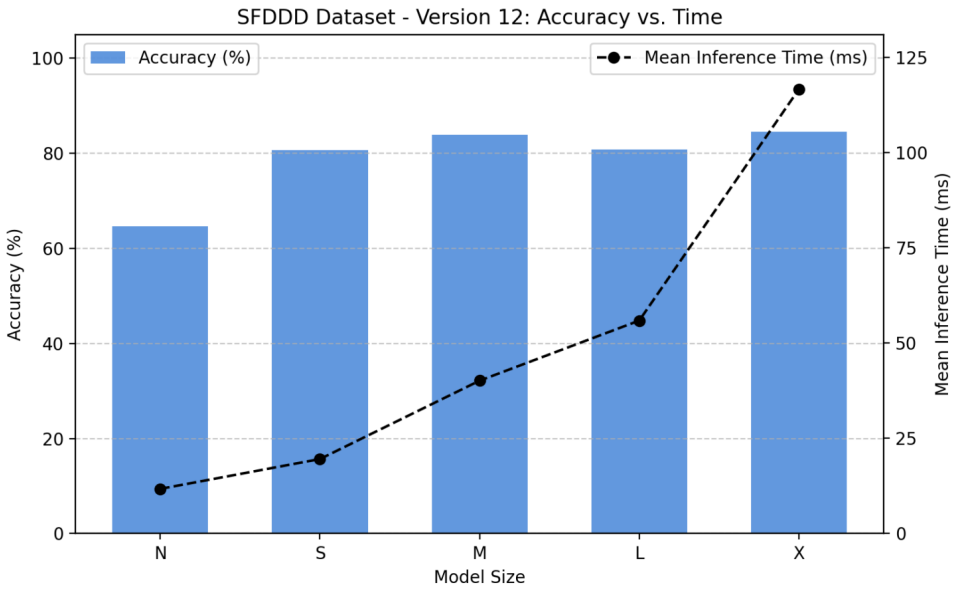


Figure 5.5: Relationship between accuracy and mean inference time for different YOLOv12 model sizes.

The Accuracy vs. Time graph (Figure 5.5) shows that the pattern of diminishing returns is even more pronounced than in the previous version. In practice, the leap in accuracy from Nano (65%) to Small (80%) represents a real advancement in distraction detection, but from the Medium model onward, the gains become minimal compared to the increase in latency. This behavior limits the use of YOLOv12 in real-time embedded applications, such as driver-assistance systems, although it remains viable for centralized fleet video analysis platforms.

5.2.3 Analysis of the Medium Granularity

The direct comparison between the extremes of the YOLOv8 and YOLOv12 versions clearly highlights the role of this intermediate granularity as a balance point between performance and computational cost.

Matrix Confusion of Dataset SFDDD for YOLO v8 Medium.
Sample size: 22424

		Predicted		
		P	N	
Ground Truth	P	TP 7362	FN 1894	Recall 79.54%
	N	FP 993	TN 12175	Specificity 92.46%
		Precision 88.11%	NPV 86.54%	Accuracy 87.13%

(a)

Matrix Confusion of Dataset SFDDD for YOLO v12 Medium.
Sample size: 22424

		Predicted		
		P	N	
Ground Truth	P	TP 6851	FN 2405	Recall 74.02%
	N	FP 1197	TN 11971	Specificity 90.91%
		Precision 85.13%	NPV 83.27%	Accuracy 83.94%

(b)

Figure 5.6: Visual representation of the confusion matrix for the Medium granularity. (a) YOLOv8 Medium. (b) YOLOv12 Medium.

The YOLOv8-Medium achieves 87.13% accuracy, with a recall of 79.54%—meaning it detects approximately 8 out of 10 drivers who are actually using a cell phone—while maintaining a precision of 88.11%. This combination makes it ideal for driver-assistance scenarios, where the system must react quickly to risky behaviors without causing "alarm fatigue" due to excessive false positives. Its mean inference time (33ms) reinforces the feasibility of embedded use, allowing for almost instantaneous alerts to the driver.

The YOLOv12-Medium, in turn, presents more modest metrics, with lower accuracy (83.94%) and recall (74.02%). Although it still operates in real-time (40ms), its overall performance is inferior to the YOLOv8-Medium, especially in situations with varying illumination or intense vehicle movement—conditions typical of the road environment.

In summary, the Medium granularity solidifies itself as the optimal point for both families, but with a clear advantage for the YOLOv8-Medium, which delivers a superior balance between precision, response time, and practical applicability in traffic.

5.2.4 Comparative Considerations

The direct comparison between the YOLOv8 and YOLOv12 families highlights substantial differences in both performance and operational behavior, especially when contextualized within the problem of detecting cell phone use while driving.

In terms of detection reliability, both families perform well, but with important nuances. YOLOv12 tends to be more conservative, achieving higher precision in its smaller granularities (95.19% in YOLOv12-Nano vs 82.04% in YOLOv8-Nano). This makes YOLOv12-Nano ideal for applications requiring maximum trust in positive predictions, such as confirmed event logging or post-processed audits. In practical road scenarios, this high precision means that virtually all alerts from the model are true—useful when false positives incur a high cost. However, the disadvantage is clear: a large portion of actual cell phone use goes unnoticed due to low recall (15.17%), making the model unsuitable for preventive or assistance systems that must actively reduce risky behaviors.

On the other hand, YOLOv8 exhibits a slightly lower precision but one that is more balanced with recall, offering a more proportional and adaptive response for continuous monitoring scenarios. In particular, YOLOv8 is clearly more aggressive in detection: its Extra-Large model achieves the highest recall (83.50%) among all evaluated models—that is, it identifies more than 8 out of 10 real occurrences of cell phone use—surpassing the YOLOv12-Extra-Large (77.98%). At the same time, YOLOv8-Extra-Large maintains high precision (83.54%). This combination ensures not only more correct detections but also a stable system, an important condition when balancing reactivity and user experience.

In terms of computational cost-benefit, YOLOv8 demonstrates a consistent advantage. The eighth-generation Medium model achieves an accuracy of 87.12% with a mean inference time of only 33 ms. YOLOv12, to reach its maximum performance, requires the Extra-Large model, which is slower (inference time 118 ms) and less energy-efficient. In vehicular environments, this difference translates into practical decisions: YOLOv8-Medium allows for processing more frames per second, shorter delays between occurrence and alert, and greater feasibility for deployment on embedded hardware.

Finally, the global efficiency analysis indicates that YOLOv8 offers a more favorable trade-off in practically all evaluated dimensions: its architecture presents a better balance between recall, precision, and inference time, making it more suitable for embedded applications and real-time alert systems. The YOLOv8-Medium, in particular, stands out as the overall optimal point.

Metric	**Nano**		**Medium**		**Extra-Large**	
	v8	**v12**	**v8**	**v12**	**v8**	**v12**
Recall	9.57%	15.17%	79.54%	74.02%	83.50%	77.98%
Specificity	98.53%	99.46%	92.46%	90.91%	90.34%	89.20%
Precision	82.04%	95.19%	88.11%	85.13%	83.54%	83.54%
NPV	60.79%	62.52%	86.54%	83.27%	88.62%	85.21%
Accuracy	61.81%	64.67%	87.13%	83.94%	87.52%	84.57%
Inference Time (ms)	9.13	11.73	32.43	40.13	88.78	116.63

Table 5.1: Comparison of Metrics by Model Granularity (v8 vs. v12) considering Nano, Medium, and Extra-Large in the SFDDD dataset.

5.3 NDS-BR ANALYSIS

The application of the YOLO family models to the Naturalistic Driving Study Brazil (NDS-BR) marks a fundamental transition in this research: from analysis in a controlled experimental environment—like the SFDDD—to a naturalistic and unsupervised context, in which the capture conditions reflect the reality of daily driving. In this scenario, the driver is not guided or monitored, the cameras are subject to vibrations, reflections, irregular angles, and abrupt variations in illumination, and human behavior occurs spontaneously. This complexity imposes new challenges for computer vision and allows for a true evaluation of the practical limits of modern architectures in detecting cell phone use in traffic.

Figure 5.7: Example of an image labeled as safe driving in the NDS-BR dataset.

In addition to the difficulties inherent to the environment, it is important to emphasize that the classification of the NDS-BR base was carried out manually, without a double-checking process. This means the annotations are subject to human error—ambiguous interpretations of gestures, partial framings, reflections that mimic devices, among others. In the example shown (see Figure 5.7), although the driver is not actively using a phone—as defined by the NDS-BR annotation protocol—two devices appear in the scene: one is clearly visible in the cabin, and another lies very close to the driver's hand. From a visual standpoint, the presence and positioning of the objects make the detection not only plausible but correct from the classifier's perspective. The discrepancy arises because the ground truth label marks the frame as "safe driving", since the mere presence of a phone does not constitute use in the dataset definition.

Thus, a significant part of the observed performance drop is not merely due to architectural limitations, but rather stems from the fundamental methodological challenge of inferring behavior solely through object detection. While the YOLO models successfully detect the physical presence of a cell phone, this detection alone is often insufficient to confirm active usage or cognitive distraction—a nuance that is the hallmark of the NDS-BR dataset. Unlike controlled datasets where the presence of a phone implies usage, the naturalistic environment presents scenarios where the device is visible but not being used (e.g., resting on a seat or dashboard). Therefore, the "error" in classification reflects a semantic gap: the model correctly identifies the object, but the behavioral label requires a contextual understanding that goes beyond simple bounding box detection. In an applied research context, this finding is valuable, as it highlights that for naturalistic data, model sophistication must be paired with behavioral logic to bridge the gap between object detection and distraction recognition.

5.3.1 YOLOv8

The YOLOv8 family is examined here through four representative granularities—Nano, Medium, Large, and Extra-Large. These variants were selected because they mark the most expressive

transitions within the model series: each step introduces a meaningful gain in predictive performance (Figure 5.8) accompanied by a measurable increase in computational cost. This selection strategy provides a clearer perspective on how accuracy and latency scale together.

Matrix Confusion of Dataset NDS for YOLO v8 Nano.
Sample size: 25546

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 3039	FN 9623	Recall 24.00%
	N	FP 439	TN 12445	Specificity 96.59%
		Precision 87.38%	NPV 56.39%	Accuracy 60.61%

(a) Nano

Matrix Confusion of Dataset NDS for YOLO v8 Medium.
Sample size: 25546

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 4737	FN 7925	Recall 37.41%
	N	FP 1028	TN 11856	Specificity 92.02%
		Precision 82.17%	NPV 59.94%	Accuracy 64.95%

(b) Medium

Matrix Confusion of Dataset NDS for YOLO v8 Large.
Sample size: 25546

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 6584	FN 6078	Recall 52.00%
	N	FP 1753	TN 11131	Specificity 86.39%
		Precision 78.97%	NPV 64.68%	Accuracy 69.35%

(c) Large

Matrix Confusion of Dataset NDS for YOLO v8 Extra-Large.
Sample size: 25546

		Predicted		
		P	**N**	
Ground Truth	**P**	TP 6946	FN 5716	Recall 54.86%
	N	FP 1884	TN 11000	Specificity 85.38%
		Precision 78.66%	NPV 65.81%	Accuracy 70.25%

(d) Extra-Large

Figure 5.8: Visual representation of the confusion matrix for different granularities of YOLOv8 on NDS-BR.

The YOLOv8-Nano (Figure 5.8), with a mean inference time of just 9.47 ms (Figure 5.9), allows for processing over 100 frames per second—excellent performance for low-power embedded hardware. However, its recall of 24% severely limits its practical utility. In a vehicle, a system based on this model would hardly detect cell phone use reliably; even if it operated with high precision, its low recall would render it ineffective in preventive contexts.

The YOLOv8-Medium (Figure 5.8) shows a significant improvement, reaching 37.41% recall and 64.95% accuracy. This balance makes it suitable for passive alert systems, such as monitoring in corporate fleets, where distraction events are logged for later analysis. Its mean inference time 37.66 ms (Figure 5.9) allows processing around 25 to 30 frames per second, sustaining real-time operations in conventional embedded units.

The YOLOv8-Large (Figure 5.8) emerges as the best representative of the family on NDS-BR. Its recall reaches 52%, with 69.35% accuracy and an inference time of 55.89 ms (Figure 5.9). Despite generating a higher number of false positives compared to smaller models, the YOLOv8-Large offers the best trade-off for practical field use: it combines operational speed and reactivity without compromising system stability.

Finally, the YOLOv8-Extra-Large (Figure 5.8) shows marginally superior performance in recall (54.86%), but with a computational cost that is almost 60% higher (86.74 ms mean latency)(Figure 5.9). This difference reduces its embedded applicability, making it more appropriate for centralized or post-processed analyses, where latency is not a critical factor.

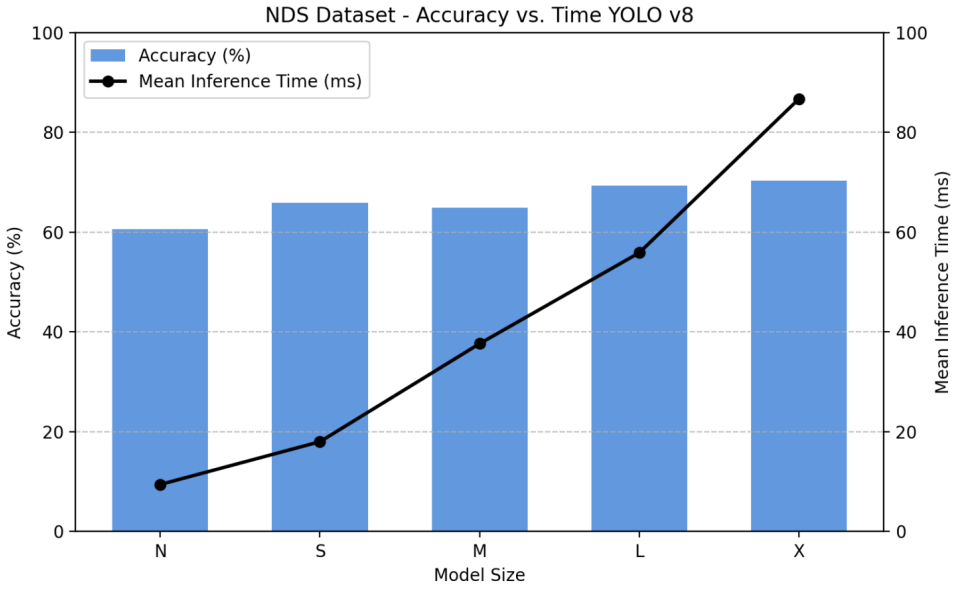


Figure 5.9: Relationship between accuracy and mean inference time for different YOLOv8 model sizes.

5.3.2 YOLOv12

As with YOLOv8, the analysis of the YOLOv12 family focuses on the Nano, Medium, Large, and Extra-Large variants. These configurations represent meaningful jumps in both accuracy and inference time, capturing the most relevant trade-offs across the architecture (Figure 5.10). By concentrating on these four granularities, the evaluation highlights how structural refinements introduced in YOLOv12 impact performance under the challenging naturalistic conditions of NDS-BR.

Matrix Confusion of Dataset NDS for YOLO v12 Nano.
Sample size: 25546

		Predicted P	Predicted N	
Ground Truth	P	TP 2592	FN 10070	Recall 20.47%
	N	FP 498	TN 12386	Specificity 96.13%
		Precision 83.88%	NPV 55.16%	Accuracy 58.63%

(a) Nano

Matrix Confusion of Dataset NDS for YOLO v12 Medium.
Sample size: 25546

		Predicted P	Predicted N	
Ground Truth	P	TP 4884	FN 7778	Recall 38.57%
	N	FP 868	TN 12016	Specificity 93.26%
		Precision 84.91%	NPV 60.71%	Accuracy 66.16%

(b) Medium

Matrix Confusion of Dataset NDS for YOLO v12 Large.
Sample size: 25546

		Predicted P	Predicted N	
Ground Truth	P	TP 7771	FN 4891	Recall 61.37%
	N	FP 1713	TN 11171	Specificity 86.70%
		Precision 81.94%	NPV 69.55%	Accuracy 74.15%

(c) Large

Matrix Confusion of Dataset NDS for YOLO v12 Extra-Large.
Sample size: 25546

		Predicted P	Predicted N	
Ground Truth	P	TP 5901	FN 6761	Recall 46.60%
	N	FP 1189	TN 11695	Specificity 90.77%
		Precision 83.23%	NPV 63.37%	Accuracy 68.88%

(d) Extra-Large

Figure 5.10: Visual representation of the confusion matrix for different granularities of YOLOv12 on NDS-BR.

The YOLOv12-Nano (Figure 5.10), while sharing the conservative profile of the previous version, is characterized by high precision but low recall (20.47%). Thus, the YOLOv12-Nano is better suited for analytical or validation scenarios, where the priority is the reliability of detections, not the total coverage of events.

The YOLOv12-Medium (Figure 5.10) proves to be more adapted to the unpredictability of traffic. With 38.57% recall and an inference time of 41.11 ms (Figure 5.11), it combines precision and speed in a balanced way. Its real-time operation is fully viable, and its performance makes it suitable for hybrid monitoring systems. In conditions such as urban roads with heavy traffic or direct sunlight reflections, the model manages to preserve the coherence of its predictions, demonstrating superior stability to the YOLOv8-Medium.

The YOLOv12-Large (Figure 5.10), in turn, represents the overall of balance in the series. With 61.37% recall and high precision (81.94%), it reconciles a good detection rate with a low false positive rate, all while maintaining a competitive inference time (54.80 ms)(Figure 5.11). In practice, this is a model capable of identifying more than half of all real-world cell phone use cases while driving, with reliability above 90%, and operating at sufficient speed for real-time embedded applications.

The YOLOv12-Extra-Large (Figure 5.10), while preserving high values of precision, shows a slight reduction in recall (46.60%) and an increase in latency (104.26 ms)(Figure 5.11). This behavior shifts it towards offline applications, such as video review and behavioral analysis in large databases.

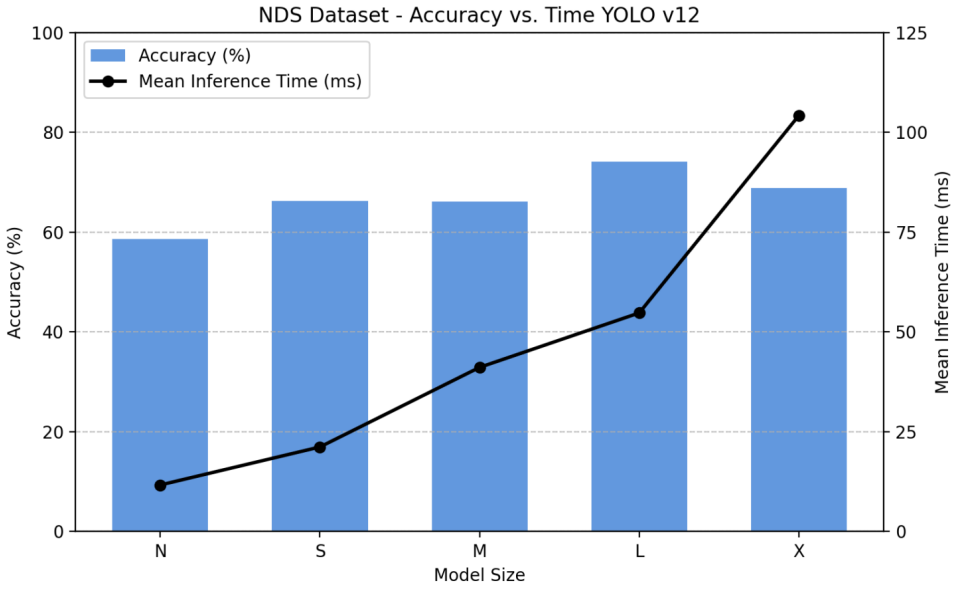


Figure 5.11: Relationship between accuracy and mean inference time for different YOLOv12 model sizes.

5.3.3 General Architectural Behavior

Overall, a reduction in absolute metrics is observed compared to the SFDDD, with accuracies ranging from approximately 58.63% (v12-Nano) to 74.15% (v12-Large) and recall rates between 20.47% (v12-Nano) and 61.37% (v12-Large). This drop was expected, given the transition from a controlled domain to a real-world environment, but it still maintains the pattern of evolution: performance improves as granularity increases, albeit with progressively smaller gains.

In the smaller granularities, such as YOLOv8-Nano and YOLOv12-Nano, the models maintain high precision (87.38% and 83.88%, respectively) but exhibit extremely low recall

(24.00% and 20.47%). In practical terms, they behave as extremely cautious classifiers—they only detect when there is high visual certainty of cell phone use. This profile makes them suitable for conservative auditing applications where the cost of false alarms is high. However, in the road context, their low reactivity makes them unviable for use in embedded alert systems, which require much higher recall to effectively warn drivers.

As we move to the medium and large granularities, a substantial gain in recall is noted. For instance, YOLOv12-Large achieves a recall of 61.37%, a significant increase from YOLOv8-Nano's 24.00%. This trend reflects the expected behavior of deep networks in noisy scenarios: more complex models begin to identify subtle patterns associated with cell phone use—such as a tilted head position or an elevated arm—but they also become more vulnerable to false alarms when neutral gestures resemble such patterns. This dialectic between "seeing more" (recall) and "erring less" (precision) is essential in traffic: in an embedded alert system, it is often preferable to detect a larger number of distractions, provided the precision remains within acceptable limits.

5.3.4 Comparative Considerations

The comparison between the YOLOv8 and YOLOv12 families on the NDS-BR reinforces a coherent picture, but with important nuances. The eighth generation maintains an advantage in efficiency and computational stability, presenting satisfactory performance even on limited hardware. The twelfth generation, resulting from architectural refinements, shows better adaptation to the real-world environment, with visible gains in recall and precision under varying conditions.

In particular, the YOLOv12-Large stands out as the overall optimal point. It combines balanced recall (61.37%) with precision above 81.94% and latency below 54.76 ms (Inference Time), establishing itself as the most suitable model for embedded applications seeking a rapid and reliable response. In practical terms, this means a detection system based on this model could operate in real-time inside commercial or private vehicles, identifying cell phone use with a good success rate and a reduced risk of false alarms—even in adverse conditions, such as vibration, heavy traffic, or nighttime illumination.

The general drop in metrics compared to the SFDDD should not be interpreted merely as a fragility of the models, but also as a reflection of the quality and consistency of the annotations. As the NDS-BR was labeled manually without double-checking, part of the discrepancies between prediction and label may derive from ambiguous human classifications—which, paradoxically, brings the results closer to reality. After all, in the field, the boundaries between "cell phone use" and "neutral behavior" are diffuse even for human observers.

Thus, the NDS-BR reveals more than just performance degradation: it exposes the fragility of human interpretation in a visually ambiguous problem, and highlights that the challenge of distraction detection in traffic is not only technical but also semantic. The performance of the YOLO models, especially in the Large granularities, shows that even between uncertainties, it is possible to reach a level of reliability high enough to support road safety decisions.

In summary, YOLOv12-Large establishes itself as the most balanced model for naturalistic scenarios (Accuracy: 74.15%), while YOLOv8-Large remains the best alternative for embedded environments with hardware constraints (Accuracy: 69.35%). Both represent different stages of the same advancement: the growing capacity of computer vision systems to understand and react to human behavior in real-time—not just as a technical exercise, but as a concrete tool to reduce risk and save lives in traffic.

Metric	Nano		Medium		Large		Extra-Large	
	v8	v12	v8	v12	v8	v12	v8	v12
Recall	24.00%	20.47%	37.41%	38.57%	52.00%	61.37%	54.86%	46.60%
Specificity	96.59%	96.13%	92.02%	93.26%	86.39%	86.70%	85.38%	90.77%
Precision	87.38%	83.88%	82.17%	84.91%	78.97%	81.94%	78.66%	83.23%
NPV	56.39%	55.16%	59.94%	60.71%	64.68%	69.55%	65.81%	63.37%
Accuracy	60.61%	58.63%	64.95%	66.16%	69.35%	74.15%	70.25%	68.88%
Inference Time (ms)	9.44	11.67	37.65	41.13	55.86	54.76	86.69	104.21

Table 5.2: Comparison of Metrics by Model Granularity (v8 vs. v12) considering Nano, Medium, Large, and Extra-Large in the NDS-BR dataset.

5.4 TEMPORAL AGGREGATION ANALYSIS ON NDS-BR

Following the methodological framework detailed in Section 4.3, the second phase of the NDS-BR experiments shifts from the analysis of static, independent images to a continuous temporal evaluation of the video sequences. This approach leverages the sequential nature of the data to mitigate the casual noise inherent in individual frame predictions, such as motion blur, lighting changes, or momentary occlusions.

5.4.1 Baseline Frame-Level Analysis

The initial evaluation was conducted on a strict frame-by-frame basis, where every individual frame in the video test set was treated as an independent sample. No temporal smoothing or aggregation was applied at this stage. This establishes a baseline performance metric for the raw model inference capabilities on the video stream. The results for the Nano and Extra-Large granularities of both YOLOv8 and YOLOv12 are summarized in Figure 5.12.

Confusion Matrix (Frames - 8n Videos)
Sample Size: 446992 frames

		Predicted P	Predicted N	
Ground Truth	P	TP 55421	FN 143714	Recall 27.83%
	N	FP 5771	TN 242086	Specificity 97.67%
		Precision 90.57%	NPV 62.75%	Accuracy 66.56%

(a) YOLOv8-Nano

Confusion Matrix (Frames - 8x Videos)
Sample Size: 446992 frames

		Predicted P	Predicted N	
Ground Truth	P	TP 116992	FN 82143	Recall 58.75%
	N	FP 29029	TN 218828	Specificity 88.29%
		Precision 80.12%	NPV 72.71%	Accuracy 75.13%

(b) YOLOv8-Extra-Large

Confusion Matrix (Frames - 12n Videos)
Sample Size: 446992 frames

		Predicted P	Predicted N	
Ground Truth	P	TP 42522	FN 156613	Recall 21.35%
	N	FP 6149	TN 241708	Specificity 97.52%
		Precision 87.37%	NPV 60.68%	Accuracy 63.59%

(c) YOLOv12-Nano

Confusion Matrix (Frames - 12x Videos)
Sample Size: 446992 frames

		Predicted P	Predicted N	
Ground Truth	P	TP 108937	FN 90198	Recall 54.71%
	N	FP 17886	TN 229971	Specificity 92.78%
		Precision 85.90%	NPV 71.83%	Accuracy 75.82%

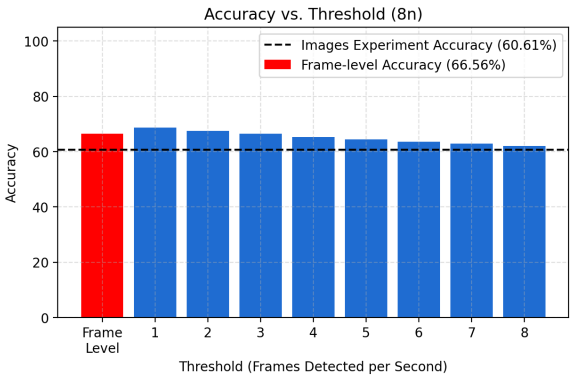
(d) YOLOv12-Extra-Large

Figure 5.12: Confusion Matrix comparing YOLOv8 and YOLOv12 in the frame-by-frame analysis.

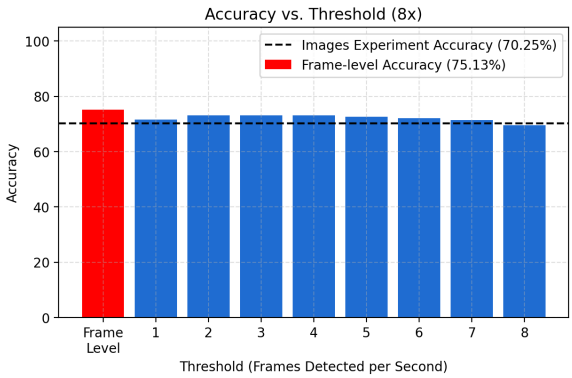
A critical observation from this baseline is the very high rate of False Negatives (FN), particularly in the Nano models. YOLOv12-Nano, for example, achieved a Recall of only 21.35%, missing the vast majority of distraction frames. Even the more robust models, such as YOLOv8-Extra-Large and YOLOv12-Extra-Large, still show modest Recalls of 58.75% and 54.71%, respectively. This suggests that while the models are highly precise when they detect an object, the frame-to-frame instability results in many "dropped" detections during a continuous distraction event.

5.4.2 Threshold Recall and Optimization

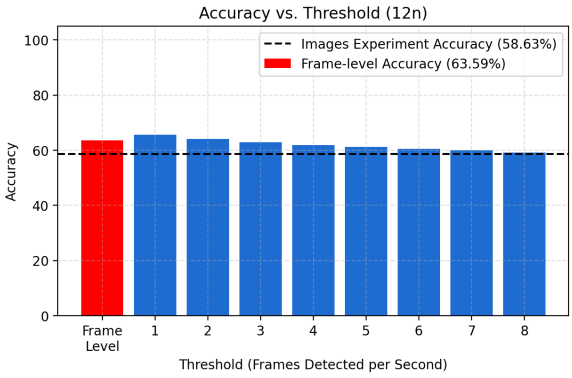
To address the instability observed in the frame-level analysis, a temporal aggregation strategy was applied. This method groups frames into one-second windows, classifying the entire second as "distracted" if the number of positive frames exceeds a threshold T. To determine the optimal operating point, a recall analysis was conducted by varying T from 1 to 8 frames. The upper bound of $T = 8$ was selected based on a preliminary analysis of the NDS-BR ground truth, which indicated that valid distraction events in the dataset contained a minimum of 8 positive frames, typically varying between 8 and 25 frames per second.



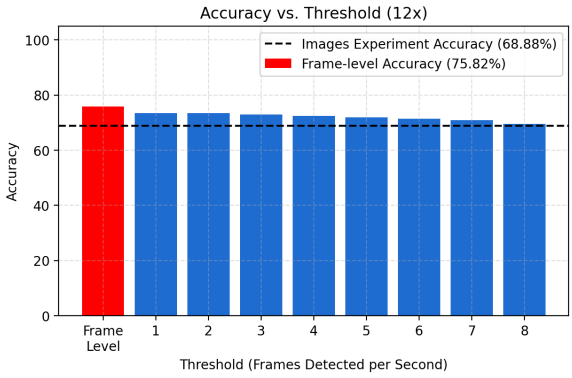
(a) YOLOv8-Nano



(b) YOLOv8-Extra-Large



(c) YOLOv12-Nano



(d) YOLOv12-Extra-Large

Figure 5.13: Accuracy comparing YOLOv8 and YOLOv12 in the threshold aggregation analysis.

The relationship between the aggregation threshold and overall accuracy is visualized in Figure 5.13. In these charts, the red bar represents the baseline frame-level accuracy established in the previous section. A crucial pattern emerges based on the model granularity.

For Nano Models (YOLOv8n and YOLOv12n), temporal aggregation with a low threshold ($T = 1$) consistently outperforms the frame-level baseline accuracy. For instance,

YOLOv8-Nano (Figure 5.13(a)) peaks at 68.71% with $T = 1$, significantly surpassing its baseline of 66.56%. Similarly, YOLOv12-Nano reaches 65.60% at $T = 1$, superior to its 63.59% baseline. This confirms that a lenient aggregation strategy is essential to smooth out the high False Negative instability inherent in faster models, leveraging temporal context to increase overall accuracy.

Conversely, in more robust Extra-Large Models (YOLOv8x and YOLOv12x), the raw disaggregated Frame Level maintains the highest accuracy. YOLOv8-Extra-Large achieves 75.13% at the frame level, which is higher than its aggregated peak of 73.20% ($T = 3$). YOLOv12-Extra-Large shows the best overall performance with 75.82% at the frame level, surpassing the aggregated peak of 73.54% ($T = 2$).

Nevertheless, for all models, both the aggregated scenario ($T = 1$) and the disaggregated (Frame Level) consistently demonstrate superior accuracy compared to the independent static image analysis (represented by black dashed line). The trend of accuracy declining as the threshold becomes stricter (requiring more frames to trigger a detection) is maintained across all architectures, highlighting that stricter thresholds tend to increase False Negatives and decrease overall accuracy.

5.4.3 Aggregation Performance with Threshold $T = 1$

Based on the recall analysis, the threshold of $T = 1$ was identified as the optimal configuration for maximizing performance in the Nano-granularity models, and serves as a strong point of comparison for the larger architectures. This setting defines a second as "distracted" if at least one frame within that second is classified as positive. This approach prioritizes recall, recovering events that might be fragmented by noise in a strict frame-level view.

Confusion Matrix (Per-Second Analysis - 8n Videos)
Sample Size: 21950 seconds
Threshold: 1 frame detected

		Predicted P	Predicted N	
Ground Truth	P	TP 4847	FN 5825	Recall 45.42%
	N	FP 1044	TN 10234	Specificity 90.74%
		Precision 82.28%	NPV 63.73%	Accuracy 68.71%

(a) YOLOv8-Nano

Confusion Matrix (Per-Second Analysis - 8x Videos)
Sample Size: 21950 seconds
Threshold: 1 frame detected

		Predicted P	Predicted N	
Ground Truth	P	TP 7556	FN 3116	Recall 70.80%
	N	FP 3109	TN 8169	Specificity 72.43%
		Precision 70.85%	NPV 72.39%	Accuracy 71.64%

(b) YOLOv8-Extra-Large

Confusion Matrix (Per-Second Analysis - 12n Videos)
Sample Size: 21950 seconds
Threshold: 1 frame detected

		Predicted P	Predicted N	
Ground Truth	P	TP 3965	FN 6707	Recall 37.15%
	N	FP 844	TN 10434	Specificity 92.52%
		Precision 82.45%	NPV 60.87%	Accuracy 65.60%

(c) YOLOv12-Nano

Confusion Matrix (Per-Second Analysis - 12x Videos)
Sample Size: 21950 seconds
Threshold: 1 frame detected

		Predicted P	Predicted N	
Ground Truth	P	TP 6553	FN 4119	Recall 61.40%
	N	FP 1691	TN 9587	Specificity 85.01%
		Precision 79.49%	NPV 69.95%	Accuracy 73.53%

(d) YOLOv12-Extra-Large

Figure 5.14: Confusion Matrix comparing YOLOv8 and YOLOv12 in the aggregation analysis (threshold=1).

The confusion matrices for this aggregated scenario ($T = 1$) are presented in Figure 5.14. For Nano Models, compared to the frame-level baseline data, YOLOv8-Nano saw an accuracy increase of 2.15 percentage points (from 66.56% to 68.71%). YOLOv12-Nano exhibited a recovery of 2.01 percentage points (from 63.59% to 65.60%). These gains confirm that temporal

aggregation with $T = 1$ is a crucial post-processing step for lighter models, as it effectively filters the instability of individual frame predictions and allows these architectures to achieve higher operational reliability.

For the Extra-Large architectures, although the accuracy with $T = 1$ is slightly lower than the frame-level baseline performance (YOLOv8x: 71.64% on $T = 1$ vs 75.13% on Frame-Level; YOLOv12x: 73.53% on $T = 1$ vs 75.82% on Frame-Level), the aggregation still surpasses the accuracy of the independent static analysis. More importantly, by utilizing the disaggregated frame level for Extra-Large models, it is possible to maintain a high baseline accuracy of 75.82% (YOLOv12x Frame Level), while ensuring that the aggregable performance ($T = 1$) serves as a strong foundation for systems requiring high recall.

5.5 FINAL COMPARATIVE SYNTHESIS

The experimental research conducted in this study—spanning from the controlled environment of the SFDDD to the unconstrained naturalistic setting of the NDS-BR, and finally to the temporal analysis of video frames—allows for a comprehensive definition of the operational limits and capabilities of the YOLOv8 and YOLOv12 architectures.

In the controlled domain (SFDDD), the results indicate that modern object detection models have reached a saturation point regarding raw feature extraction. The YOLOv8-Medium emerged as the most efficient scenario, offering a detection accuracy of 87.13% with a latency of only 32.43 ms. This confirms that, in ideal lighting and camera positioning, excessive model depth (Large/Extra-Large) yields constant returns.

However, the transition to the NDS-BR dataset exposed the fragility of these models when facing the domain gap. The drop in accuracy to the 58%–74% range highlights that generalization without fine-tuning is limited by the complexity of real-world driving behavior. Here, the architectural refinements of YOLOv12 proved valuable, with the Large version offering the best balance of robustness (74.15% accuracy) against environmental noise in the static analysis.

Finally, the temporal analysis demonstrated that the random nature of frame-by-frame inference can be effectively mitigated. For models of Nano granularity, applying a low aggregation threshold ($T = 1$) proved crucial, recovering performance and surpassing the disaggregated frame scenario (YOLOv8n reaching 68.71%).

For models of Extra-Large granularity, the disaggregated Frame Level performance achieved the highest accuracy (75.82% with YOLOv12-Extra-Large), outperforming both the aggregated scenario and the independent static analysis. In all cases, the use of video data, whether aggregated or disaggregated, leads to superior performance compared to the analysis of independent images.

This suggests that for video-based applications, temporal context—either through the inherent stability of larger models or through aggregation post-processing in smaller models—is fundamental. Table 5.3 summarizes the best-performing configurations for each experimental scenario, establishing a reference baseline for future developments.

Scenario	**Dataset**	**Best Model**	**Acc (%)**	**Recall (%)**	**Time (ms)**	**Operational Profile**
Controlled Static	SFDDD	YOLOv8-Medium	87.13	79.54	32.43	**Optimal Efficiency**: Best trade-off in an ideal environment.
Naturalistic Static	NDS-BR	YOLOv12-Large	74.15	61.37	54.76	**Robustness**: Best adaptation to lighting/noise.
Naturalistic Video	NDS-BR	YOLOv12-Extra-Large (Frame Level)	75.82	54.71	104.21	**Peak Accuracy**: Best performance in video analysis.

Table 5.3: Summary of the best-performing models across all experimental conditions.

6 CONCLUSION

This study contributes to the field of computer vision applied to intelligent transportation systems (ITS) primarily through a zero-shot evaluation of state-of-the-art architectures. By comparing the recently released YOLOv12 against the established YOLOv8 without task-specific fine-tuning, we highlighted the raw generalization capabilities of these models. The analysis demonstrated that while architectural improvements in version 12 provide gains in precision and specificity, the lack of domain-specific training remains a bottleneck for recall in naturalistic environments.

Additionally, the work establishes a crucial benchmark on the NDS-BR dataset, exposing the semantic gap present in current data sources. We identified that the discrepancy between visual detection—such as the physical presence of a phone—and the ground truth labeling, which may classify the scene as "safe driving" despite the object's presence, constitutes a major challenge. This insight provides a realistic baseline for researchers working with non-curated, real-world driving data, emphasizing that data quality is as determinant as model sophistication.

Finally, the research empirically validates temporal aggregation strategies. We demonstrate that a low-threshold aggregation ($T = 1$) is sufficient to stabilize and increase the accuracy of Nano models. Conversely, for more robust models, the disaggregated frame-level performance achieves maximum accuracy, but aggregation ($T = 1$) still surpasses static analysis. This finding is particularly relevant for resource-constrained embedded systems, as it enables the use of smaller, faster models to achieve reliability comparable to larger architectures by leveraging temporal redundancy.

The results presented in this chapter establish a baseline for the capabilities of pre-trained models, yet the performance gap observed between the controlled SFDDD environment and the naturalistic NDS-BR suggests that the next evolutionary step lies in specialized training. Future investigations should focus on fine-tuning and transfer learning specifically on the NDS-BR dataset.

The central hypothesis is that by retraining the weights of the YOLOv8 and YOLOv12 backbones using the specific visual distribution of the NDS-BR—including cabin geometry, local lighting conditions, and camera angles—it is possible to overcome the current performance ceiling. Specifically, future efforts should aim to supervise the fine-tuning of the best-performing architectures identified in this study, namely YOLOv8-Medium and YOLOv12-Large, using the NDS-BR training partition to align the model's feature extraction with the specific definitions of distraction in this dataset.

Concurrent with this technical adjustment, it is crucial to address label ambiguity by implementing a data cleaning or "soft-labeling" pipeline. This would handle the ambiguities identified in the analysis, potentially improving the model's ability to distinguish between mere object presence and actual usage. Ultimately, these steps aim to surpass the zero-shot benchmarks presented in this chapter, quantifying the gain provided by transfer learning to develop a specialized model that exceeds the 75% accuracy barrier observed.

REFERENCES

Abouelnaga, Y., Eraqi, H. M., and Moustafa, M. N. (2018). Real-time distracted driver posture classification.

Baheti, B., Gajre, S., and Talbar, S. (2018). Detection of distracted driver using convolutional neural network. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1145–11456.

Bastos, J. T., Santos, P. A. B. d., Amancio, E. C., Gadda, T. M. C., Ramalho, J. A., King, M. J., and Oviedo-Trespalacios, O. (2020). Naturalistic driving study in brazil: An analysis of mobile phone use behavior while driving. International Journal of Environmental Research and Public Health, 17(17).

Cabot, J. H. and Ross, E. G. (2023). Evaluating prediction model performance. Surgery, 174(3):723–726.

CDC (2025). Distracted driving. `https://www.cdc.gov/distracted-driving/about/index.html`. Acessado em 31/07/2025.

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., and Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences, 113(10):2636–2641.

Fu, Y., Ran, T., Xiao, W., Yuan, L., Zhao, J., He, L., and Mei, J. (2024). Gd-yolo: An improved convolutional neural network architecture for real-time detection of smoking and phone use behaviors. Digital Signal Processing, 151:104554.

Ge, T., Ning, B., and Xie, Y. (2025). Yolo-afr: An improved yolov12-based model for accurate and real-time dangerous driving behavior detection. Applied Sciences, 15(11).

GTSC (2025). Distracted driving the number one contributing factor of motor vehicle crashes in new york state. `https://trafficsafety.ny.gov/distracted-driving`. Acessado em 31/07/2025.

Hasan, M. Z., Chen, J., Wang, J., Rahman, M. S., Joshi, A., Velipasalar, S., Hegde, C., Sharma, A., and Sarkar, S. (2024). Vision-language models can identify distracted driver behavior from naturalistic videos. IEEE Transactions on Intelligent Transportation Systems, 25(9):11602–11616.

Hossain, M. U., Rahman, M. A., Islam, M. M., Akhter, A., Uddin, M. A., and Paul, B. K. (2022). Automatic driver distraction detection using deep convolutional neural networks. Intelligent Systems with Applications, 14:200075.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2):1.

Huang, C., Wang, X., Cao, J., Wang, S., and Zhang, Y. (2020). Hcf: A hybrid cnn framework for behavior detection of distracted drivers. IEEE Access, 8:109335–109349.

Jocher, G. (2024). Models supported by ultralytics. https://docs.ultralytics.com/models/. Acessado em 26/10/2025.

Jocher, G. (2025). Yolov5 vs. yolov8: A detailed comparisons. https://docs.ultralytics.com/compare/yolov5-vs-yolov8/. Acessado em 26/10/2025.

Kateb, F., Monowar, M. M., Hamid, M. A., Ohi, A., and Ph. D., M. (2021). Fruitdet: Attentive feature aggregation for real-time fruit detection in orchards. Agronomy, 11:2440.

Khalil, H. A., Hammad, S. A., Abd El Munim, H. E., and Maged, S. A. (2025). Low-cost driver monitoring system using deep learning. IEEE Access, 13:14151–14164.

Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements.

Li, L., Zhong, B., Hutmacher, C., Liang, Y., Horrey, W. J., and Xu, X. (2020). Detection of driver manual distraction via image-based hand and ear recognition. Accident Analysis & Prevention, 137:105432.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Liu, S., Wang, Y., Yu, Q., Liu, H., and Peng, Z. (2022). Ceam-yolov7: Improved yolov7 based on channel expansion and attention mechanism for driver distraction behavior detection. IEEE Access, 10:129116–129124.

Liu, T., Yang, Y., Huang, G.-B., Yeo, Y. K., and Lin, Z. (2016). Driver distraction detection using semi-supervised machine learning. IEEE Transactions on Intelligent Transportation Systems, 17(4):1108–1120.

Liu, Z., Zhu, X., Pang, L., Fu, X., Zhu, H., and Liu, X. (2024). Ava-yolo: image-based multiscale feature fusion enhanced perception model for snow avalanche detection. Measurement Science and Technology, 35.

Ma, Yunsheng, Wang, and Ziran (2024a). Vit-dd: Multi-task vision transformer for semi-supervised driver distraction detection. In 2024 IEEE Intelligent Vehicles Symposium (IV), pages 417–423.

Ma, B., Fu, Z., Rakheja, S., Zhao, D., He, W., Ming, W., and Zhang, Z. (2024b). Distracted driving behavior and driver's emotion detection based on improved yolov8 with attention mechanism. IEEE Access, 12:37983–37994.

Masood, S., Rai, A., Aggarwal, A., Doja, M., and Ahmad, M. (2020). Detecting distraction of drivers using convolutional neural network. Pattern Recognition Letters, 139:79–85.

Meng, S., Yue, Y., and Xu, T. (2025). Enhanced yolov11 framework for accurate multi-fault detection in uav photovoltaic inspection. Sensors, 25(17).

Mohammed, A. A., Geng, X., Wang, J., and Ali, Z. (2024). Driver distraction detection using semi-supervised lightweight vision transformer. Engineering Applications of Artificial Intelligence, 129:107618.

Montoya, A., Holman, D., Smith, T., and Kan, W. (2016). State farm distracted driver detection. https://kaggle.com/competitions/state-farm-distracted-driver-detection. Acessado em 26/10/2025.

Neamah, O. N., Almohamad, T. A., and Bayir, R. (2024). Enhancing road safety: Real-time distracted driver detection using nvidia jetson nano and yolov8. In 2024 Zooming Innovation in Consumer Technologies Conference (ZINC), pages 194–198.

NHTSA (2025). Nhtsa estimates 39,345 traffic fatalities in 2024. https://www.nhtsa.gov/press-releases/nhtsa-2023-traffic-fatalities-2024-estimates. Acessado em 31/07/2025.

Poon, Y.-S., Lin, C.-C., Liu, Y.-H., and Fan, C.-P. (2022). Yolo-based deep learning design for in-cabin monitoring system with fisheye-lens camera. In 2022 IEEE International Conference on Consumer Electronics (ICCE), pages 1–4.

PRF (2025). Anuário 2024. https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/diest-arquivos/anuario-2024_final.html. Acessado em 31/07/2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788.

Sajid, F., Javed, A. R., Basharat, A., Kryvinska, N., Afzal, A., and Rizwan, M. (2021). An efficient deep learning framework for distracted driver detection. IEEE Access, 9:169270–169280.

Tango, F. and Botta, M. (2013). Real-time detection system of driver distraction using machine learning. IEEE Transactions on Intelligent Transportation Systems, 14(2):894–905.

Tian, Y., Ye, Q., and Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., and Ding, G. (2024). Yolov10: Real-time end-to-end object detection.

WHO (2023). Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Acessado em 26/06/2025.

Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., and Wang, F.-Y. (2019). Driver activity recognition for intelligent vehicles: A deep learning approach. IEEE Transactions on Vehicular Technology, 68(6):5379–5390.