applied sciences

MDPI

Systematic Review

Drowsiness Detection in Drivers: A Systematic Review of Deep Learning-Based Models

Tiago Fonseca [1] **and Sara Ferreira** [2,*]

[1] Department of Civil and Georesources Engineering, Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal; tiago.fdafonseca@gmail.com
[2] CITTA—Research Centre for Territory, Transports and Environment, Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
* Correspondence: sara@fe.up.pt; Tel.: +351-22-508-1968

Abstract

Deep learning (DL) models show considerable promise in detecting driver drowsiness, a major contributor to road traffic crashes. This systematic review evaluates the performance, contexts of application, and implementation challenges of DL-based drowsiness detection systems. Conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines, the review includes peer-reviewed empirical studies published between 2015 and 2025 that develop and validate DL models using data collected in real or simulated driving environments. Studies were identified through systematic searches in PubMed, Scopus, Web of Science, ScienceDirect, and IEEE Xplore, last updated in March 2025. Due to methodological heterogeneity, findings are synthesized narratively. Eighty-one studies meet the inclusion criteria. Most employ Convolutional Neural Networks, Recurrent Neural Networks, or hybrid architectures and use behavioral, physiological, or multimodal inputs. Reported median values for accuracy and F1-score exceed 0.95 under both simulated and real-world conditions. However, studies frequently lack demographic diversity, standardized performance reporting, and robust validation protocols. Key limitations include limited dataset transparency, inconsistent evaluation metrics, and insufficient attention to ethical and privacy considerations. While DL models exhibit strong predictive performance, their real-world deployment remains limited by practical and methodological constraints. Future research should place emphasis on the development of inclusive datasets, the conduct of multi-context evaluations, the advancement of real-world deployment strategies, and the rigorous adherence to ethical standards.

Keywords: monitoring system; driver assistance; road safety; fatigue; artificial intelligence; machine learning

Academic Editor: Christos Bouras

Received: 12 July 2025
Revised: 11 August 2025
Accepted: 12 August 2025
Published: 15 August 2025

Citation: Fonseca, T.; Ferreira, S. Drowsiness Detection in Drivers: A Systematic Review of Deep Learning-Based Models. *Appl. Sci.* **2025**, *15*, 9018. https://doi.org/10.3390/app15169018

1. Introduction

Driver drowsiness is a major contributor to road traffic incidents and is widely recognized as a critical public safety issue. Fatigue impairs attentional control, reaction time, and decision-making ability, thereby increasing the risk of crashes and injury severity [1,2]. In the United States alone, drowsy driving was estimated to be a factor in 17.6% of all fatal motor vehicle crashes between 2017 and 2021, resulting in nearly 30,000 deaths [3]. Internationally, the E-Survey of Road Users' Attitudes (ESRA) reported that 18–20% of drivers in Europe, North America, and Asia-Oceania experienced difficulty staying awake while driving in the past month, underscoring the global scale of the problem [4].

To address this risk, researchers have increasingly explored driver monitoring systems designed to detect early signs of drowsiness. Traditional approaches, including self-reported sleepiness, vehicle-based indicators, and isolated physiological signals, often lack the sensitivity, scalability, and adaptability needed for real-world applications [5]. In contrast, deep learning (DL) has emerged as a promising alternative, capable of extracting complex features from diverse input sources such as facial expressions, eye movements, and biosignals [6].

Despite growing interest and progress, the literature remains fragmented. A variety of DL architectures have been proposed for drowsiness detection, but few studies offer comparative evaluations, and no consensus has emerged on optimal model design. Reporting practices also vary widely, with inconsistencies in performance metrics and evaluation protocols limiting comparability. In addition, the datasets used for model development differ in size, signal type, contextual realism, and demographic diversity, raising concerns about generalizability. Practical challenges such as high computational demands, limited integration with vehicle systems, and unresolved privacy issues also hinder real-world deployment.

In light of these challenges, this systematic review synthesizes research on the use of deep learning for driver drowsiness detection. The review is structured around the following research questions (RQ):

RQ1: Which deep learning models are used to detect drowsiness in drivers?

RQ2: How precise and reliable are these deep learning models in detecting drowsiness?

RQ3: What types of datasets are used for training and validating deep learning models?

RQ4: What are the main challenges and limitations in developing deep learning-based drowsiness detection systems?

By addressing these questions, the review aims to provide a comprehensive overview of the field, highlight persistent challenges, and support the development of scalable solutions to enhance driver safety in real-world transportation systems.

This article is structured to provide a clear and systematic presentation of the review process and findings. Section 2 details the methodology, including eligibility criteria, search strategy, screening procedures, and data extraction methods. Section 3 presents the main findings, organized into subsections addressing the characteristics of included studies, evaluation contexts, datasets used, model architectures, performance metrics, and real-world feasibility. Section 4 offers an in-depth discussion of these findings, considers their implications, and identifies gaps and limitations in the existing literature. Section 5 concludes the review by summarizing its key contributions and outlining future research directions to improve the design, evaluation, and implementation of deep learning-based drowsiness detection systems.

2. Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [7]. The PRISMA framework ensures methodological rigor by promoting transparency, reproducibility, and structured reporting. Additional details are available in the PRISMA 2020 checklist (see Supplementary Materials, Document S1). The procedures applied in this review are outlined below.

2.1. Protocol

Prior to initiating the review, a structured protocol was developed to define the objectives, research questions, eligibility criteria, and methodological approach. The protocol was prospectively registered in the International Prospective Register of Systematic Re-

views (PROSPERO) under the reference CRD420251078841 [8], establishing a transparent foundation to guide the review process and minimize bias.

The review was conducted in four phases: identification of relevant studies through database searches; screening of titles and abstracts; full-text evaluation to assess eligibility; and inclusion of studies meeting all predefined criteria. Inclusion and exclusion criteria were applied consistently throughout, with decisions documented to ensure methodological transparency and alignment with the review objectives.

2.2. Eligibility Criteria

This review focused on studies investigating deep learning-based systems for detecting drowsiness in drivers. To be eligible, studies were required to meet the following conditions.

First, the study had to target drivers—commercial, private, or professional—as the population of interest. Studies focusing on non-driving populations such as students, healthcare workers, or pilots were excluded due to differing contextual demands.

Second, the study needed to develop, apply, or evaluate deep learning models specifically designed for drowsiness detection. Accepted model types included Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures. Studies based solely on conventional machine learning methods, such as Support Vector Machines or Decision Trees, were excluded.

Third, eligible studies were required to report performance metrics relevant to drowsiness detection, including accuracy, precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC), enabling meaningful comparisons across studies.

Fourth, model validation had to be conducted using data obtained under real or simulated driving conditions to ensure applicability to the target context.

Fifth, only original empirical research articles published in peer-reviewed journals were considered. Studies had to be written in English and published between January 2015 and February 2025 to ensure the inclusion of contemporary, high-quality evidence.

Studies were excluded if they lacked deep learning components, omitted key performance metrics, failed to validate models using driving-related data, or focused solely on hardware without evaluating model performance. In addition, review articles, theoretical papers, editorials, opinion pieces, conference abstracts, non-peer-reviewed publications, and studies not published in English were excluded.

These criteria were established to ensure the inclusion of scientifically validated and contextually relevant studies, allowing for a robust synthesis of evidence on the performance and real-world applicability of deep learning models for driver drowsiness detection.

2.3. Search Strategy

A systematic literature search was conducted in March 2025 across five major electronic databases: PubMed, Scopus, Web of Science, ScienceDirect, and IEEE Xplore. These databases were selected for their broad and complementary coverage of health, engineering, and transportation research domains. The search strategy employed Boolean operators and the following combination of keywords: “driver” AND (“drowsiness” OR “sleepiness”) AND “detection” AND “deep learning”.

No supplementary search techniques, such as citation tracking or snowballing, were used. This focused approach ensured a transparent, replicable, and methodologically sound identification of relevant studies while reducing the risk of bias associated with non-systematic retrieval methods.

2.4. Data Collection and Extraction

Data collection and extraction were conducted by a single reviewer following a standardized protocol to ensure methodological consistency and reduce potential bias. After completing the database searches, all retrieved records were imported into the Rayyan platform for systematic reviews (Rayyan Systems Inc., Cambridge, MA, USA, 2025), which enabled automatic duplicate removal. Title and abstract screening was performed in Rayyan based on predefined inclusion and exclusion criteria to assess each record's relevance. Studies that met these initial criteria underwent full-text review to determine final inclusion. This systematic and transparent procedure ensured that only studies directly addressing the review's research questions were retained for synthesis.

For data extraction, a structured spreadsheet was developed using Microsoft Excel (Microsoft Corporation, Redmond, WA, USA, Version 16.77.1, 2025) to collect key information from each included study. The extraction process was supported by ChatGPT (OpenAI, San Francisco, CA, USA, GPT-4, 2025), which assisted in identifying and organizing study details such as driving context, deep learning models, performance metrics, and dataset characteristics. All extracted data were manually reviewed and cross-verified against the original study reports to ensure accuracy and consistency.

2.5. Data Synthesis

Data synthesis is conducted using a structured narrative approach to address the four research questions that guide this review. In light of the marked heterogeneity across the included studies—particularly in terms of modeling objectives, input modalities, evaluation settings, and outcome measures—a quantitative meta-analysis is not feasible. Instead, findings are organized thematically to capture recurring patterns, methodological trends, and relevant contrasts in the implementation and performance of deep learning-based drowsiness detection systems.

The synthesis is presented across dedicated subsections in the results, encompassing study characteristics, model architectures, reported performance metrics, dataset properties, and implementation challenges. This organization facilitates direct alignment between the extracted evidence and each research question, while allowing for comparative interpretation across studies.

To support transparency and reproducibility, three supplementary tables are provided. Table A1 outlines core study attributes, including authorship, country of origin, driving context (simulated or real-world), type of deep learning architecture employed, and the model's operational objective (real-time or offline). Table A2 consolidates the main performance metrics reported by the included studies—accuracy, precision, recall, F1-score, and AUC-ROC—enabling descriptive comparisons across modeling approaches. Table A3 summarizes dataset sources, data types (behavioral, physiological, vehicle-based, or multimodal), reported technical challenges, and recommendations proposed to improve model robustness and applicability.

Where applicable, summary statistics (e.g., median and interquartile range) are calculated to provide an aggregated view of model performance under different testing conditions. These descriptive insights are integrated into the narrative to contextualize the reported results and to identify common benchmarks across simulation-based and real-world evaluations.

This synthesis strategy enables a comprehensive overview of the current state of the literature, emphasizing methodological diversity, implementation barriers, and critical gaps. It also provides an evidence-informed basis for interpreting the reliability, scalability, and translational potential of deep learning models for drowsiness detection in driving contexts.

3. Results

This section presents the main findings of the systematic review, structured to address the four research questions that guided the study. It explores how deep learning models have been used to detect drowsiness in drivers, focusing on their architectures, performance, data sources, and implementation challenges.

Section 3.1 describes the study selection process based on PRISMA 2020 guidelines, outlining how records were identified, screened, and assessed for eligibility. Section 3.2 offers an overview of the included studies, covering publication trends, geographic distribution, journal quartiles, and evaluation contexts.

Section 3.3 analyzes the deep learning models employed in drowsiness detection, examining architectural types, operational objectives, and computational feasibility. Section 3.4 evaluates model accuracy and reliability, discussing performance metrics across different testing conditions, error handling approaches, and adaptability to demographic and environmental variability.

Section 3.5 focuses on the datasets used for model training and validation, addressing their sources, data modalities, diversity, and ground truth annotation methods. Section 3.6 highlights the main technical, practical, and ethical challenges identified across the studies and discusses strategies proposed to enhance model robustness and deployment readiness.

3.1. Study Selection

The selection process for this systematic review followed the PRISMA 2020 guidelines and involved multiple phases: identification, screening, eligibility assessment, and inclusion. A total of 1606 records were identified through comprehensive searches across five major electronic databases: PubMed, Scopus, Web of Science, IEEE Xplore, and ScienceDirect. No additional records were obtained through other sources or registers.

In the identification phase, 899 records were excluded based on predefined exclusion criteria. These exclusions included 3 records published outside the selected date range, 544 that did not meet the document type requirements, 349 excluded due to non-journal source types, and 3 articles that were not published in English. Additionally, 172 duplicate records were removed, resulting in 535 unique records eligible for the screening phase.

During the screening phase, titles and abstracts of the 535 remaining records were evaluated to determine their alignment with the inclusion criteria. A total of 374 records were excluded at this stage for reasons such as lack of focus on drivers, absence of deep learning models for drowsiness detection trained on real or simulated driving conditions, or failure to report model performance metrics such as accuracy, precision, recall, F1-score, or AUC-ROC. This process yielded 161 records deemed relevant for full-text assessment.

Out of the 161 records selected for full-text review, 158 full-text articles were successfully retrieved. Despite exhaustive attempts using institutional databases and interlibrary services, 3 records could not be accessed and were therefore excluded. Each of the 158 full-text articles was then subjected to an eligibility assessment. Of these, 77 articles were excluded: 1 for not focusing on drivers, 62 for not presenting deep learning models trained on real or simulated driving data, and 14 for not reporting relevant performance metrics.

Following this screening and eligibility assessment process, a total of 81 studies met all inclusion criteria and were included in the final qualitative synthesis. These studies reflect a broad spectrum of methodological approaches, geographic regions, and deep learning architectures applied to the task of drowsiness detection in driving contexts. The PRISMA flow diagram (see Figure 1) visually summarizes each stage of the selection process and the number of records included and excluded at each step.

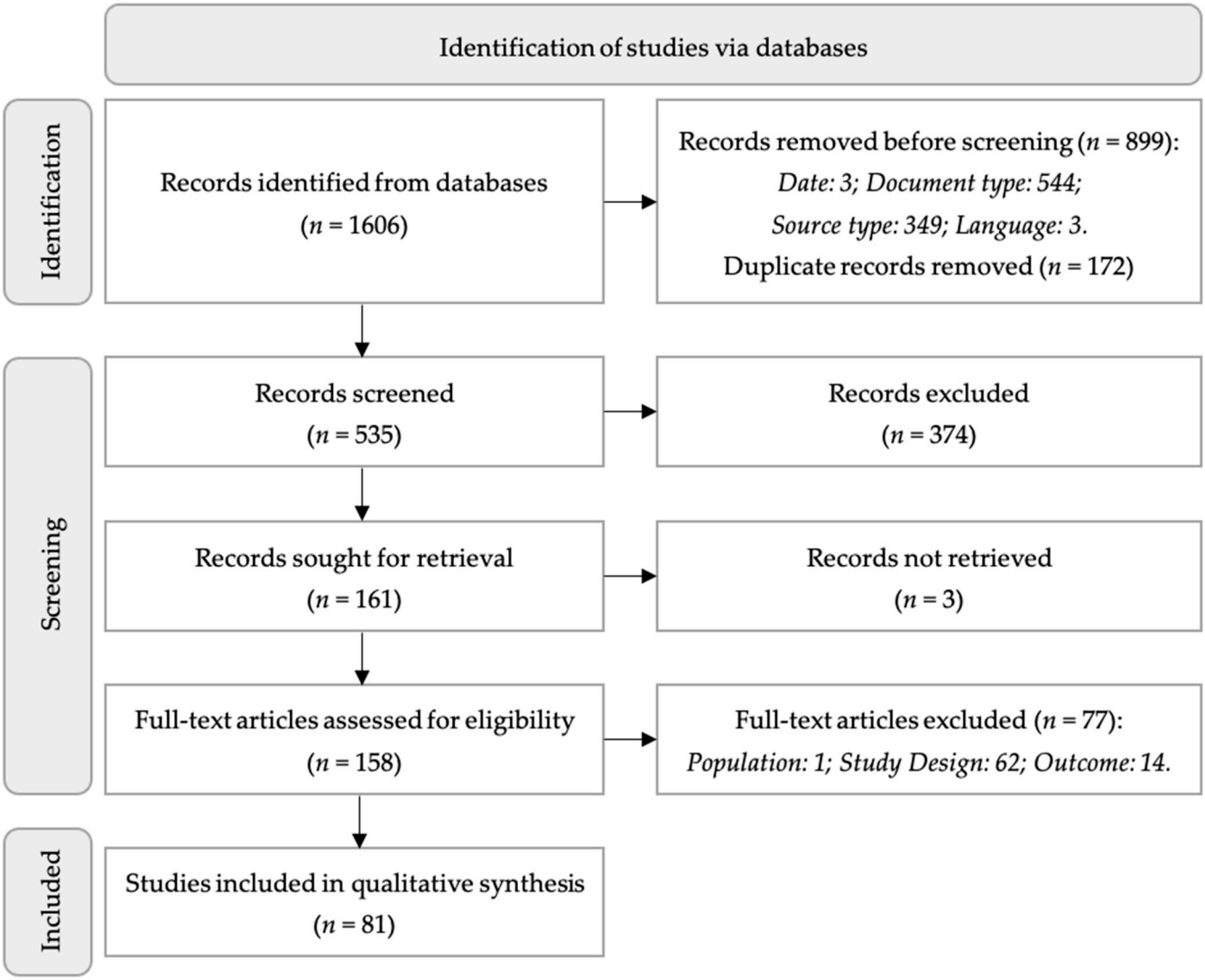

Figure 1. The identification, screening, and inclusion process of eligible studies using PRISMA 2020.

3.2. Overview of Included Studies

This section introduces the main features of the 81 studies selected for this systematic review. The studies reflect a wide range of geographic origins, evaluation contexts, and dataset types. Collectively, they form the empirical foundation for assessing how deep learning has been applied to driver drowsiness detection.

The following subsections detail three major aspects of the included studies. Section 3.2.1 presents trends in publication output, country distribution, journal quartiles, and general methodological traits. Section 3.2.2 discusses where and how models were tested, distinguishing between simulated and real-world driving contexts. Section 3.2.3 reviews the characteristics of the datasets employed, including their sources, modalities, and demographic representation. These components provide essential context for interpreting the results in subsequent sections.

3.2.1. Characteristics of Studies

The temporal distribution of studies indicates a sharp and sustained increase in scholarly output over the last six years (see Figure 2). From a modest base of 3 cumulative studies by 2019, the number rose to 7 by 2020, then more than doubled by 2021 with 16 publications. The trend accelerated further, reaching 30 by 2022 and 49 by 2023. By the end of 2024, the cumulative total of studies had surged to 75, and by early 2025, the full set

of 81 studies was identified. This pattern reflects not only the maturation of deep learning technologies but also a heightened global emphasis on improving road safety through intelligent monitoring systems. Advances in computing power, algorithmic development, and the availability of annotated datasets likely contributed to this expansion.

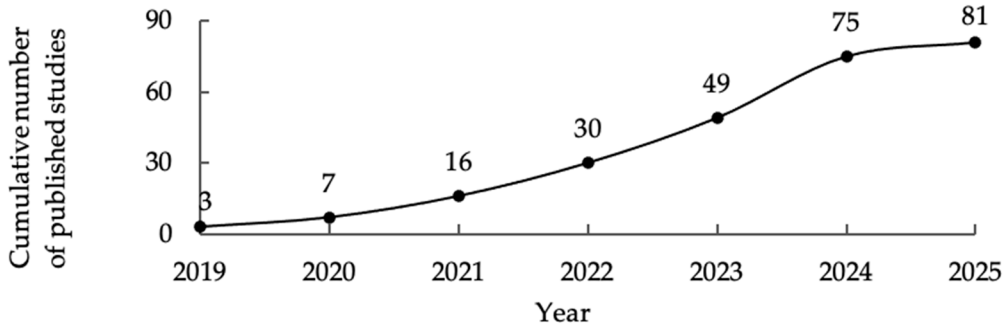

Figure 2. Cumulative number of published studies on deep learning-based models for drowsiness detection.

In terms of geographical distribution, the studies were carried out across multiple countries, with a significant concentration in Asia (see Figure 3). China was the most prolific contributor, responsible for 29 studies (36%), followed by India with 12 (15%), and Saudi Arabia with 6 (7%). South Korea and Pakistan each contributed 3 studies (4% each), while the remaining 28 studies (34%) were conducted across diverse regions including Europe, America, Oceania, and Africa. Although this distribution underscores the global relevance of drowsiness detection research, it also reveals a potential regional imbalance in dataset characteristics and driving conditions, which could affect the generalizability of findings.

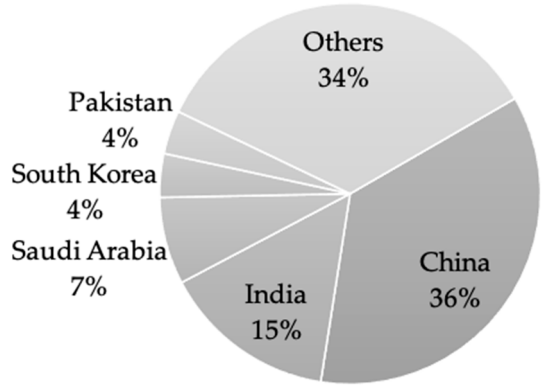

Figure 3. Country wise distribution of selected studies.

The scientific impact of the included studies is reflected in their publication venues (see Table 1). A majority of 48 out of 81 studies (59%) appeared in Q1 journals, indicating strong peer review standards and recognition in high impact outlets.

Table 1. Number of selected studies per quartile.

Quartile Ranking	Number of Studies
Q1	48
Q2	20
Q3	11
Q4	2

An additional 20 studies (25%) were published in Q2 journals, with 11 (14%) and 2 (2%) published in Q3 and Q4 journals, respectively. This suggests that the topic is well integrated into mainstream scientific discourse, particularly within fields such as artificial intelligence, neuroscience, and transportation safety.

From a methodological standpoint, the studies exhibit substantial heterogeneity. Most were empirical investigations centered on the design, training, and validation of deep learning models using either real-world or simulated driving data. Simulated driving was commonly employed to induce drowsiness under controlled conditions, while real-world setups often involved naturalistic driving scenarios with in vehicle sensor data. Behavioral data such as eye images and facial expressions were the most frequently used inputs, followed by physiological signals like electroencephalogram (EEG) and electrocardiogram (ECG). A growing number of studies adopted multimodal approaches that integrate vehicle-based, physiological, and behavioral cues to improve prediction performance.

Supervised learning frameworks dominated the modeling strategies, with binary classification between drowsy and alert states as the most common output. Ground truth labeling was typically derived from physiological thresholds, expert annotations, or driver task performance. While many studies focused on achieving real-time detection suitable for Advanced Driver Assistance Systems (ADASs), a subset prioritized offline analysis for performance benchmarking. Validation methods ranged from internal cross validation to external testing using publicly available datasets, with varying degrees of transparency and reproducibility.

In sum, the included studies reflect a dynamic and methodologically diverse research landscape. The upward trajectory in publication volume, concentration of studies in high impact journals, and increasing reliance on multimodal and real time systems signal a growing maturity in the field. These characteristics form a critical foundation for the subsequent analyses of deep learning models, performance metrics, datasets, and implementation challenges presented in the following sections.

3.2.2. Contexts of Model Evaluation

Understanding the context in which deep learning models are evaluated is essential to assess their applicability to real-world driving scenarios. Among the studies included in this review, a clear distinction emerged between those conducted in simulated environments and those implemented in real world settings. This division is critical because it influences not only the quality and type of data collected but also the ecological validity of the model evaluation.

Simulated driving environments were widely adopted as experimental testbeds due to their ability to replicate drowsiness-inducing conditions in a safe and controlled manner. These setups allowed researchers to standardize scenarios, manipulate fatigue levels, and collect high-resolution multimodal data with minimal ethical concerns. Studies using driving simulators often induced drowsiness through prolonged tasks, monotonous roads, or night-time scenarios. Driver responses were measured using combined datasets that included physiological signals, behavioral data, and vehicle-based metrics. Simulations also facilitated consistent labeling of ground truth data based on controlled experimental events or physiological thresholds.

In contrast, real world evaluations involved testing models during actual driving, either in private vehicles or fleet operations. These studies provided data that more accurately reflected the complexity of real-life driving, including variations in road conditions, lighting, driver behavior, and vehicle dynamics. Real-world contexts introduced significant challenges, such as uncontrolled environmental factors and limitations in participant monitoring, but also yielded findings with greater external validity. Many studies in this

category used unobtrusive sensors or embedded systems that collected data during regular driving sessions, enabling the models to be assessed under realistic operational constraints.

A smaller subset of studies combined both contexts, training models on simulated data and then testing them in real driving conditions to evaluate generalization performance. This approach highlighted the importance of model adaptability and robustness across environments, especially for applications aimed at integration into commercial ADAS.

Overall, while simulated environments continue to play a crucial role in initial model development and benchmarking, real-world testing is indispensable for verifying operational feasibility and practical effectiveness. The coexistence of both evaluation contexts in the literature underscores the complementary nature of these approaches and the ongoing need to balance experimental control with ecological validity in drowsiness detection research.

3.2.3. Summary of Datasets Used

The characteristics and sources of datasets used for training and validating deep learning models represent an important foundation for understanding model performance, generalizability, and limitations. Across the reviewed studies, there was substantial variation in dataset size, composition, modality, and accessibility, reflecting both the opportunities and challenges inherent in collecting data relevant to drowsiness detection.

A significant number of studies relied on publicly available datasets, including commonly used benchmark collections such as the NTHU-DDD [9], YawDD [10], SEED-VIG [11], and the SADT [12] dataset. These resources provided behavioral and physiological data, including eye state images, head poses, and EEG signals, and were widely used for algorithm benchmarking and comparative analysis. Public datasets facilitated reproducibility and model validation across studies, but they also exhibited notable limitations such as narrow demographic representation, limited driving conditions, or lack of multimodal input.

In contrast, some studies developed proprietary datasets tailored to specific research goals. These datasets were often collected through controlled laboratory experiments or on-road vehicle trials and tended to include richer and more diverse data streams. Multimodal datasets combining eye tracking, facial expression, heart rate, EEG signals, and vehicle telemetry were increasingly employed to improve model robustness. However, such datasets were frequently restricted in size and rarely shared openly, which limited transparency and hindered replication efforts.

Dataset diversity was a key concern identified in this review. A large proportion of datasets focused on specific age groups, geographic regions, or experimental conditions, resulting in potential biases. The limited representation of gender, ethnicity, and environmental variability in many datasets raised concerns about the fairness and generalizability of the resulting models. Furthermore, several studies did not report sufficient details regarding sample size, annotation protocols, or data balancing strategies, complicating the evaluation of model training quality.

Despite these limitations, some recent efforts have aimed to create larger and more inclusive datasets by aggregating data from multiple sources or conducting multi-phase data collection campaigns. A few studies also incorporated longitudinal data, enabling temporal modeling of driver fatigue over extended periods.

Overall, the review revealed an increasing reliance on multimodal data sources for drowsiness detection, although concerns remain regarding standardization and accessibility. The heterogeneity of datasets used across studies underscores the importance of clear documentation and open sharing practices, particularly for advancing cross-study comparisons and collaborative model development.

3.3. Deep Learning Models for Drowsiness Detection (RQ1)

This section addresses the first research question of the review: which deep learning models are used to detect drowsiness in drivers? The included studies demonstrate a variety of architectures, design goals, and implementation strategies aimed at identifying driver fatigue based on behavioral, physiological, or multimodal inputs. Understanding the specific types of models used, their intended deployment scenarios, and their technical demands is essential for evaluating their suitability in real-world applications.

The following subsections provide a structured analysis of the deep learning models applied in the reviewed literature. Section 3.3.1 identifies and quantifies the main types of architectures employed, such as CNNs, RNNs, and hybrid configurations. Section 3.3.2 classifies the models according to their objective, whether designed for real-time operation or offline analysis. Section 3.3.3 explores computational requirements and the practical feasibility of deploying these models in vehicle environments. Together, these analyses form a comprehensive overview of how deep learning architectures are applied in drowsiness detection and provide the basis for further evaluation in subsequent sections.

3.3.1. Types and Frequencies of Architectures

The studies reviewed in this systematic analysis utilized a diverse set of deep learning architectures to detect driver drowsiness. The most frequently adopted model type was the Convolutional Neural Network, which appeared in 35 studies. CNNs were primarily used for processing behavioral data, including eye images, facial expressions, and head pose, due to their strong performance in image recognition and spatial pattern extraction.

Recurrent Neural Networks, including LSTM variants, were used in 16 studies. These architectures were applied when temporal dynamics were central to detection, particularly with time-series physiological data such as EEG and ECG signals. LSTMs were favored in cases where the models needed to retain memory of prior inputs across sequences, making them suitable for capturing evolving patterns of fatigue.

Hybrid models that combined different architectural components appeared in 12 studies. These configurations integrated CNN layers for spatial feature extraction with RNN components to process sequential data, or combined multiple deep learning streams for multimodal input fusion. This design aimed to leverage the strengths of different architectures for enhanced performance.

Transformer-based architectures were used in 6 studies. These models were primarily applied in experimental settings and targeted sequence modeling or multimodal attention mechanisms. While their application remains limited, they reflect an emerging interest in applying attention-based frameworks to the problem of drowsiness detection.

Other architectures, including autoencoders and fully connected deep neural networks, were reported in 12 studies and were generally used as baseline models or in combination with other techniques. Overall, the architecture choices reflect a trade-off between model complexity, input modality, and the specific temporal or spatial characteristics of the data being analyzed.

3.3.2. Model Objective: Real-Time vs. Offline Detection

Another key aspect of model design concerns the intended use case—whether the system is developed for real-time detection during active driving or for offline analysis based on previously collected data. Among the reviewed studies, 35 explicitly stated that their models were designed for real-time detection. These models typically aimed to provide immediate feedback or trigger alerts in ADAS. Real-time models prioritized fast inference times, minimal latency, and low computational overhead, and were often deployed on embedded platforms, mobile devices, or in-vehicle systems.

In contrast, 46 studies implemented models intended for offline analysis. These models were primarily used for post hoc evaluation, performance benchmarking, or retrospective monitoring of drowsiness-related behavior. Offline models benefited from fewer constraints in terms of execution time and hardware requirements, allowing for more complex architectures and detailed processing pipelines. They were commonly applied in research environments or simulation-based experiments where real-time performance was not a priority.

The distribution between real-time and offline objectives reveals distinct trade-offs in model design. Real-time models are more constrained by computational efficiency and robustness, while offline models tend to focus on accuracy and interpretability. Understanding these distinctions is essential for evaluating the practical implications of each study's findings and for informing future system development aimed at real-world implementation.

3.3.3. Computational Requirements and Feasibility of Deployment

The computational demands and deployment feasibility of deep learning models are critical considerations for their adoption in real-world driving scenarios. Among the reviewed studies, only 34 explicitly discussed the computational requirements of their models, and even fewer provided concrete details regarding hardware specifications or inference performance.

Studies that targeted real-time implementation typically reported lightweight architectures optimized for embedded platforms, such as Raspberry Pi, NVIDIA Jetson, or Android-based systems. These models were characterized by low latency and modest memory requirements, often achieved through model pruning, quantization, or the use of shallow network structures. Deployment feasibility in such cases was closely linked to the ability to run inference locally without relying on cloud infrastructure.

In contrast, offline models frequently involved more complex or deeper architectures that required greater computational power, typically evaluated using desktop GPUs or cloud-based environments. These models prioritized accuracy and robustness over real-time efficiency and were commonly used in simulation studies or retrospective data analyses. Some studies applied transfer learning from large-scale pre-trained networks, which further increased resource demands during training but not necessarily during inference.

Only a limited number of studies assessed energy consumption, model scalability, or compatibility with automotive-grade processors. Similarly, few studies evaluated the resilience of models to varying hardware conditions or the impact of real-time data stream interruptions. These omissions highlight a gap in practical deployment research, particularly in operational settings such as commercial fleets or consumer vehicles.

3.4. Model Accuracy and Reliability (RQ2)

This section addresses the second research question of the review: how precise and reliable are deep learning models in detecting drowsiness in drivers? The studies included in this review reported various performance metrics under different conditions and constraints. In addition to accuracy, precision, recall, F1-score, and AUC-ROC, some studies explored how these metrics varied across testing environments and population subgroups, and how errors such as false positives and false negatives were managed.

The following subsections provide a structured synthesis of these findings. Section 3.4.1 summarizes the reported values of standard performance metrics. Section 3.4.2 examines model performance in different evaluation settings, comparing simulated and real-world testing. Section 3.4.3 discusses strategies used to address misclassification, including the handling of false positives and false negatives. Section 3.4.4 explores model adaptability to variations in demographic characteristics and driving scenarios. Together, these analyses

help characterize the robustness, limitations, and practical readiness of the reviewed models for deployment in diverse real-world conditions.

3.4.1. Accuracy, Precision, Recall, F1-Score, AUC-ROC Benchmarks

This section presents a summary of the main performance metrics used to evaluate deep learning models for drowsiness detection. These metrics include accuracy, precision, recall (sensitivity), F1-score, and AUC-ROC. Table 2 reports the median, standard deviation, and interquartile range (Q1–Q3) for each metric based on the values extracted from the included studies.

Table 2. Summary statistics of model performance metrics across included studies.

Metric	Median	Std. Dev.	Q1 (25%)	Q3 (75%)
Accuracy	0.952	0.072	0.904	0.979
Precision	0.956	0.077	0.912	0.980
Recall	0.953	0.077	0.918	0.980
F1-score	0.953	0.083	0.903	0.976
AUC-ROC	0.975	0.101	0.957	0.990

Among the 81 included studies, accuracy was the most frequently reported metric, appearing in 77 studies. The median accuracy was 0.952. Over 75% of models achieved accuracy scores above 0.904, particularly those employing multimodal inputs or CNNs based on behavioral analysis. These results indicate that the majority of models demonstrated strong classification performance under the conditions tested.

Precision was reported in 42 studies. The median precision was 0.956, indicating a generally high rate of correctly predicted drowsiness cases among all positive predictions. High precision was often associated with models trained using curated and balanced datasets or those leveraging sensor fusion techniques.

Recall, also known as sensitivity, was reported in 51 studies and yielded a median of 0.953. High recall values suggest that the models were effective at detecting true positive cases of drowsiness. This was especially evident in studies using physiological signals, such as EEG or ECG, processed through LSTM or hybrid sequential architectures capable of capturing temporal dynamics of fatigue onset.

F1-score, available in 52 studies, provided a harmonic mean between precision and recall. It showed a slightly broader variability, with a median of 0.953 and a standard deviation of 0.083. This metric proved useful in evaluating models under class imbalance, where accuracy alone could be misleading. In some studies, notably those conducted under real-world or noisy conditions, F1-score values dropped relative to accuracy, highlighting increased rates of false positives or false negatives.

Only 12 studies reported the AUC-ROC, a metric that reflects a model's ability to discriminate between classes independent of the decision threshold. Despite the smaller sample, AUC-ROC values were notably high, with a median of 0.975 and interquartile values between 0.957 and 0.990. Models incorporating EEG data or attention-based layers consistently achieved higher AUCs, confirming their strong discriminatory power across varying detection thresholds.

Collectively, the benchmark values presented above underscore that deep learning models for drowsiness detection have achieved high levels of predictive performance. Nonetheless, inconsistencies in reporting were observed. Several studies failed to report standard deviation or confidence intervals, and many omitted important contextual information, such as dataset composition or evaluation protocol. These omissions hindered direct comparisons between models and limited the assessment of their generalizability.

Furthermore, the exclusive reporting of accuracy in some studies—without complementary metrics such as recall or F1-score—may mask limitations in model robustness, especially under imbalanced or real-world data.

3.4.2. Performance in Different Testing

Testing conditions varied considerably across the reviewed studies, with 63 models evaluated exclusively in simulated environments, 15 assessed using real-world data, and 3 tested in both settings. Simulated testing was typically conducted in controlled environments using driving simulators or pre-recorded datasets, allowing for precise control of drowsiness-inducing variables and standardized data collection.

Among the studies that reported both accuracy and F1-score, performance outcomes differed between simulation and real-world contexts (see Table 3). Simulation-based models showed a median accuracy of 0.958 and a median F1-score of 0.948. In comparison, real-world models achieved higher median values—0.977 for accuracy and 0.972 for F1-score. These results suggest that while both settings yield strong model performance, real-world validation is associated with more robust outcomes.

Table 3. Performance comparison between simulated and real-world testing.

Metric	Simulated	Real-World
Accuracy (median)	0.958	0.977
F1-score (median)	0.948	0.972

Several factors may contribute to this difference. Real-world studies often employ multimodal inputs and are tuned for greater adaptability to operational conditions. Additionally, data collected in real driving scenarios may promote better generalization and model refinement. However, publication bias cannot be ruled out, as real-world studies with lower performance may be underreported.

Only a few studies directly compared simulated and real-world testing using the same models. In these cases, performance typically declined in real-world settings, indicating that simulation-trained models may require adaptation to maintain accuracy under authentic conditions. This observation highlights the importance of incorporating real-world testing to ensure practical applicability.

Real-world studies also tended to discuss practical deployment constraints more thoroughly, such as latency, sensor performance, and usability challenges. In contrast, simulated studies emphasized reproducibility and benchmarking under controlled conditions, which remain valuable for early-stage development and algorithm validation.

In summary, both testing environments offer unique advantages. Simulated evaluations support standardized experimentation and rapid iteration, while real-world testing is essential to validate system performance under realistic conditions. The consistent advantage observed in real-world median performance underscores the importance of field validation when assessing readiness for deployment.

3.4.3. Handling of False Positives and Negatives

Effective management of misclassification, particularly false positives and false negatives, is essential for the reliability and acceptance of drowsiness detection systems.

False positives—cases where an alert is incorrectly triggered in a non-drowsy driver—were a central concern for studies focused on real-time deployment. Frequent or unjustified alerts were seen as a source of driver annoyance and could contribute to warning fatigue, ultimately undermining user trust. To mitigate this, several studies employed post-processing techniques such as temporal smoothing, threshold calibration, or majority-voting mecha-

nisms over sequential predictions to reduce spurious alerts. Models trained with balanced datasets or using weighted loss functions were also reported to exhibit improved resistance to false positive inflation.

False negatives—instances where a drowsy state goes undetected—were regarded as a more critical safety risk. Several studies prioritized minimizing false negatives by tuning model sensitivity, even at the expense of slightly lower precision. In particular, studies based on physiological signals (e.g., EEG, ECG) showed a stronger focus on minimizing false negatives through signal segmentation, dynamic windowing, or high-resolution feature extraction.

Few studies quantitatively reported false positive and false negative rates, typically presenting them through confusion matrices or derived metrics such as specificity and sensitivity. In the limited cases where both error types were directly compared, a trade-off was evident: models that aggressively minimized false positives often exhibited higher false negative rates, and vice versa.

Hybrid models and those using multimodal data sources (e.g., combining behavioral and physiological features) demonstrated more balanced error profiles. These approaches allowed for complementary error compensation across modalities, resulting in greater stability in classification under real-world conditions. Attention-based mechanisms and adaptive thresholding were also used to modulate predictions based on contextual cues, further reducing error volatility.

In summary, handling of misclassification remains a critical and nuanced challenge. While technical strategies such as data balancing, adaptive thresholds, and ensemble learning show promise, explicit and transparent reporting of error trade-offs is still limited. Greater attention to the practical implications of false alerts versus missed detections is necessary to improve both the safety and usability of deployed systems.

3.4.4. Model Adaptability to Demographics and Driving Scenarios

Adaptability to different demographic groups and driving scenarios is critical for ensuring fairness, inclusiveness, and consistent model performance across real-world applications. Some studies in this review considered variations in demographics, such as age, gender, or ethnicity, while others explored model robustness under different driving conditions.

Approaches to demographic adaptability included performance testing across user subgroups and incorporating demographic features during training. Although explicit reporting of demographic-specific results was limited, some studies noted potential performance differences linked to age or gender imbalances in the training data. In response, a few models applied techniques such as data augmentation or the inclusion of auxiliary demographic variables to improve generalization.

Regarding driving scenarios, studies explored model behavior across varying road conditions, lighting environments, and traffic patterns. Multimodal models—particularly those incorporating physiological and behavioral signals in addition to vehicle-based features—were more frequently associated with consistent performance across these conditions. Context-aware designs and architectures employing adaptive mechanisms or transfer learning further contributed to environmental robustness.

Despite these advances, long-term adaptability remains underexplored. Most studies were limited to single-session evaluations, often under constrained conditions, without longitudinal or multi-context validation. Additionally, few studies addressed the dynamic interplay between driver characteristics and contextual factors, such as how age or stress levels might interact with night driving or congested environments.

In summary, while selected models demonstrated strategies to enhance adaptability, the evidence base remains fragmented and inconsistently reported. Available data suggest

that leveraging multimodal inputs and designing models with contextual sensitivity may support greater robustness across populations and operational scenarios. However, the lack of standardized evaluation protocols and limited subgroup reporting continue to challenge broad conclusions about model generalizability in real-world use cases.

3.5. Datasets and Data Characteristics (RQ3)

This section addresses the third research question of the review: what types of datasets are used to train and validate deep learning models for drowsiness detection in drivers? Understanding dataset characteristics is fundamental to evaluating model performance, generalizability, and fairness.

The following subsections provide a structured overview of the datasets employed in the reviewed studies. Section 3.5.1 distinguishes between open-source and proprietary datasets, highlighting trends in accessibility and usage. Section 3.5.2 examines the types of data modalities used for model input, including vehicle-based, physiological, behavioral, and multimodal sources. Section 3.5.3 explores dataset size and diversity, with a focus on demographic coverage and environmental variability. Section 3.5.4 summarizes the methods used to annotate data and establish ground truth labels for model training. Together, these aspects offer a comprehensive perspective on the empirical foundations underlying deep learning-based drowsiness detection systems.

3.5.1. Dataset Sources

Dataset accessibility significantly influences the reproducibility, comparability, and scalability of deep learning models. Among the reviewed studies, two primary categories of datasets were observed: open-source datasets that are publicly accessible and proprietary datasets developed for specific research purposes.

Open-source datasets were frequently adopted due to their availability, standardized formats, and established benchmarks. These datasets enabled researchers to validate models against previously reported results, contributing to comparability across studies. Prominent examples include the NTHU-DDD, YawDD, SEED-VIG, and SADT datasets. These collections typically offered annotated facial images, eye states, head poses, or physiological signals such as EEG, and were collected under either controlled or semi-naturalistic conditions. The availability of such datasets supports benchmarking and facilitates incremental advancements in algorithm development.

However, these public datasets are not without limitations. They often exhibit constraints in demographic diversity, recording environments, and sensor variety, which can reduce their applicability to broader real-world contexts. Additionally, the reuse of the same datasets across multiple studies may introduce risks of overfitting or overestimation of model performance when not carefully managed.

In contrast, proprietary datasets were developed to address specific experimental needs and often featured richer multimodal content, higher temporal resolution, or tailored task scenarios. These datasets allowed for more nuanced exploration of behavioral and physiological signals and enabled the inclusion of emerging sensor technologies not yet captured in public datasets. Nonetheless, the restricted accessibility of proprietary data hinders replication and reduces the transparency of validation procedures. In many cases, critical information regarding participant demographics, experimental protocols, or labeling strategies was only partially reported, limiting the interpretability and generalizability of findings.

Some studies adopted hybrid approaches by pretraining models on large-scale public datasets and fine-tuning them using smaller proprietary collections. This strategy aimed to

balance the strengths of open datasets with the context-specific richness of private ones, especially when adapting models to new populations or deployment scenarios.

In summary, while open-source datasets remain essential for promoting transparency and fostering collaboration, proprietary datasets contribute to innovation through tailored design and sensor integration. Future progress will depend on more systematic reporting, broader demographic coverage, and increased efforts to share high-quality data under standardized ethical and technical frameworks.

3.5.2. Data Modalities

The type of input data used in deep learning models significantly influences their capacity to detect drowsiness with accuracy and reliability. The reviewed studies employed a range of data modalities, either individually or in combination, to capture relevant indicators of driver alertness. These modalities were broadly classified into behavioral, physiological, vehicle-based, and multimodal categories.

Behavioral data was the most commonly employed modality. Studies utilizing behavioral input typically relied on facial imagery, particularly eye region monitoring, blink rates, gaze direction, and head pose estimation. These features were extracted from in-cabin video streams and analyzed to infer states of drowsiness. The non-invasive nature of behavioral monitoring and its compatibility with real-time deployment contributed to its widespread adoption. However, behavioral methods can be sensitive to lighting conditions, occlusions, and camera positioning, which may impact their robustness in uncontrolled environments.

Physiological data constituted the second most frequent modality. This category included biosignals such as EEG, ECG, and electrooculography (EOG). These signals are known to provide direct insight into the neurological and cardiac states associated with fatigue. Studies employing physiological inputs often reported high accuracy levels due to the objective nature of these measures. Nonetheless, the need for contact-based sensors and the potential discomfort for drivers pose significant barriers to widespread adoption in operational settings.

Vehicle-based data, although less commonly used in isolation, encompassed indicators such as steering patterns, lane deviation, and pedal activity. These features are typically derived from vehicle telemetry and are valuable for long-term monitoring in real-world contexts. Vehicle-based cues may reflect both driver-specific traits and context-dependent variations in alertness, but they can also be influenced by factors unrelated to fatigue, such as road geometry or traffic density.

A growing number of studies employed multimodal approaches that integrate two or more of the aforementioned modalities. These hybrid systems aimed to enhance detection robustness by compensating for the limitations of individual data streams. For example, combining facial analysis with EEG signals allowed for both external and internal indicators of drowsiness to be captured. Multimodal systems were particularly prevalent in simulator-based experiments where sensor placement and data synchronization could be tightly controlled. Although these setups demonstrated superior performance metrics, they also introduced higher complexity in terms of hardware, data processing, and model integration.

In summary, the reviewed studies reveal a diverse and evolving landscape of data modalities, each offering unique advantages and challenges. While behavioral data remains the most practical for real-time deployment, physiological and vehicle-based signals contribute valuable depth when feasible. Multimodal strategies are gaining traction as the preferred approach for improving detection accuracy and reliability, particularly in research contexts aiming for comprehensive modeling of driver state.

3.5.3. Dataset Size and Diversity

The size and diversity of datasets are critical in shaping the generalizability and robustness of deep learning models. In this review, reported dataset sizes varied considerably, ranging from a few hundred to several thousand samples. However, precise information regarding the number of subjects and data points was frequently missing. Many studies failed to provide a clear breakdown of how data were distributed across individuals or conditions, making it difficult to assess data representativeness.

Section 3.4.4 previously discussed how model adaptability is influenced by demographic and environmental factors. While that subsection focused on model-level strategies and observed performance across contexts, the present subsection turns to the composition of the datasets themselves. It examines the foundational data that enable—or limit—model generalizability.

Regarding demographic coverage, most datasets were developed from relatively uniform populations, with limited variation in participant characteristics. Age, gender, and ethnicity were often underreported or not mentioned at all. Where mentioned, datasets typically reflected young, local populations, often comprising university students. As a result, there is little empirical basis for evaluating model fairness or bias across demographic groups, which limits conclusions about their potential deployment in real-world applications.

Environmental diversity in data collection was also inconsistently documented. Some datasets captured driving in varied road types, lighting conditions, and traffic scenarios, while others were strictly confined to controlled environments. However, the extent and nature of these conditions were often vaguely described, with minimal detail about how environmental variability was captured or quantified. This lack of clarity makes it difficult to determine whether models trained on such data would perform reliably in more dynamic or unpredictable driving environments.

A small number of studies indicated that data collection occurred over multiple days or tasks, which may contribute to temporal variability in the dataset. Nonetheless, longitudinal data were rare, and few datasets explicitly stated whether the same subjects were recorded under different conditions. The absence of such temporal dimension reduces the opportunity to train models capable of adapting to changes in driver state or behavior over time.

While some datasets may inherently contain diverse scenarios or participant characteristics, the general lack of standardized reporting undermines the ability to compare datasets or assess the suitability of specific data sources for model training. Without detailed metadata or structured documentation, it remains unclear to what extent dataset composition supports generalizable and inclusive drowsiness detection systems.

In summary, although the reviewed studies employed datasets of varying scales and contexts, inconsistencies in reporting demographic and environmental characteristics limit the interpretability and transferability of their findings. This subsection complements the discussion in Section 3.4.4 by emphasizing how dataset-level limitations—not just model-level design—contribute to gaps in model robustness. More transparent and systematic dataset documentation is needed to support the development of equitable deep learning models in this domain.

3.5.4. Annotation and Ground Truth Methods

The reliability of deep learning models for drowsiness detection is closely tied to how training data are annotated and how ground truth labels are established. The reviewed studies employed a range of strategies reflecting both the type of input data and the context of data collection.

Physiological thresholds were a common approach in studies using EEG, EOG, or ECG data. Labels were typically assigned based on defined metrics, such as increases in theta or alpha power for EEG or variability in heart rate signals for ECG. These methods offer objective criteria grounded in neurophysiology but are sensitive to inter-individual variability and signal quality. Additionally, different thresholding criteria were applied across studies, complicating direct comparisons.

Studies using behavioral data—such as video of facial expressions and eye states—often relied on manual annotation by human raters. Labeling involved identifying observable signs of drowsiness, including prolonged eye closure, yawning, gaze deviation, or head movements. While these annotations were typically guided by pre-established rules or clinical indicators, few studies reported inter-rater agreement or validation procedures to ensure consistency and reliability. This lack of methodological transparency hinders the reproducibility and comparability of results.

Subjective self-assessments using validated instruments like the Karolinska Sleepiness Scale (KSS) were also used, particularly in simulator-based studies. Some studies combined subjective reports with objective physiological or behavioral indicators to refine label quality. However, differences in administration timing, scale thresholds, and respondent interpretation introduced variability. Only a limited number of studies described calibration efforts to align subjective and objective drowsiness markers.

Experimental task design also played a role in ground truth definition. Studies frequently segmented driving sessions into predefined time blocks under the assumption that drowsiness increased with prolonged task duration or exposure to monotonous stimuli. In these cases, labels were often assigned based on elapsed time, with drowsiness assumed in later stages of the experiment. While this method provided structured data labeling, it risked overgeneralizing individual fatigue responses.

Some studies employed hybrid approaches, integrating multiple data sources—physiological signals, expert observations, and self-reports—to triangulate driver state and establish a more robust ground truth. These studies generally offered better alignment between physiological and behavioral manifestations of drowsiness but often lacked detailed protocols describing how conflicts between sources were resolved.

Automation in annotation was occasionally implemented through rule-based algorithms or real-time signal monitoring. However, few studies detailed the validation of automated labeling procedures or provided benchmarks comparing them to manual annotation or ground truth standards. In some cases, labeling may have occurred post hoc, without ensuring blinding to model predictions, potentially introducing confirmation bias.

Across the reviewed literature, there was a consistent lack of standardized documentation regarding annotation workflows. Key elements such as annotation toolkits, rater training, validation checks, and labeling consistency were rarely reported. This absence of methodological rigor limits the interpretability of performance metrics derived from the annotated data.

In summary, annotation and labeling practices in deep learning-based drowsiness detection research vary widely in terms of rigor, transparency, and methodological alignment. While some studies demonstrated innovative multi-source labeling strategies, inconsistent documentation and limited validation reporting remain pervasive. Addressing these gaps is critical for enabling reproducibility, enhancing model comparability, and strengthening the empirical basis of drowsiness detection systems.

3.6. Challenges and Limitations (RQ4)

While deep learning has shown promise in enhancing drowsiness detection systems, several challenges and limitations continue to affect the development, validation, and

deployment of such models. The studies reviewed in this systematic analysis reveal a wide range of technical, operational, ethical, and methodological constraints that must be addressed to ensure the safe and effective implementation of these systems in real-world contexts.

This section responds to the fourth research question: what are the main challenges and limitations in developing deep learning-based drowsiness detection systems? It provides a detailed synthesis of the major obstacles encountered in the research and development of such systems. Section 3.6.1 focuses on technical issues related to model training and generalization, including overfitting, limited dataset size, and challenges in achieving robustness across diverse users and driving conditions. Section 3.6.2 examines practical barriers to real-world deployment, such as hardware constraints, cost, and environmental variability. Section 3.6.3 explores ethical and privacy concerns associated with the use of personal and biometric data in driver monitoring technologies. Finally, Section 3.6.4 reviews strategies proposed or adopted by researchers to enhance the robustness and applicability of deep learning models, including architectural innovations, data augmentation, and domain adaptation techniques.

By organizing these findings into distinct yet interconnected categories, this section aims to clarify the multifaceted limitations that hinder progress in the field and to contextualize the performance results discussed earlier. These insights are essential for guiding future research directions and informing the design of more reliable and inclusive drowsiness detection technologies.

3.6.1. Technical Challenges

Technical limitations were among the most frequently cited challenges across the reviewed studies. Overfitting was a recurrent issue, particularly in models trained on datasets with limited size or diversity. These models often demonstrated high accuracy on training data but failed to generalize to new subjects or driving contexts, indicating poor robustness. Overfitting was exacerbated in studies that lacked cross-validation or did not apply regularization techniques, leading to models that captured noise or spurious correlations specific to the training set.

Closely linked to overfitting was the broader challenge of generalization. A number of studies reported diminished model performance when exposed to variations in driver demographics, vehicle environments, or data acquisition conditions. For instance, changes in lighting, background, camera angles, or driver posture introduced inconsistencies that degraded the predictive accuracy of models not trained with sufficient variability. These findings underscore the limitations of models developed and validated within narrowly controlled experimental setups.

Data scarcity also emerged as a central obstacle. Although public benchmark datasets such as NTHU-DDD and YawDD were frequently used, they typically featured constrained scenarios and homogeneous populations, limiting their applicability to broader use cases. Studies that collected proprietary data often produced richer multimodal inputs, but these datasets were rarely made publicly available, thereby impeding replicability and cross-study validation. Furthermore, many studies failed to report detailed information on dataset balance, subject diversity, or data preprocessing pipelines, which further complicates the assessment of generalizability.

A compounding issue was the inconsistency in sample sizes and annotation quality. Several studies did not provide sufficient detail about the number of subjects, the volume of data collected per individual, or the annotation strategy, leading to uncertainty about the statistical reliability of the reported results. The absence of standardized evaluation

protocols made it difficult to benchmark performance across studies or to compare outcomes between models trained on different datasets.

In sum, the technical challenges identified in this review reveal systemic gaps in dataset diversity, methodological rigor, and evaluation transparency. Addressing these gaps will require coordinated efforts to develop and share inclusive datasets, adopt more rigorous validation frameworks, and ensure that models are trained under conditions that reflect the variability inherent in real-world driving.

3.6.2. Real-World Implementation Barriers

In addition to technical concerns, a set of practical and operational challenges continue to hinder the successful implementation of deep learning-based drowsiness detection systems in real-world driving environments. These barriers often arise from constraints in computational capacity, hardware integration, environmental variability, and user acceptance.

A primary challenge lies in the computational and financial cost of deployment. High-performing models, particularly those employing deep convolutional or multimodal architectures, often require processing capabilities that exceed what is available in typical vehicle onboard units. While GPUs or AI-accelerated edge devices can handle such models, their cost, size, and energy requirements limit feasibility in production-scale deployment. These constraints are especially critical in the commercial transport sector, where economic considerations dominate technology adoption decisions.

Hardware limitations compound this issue. Several reviewed studies relied on laboratory-grade sensors such as multi-channel EEG systems or infrared cameras, which are impractical for continuous in-vehicle use. When researchers substituted these for more accessible alternatives—such as monocular cameras or wearables—the resulting drop in data resolution and signal reliability often impacted model performance. Moreover, discrepancies in sensor calibration, alignment, and synchronization between training and deployment environments posed additional challenges for replicability and robustness.

Environmental conditions further introduce unpredictability into model behavior. Studies showed that variations in ambient lighting, road surface, cabin configuration, or external weather could significantly distort behavioral or physiological signals. For instance, changes in natural light can mask facial landmarks, while vehicle motion artifacts affect sensor stability. These variations are difficult to simulate fully in experimental setups and remain a primary reason for performance degradation when transitioning from controlled to real-world contexts.

System integration also emerged as a recurring barrier. For a model to be embedded in a vehicle system, it must operate within real-time processing constraints, comply with automotive safety regulations, and interface seamlessly with existing infotainment or telematics platforms. Only a limited number of studies engaged with these engineering requirements, indicating a disconnect between algorithm development and deployment readiness. The absence of modular design standards or compatibility with automotive-grade hardware further inhibits broader adoption.

Usability and driver acceptance are equally critical. Systems perceived as intrusive—such as those requiring facial monitoring at close range or prolonged wearable use—may lead to discomfort or resistance. Moreover, false positives and false alarms can reduce driver trust, particularly if corrective alerts occur too frequently or under benign conditions. Surprisingly, few studies systematically evaluated user feedback, ergonomic design, or post-deployment maintenance, all of which are vital for sustainable adoption.

Together, these real-world barriers illustrate the complexity of moving from lab-validated models to practical, in-vehicle solutions. They highlight the importance of cross-disciplinary collaboration involving data scientists, automotive engineers, human

factors specialists, and regulatory bodies to co-design systems that are not only technically sound but also feasible, acceptable, and scalable in diverse transportation contexts.

3.6.3. Ethical and Privacy Considerations

Ethical and privacy concerns are increasingly relevant in the deployment of deep learning-based drowsiness detection systems, particularly due to the sensitive nature of the data collected and processed. Many of the reviewed studies involved the use of physiological signals or behavioral metrics, all of which may reveal deeply personal information about the driver. Despite this, few studies explicitly addressed the ethical implications or compliance measures related to data handling.

One of the primary ethical concerns lies in the collection of biometric data such as facial images, EEG signals, or heart rate information. These data types, especially when continuously collected, raise important questions regarding informed consent, data ownership, and the potential misuse of personal information. In some cases, studies did not provide detailed information about consent protocols or the safeguarding of participant anonymity, which limits transparency and raises concerns about ethical oversight.

Privacy protection is also challenged by the real-time transmission and processing of driver data, particularly in systems integrated into cloud-based infrastructures or fleet management platforms. Without clear encryption standards or access controls, there is a risk of unauthorized data access or leakage. Furthermore, systems that store historical driver behavior could be susceptible to profiling or surveillance, especially in commercial driving settings where monitoring is often tied to performance evaluation.

A related concern is the lack of standardization in data governance practices across studies. The absence of consistent ethical review processes, data retention policies, or impact assessments makes it difficult to evaluate the broader societal implications of these technologies. Only a small number of studies referred to institutional review board approval or compliance with frameworks such as the General Data Protection Regulation (GDPR), despite the legal and ethical importance of these protocols.

Moreover, algorithmic bias remains a largely unaddressed issue. Models trained on demographically skewed datasets may inadvertently encode and amplify inequalities, resulting in differential performance across subgroups. Without deliberate bias mitigation strategies or subgroup analysis, these systems risk reinforcing systemic disparities rather than promoting equitable safety outcomes.

In sum, while the technical promise of deep learning in drowsiness detection is significant, its ethical deployment requires stronger commitments to privacy, transparency, and fairness. More consistent reporting on consent procedures, data anonymization methods, and ethical compliance is needed to align research practices with societal expectations. Addressing these considerations is essential for fostering trust and ensuring the responsible integration of such technologies into everyday driving contexts.

3.6.4. Strategies to Enhance Model Robustness and Real-World Applicability

In response to the diverse challenges identified throughout this review, numerous studies proposed or adopted strategies to improve the robustness and applicability of deep learning-based drowsiness detection systems. These strategies aimed to address limitations related to data variability, model generalization, hardware compatibility, and operational stability, reflecting a growing effort to transition from experimental models to real-world applications.

One widely adopted approach involved the use of multimodal data inputs. By integrating behavioral, physiological, and vehicle-based signals, researchers sought to enhance the model's ability to detect drowsiness under a broader range of conditions. Multimodal

systems were generally more resilient to noise and more capable of capturing nuanced indicators of fatigue that may be missed by single-source approaches. This design also allowed models to compensate for missing or degraded data from one modality by leveraging signals from another.

Architectural innovations also played an important role. Several studies experimented with hybrid architectures, combining CNNs with recurrent layers such as LSTM or Gated Recurrent Unit (GRU) to capture both spatial and temporal patterns. These hybrid designs helped improve model sensitivity to dynamic driver behaviors over time. Others introduced attention mechanisms or lightweight transformer models, aiming to balance accuracy with computational efficiency.

To address overfitting and generalization, a number of studies employed data augmentation techniques. Synthetic variations in training samples—such as flipped images, altered brightness, or time-series perturbations—were used to increase training diversity and reduce sensitivity to noise. Some studies also incorporated transfer learning, leveraging pre-trained models on related tasks and fine-tuning them on driver-specific data to improve learning efficiency and performance with smaller datasets.

Domain adaptation methods were another strategy used to bridge the gap between training and testing conditions. These techniques included adversarial training, feature alignment, and normalization strategies to reduce discrepancies between source and target domains. While still relatively novel in this field, such approaches show potential for improving model portability across different vehicles, environments, or demographic groups.

On the deployment front, efforts were made to reduce model size and complexity for edge-device compatibility. Techniques such as model pruning, quantization, and distillation enabled researchers to compress deep learning models without substantial loss in performance. This adaptation is especially relevant for integration into real-time vehicle systems where memory and processing power are constrained.

In addition, a few studies emphasized user-centered strategies to promote system acceptability and practical use. These included calibration routines that personalize detection thresholds, as well as explainable AI techniques that enhance transparency in model decisions. However, such approaches remain underexplored and warrant further investigation.

In summary, while challenges in real-world deployment persist, the reviewed literature highlights a wide array of strategies to enhance model robustness and applicability. These efforts represent a step toward building more adaptive, reliable, and scalable systems that can be safely and effectively implemented in diverse driving environments.

4. Discussion

This section provides an integrative analysis of the main findings presented in the results and offers a broader perspective on their implications. While the previous sections systematically addressed each research question through descriptive and comparative summaries, this discussion aims to interpret these results in light of existing knowledge, methodological strengths and gaps, and the broader context of intelligent transportation and public health.

The discussion is organized into four subsections. Section 4.1 synthesizes the key insights obtained from the review, addressing each research question and highlighting consistent patterns, divergences, and open issues. Section 4.2 outlines the strengths and limitations of the systematic review itself, including methodological decisions, data availability, and scope constraints. Section 4.3 explores the potential implications of the findings for road safety policy and regulation, particularly in relation to professional driving and fatigue management strategies. Lastly, Section 4.4 identifies priorities for future research,

drawing attention to knowledge gaps, underexplored themes, and promising directions for technological and methodological development.

Together, these components provide a critical bridge between the review's empirical findings and its broader contributions to the field of drowsiness detection and driver safety using deep learning models.

4.1. Key Findings

This review presents a comprehensive analysis of how deep learning models have been applied to detect driver drowsiness, revealing a combination of technical advancement, methodological diversity, and ongoing limitations. Convolutional Neural Networks emerged as the predominant architecture, adopted in 35 studies, owing to their proficiency in extracting spatial features from facial imagery and behavioral data. Recurrent Neural Networks, especially LSTM models, were used in 16 studies to process time-series physiological signals. Hybrid architectures combining CNN and RNN components were featured in 12 studies, offering advantages in modeling both spatial and temporal dependencies. Transformer-based models, though less common, appeared in six studies, pointing to a growing interest in more flexible sequence modeling frameworks. These model choices reflected the type of data used and the need to classify drowsy versus alert states in a supervised learning context. While many studies targeted real-time deployment scenarios, others focused on offline analysis for benchmarking purposes.

Performance outcomes varied across the studies, with reported accuracies ranging from 0.732 to 0.997. Among those that disclosed both accuracy and F1-score, studies conducted in real-world driving environments demonstrated stronger average results—0.972 for accuracy and 0.950 for F1-score. Simulated driving environments, although more controllable, yielded lower average metrics of 0.927 and 0.912, respectively. This discrepancy underlines the relevance of testing under real operational conditions, despite inherent variability. However, performance metrics were not consistently reported, with many studies omitting precision, recall, or AUC-ROC values. Furthermore, only a small subset addressed false positive and false negative outcomes in sufficient detail, which are critical considerations for system reliability and driver trust.

The review also highlighted considerable variation in dataset characteristics, which significantly influenced model development and evaluation. Publicly available datasets such as NTHU-DDD, YawDD, SEED-VIG, and SADT were widely utilized for benchmarking purposes. While they enabled cross-study comparisons, they often lacked demographic and contextual diversity. Proprietary datasets, typically used in controlled experiments or vehicle trials, provided more complex multimodal data—such as behavioral, physiological, and vehicle-based inputs—but were seldom made available for replication. Behavioral data, particularly facial and eye-related features, constituted the most common input modality, followed by EEG and ECG signals. Some studies combined modalities to increase robustness, though the variability in dataset composition and annotation limited direct comparison of model outcomes.

Generalization challenges emerged as a recurrent issue, especially among models trained on narrow datasets. Overfitting was common in studies with limited sample diversity. Several papers cited the need to optimize algorithms for embedded, low-power devices in order to support real-time deployment. Environmental variability—including changes in lighting, road conditions, and driver behavior—posed additional challenges for robust implementation. Ethical and privacy-related concerns received minimal attention, with only a few studies addressing data protection, informed consent, or regulatory compliance such as GDPR. This highlights a pressing need to strengthen ethical frameworks for biometric monitoring.

Despite these challenges, a number of studies proposed strategies to improve model robustness and applicability. Approaches included the development of lightweight architectures, use of transfer learning, data augmentation techniques, and early explorations of personalized modeling. While explainable AI and user-centered design were mentioned sporadically, these areas remain underdeveloped. Most efforts focused on technical optimization rather than systemic integration or user experience.

In summary, the review captures an evolving research landscape marked by technical progress and methodological variation, but also constrained by gaps in reporting standards, dataset representativeness, and ethical transparency. The successful deployment of deep learning-based drowsiness detection systems will depend on balancing performance optimization with fairness, usability, and adaptability to real-world conditions.

4.2. Strengths and Limitations of the Review

The systematic approach adopted in this review presents several strengths that enhance the reliability and relevance of its findings. The study adhered to PRISMA guidelines, incorporated a broad and well-defined search strategy, and used explicit inclusion and exclusion criteria to ensure transparency and reproducibility. Only peer-reviewed journal articles were considered, and a structured data extraction process was applied across all included studies. The analysis of 81 studies over a ten-year period (2015–2025) enabled the identification of key developments and emerging trends in the field of deep learning-based drowsiness detection.

A further strength lies in the comprehensive categorization of model architectures, dataset characteristics, evaluation settings, and performance metrics. The comparative analysis between simulated and real-world testing contexts provided insights into the operational feasibility of the models under different conditions. The review also integrated findings on annotation practices, computational constraints, and real-world deployment issues, contributing to a more holistic understanding of the research landscape.

However, the review has important limitations that should be acknowledged. First, by excluding studies not published in English, relevant regional literature—particularly from countries with active local research initiatives—may have been omitted. This linguistic restriction could skew the geographical representativeness of the evidence base.

Second, the exclusion of conference papers, technical reports, and gray literature may have limited the scope of technological innovations captured, especially given that many advances in machine learning are initially reported outside traditional journals. This decision, while ensuring peer review quality, may also have contributed to a publication bias that favors mature or successful models.

Third, the reliance on explicitly reported information significantly constrained the depth of synthesis. Many studies did not provide key methodological details such as sample size, demographic breakdown, annotation procedures, or computational specifications. In several cases, essential information on model architectures, hyperparameter configurations, and data preprocessing steps was also missing, which hinders reproducibility and makes it difficult to compare approaches on equal terms. Inconsistent reporting of evaluation metrics—including the omission of precision, recall, and AUC-ROC—further hampered efforts to conduct quantitative comparisons or establish benchmarks across architectures.

Furthermore, several critical topics were underexplored in the primary literature, including ethical safeguards, data protection protocols, and regulatory compliance. This limited the ability of the review to draw strong conclusions about privacy risks, user consent frameworks, or the social acceptability of drowsiness monitoring systems. Real-world implementation factors such as long-term reliability, maintenance, and user feedback

were also insufficiently addressed across studies, narrowing the practical applicability of the findings.

Finally, while the review aimed to assess model adaptability across demographic groups and driving scenarios, the lack of standardized reporting and subgroup analyses in the original studies made it difficult to evaluate fairness or inclusiveness in a systematic way. As a result, several findings remain interpretative rather than conclusive.

Taken together, these limitations highlight the need for improved methodological transparency, broader inclusion criteria, and standardized evaluation protocols in future research. Despite these constraints, the review offers a structured and comprehensive synthesis of the current landscape, establishing a valuable foundation for guiding future technological development, policy design, and scholarly inquiry in the field of AI-based driver drowsiness detection.

4.3. Policy Implications

The findings of this review hold several important implications for public policy and regulatory frameworks aimed at enhancing road safety through the integration of intelligent driver monitoring systems. Drowsiness remains a critical and underreported factor in road traffic incidents, particularly among long-haul and commercial drivers. The application of deep learning models for timely and automated detection represents a promising tool for reducing these risks.

One major implication concerns the role of regulatory bodies in facilitating the adoption of fatigue detection technologies. Given the real-time detection capabilities demonstrated in many studies, public policies could play a central role by encouraging or mandating the implementation of certified AI-based fatigue monitoring systems in commercial transport fleets. Regulatory incentives, such as tax credits or insurance discounts for companies that adopt approved systems, may accelerate uptake. These measures could be integrated with broader occupational health regulations, including mandatory rest periods, shift duration limits, and wellness programs.

Another key policy consideration is the need for standardized certification frameworks for drowsiness detection technologies. The substantial heterogeneity observed in model validation practices, input modalities, and performance metrics poses a challenge to interoperability, comparability, and public trust. Establishing harmonized benchmarks for model evaluation—aligned with industry safety standards and regulatory norms—would provide clarity to manufacturers, fleet operators, and enforcement agencies. These standards should encompass not only performance thresholds but also data integrity, fail-safety requirements, and user interface criteria.

The review also highlights important equity concerns that must inform policy development. The predominance of datasets with narrow demographic or geographic scope raises the risk of algorithmic bias and unequal safety benefits. Policymakers must advocate for inclusivity in both data collection and system evaluation, ensuring that fatigue detection models are representative of the diversity in age, gender, ethnicity, and driving environments. Public investment in open, diverse, and ethically collected datasets could play a pivotal role in improving fairness and performance across populations.

Ethical oversight and data governance also require greater policy attention. The use of biometric and behavioral data for driver monitoring poses risks related to privacy, consent, and surveillance. Policies should require explicit, informed consent for data collection and define strict limitations on data use, retention, and sharing. Guidelines should mandate transparency regarding how models operate, what data they collect, and how outcomes are used in employment or legal contexts. These protections are particularly critical for professional drivers whose livelihoods may be affected by system outputs.

Lastly, effective policy implementation will depend on fostering collaboration between governments, academia, technology developers, and transportation stakeholders. Pilot programs supported by public agencies could facilitate the validation of emerging models in real-world environments and generate evidence for regulatory refinement. Moreover, public education campaigns and driver training initiatives can help increase awareness and acceptance of monitoring technologies, ensuring they are perceived as tools for support rather than control.

In sum, the review supports a multi-faceted policy response that balances innovation with fairness, accountability, and transparency. Aligning technological development with regulatory foresight will be essential to unlocking the full potential of deep learning systems in promoting safer roads and healthier work conditions for professional drivers.

4.4. Future Research Directions

The findings of this review highlight several opportunities for advancing deep learning–based drowsiness detection research. While notable technical progress has been achieved, persistent gaps remain in generalizability, transparency, and real-world applicability.

Dataset diversity and representativeness remain critical priorities. Many high-performing models have been trained on datasets with limited demographic or environmental coverage, such as the EEG-based systems of Gao et al. [13] and Jiao et al. [14], which relied on small, homogeneous participant groups. Broader demographic representation—encompassing age, gender, ethnicity, and driving contexts—is essential to mitigate algorithmic bias. Recent multimodal datasets, such as those collected by C. He et al. [15], show the potential for more inclusive data by combining physiological, behavioral, and vehicle-based signals in operational settings. Public investment in large-scale, diverse, ethically sourced, and open-access datasets could accelerate progress in this area.

Longitudinal and context-rich evaluations are also underexplored. Most studies in this review, including the real-world system by Chew et al. [16], assessed performance over short sessions. Extending evaluations over days or weeks would enable the development of adaptive thresholds tailored to individual drivers, potentially reducing false positives and enhancing trust. Moreover, Hu et al. [17] demonstrated that incorporating varied lighting and road conditions into testing can support multi-context validation, a practice that should be expanded.

Real-world deployment and edge optimization need further attention. Lightweight implementations, such as those by Nguyen et al. [18] and Soman et al. [19], show that real-time inference on embedded devices is achievable. However, large-scale validation in operational fleets remains rare. Field trials, like those conducted by Yu et al. [20], provide partial insight into performance under operational constraints but highlight the need for systematic evaluation of model stability under hardware variability, network interruptions, and environmental noise.

Finally, ethical safeguards and privacy-preserving approaches must be embedded from the outset. Few studies, including Florez et al. [21] and Hu et al. [17], explicitly described consent protocols or secure data handling despite collecting sensitive biometric information. Advances in privacy-preserving machine learning, such as federated learning, could be adapted for drowsiness detection to minimize raw data transmission without compromising model performance.

In summary, future research should prioritize (i) inclusive and diverse datasets, (ii) longitudinal, multi-context evaluations, (iii) real-world deployment strategies, and (iv) built-in ethical safeguards. Achieving these goals will require anchoring technical innovation in human-centered design to ensure that experimental models evolve into scalable, trustworthy systems capable of improving road safety worldwide.

5. Conclusions

This systematic review provides a comprehensive synthesis of peer-reviewed studies investigating the use of deep learning models for driver drowsiness detection in both simulated and real-world contexts. Drawing on 81 eligible studies published between 2015 and 2025, the review analyzes key aspects of model development, evaluation, and applicability, offering critical insights into the current state of research and practice in this field.

The findings indicate that a wide range of DL architectures—most notably CNNs, RNNs, LSTM networks, and hybrid models—have been employed to detect drowsiness based on behavioral, physiological, and multimodal inputs. While many models report high accuracy and F1-score values, especially in controlled environments, real-world performance remains dependent on data diversity, input robustness, and contextual adaptability.

Datasets used to train and validate these models vary considerably in size, modality, and demographic representation. Public datasets support comparability but often lack environmental realism and diversity. Proprietary datasets offer richer data streams but are seldom shared, limiting reproducibility. Moreover, inconsistencies in ground truth labeling, limited reporting on annotation procedures, and insufficient subgroup analysis constrain the generalizability and fairness of many models.

Technical and operational challenges persist, including overfitting, limited model interpretability, computational constraints, and ethical concerns related to privacy and user consent. Although several studies propose solutions such as multimodal fusion, lightweight architectures, transfer learning, and attention mechanisms, deployment in real-world driving remains limited. Ethical safeguards, user-centered design, and standardized evaluation protocols are still underdeveloped.

In conclusion, while deep learning holds considerable promise for enhancing driver safety through automated drowsiness detection, further progress depends on addressing key methodological, practical, and ethical barriers. Future research should prioritize the development of diverse and inclusive datasets, longitudinal and real-world validations, explainable AI strategies, and privacy-preserving frameworks. Advancing in these areas is essential to translate experimental models into effective, trustworthy, and scalable systems for use in intelligent transportation and occupational safety.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app15169018/s1. Document S1: PRISMA 2020 checklist.

Author Contributions: Conceptualization, T.F. and S.F.; methodology, T.F.; validation, S.F.; formal analysis, T.F.; investigation, T.F.; resources, S.F.; data curation, T.F.; writing—original draft prepa ration, T.F.; writing—review and editing, S.F.; visualization, T.F. and S.F.; supervision, S.F.; project administration, T.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financially supported by national funds through the FCT/MCTES, under the project 2023.15776.PEX-uRisK—Understanding risky behaviors at the wheel using a driving simulator.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to acknowledge the Doctoral Program in Occupational Safety and Health at the University of Porto for providing access to digital library resources, which enabled the retrieval of the studies included in this review. This study was also supported by the exploratory research project uRisk: Understanding risky behaviors at the wheel using a driving simulator, led by the Faculty of Engineering of the University of Porto. During the preparation of this review, the authors used ChatGPT for the purposes of preliminary data extraction and text drafting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADAS	Advanced driver assistance system
AI	Artificial intelligence
AUC-ROC	Area under the receiver operating characteristic curve
CNN	Convolutional neural network
DL	Deep learning
ECG	Electrocardiogram
EEG	Electroencephalogram
EOG	Electrooculography
ESRA	E-survey of road users' attitudes
GDPR	General data protection regulation
GPU	Graphics processing unit
GRU	Gated recurrent unit
KSS	Karolinska sleepiness scale
LSTM	Long short-term memory
NTHU-DDD	National Tsinghua university driver drowsiness detection
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
PROSPERO	International prospective register of systematic reviews
RNN	Recurrent neural network
RQ	Research question
SADT	Sustained-attention driving task
YawDD	Yawning detection dataset

Appendix A

The Appendix A includes three supplementary tables that provide additional detail on the reviewed studies. Table A1 summarizes each study's authorship, country of origin, driving context, deep learning model type, and inference mode (real-time or offline). Table A2 compiles reported performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, enabling comparisons across studies. Table A3 details the datasets used, the type of data collected (e.g., behavioral, physiological), technical challenges encountered, and recommendations proposed to improve model robustness and real-world applicability.

Table A1. Summary of study contexts and model attributes.

Study	Country	Driving Context	DL Model Tye(s)	Inference Mode
Adhithyaa et al. (2023) [22]	India	Simulated	Multistage Adaptive 3D-CNN	Real-time
Ahmed et al. (2022) [23]	India	Simulated	Ensemble CNN with two InceptionV3 modules	Real-time
Akrout & Fakhfakh (2023) [24]	Saudi Arabia	Simulated	MobileNetV3 + Deep LSTM	Real-time
Alameen & Alhothali (2023) [25]	Saudi Arabia	Simulated	3D-CNN + LSTM	Real-time
Alghanim et al. (2024) [26]	Jordan	Simulated	Inception-ResNetV2 (hybrid CNN with dilated convolutions)	Offline
Alguindigue et al. (2024) [27]	Canada	Simulated	SNN, 1D-CNN, CRNN	Real-time
Almazroi et al. (2023) [28]	Saudi Arabia	Simulated	MobileNetV3 + SSD + CNN	Real-time

Table A1. *Cont.*

Study	Country	Driving Context	DL Model Tye(s)	Inference Mode
Anber et al. (2022) [29]	Saudi Arabia	Simulated	AlexNet (fine-tuned and feature extractor) + SVM + NMF	Offline
Ansari et al. (2022) [30]	Australia	Simulated	reLU-BiLSTM	Offline
Arefnezhad et al. (2020) [31]	Austria	Simulated	CNN, CNN-LSTM, CNN-GRU	Offline
Bearly & Chitra (2024) [32]	India	Simulated	3DDGAN-TLALSTM (3D Dependent GAN + Three-Level Attention LSTM)	Offline
Bekhouche et al. (2022) [33]	France	Simulated	ResNet-50 + FCFS + SVM	Offline
Benmohamed & Zarzour (2024) [34]	Algeria	Simulated	AlexNet (global features) + LSTM + handcrafted structural features	Offline
J. Chen, Wang, Wang et al. (2022) [35]	China	Simulated	CNN with transfer learning (AlexNet, ResNet18)	Offline
J. Chen et al. (2021) [36]	China	Real-world	12-layer ConvNet (end-to-end CNN)	Offline
J. Chen, Wang, He et al. (2022) [37]	China	Real-world	CNN (6 architectures tested) using PLI-based functional brain network images	Offline
C. Chen et al. (2023) [38]	China	Simulated	SACC-CapsNet (Capsule Network with temporal-channel and channel-connectivity attention)	Offline
Chew et al. (2024) [16]	Malaysia	Real-world	CNN (ResNet + DenseNet) + rPPG (HR monitoring)	Real-time
Civik & Yuzgec (2023) [39]	Turkey	Real-world	CNN (eye model + mouth model)	Real-time
Cui et al. (2022) [40]	Singapore	Simulated	CNN (custom, compact)	Offline
Ding et al. (2024) [41]	United States	Simulated	Few-shot attention-based neural network	Offline
Dua et al. (2021) [42]	India	Simulated	AlexNet, VGG-FaceNet, FlowImageNet, ResNet	Offline
Ebrahimian et al. (2022) [43]	Finland	Simulated	CNN, CNN-LSTM	Offline
Fa et al. (2023) [44]	China	Simulated	MS-STAGCN (Multiscale Spatio-Temporal Attention Graph Convolutional Network)	Real-time
X. Feng, Guo et al. (2024) [45]	China	Simulated	ID3RSNet (interpretable residual shrinkage network)	Real-time
X. Feng, Dai et al. (2025) [46]	China	Simulated	PASAN-CA (Pseudo-label-assisted subdomain adaptation network with coordinate attention)	Real-time
W. Feng et al. (2025) [47]	China	Simulated	Separable CNN + Gumbel-Softmax channel selection	Real-time
Florez et al. (2023) [21]	Peru	Real-world	InceptionV3, VGG16, ResNet50V2 (Transfer Learning)	Real-time
Gao et al. (2019) [13]	China	Simulated	Recurrence Network + CNN (RN-CNN)	Offline

Table A1. *Cont.*

Study	Country	Driving Context	DL Model Tye(s)	Inference Mode
Guo & Markoni (2019) [48]	Taiwan	Simulated	Hybrid CNN + LSTM	Offline
C. He et al. (2024) [15]	China	Real-world	Attention-BiLSTM	Real-time
H. He et al. (2020) [49]	China	Real-world	Two-Stage CNN (YOLOv3-inspired detection + State Recognition Network)	Real-time
L. He et al. (2024) [50]	China	Simulated	ARMFCN-LSTM, GARMFCN-LSTM (CNN + LSTM + attention + WGAN-GP)	Offline
Nguyen et al. (2023) [18]	South Korea	Simulated	MLP, CNN	Real-time
Hu et al. (2024) [17]	China	Real-world	STFN-BRPS (CNN-BiLSTM + GCN + Channel Attention Fusion)	Offline and pseudo-online
Huang et al. (2022) [51]	China	Simulated	RF-DCM (CNN with Feature Recalibration and Fusion + LSTM)	Real-time
Hultman et al. (2021) [52]	Sweden	Real-world and simulated	CNN-LSTM	Offline
Iwamoto et al. (2021) [53]	Japan	Simulated	LSTM-Autoencoder	Offline
Jamshidi et al. (2021) [54]	Iran	Simulated	Hierarchical Deep Neural Network (ResNet + VGG16 + LSTM)	Real-time
Jarndal et al. (2025) [55]	United Arab Emirates	Real-world	Vision Transformers (ViT)	Real-time
Jia et al. (2022) [56]	China	Simulated	M1-FDNet + M2-PENet + M3-SJNet + MF-Algorithm	Real-time
Jiao & Jiang (2022) [57]	China	Simulated	Bimodal-LSTM	Offline
Jiao et al. (2020) [14]	China	Simulated	LSTM + CWGAN	Offline
Jiao et al. (2023) [58]	China	Simulated	MS-1D-CNN (Multi-scale 1D CNN)	Offline
Kielty et al. (2023) [59]	Ireland	Simulated	CNN + Self-Attention + BiLSTM	Real-time
Kır Savaş & Becerikli (2022) [60]	Turkey	Real-world and simulated	Deep Belief Network (DBN)	Offline
Kumar et al. (2023) [61]	India	Simulated	Modified InceptionV3 + LSTM	Offline
Lamaazi et al. (2023) [62]	United Arab Emirates	Real-world	VGG16-based CNN (face/eye/mouth) + two-layer LSTM (driving behavior)	Real-time
Latreche et al. (2025) [63]	Algeria	Simulated	CNN (optimized) + Hybrid ML classifiers (CNN-SVM, CNN-RF, etc.)	Offline
Q. Li et al. (2024) [64]	United States	Simulated	FD-LiteNet (NAS-derived CNN)	Offline
T. Li & Li (2024) [65]	China	Simulated	PFLD + ViT + LSTM (multi-granularity temporal model)	Offline
Lin et al. (2025) [66]	China	Simulated	CNN	Offline

Table A1. *Cont.*

Study	Country	Driving Context	DL Model Tye(s)	Inference Mode
Majeed et al. (2023) [67]	Pakistan	Simulated	CNN, CNN-RNN	Offline
Mate et al. (2024) [68]	India	Simulated	VGG19, ResNet50V2, MobileNetV2, Xception, InceptionV3, DenseNet169, InceptionResNetV2	Offline
Min et al. (2023) [69]	China	Simulated	SVM (linear, RBF), BiLSTM	Real-time
Mukherjee & Roy (2024) [70]	India	Simulated	Stacked Autoencoder + TLSTM + Attention mechanism	Offline
Nandyal & Sharanabasappa (2024) [71]	India	Simulated	CNN-EFF-ResNet 18	Offline
Obaidan et al. (2024) [72]	Saudi Arabia	Simulated	Deep Multi-Scale CNN	Real-time
Paulo et al. (2021) [73]	Portugal	Simulated	CNN (custom, shallow, single conv. layer)	Offline
Peng et al. (2024) [74]	China	Simulated	3D-CNN (video) + 1D-CNN (signals) + Fusion network	Real-time
Priyanka et al. (2024) [75]	India	Simulated	CNN + LSTM	Offline
Quddus et al. (2021) [76]	Canada	Simulated	R-LSTM and C-LSTM (Recurrent and Convolutional LSTM)	Offline
Ramzan et al. (2024) [77]	Pakistan	Real-world	Custom 30-layer CNN (CDLM) + PCA + HOG + ML classifiers (XGBoost, SVM, RF)	Real-time
Sedik et al. (2023) [78]	Saudi Arabia	Real-world	3D CNN, 2D CNN, SVM, RF, DT, KNN, QDA, MLP, LR	Offline
Shalash (2021) [79]	Egypt	Simulated	CNN (custom, 18-layer architecture)	Offline
Sharanabasappa & Nandyal (2022) [80]	India	Simulated	Ensemble Learning (DT, KNN, ANN, SVM) with handcrafted features and ReliefF, Infinite, Correlation, Term Variance	Offline
Sohail et al. (2024) [81]	Pakistan	Real-world	Custom CNN architecture	Real-time
Soman et al. (2024) [19]	India	Simulated	CNN-LSTM hybrid	Real-time
Sun et al. (2023) [82]	South Korea	Simulated	Facial Feature Fusion CNN (FFF-CNN)	Real-time
Tang et al. (2024) [83]	China	Simulated	MSCNN + CAM (Attention-Guided Multiscale CNN)	Offline
Turki et al. (2024) [84]	Tunisia	Real-world	VGG16, VGG19, ResNet50 (Transfer Learning)	Real-time
Vijaypriya & Uma (2023) [85]	India	Real-world and simulated	Multi-Scale CNN with Flamingo Search Optimization (MCNN + FSA)	Real-time
Wang et al. (2025) [86]	China	Simulated	CNN-LSTM with multi-feature fusion (RWECN + DE + SQ)	Offline
Wijnands et al. (2020) [87]	Australia	Simulated	Depthwise separable 3D CNN	Real-time
H. Yang et al. (2021) [88]	China	Simulated	3D Convolution + BiLSTM (3D-LTS)	Offline

Table A1. *Cont.*

Study	Country	Driving Context	DL Model Tye(s)	Inference Mode
E. Yang & Yi (2024) [89]	South Korea	Simulated	ShuffleNet + ELM	Real-time
K. Yang et al. (2025) [90]	China	Simulated	Adaptive multi-branch CNN (adMBCNN: CNN + handcrafted features + functional network)	Offline
You et al. (2019) [91]	China	Simulated	Deep Cascaded CNN (DCCNN) + SVM classifier (custom EAR-based)	Real-time
Yu et al. (2024) [20]	China	Real-world	LSTM	Offline
Zeghlache et al. (2022) [92]	France	Simulated	Bayesian LSTM Autoencoder + XGBoost	Offline
Zhang et al. (2023) [93]	China	Simulated	Multi-granularity CNN + LSTM (LMDF)	Real-time

Table A2. Reported model performance metrics.

Study	Accuracy	Precision (PPV)	Recall (Sensitivity)	F1-Score	AUC-ROC
Adhithyaa et al. (2023) [22]	0.774	NR	NR	0.781	0.8005
Ahmed et al. (2022) [23]	0.971	NR	NR	NR	NR
Akrout & Fakhfakh (2023) [24]	0.984	0.929 (YawDD), 0.956 (DEAP), 0.984 (MiraclHB)	0.933 (YawDD), 0.962 (DEAP), 0.984 (MiraclHB)	0.931 (YawDD), 0.959 (DEAP), 0.984 (MiraclHB)	NR
Alameen & Alhothali (2023) [25]	0.96 (YawDD), 0.93 (Side-3MDAD), 0.90 (Front-3MDAD)	0.93 (YawDD), 0.90 (Side-3MDAD), 0.90 (Front-3MDAD)	1.00 (YawDD), 0.95 (Side-3MDAD), 0.90 (Front-3MDAD)	0.96 (YawDD), 0.93 (Side-3MDAD), 0.90 (Front-3MDAD)	NR
Alghanim et al. (2024) [26]	0.9887 (Figshare), 0.8273 (SEED-VIG)	NR	NR	NR	NR
Alguindigue et al. (2024) [27]	0.9828 (HRV), 0.9632 (EDA), 0.90 (Eye)	0.9828 (HRV), 0.9632 (EDA), 0.90 (Eye)	0.98 (HRV), 0.96 (EDA), 0.90 (Eye)	0.98 (HRV), 0.96 (EDA)	NR
Almazroi et al. (2023) [28]	0.97	0.992	0.994	0.997	NR
Anber et al. (2022) [29]	0.957 (Transfer Learning), 0.9965 (Hybrid)	0.957 (Transfer Learning), 0.9965 (Hybrid)	0.958 (Transfer Learning), 0.9965 (Hybrid)	0.958 (Transfer Learning), 0.9965 (Hybrid)	NR
Ansari et al. (2022) [30]	0.976 (Subject1), 0.979 (Subject2)	0.9738	0.9754	0.9746	NR
Arefnezhad et al. (2020) [31]	0.9504 (CNN-LSTM)	0.95 (CNN-LSTM)	0.94 (CNN-LSTM)	0.94 (CNN-LSTM)	NR
Bearly & Chitra (2024) [32]	0.9182	NR	0.913	NR	NR
Bekhouche et al. (2022) [33]	0.8763	NR	NR	0.8641	NR
Benmohamed & Zarzour (2024) [34]	0.9012	NR	NR	NR	NR
J. Chen, Wang, Wang et al. (2022) [35]	0.9844 (AlexNet relu4), 0.9313 (ResNet18)	0.9680 (AlexNet relu4), 0.9114 (ResNet18)	1.000 (AlexNet relu4), 0.9474 (ResNet18)	NR	NR
J. Chen et al. (2021) [36]	0.9702	0.9674	0.9776	0.9719	NR
J. Chen, Wang, He et al. (2022) [37]	0.954 (Model 4)	0.955 (Model 4)	0.939 (Model 4)	0.947 (Model 4)	0.9953 (Model 4)
C. Chen et al. (2023) [38]	0.9417 (session 1), 0.9059 (session 2)	NR	0.9591 (session 1), 0.9382 (session 2)	0.9594 (session 1), 0.9399 (session 2)	NR
Chew et al. (2024) [16]	0.9421	NR	NR	0.97	NR
Civik & Yuzgec (2023) [39]	0.96	0.8333	1.00	0.9091	NR
Cui et al. (2022) [40]	0.7322	NR	NR	NR	NR
Ding et al. (2024) [41]	0.86	NR	NR	0.86	NR

Table A2. *Cont.*

Study	Accuracy	Precision (PPV)	Recall (Sensitivity)	F1-Score	AUC-ROC
Dua et al. (2021) [42]	0.8500	0.8630	0.8200	0.8409	NR
Ebrahimian et al. (2022) [43]	0.91 (3-level), 0.67 (5-level)	0.87 (3-level), 0.66 (5-level)	0.87 (3-level), 0.67 (5-level)	NR	NR
Fa et al. (2023) [44]	0.924	0.924	0.924	0.924	NR
X. Feng, Guo et al. (2024) [45]	0.7472 (unbalanced), 0.7716 (balanced)	NR	NR	0.7127 (unbalanced), 0.7717 (balanced)	NR
X. Feng, Dai et al. (2025) [46]	0.8585 (SADT), 0.9465 (SEED-VIG)	NR	NR	NR	NR
W. Feng et al. (2025) [47]	0.8084	0.8601	0.7448	0.7965	NR
Florez et al. (2023) [21]	0.9927 (InceptionV3), 0.9939 (VGG16), 0.9971 (ResNet50V2)	0.9957 (InceptionV3), 0.9941 (VGG16), 0.9994 (ResNet50V2)	0.9908 (InceptionV3), 0.9937 (VGG16), 0.9947 (ResNet50V2)	0.9927 (InceptionV3), 0.9939 (VGG16), 0.9971 (ResNet50V2)	NR
Gao et al. (2019) [13]	0.9295	NR	NR	NR	NR
Guo & Markoni (2019) [48]	0.8485	NR	NR	NR	NR
C. He et al. (2024) [15]	0.9921	0.8444	0.8201	0.8321	NR
H. He et al. (2020) [49]	0.947	NR	NR	NR	NR
L. He et al. (2024) [50]	0.9584 (ARMFCN-LSTM), 0.8470 (GARMFCN-LSTM)	0.98 (GARMFCN-LSTM), 0.93 (ARMFCN-LSTM)	0.97 (GARMFCN-LSTM), 0.81 (ARMFCN-LSTM)	0.97 (GARMFCN-LSTM), 0.85 (ARMFCN-LSTM)	1.00 (GARMFCN-LSTM), 0.96 (ARMFCN-LSTM)
Nguyen et al. (2023) [18]	0.9487 (MLP), 0.9624 (CNN)	NR	0.9624 (MLP), 0.9718 (CNN)	0.9497 (MLP), 0.9626 (CNN)	NR
Hu et al. (2024) [17]	0.9243	0.9152	0.9289	0.927	0.957
Huang et al. (2022) [51]	NR	NR	NR	0.8942	NR
Hultman et al. (2021) [52]	0.82	NR	NR	NR	NR
Iwamoto et al. (2021) [53]	NR	NR	0.81	NR	0.88
Jamshidi et al. (2021) [54]	0.8719	NR	NR	NR	NR
Jarndal et al. (2025) [55]	0.9889 (NTHU-DDD), 0.994 (UTA-RLDD)	NR	NR	NR	NR
Jia et al. (2022) [56]	0.978	NR	NR	NR	NR
Jiao & Jiang (2022) [57]	0.969 (HEOG + HSUM)	0.634 (HEOG + HSUM)	0.978 (HEOG + HSUM)	0.765 (HEOG + HSUM)	NR
Jiao et al. (2020) [14]	0.9814	NR	NR	0.946	NR
Jiao et al. (2023) [58]	0.993	0.989	0.995	0.991	NR
Kielty et al. (2023) [59]	NR	0.959 (internal test), 0.899 (YawDD)	0.947 (internal test), 0.910 (YawDD)	0.953 (internal test), 0.904 (YawDD)	NR
Kır Savaş & Becerikli (2022) [60]	0.86	NR	NR	NR	NR
Kumar et al. (2023) [61]	0.9136	0.74	0.92	0.82	NR
Lamaazi et al. (2023) [62]	0.973 (CNN-eye), 0.982 (CNN-mouth), 0.93 (LSTM)	0.93 (LSTM)	0.89 (LSTM)	0.91 (LSTM)	NR
Latreche et al. (2025) [63]	0.99 (CNN-SVM)	0.98 (CNN-SVM)	0.99 (CNN-SVM)	0.99 (CNN-SVM)	0.99 (CNN-SVM)
Q. Li et al. (2024) [64]	0.9705 (FD-LiteNet1), 0.9972 (FD-LiteNet2)	NR	NR	NR	NR
T. Li & Li (2024) [65]	0.9315	NR	NR	NR	NR
Lin et al. (2025) [66]	0.8756	0.9169	0.8995	0.9016	NR
Majeed et al. (2023) [67]	0.9669 (CNN-2), 0.9564 (CNN-RNN)	0.9569 (CNN-2), 0.9541 (CNN-RNN)	0.9558 (CNN-2), 0.9186 (CNN-RNN)	0.9563 (CNN-2), 0.9360 (CNN-RNN)	NR
Mate et al. (2024) [68]	0.9651 (VGG19)	0.9814 (VGG19)	0.9536 (VGG19)	0.9673 (VGG19)	NR
Min et al. (2023) [69]	0.9941 (linear), 0.7449 (RBF), 0.7365 (BiLSTM)	NR	0.9897 (linear), 0.6170 (RBF), 0.7404 (BiLSTM)	0.9900 (linear), 0.5933 (RBF), 0.6630 (BiLSTM)	0.692 (RBF), 0.704 (BiLSTM)

Table A2. *Cont.*

Study	Accuracy	Precision (PPV)	Recall (Sensitivity)	F1-Score	AUC-ROC
Mukherjee & Roy (2024) [70]	0.9863	0.986	0.986	0.983	NR
Nandyal & Sharanabasappa (2024) [71]	0.9130	NR	0.9210	NR	NR
Obaidan et al. (2024) [72]	0.9703	0.9553	0.9554	0.9553	NR
Paulo et al. (2021) [73]	0.7587	NR	NR	NR	NR
Peng et al. (2024) [74]	0.9315	NR	0.9171	NR	NR
Priyanka et al. (2024) [75]	0.9600	0.9500	0.9500	0.9500	NR
Quddus et al. (2021) [76]	0.8797 (R-LSTM), 0.9787 (C-LSTM)	NR	0.9531 (R-LSTM), 0.9941 (C-LSTM)	NR	NR
Ramzan et al. (2024) [77]	0.997 (CNN), 0.982 (Hybrid), 0.984 (ML with XGBoost)	0.998 (CNN), 0.726 (Hybrid), 0.985 (ML with XGBoost)	0.999 (CNN), 0.861 (Hybrid), 0.985 (ML with XGBoost)	0.992 (CNN), 0.787 (Hybrid), 0.985 (ML with XGBoost)	NR
Sedik et al. (2023) [78]	0.98 (3D CNN on Combined Dataset)	0.98 (3D CNN on Combined Dataset)	0.98 (3D CNN on Combined Dataset)	0.98 (3D CNN on Combined Dataset)	NR
Shalash (2021) [79]	0.9436 (FP1), 0.9257 (T3), 0.9302 (Oz)	NR	NR	NR	0.9798 (FP1), 0.97 (T3), 0.9746 (Oz)
Sharanabasappa & Nandyal (2022) [80]	0.9419	NR	0.9858	0.9764	NR
Sohail et al. (2024) [81]	0.950	NR	0.940	NR	NR
Soman et al. (2024) [19]	0.98	0.95	0.93	0.94	0.99
Sun et al. (2023) [82]	0.9489 (DFD), 0.9835 (CEW)	NR	NR	0.9479 (DFD), 0.9832 (CEW)	0.9882 (DFD), 0.9979 (CEW)
Tang et al. (2024) [83]	0.905	0.867	NR	0.824	NR
Turki et al. (2024) [84]	0.9722 (VGG16), 0.9630 (VGG19), 0.9838 (ResNet50)	0.9724 (VGG16), 0.9658 (VGG19), 0.9842 (ResNet50)	0.9720 (VGG16), 0.9624 (VGG19), 0.9837 (ResNet50)	0.9722 (VGG16), 0.9641 (VGG19), 0.9839 (ResNet50)	NR
Vijaypriya & Uma (2023) [85]	0.9838 (YAWDD), 0.9826 (NTHU-DDD)	0.9707 (YAWDD), 0.9945 (NTHU-DDD)	0.9785 (YAWDD), 0.9811 (NTHU-DDD)	0.9746 (YAWDD), 0.9878 (NTHU-DDD)	NR
Wang et al. (2025) [86]	0.9657 (SEED-VIG); 0.9923 (Mendeley)	0.9601	0.9512	0.9554	NR
Wijnands et al. (2020) [87]	0.739	NR	NR	NR	NR
H. Yang et al. (2021) [88]	0.834 (YawDDR); 0.805 (MFAY)	NR	NR	NR	NR
E. Yang & Yi (2024) [89]	0.9705	0.9587	0.9269	0.9553	0.9705
K. Yang et al. (2025) [90]	0.9602 (SEED-VIG), 0.9184 (Cui)	0.9514 (SEED-VIG), 0.9250 (Cui)	0.9241 (SEED-VIG), 0.9117 (Cui)	0.9351 (SEED-VIG), 0.9179 (Cui)	NR
You et al. (2019) [91]	0.948	NR	NR	NR	NR
Yu et al. (2024) [20]	0.9736	0.9781	0.9778	0.9780	0.99
Zeghlache et al. (2022) [92]	NR	0.84	0.75	0.76	NR
Zhang et al. (2023) [93]	0.9005	NR	NR	NR	NR

Note: NR represents not reported.

Table A3. Dataset characteristics and implementation challenges.

Study	Dataset Source	Data Type	Technical Challenges	Recommendations
Adhithyaa et al. (2023) [22]	Open-source + proprietary	Behavioral (facial landmarks)	Hardware constraints, facial region variability, lighting changes	Adaptive architecture (fusion and sub-models) reduces overfitting; augmentation and pyramid input improve stability

Table A3. *Cont.*

Study	Dataset Source	Data Type	Technical Challenges	Recommendations
Ahmed et al. (2022) [23]	Open-source	Behavioral (facial landmarks: eyes and mouth regions)	Lighting variation, facial occlusion, smiling confused with eye closure	Facial subsampling and weighted ensemble improved robustness and reduced false detections
Akrout & Fakhfakh (2023) [24]	Open-source + proprietary	Behavioral (facial landmarks: iris, eyelids, head pose)	Lighting variation, facial occlusion, fatigue state subjectivity	Iris normalization, MediaPipe landmarks, multi-source feature fusion improved robustness
Alameen & Alhothali (2023) [25]	Open-source	Behavioral (RGB video)	Illumination variance; distractions from background	BN layer placement affects generalization per dataset
Alghanim et al. (2024) [26]	Open-source	Physiological (EEG spectrogram images)	Nonstationarity of EEG; data augmentation not fully effective; high training time	Use of Inception and dilated ResNet blocks; 30–50% overlap improves robustness
Alguindigue et al. (2024) [27]	Proprietary	Physiological (HRV, EDA), Behavioral (Eye tracking)	Class imbalance (especially in Eye model); device calibration	Use ensemble methods; improve minority class detection
Almazroi et al. (2023) [28]	Proprietary	Behavioral (facial landmarks: eye, mouth; objects; seatbelt use)	Occlusion, mouth covered by hand, low-light performance, alert timing threshold sensitivity	Eye/mouth ratio and MobileNet SSD integration improves accuracy and speed
Anber et al. (2022) [29]	Open-source	Behavioral (head position, mouth movement; face-based behavioral cues)	Lighting sensitivity; eye occlusion; limited generalizability	Combining AlexNet with NMF and SVM improves performance over transfer learning alone
Ansari et al. (2022) [30]	Proprietary	Behavioral (head posture)	Limited dataset; subjectivity in labeling; variability in fatigue behavior	Future use of smart seats and clothing; address data limitations with unsupervised clustering
Arefnezhad et al. (2020) [31]	Proprietary	Vehicle-based (steering wheel angle, velocity, yaw rate, lateral deviation, acceleration)	Noisy signals; high intra-class variability; difficult to differentiate moderate vs. extreme drowsiness	Use of CNN and RNN improves temporal modeling and detection accuracy
Bearly & Chitra (2024) [32]	Open-source	Behavioral (face: eyes, mouth, head position; RGB/NIR)	Noise in individual frame decisions; resolved using temporal smoothing	Combining GAN with multilevel attention improved robustness and reduced false alarms
Bekhouche et al. (2022) [33]	Open-source	Behavioral (video frames)	Class imbalance; scenario dependency; facial variation at night	Use of FCFS reduced features from 4096 to ~253 with better performance
Benmohamed & Zarzour (2024) [34]	Open-source	Behavioral (facial structural metrics and CNN features)	Low quality of IR video at night reduces feature extraction reliability	Combining CNN and structural fusion; frame aggregation improved detection
J. Chen, Wang, Wang et al. (2022) [35]	Proprietary	Physiological (EEG phase coherence images)	Small dataset, inter-subject variability, data imbalance	Use of relu4 layer and SVM improves results; extract features from shallow layers
J. Chen et al. (2021) [36]	Proprietary	Physiological (EEG—14-channel Emotiv EPOC, 128 Hz)	Signal noise, inter-subject variability, small sample size	End-to-end learning on raw EEG improved generalization and accuracy versus handcrafted features
J. Chen, Wang, He et al. (2022) [37]	Proprietary	Physiological (EEG—14 channels, Emotiv EPOC, 128 Hz)	EEG drift, signal noise, intra-subject variability, small dataset	Use of PLI adjacency matrices and multi-frequency band fusion improved CNN performance
C. Chen et al. (2023) [38]	Proprietary	Physiological (EEG—24-channel wireless dry EEG, 250 Hz)	EEG noise, inter-subject variability, data correlation due to continuous signals	Combining temporal-channel attention, covariance matrix, and capsule routing improved generalization and interpretability

Table A3. *Cont.*

Study	Dataset Source	Data Type	Technical Challenges	Recommendations
Chew et al. (2024) [16]	Open-source + proprietary	Multimodal (behavioral: face images; physiological: rPPG, HR)	Lighting sensitivity; camera angle dependency; noise in rPPG signal	Use of top-center camera angle and 4K webcam; optimize for embedded deployment
Civik & Yuzgec (2023) [39]	Open-source	Behavioral (facial landmarks: eyes and mouth)	Lighting variation, eye occlusion, mouth coverage, system delay under low light	Separate CNNs for eye and mouth improve detection of complex fatigue states
Cui et al. (2022) [40]	Open-source	Physiological (single-channel EEG from Oz)	EEG variability, low SNR, inter-subject drift	Use of GAP layer, CAM visualization to enhance interpretability
Ding et al. (2024) [41]	Open-source	Physiological (EEG)	Limited training data, subject variability, anomaly detection, similarity measures	Use of attention-based feature extraction and anomaly detection block improves robustness
Dua et al. (2021) [42]	Open-source	Behavioral (RGB video frames and optical flow)	Class imbalance; occlusion from sunglasses or hand gestures	Use of ensemble to balance strengths of each model
Ebrahimian et al. (2022) [43]	Proprietary	Physiological (ECG, respiration via thermal imaging)	Signal variability, mechanical latency in physiological response, labeling subjectivity	Multi-signal fusion (HRV, PSD, RR); CNN-LSTM superior to CNN in most tasks
Fa et al. (2023) [44]	Open-source	Behavioral (facial landmarks via OpenPose)	Occlusion, inter-subject variation, facial landmark misalignment	Multi-scale graph aggregation and coordinate attention improved spatial-temporal robustness
X. Feng, Guo et al. (2024) [45]	Open-source	Physiological (EEG—Oz channel)	Signal noise, inter-subject variability, cross-subject generalization challenge	Adaptive thresholding, GAP, and ECAM modules enhanced robustness and interpretability
X. Feng, Dai et al. (2025) [46]	Open-source	Physiological (EEG—30 channels for SADT, 17 channels for SEED-VIG)	EEG noise, EMG interference, inter-subject variability	Coordinate attention, LMMD, and curriculum pseudo labeling improved generalization across subjects
W. Feng et al. (2025) [47]	Open-source	Physiological (EEG—30 channels, 128 Hz)	Channel duplication, subject variability, noisy signals	Gumbel-Softmax improves channel selection; separable CNN reduces complexity and increases accuracy
Florez et al. (2023) [21]	Open-source	Behavioral (eye region video frames)	Small dataset; limited generalization; image redundancy	ROI correction using MediaPipe and CNN; Grad-CAM for interpretability
Gao et al. (2019) [13]	Proprietary	Physiological (multichannel EEG)	EEG autocorrelation; subject variability; feature interpretability	Uses multiplex recurrence network and mutual information matrix for CNN input
Guo & Markoni (2019) [48]	Open-source	Behavioral (RGB facial landmarks)	Limited demographic diversity; expression similarity among classes	Use temporal context (LSTM) improved stability over single-frame CNNs
C. He et al. (2024) [15]	Proprietary	Multimodal (physiological: PPG, heart rate, GSR, wrist acceleration; vehicle-based: velocity, acceleration, direction, slope, load; temporal: time, rest/work duration)	Class imbalance, subjectivity in video labeling, small sample	Fusion of diverse features improved robustness; dropout and focal loss used to stabilize training
H. He et al. (2020) [49]	Open-source + proprietary	Behavioral (facial regions: eyes, mouth)	Illumination variation; need for gamma correction	Use of gamma correction, lightweight CNN design, real-time test
L. He et al. (2024) [50]	Open-source	Physiological (EEG, EOG)	Class imbalance, overfitting, small datasets	GAN-based data augmentation (WGAN-GP); adaptive convolution; attention mechanisms

Table A3. *Cont.*

Study	Dataset Source	Data Type	Technical Challenges	Recommendations
Nguyen et al. (2023) [18]	Proprietary	Physiological (wireless EEG: behind-the-ear, single-channel, real-time)	Low resolution EEG; motion artifacts; noise filtering challenges	Dropout, quantization, batch normalization improve lightweight model performance in embedded setup
Hu et al. (2024) [17]	Proprietary	Physiological (EEG, 14 channels)	EEG artifacts, subjectivity in annotation, class imbalance	Functional brain region partitioning, multi-branch fusion, focal loss for imbalance
Huang et al. (2022) [51]	Open-source	Behavioral (facial landmarks: global face, eyes, mouth, glabella)	Pose variation, occlusion, ambiguous yawning vs. speaking	Combining multi-granularity input, FRN, and FFN improved accuracy and robustness in head pose variation
Hultman et al. (2021) [52]	Proprietary	Physiological (EEG, EOG, ECG)	Sensor noise; inter-subject variability; class imbalance	Combining EEG and ECG features helps accuracy; early fusion preferred
Iwamoto et al. (2021) [53]	Proprietary	Physiological (ECG—RRI sequences)	Ambiguity in expert scoring, inter-participant variability, simulator realism gap	LSTM-AE improved anomaly detection over PCA and HRV-based models
Jamshidi et al. (2021) [54]	Open-source	Behavioral (facial landmarks)	Occlusion, overfitting on training data, temporal labeling noise	Combination of spatial and temporal phases improved detection; situation-specific training helped generalization
Jarndal et al. (2025) [55]	Open-source	Behavioral (face video; full facial image)	Hardware resource constraints; lighting variation; face obstruction	Uses entire face with ViT, improving robustness in occluded or dark scenes
Jia et al. (2022) [56]	Proprietary	Behavioral (facial landmarks: eyes, mouth, head pose)	Facial occlusion (glasses, masks), lighting variability, real-time inference delay	Multi-module design with feature fusion increased robustness to occlusion and inconsistent signals
Jiao & Jiang (2022) [57]	Proprietary	Physiological (EOG, EEG via O2 and HSUM)	Data imbalance; low SEM frequency; limited samples for DL	Combine with SMOTE or GAN to improve sample size and reduce FP
Jiao et al. (2020) [14]	Proprietary	Physiological (EEG, EOG)	Temporal imprecision in labeling; signal variability; class imbalance	CWGAN improved data balance; sliding window settings boosted stability
Jiao et al. (2023) [58]	Open-source	Physiological (EOG)	Data imbalance; noise in physiological signals; variability across subjects	Multi-scale convolution improves feature representation
Kielty et al. (2023) [59]	Open-source + proprietary	Behavioral (event-based facial sequences, seatbelt motion)	Event sparsity in static scenes; hand-over-mouth occlusion; class imbalance in seatbelt transitions	Event fusion, attention maps, and recurrent layers improve robustness under occlusion and motion variance
Kır Savaş & Becerikli (2022) [60]	Open-source + proprietary	Behavioral (RGB face images: eyes and mouth)	Reconstruction error and model depth tuning; sensitivity to lighting and occlusion	DBN used for unsupervised feature learning; performance improved with deeper layers and symptom-specific models
Kumar et al. (2023) [61]	Open-source	Behavioral (RGB facial landmarks)	Overfitting risk; image quality issues; lighting variation	Dropout and global average pooling layers added for better generalization
Lamaazi et al. (2023) [62]	Open-source + proprietary	Multimodal (behavioral: eyes, mouth; vehicle-based: acceleration x/y/z)	Class confusion between yawning vs. mouth open, head tilt, lighting variation, accelerometer sequence noise	Multistage detection (vision and sensors) reduced false positives and improved detection latency
Latreche et al. (2025) [63]	Open-source	Physiological (EEG 32-channel)	Small sample size; lack of generalization; manual label noise	Optimization with RS and Optuna improved precision, reduced overfitting

Table A3. *Cont.*

Study	Dataset Source	Data Type	Technical Challenges	Recommendations
Q. Li et al. (2024) [64]	Proprietary	Physiological (EEG from 32-channel cap, regions A–D)	High search cost; model performance sensitive to EEG region used	NAS enables optimal trade-off between performance and efficiency
T. Li & Li (2024) [65]	Open-source	Behavioral (face video; EAR, MAR, head pose, ViT output)	Occlusion (glasses), pose variation, detection failures, dataset limitations	ViT adds global semantics; LSTM captures drowsiness temporal trends
Lin et al. (2025) [66]	Open-source + proprietary	Physiological (EEG)	EEG noise, individual variability, fine-grained imbalance	Combining attention, fusion, and channel selection improve generalization; focal loss addresses imbalance
Majeed et al. (2023) [67]	Open-source	Behavioral (video frames)	Potential overfitting; effect of data augmentation; occlusion challenges	Augmentation helps generalization; CNN-RNN for spatiotemporal features
Mate et al. (2024) [68]	Open-source	Behavioral (images extracted from videos)	Overfitting risk; limited generalization; tuning challenges	Use of multiple architectures improves comparability; augmentation used
Min et al. (2023) [69]	Open-source	Multimodal (physiological: EEG; behavioral: eye images via video)	Facial and muscle artifacts; low-channel EEG; inter-subject generalization	Fusion of EEG and SIFT-based eye features improves detection robustness
Mukherjee & Roy (2024) [70]	Proprietary	Multimodal (physiological: EEG, EMG, pulse, respiration, GSR; behavioral: head movement)	Signal noise, subjectivity in labeling; short windows improve detection	Use of TLSTM and attention for temporal relevance; 250 ms window practical
Nandyal & Sharanabasappa (2024) [71]	Open-source	Behavioral (video/image)	Class imbalance; vanishing gradient; overfitting risks	Use of optimization algorithm (FA) to avoid local optima
Obaidan et al. (2024) [72]	Open-source	Physiological (EEG)	Limited training data, inter-subject variability, non-stationary EEG signals	Multi-scale CNN architecture, DE preprocessing, spatial-spectral learning improves robustness
Paulo et al. (2021) [73]	Open-source	Physiological (EEG)	Inter-subject variability, low SNR, generalization challenges	Explore RNNs, reduce channels, apply attention mechanisms
Peng et al. (2024) [74]	Proprietary	Multimodal (behavioral: RGB face video; physiological: HR, EDA, BVP)	Signal noise, subjectivity in self-labeling, imbalance of fatigue levels	Combines facial and physiological data; short-window detection; attention maps
Priyanka et al. (2024) [75]	Proprietary	Multimodal (behavioral, physiological, vehicle-based)	Imbalanced dataset; need for SMOTE	Personalized models; real-world testing suggested
Quddus et al. (2021) [76]	Proprietary	Behavioral (eye image patches 48×48 from 2 cameras)	Facial occlusion, lighting variation, mismatch between EEG and video timestamps	C-LSTM outperforms eye-tracking methods with lower error and better generalizability
Ramzan et al. (2024) [77]	Open-source	Behavioral (video: eye, face, mouth regions)	Training time (CNN: ~603s/epoch, Hybrid: ~206s); overfitting managed by dropout	Combining hybrid CNN, PCA, and HOG boosts accuracy and training efficiency
Sedik et al. (2023) [78]	Open-source	Behavioral (RGB facial landmarks, eye and mouth region, NIR for DROZY)	DROZY limitations (lighting, occlusion); overfitting risk mitigated by augmentation	Combining image and video datasets improves robustness across symptoms
Shalash (2021) [79]	Open-source	Physiological (single-channel EEG converted to spectrogram)	Overfitting; small dataset; high computational cost	Use of reassignment spectrogram, dropout, and L2 regularization
Sharanabasappa & Nandyal (2022) [80]	Open-source	Behavioral (RGB images: eye, face, mouth regions)	Manual annotation burden; subjectivity in labeling; variance in image quality	Feature selection improves accuracy over deep CNNs; ensemble stabilizes prediction

Table A3. *Cont.*

Study	Dataset Source	Data Type	Technical Challenges	Recommendations
Sohail et al. (2024) [81]	Open-source	Behavioral (face images: eyes open/closed, yawn/no yawn)	Lighting conditions; camera placement; lack of occlusion handling	CNN and SMOTE used for balance; MaxPooling and Softmax activation
Soman et al. (2024) [19]	Open-source + proprietary	Behavioral (facial landmarks: EAR, MAR, PUC, MOE from camera)	Pupil detection errors, lighting variation, cultural facial trait diversity	Facial ratio fusion (EAR, MAR, PUC, MOE), Jetson Nano optimization, dropout and early stopping improve robustness
Sun et al. (2023) [82]	Open-source + proprietary	Behavioral (facial landmarks)	Low-quality inputs, occlusion, inconsistent eye states, data balance	FAM and SIM improve feature fusion on noisy inputs
Tang et al. (2024) [83]	Open-source	Physiological (21-channel EEG: forehead, temporal, posterior)	EEG noise, inter-subject variability, impact of CAM module placement	CAM placement after MSCNN improves feature quality and channel weighting
Turki et al. (2024) [84]	Open-source	Behavioral (face video; eye/mouth landmarks)	Image quality, illumination, face rotation, occlusion, false positives	Ensemble of CNNs and Chebyshev distance improves robustness and reduces false alerts
Vijaypriya & Uma (2023) [85]	Open-source	Behavioral (facial landmarks)	Low data variety; synthetic augmentation mentioned but not detailed	Use of Flamingo Search Optimization and wavelet feature fusion improves accuracy
Wang et al. (2025) [86]	Open-source	Physiological (EEG)	EEG signal noise, incomplete data, zero padding effects	Fused RWECN, DE, SQ features improved accuracy; filling missing values
Wijnands et al. (2020) [87]	Open-source	Behavioral (facial video—yawning, blinking, nodding, head pose)	Small dataset size; weak segment-level labels; slow inference on mobile devices	Temporal fusion and 3D depthwise convolutions improved robustness to occlusion and blinking patterns
H. Yang et al. (2021) [88]	Open-source + proprietary	Behavioral (RGB video frames)	Low resolution; camera vibration; similar facial actions	Use additional features for diverse lighting; improve resolution/deblurring
E. Yang & Yi (2024) [89]	Open-source	Behavioral (facial landmarks)	Small dataset size, simplified binary labeling, generalization challenges	Use of NGO for hyperparameter tuning and ShuffleNet for efficient extraction
K. Yang et al. (2025) [90]	Open-source	Physiological (EEG: DE, PSD, FE, SCC functional network)	EEG noise; differences in feature impact by dataset	Feature-level fusion improves classification versus single-feature GCNs
You et al. (2019) [91]	Open-source + proprietary	Behavioral (RGB facial video; EAR from eye landmarks)	Illumination variation, landmark errors, small eye sizes	EAR individualized classifier; uses PERCLOS and head position fallback
Yu et al. (2024) [20]	Proprietary	Multimodal (physiological: PPG; Behavioral: facial; behavioral: head pose)	Signal noise, alignment of video and PPG, fusion across modalities	Fusion of PPG, facial, and head pose improves accuracy; ensemble needed
Zeghlache et al. (2022) [92]	Open-source	Physiological (EEG, EOG)	Encoding loss tradeoff; optimal zdim tuning; noisy EEG channels	Dimensionality reduction improves classification; LSTM-VAE offers balanced features
Zhang et al. (2023) [93]	Open-source	Behavioral (facial landmarks: local patches, eyes, mouth, nose, full face)	Head pose variation, occlusion, ambiguous blinking vs. closing eyes	Multi-granularity representation and LSTM fusion improve spatial-temporal robustness

References

1. Grandjean, E. Fatigue in industry. *Br. J. Ind. Med.* **1979**, *36*, 175–186. [CrossRef]
2. Williamson, A.; Lombardi, D.A.; Folkard, S.; Stutts, J.; Courtney, T.K.; Connor, J.L. The link between fatigue and safety. *Accid. Anal. Prev.* **2011**, *43*, 498–515. [CrossRef]
3. Tefft, B.C. *Drowsy Driving in Fatal Crashes, United States, 2017–2021*; AAA Foundation for Traffic Safety: Washington, DC, USA, 2024.

4. Areal, A.; Pires, C.; Pita, R.; Marques, P.; Trigoso, J. *Distraction (Mobile Phone Use) & Fatigue*; ESRA3 Thematic report Nr. 3; ESRA: Utrecht, The Netherlands, 2024.
5. Ebrahim Shaik, M. A systematic review on detection and prediction of driver drowsiness. *Transp. Res. Interdiscip. Perspect.* **2023**, *21*, 100864. [CrossRef]
6. El-Nabi, S.A.; El-Shafai, W.; El-Rabaie, E.S.M.; Ramadan, K.F.; Abd El-Samie, F.E.; Mohsen, S. Machine learning and deep learning techniques for driver fatigue and drowsiness detection: A review. *Multimed. Tools Appl.* **2024**, *83*, 9441–9477. [CrossRef]
7. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
8. Fonseca, T.; Ferreira, S. Drowsiness Detection in Drivers: A Systematic Review of Deep Learning-Based Models. Available online: https://www.crd.york.ac.uk/PROSPERO/view/CRD420251078841 (accessed on 12 July 2025).
9. Weng, C.-H.; Lai, Y.-H.; Lai, S.-H. Driver Drowsiness Detection via a Hierarchical Temporal Deep Belief Network. In *Asian Conference on Computer Vision Workshop on Driver Drowsiness Detection from Video*; Springer: Taipei, Taiwan, 2016.
10. Abtahi, S.; Omidyeganeh, M.; Shirmohammadi, S.; Hariri, B. YawDD: A yawning detection dataset. In Proceedings of the 5th ACM Multimedia Systems Conference, Singapore, 19 March 2014; Association for Computing Machinery: New York, NY, USA; pp. 24–28.
11. Zheng, W.L.; Lu, B.L. A multimodal approach to estimating vigilance using EEG and forehead EOG. *J. Neural Eng.* **2017**, *14*, 026017. [CrossRef]
12. Cao, Z.; Chuang, C.H.; King, J.K.; Lin, C.T. Multi-channel EEG recordings during a sustained-attention driving task. *Sci. Data* **2019**, *6*, 19. [CrossRef] [PubMed]
13. Gao, Z.K.; Li, Y.L.; Yang, Y.X.; Ma, C. A recurrence network-based convolutional neural network for fatigue driving detection from EEG. *Chaos* **2019**, *29*, 113126. [CrossRef] [PubMed]
14. Jiao, Y.; Deng, Y.; Luo, Y.; Lu, B.L. Driver sleepiness detection from EEG and EOG signals using GAN and LSTM networks. *Neurocomputing* **2020**, *408*, 100–111. [CrossRef]
15. He, C.; Xu, P.; Pei, X.; Wang, Q.; Yue, Y.; Han, C. Fatigue at the wheel: A non-visual approach to truck driver fatigue detection by multi-feature fusion. *Accid. Anal. Prev.* **2024**, *199*, 107511. [CrossRef]
16. Chew, Y.X.; Razak, S.F.A.; Yogarayan, S.; Ismail, S.N.M.S. Dual-Modal Drowsiness Detection to Enhance Driver Safety. *Comput. Mater. Contin.* **2024**, *81*, 4397–4417. [CrossRef]
17. Hu, F.; Zhang, L.; Yang, X.; Zhang, W.A. EEG-Based Driver Fatigue Detection Using Spatio-Temporal Fusion Network with Brain Region Partitioning Strategy. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 9618–9630. [CrossRef]
18. Nguyen, H.T.; Mai, N.D.; Lee, B.G.; Chung, W.Y. Behind-the-Ear EEG-Based Wearable Driver Drowsiness Detection System Using Embedded Tiny Neural Networks. *IEEE Sens. J.* **2023**, *23*, 23875–23892. [CrossRef]
19. Soman, S.P.; Kumar, G.S.; Nuthalapati, S.B.; Zafar, S.; Abubeker, K.M. Internet of things assisted deep learning enabled driver drowsiness monitoring and alert system using CNN-LSTM framework. *Eng. Res. Express* **2024**, *6*, 045239. [CrossRef]
20. Yu, L.; Yang, X.; Wei, H.; Liu, J.; Li, B. Driver fatigue detection using PPG signal, facial features, head postures with an LSTM model. *Heliyon* **2024**, *10*, e39479. [CrossRef] [PubMed]
21. Florez, R.; Palomino-Quispe, F.; Coaquira-Castillo, R.J.; Herrera-Levano, J.C.; Paixão, T.; Alvarez, A.B. A CNN-Based Approach for Driver Drowsiness Detection by Real-Time Eye State Identification. *Appl. Sci.* **2023**, *13*, 7849. [CrossRef]
22. Adhithyaa, N.; Tamilarasi, A.; Sivabalaselvamani, D.; Rahunathan, L. Face Positioned Driver Drowsiness Detection Using Multistage Adaptive 3D Convolutional Neural Network. *Inf. Technol. Control* **2023**, *52*, 713–730. [CrossRef]
23. Ahmed, M.; Masood, S.; Ahmad, M.; Abd El-Latif, A.A. Intelligent Driver Drowsiness Detection for Traffic Safety Based on Multi CNN Deep Model and Facial Subsampling. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19743–19752. [CrossRef]
24. Akrout, B.; Fakhfakh, S. How to Prevent Drivers before Their Sleepiness Using Deep Learning-Based Approach. *Electronics* **2023**, *12*, 965. [CrossRef]
25. Alameen, S.A.; Alhothali, A.M. A Lightweight Driver Drowsiness Detection System Using 3DCNN with LSTM. *Comput. Syst. Sci. Eng.* **2022**, *44*, 895–912. [CrossRef]
26. Alghanim, M.; Attar, H.; Rezaee, K.; Khosravi, M.; Solyman, A.; Kanan, M.A. A Hybrid Deep Neural Network Approach to Recognize Driving Fatigue Based on EEG Signals. *Int. J. Intell. Syst.* **2024**, *2024*, 9898333. [CrossRef]
27. Alguindigue, J.; Singh, A.; Narayan, A.; Samuel, S. Biosignals Monitoring for Driver Drowsiness Detection Using Deep Neural Networks. *IEEE Access* **2024**, *12*, 93075–93086. [CrossRef]
28. Almazroi, A.A.; Alqarni, M.A.; Aslam, N.; Shah, R.A. Real-Time CNN-Based Driver Distraction & Drowsiness Detection System. *Intell. Autom. Soft Comput.* **2023**, *37*, 2153–2174. [CrossRef]
29. Anber, S.; Alsaggaf, W.; Shalash, W. A Hybrid Driver Fatigue and Distraction Detection Model Using AlexNet Based on Facial Features. *Electronics* **2022**, *11*, 285. [CrossRef]

30. Ansari, S.; Naghdy, F.; Du, H.; Pahnwar, Y.N. Driver Mental Fatigue Detection Based on Head Posture Using New Modified reLU-BiLSTM Deep Neural Network. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 10957–10969. [CrossRef]
31. Arefnezhad, S.; Samiee, S.; Eichberger, A.; Frühwirth, M.; Kaufmann, C.; Klotz, E. Applying deep neural networks for multi-level classification of driver drowsiness using Vehicle-based measures. *Expert. Syst. Appl.* **2020**, *162*, 113778. [CrossRef]
32. Bearly, E.M.; Chitra, R. Automatic drowsiness detection for preventing road accidents via 3dgan and three-level attention. *Multimed. Tools Appl.* **2024**, *83*, 48261–48274. [CrossRef]
33. Bekhouche, S.E.; Ruichek, Y.; Dornaika, F. Driver drowsiness detection in video sequences using hybrid selection of deep features. *Knowl. Based Syst.* **2022**, *252*, 109436. [CrossRef]
34. Benmohamed, A.; Zarzour, H. A Deep Learning-Based System for Driver Fatigue Detection. *Ing. Des Syst. D'information* **2024**, *29*, 1779–1788. [CrossRef]
35. Chen, J.; Wang, H.; Wang, S.; He, E.; Zhang, T.; Wang, L. Convolutional neural network with transfer learning approach for detection of unfavorable driving state using phase coherence image. *Expert. Syst. Appl.* **2022**, *187*, 116016. [CrossRef]
36. Chen, J.; Wang, S.; He, E.; Wang, H.; Wang, L. Recognizing drowsiness in young men during real driving based on electroencephalography using an end-to-end deep learning approach. *Biomed. Signal. Process. Control* **2021**, *69*, 102792. [CrossRef]
37. Chen, J.; Wang, S.; He, E.; Wang, H.; Wang, L. Two-dimensional phase lag index image representation of electroencephalography for automated recognition of driver fatigue using convolutional neural network. *Expert Syst. Appl.* **2022**, *191*, 116339. [CrossRef]
38. Chen, C.; Ji, Z.; Sun, Y.; Bezerianos, A.; Thakor, N.; Wang, H. Self-Attentive Channel-Connectivity Capsule Network for EEG-Based Driving Fatigue Detection. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 3152–3162. [CrossRef]
39. Civik, E.; Yuzgec, U. Real-time driver fatigue detection system with deep learning on a low-cost embedded system. *Microprocess. Microsyst.* **2023**, *99*, 104851. [CrossRef]
40. Cui, J.; Lan, Z.; Liu, Y.; Li, R.; Li, F.; Sourina, O.; Müller-Wittig, W. A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods* **2022**, *202*, 173–184. [CrossRef]
41. Ding, N.; Zhang, C.; Eskandarian, A. EEG-fest: Few-shot based attention network for driver's drowsiness estimation with EEG signals. *Biomed. Phys. Eng. Express* **2024**, *10*, 015008. [CrossRef] [PubMed]
42. Dua, M.; Shakshi Singla, R.; Raj, S.; Jangra, A. Deep CNN models-based ensemble approach to driver drowsiness detection. *Neural Comput. Appl.* **2021**, *33*, 3155–3168. [CrossRef]
43. Ebrahimian, S.; Nahvi, A.; Tashakori, M.; Salmanzadeh, H.; Mohseni, O.; Leppänen, T. Multi-Level Classification of Driver Drowsiness by Simultaneous Analysis of ECG and Respiration Signals Using Deep Neural Networks. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10736. [CrossRef]
44. Fa, S.; Yang, X.; Han, S.; Feng, Z.; Chen, Y. Multi-scale spatial–temporal attention graph convolutional networks for driver fatigue detection. *J. Vis. Commun. Image Represent.* **2023**, *93*, 103826. [CrossRef]
45. Feng, X.; Guo, Z.; Kwong, S. ID3RSNet: Cross-subject driver drowsiness detection from raw single-channel EEG with an interpretable residual shrinkage network. *Front. Neurosci.* **2024**, *18*, 1508747. [CrossRef] [PubMed]
46. Feng, X.; Dai, S.; Guo, Z. Pseudo-label-assisted subdomain adaptation network with coordinate attention for EEG-based driver drowsiness detection. *Biomed. Signal. Process. Control* **2025**, *101*, 107132. [CrossRef]
47. Feng, W.; Wang, X.; Xie, J.; Liu, W.; Qiao, Y.; Liu, G. Real-Time EEG-Based Driver Drowsiness Detection Based on Convolutional Neural Network With Gumbel-Softmax Trick. *IEEE Sens. J.* **2025**, *25*, 1860–1871. [CrossRef]
48. Guo, J.M.; Markoni, H. Driver drowsiness detection using hybrid convolutional neural network and long short-term memory. *Multimed. Tools Appl.* **2019**, *78*, 29059–29087. [CrossRef]
49. He, H.; Zhang, X.; Jiang, F.; Wang, C.; Yang, Y.; Liu, W.; Peng, J. A Real-time Driver Fatigue Detection Method Based on Two-Stage Convolutional Neural Network. *IFAC-PapersOnLine* **2020**, *53*, 15374–15379. [CrossRef]
50. He, L.; Zhang, L.; Sun, Q.; Lin, X.T. A generative adaptive convolutional neural network with attention mechanism for driver fatigue detection with class-imbalanced and insufficient data. *Behav. Brain Res.* **2024**, *464*, 114898. [CrossRef]
51. Huang, R.; Wang, Y.; Li, Z.; Lei, Z.; Xu, Y. RF-DCM: Multi-Granularity Deep Convolutional Model Based on Feature Recalibration and Fusion for Driver Fatigue Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 630–640. [CrossRef]
52. Hultman, M.; Johansson, I.; Lindqvist, F.; Ahlström, C. Driver sleepiness detection with deep neural networks using electrophysiological data. *Physiol. Meas.* **2021**, *42*, 034001. [CrossRef] [PubMed]
53. Iwamoto, H.; Hori, K.; Fujiwara, K.; Kano, M. Real-driving-implementable drowsy driving detection method using heart rate variability based on long short-term memory and autoencoder. *IFAC-PapersOnLine* **2021**, *54*, 526–531. [CrossRef]
54. Jamshidi, S.; Azmi, R.; Sharghi, M.; Soryani, M. Hierarchical deep neural networks to detect driver drowsiness. *Multimed. Tools Appl.* **2021**, *80*, 16045–16058. [CrossRef]
55. Jarndal, A.; Tawfik, H.; Siam, A.I.; Alsyouf, I.; Cheaitou, A. A Real-Time Vision Transformers-Based System for Enhanced Driver Drowsiness Detection and Vehicle Safety. *IEEE Access* **2025**, *13*, 1790–1803. [CrossRef]
56. Jia, H.; Xiao, Z.; Ji, P. Real-time fatigue driving detection system based on multi-module fusion. *Comput. Graph.* **2022**, *108*, 22–33. [CrossRef]

57. Jiao, Y.; Jiang, F. Detecting slow eye movements with bimodal-LSTM for recognizing drivers' sleep onset period. *Biomed. Signal. Process. Control* **2022**, *75*, 103608. [CrossRef]
58. Jiao, Y.; He, X.; Jiao, Z. Detecting slow eye movements using multi-scale one-dimensional convolutional neural network for driver sleepiness detection. *J. Neurosci. Methods* **2023**, *397*, 109939. [CrossRef]
59. Kielty, P.; Dilmaghani, M.S.; Shariff, W.; Ryan, C.; Lemley, J.; Corcoran, P. Neuromorphic Driver Monitoring Systems: A Proof-of-Concept for Yawn Detection and Seatbelt State Detection using an Event Camera. *IEEE Access* **2023**, *11*, 96363–96373. [CrossRef]
60. Savaş, B.K.; Becerikli, Y. Behavior-based driver fatigue detection system with deep belief network. *Neural Comput. Appl.* **2022**, *34*, 14053–14065. [CrossRef]
61. Kumar, V.; Sharma, S. Ranjeet: Driver drowsiness detection using modified deep learning architecture. *Evol. Intell.* **2023**, *16*, 1907–1916. [CrossRef]
62. Lamaazi, H.; Alqassab, A.; Fadul, R.A.; Mizouni, R. Smart Edge-Based Driver Drowsiness Detection in Mobile Crowdsourcing. *IEEE Access* **2023**, *11*, 21863–21872. [CrossRef]
63. Latreche, I.; Slatnia, S.; Kazar, O.; Harous, S. An optimized deep hybrid learning for multi-channel EEG-based driver drowsiness detection. *Biomed. Signal. Process. Control* **2025**, *99*, 106881. [CrossRef]
64. Li, Q.; Luo, Z.; Qi, R.; Zheng, J. Automatic Searching of Lightweight and High-Performing CNN Architectures for EEG-Based Driving Fatigue Detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–11. [CrossRef]
65. Li, T.; Li, C. A Deep Learning Model Based on Multi-Granularity Facial Features and LSTM Network for Driver Drowsiness Detection. *J. Appl. Sci. Eng.* **2024**, *27*, 2799–2811. [CrossRef]
66. Lin, X.; Huang, Z.; Ma, W.; Tang, W. EEG-based driver drowsiness detection based on simulated driving environment. *Neurocomputing* **2025**, *616*, 128961. [CrossRef]
67. Majeed, F.; Shafique, U.; Safran, M.; Alfarhood, S.; Ashraf, I. Detection of Drowsiness among Drivers Using Novel Deep Convolutional Neural Network Model. *Sensors* **2023**, *23*, 8741. [CrossRef] [PubMed]
68. Mate, P.; Apte, N.; Parate, M.; Sharma, S. Detection of driver drowsiness using transfer learning techniques. *Multimed. Tools Appl.* **2024**, *83*, 35553–35582. [CrossRef]
69. Min, J.; Cai, M.; Gou, C.; Xiong, C.; Yao, X. Fusion of forehead EEG with machine vision for real-time fatigue detection in an automatic processing pipeline. *Neural Comput. Appl.* **2023**, *35*, 8859–8872. [CrossRef]
70. Mukherjee, P.; Roy, A.H. A novel deep learning-based technique for driver drowsiness detection. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2024**, *34*, 667–684. [CrossRef]
71. Nandyal, S.; Sharanabasappa, S. Deep ResNet 18 and enhanced firefly optimization algorithm for on-road vehicle driver drowsiness detection. *J. Auton. Intell.* **2024**, *7*. [CrossRef]
72. Bin Obaidan, H.; Hussain, M.; AlMajed, R. EEG_DMNet: A Deep Multi-Scale Convolutional Neural Network for Electroencephalography-Based Driver Drowsiness Detection. *Electronics* **2024**, *13*, 2084. [CrossRef]
73. Paulo, J.R.; Pires, G.; Nunes, U.J. Cross-Subject Zero Calibration Driver's Drowsiness Detection: Exploring Spatiotemporal Image Encoding of EEG Signals for Convolutional Neural Network Classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 905–915. [CrossRef] [PubMed]
74. Peng, Y.; Deng, H.; Xiang, G.; Wu, X.; Yu, X.; Li, Y.; Yu, T. A Multi-Source Fusion Approach for Driver Fatigue Detection Using Physiological Signals and Facial Image. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 16614–16624. [CrossRef]
75. Priyanka, S.; Shanthi, S.; Saran Kumar, A.; Praveen, V. Data fusion for driver drowsiness recognition: A multimodal perspective. *Egypt. Inform. J.* **2024**, *27*, 100529. [CrossRef]
76. Quddus, A.; Shahidi Zandi, A.; Prest, L.; Comeau, F.J.E. Using long short term memory and convolutional neural networks for driver drowsiness detection. *Accid. Anal. Prev.* **2021**, *156*, 106107. [CrossRef]
77. Ramzan, M.; Abid, A.; Fayyaz, M.; Alahmadi, T.J.; Nobanee, H.; Rehman, A. A Novel Hybrid Approach for Driver Drowsiness Detection Using a Custom Deep Learning Model. *IEEE Access* **2024**, *12*, 126866–126884. [CrossRef]
78. Sedik, A.; Marey, M.; Mostafa, H. An Adaptive Fatigue Detection System Based on 3D CNNs and Ensemble Models. *Symmetry* **2023**, *15*, 1274. [CrossRef]
79. Shalash, W.M. A Deep Learning CNN Model for Driver Fatigue Detection Using Single EEG Channel. *J. Theor. Appl. Inf. Technol.* **2021**, *31*, 462–477.
80. Sharanabasappa; Nandyal, S. An ensemble learning model for driver drowsiness detection and accident prevention using the behavioral features analysis. *Int. J. Intell. Comput. Cybern.* **2022**, *15*, 224–244. [CrossRef]
81. Sohail, A.; Shah, A.A.; Ilyas, S.; Alshammry, N. A CNN-based Deep Learning Framework for Driver's Drowsiness Detection. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 169–178. [CrossRef]
82. Sun, Z.; Miao, Y.; Jeon, J.Y.; Kong, Y.; Park, G. Facial feature fusion convolutional neural network for driver fatigue detection. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106981. [CrossRef]

83. Tang, J.; Zhou, W.; Zheng, W.; Zeng, Z.; Li, J.; Su, R.; Adili, T.; Chen, W.; Chen, C.; Luo, J. Attention-Guided Multiscale Convolutional Neural Network for Driving Fatigue Detection. *IEEE Sens. J.* **2024**, *24*, 23280–23290. [CrossRef]
84. Turki, A.; Kahouli, O.; Albadran, S.; Ksantini, M.; Aloui, A.; Amara, M. Ben: A sophisticated Drowsiness Detection System via Deep Transfer Learning for real time scenarios. *AIMS Math.* **2024**, *9*, 3211–3234. [CrossRef]
85. Vijaypriya, V.; Uma, M. Facial Feature-Based Drowsiness Detection with Multi-Scale Convolutional Neural Network. *IEEE Access* **2023**, *11*, 63417–63429. [CrossRef]
86. Wang, K.; Mao, X.; Song, Y.; Chen, Q. EEG-based fatigue state evaluation by combining complex network and frequency-spatial features. *J. Neurosci. Methods* **2025**, *416*, 110385. [CrossRef]
87. Wijnands, J.S.; Thompson, J.; Nice, K.A.; Aschwanden, G.D.P.A.; Stevenson, M. Real-time monitoring of driver drowsiness on mobile platforms using 3D neural networks. *Neural Comput. Appl.* **2020**, *32*, 9731–9743. [CrossRef]
88. Yang, H.; Liu, L.; Min, W.; Yang, X.; Xiong, X. Driver Yawning Detection Based on Subtle Facial Action Recognition. *IEEE Trans. Multimed.* **2021**, *23*, 572–583. [CrossRef]
89. Yang, E.; Yi, O. Enhancing Road Safety: Deep Learning-Based Intelligent Driver Drowsiness Detection for Advanced Driver-Assistance Systems. *Electronics* **2024**, *13*, 708. [CrossRef]
90. Yang, K.; Zhang, K.; Hu, Y.; Xu, J.; Yang, B.; Kong, W.; Zhang, J. Adaptive multi-branch CNN of integrating manual features and functional network for driver fatigue detection. *Biomed. Signal. Process. Control* **2025**, *102*, 107262. [CrossRef]
91. You, F.; Li, X.; Gong, Y.; Wang, H.; Li, H. A Real-time Driving Drowsiness Detection Algorithm with Individual Differences Consideration. *IEEE Access* **2019**, *7*, 179396–179408. [CrossRef]
92. Zeghlache, R.; Labiod, M.A.; Mellouk, A. Driver vigilance estimation with Bayesian LSTM Auto-encoder and XGBoost using EEG/EOG data. *IFAC-PapersOnLine* **2022**, *55*, 89–94. [CrossRef]
93. Zhang, H.; Liu, T.; Lyu, J.; Chen, D.; Yuan, Z. Integrate memory mechanism in multi-granularity deep framework for driver drowsiness detection. *Intell. Robot.* **2023**, *3*, 614–631. [CrossRef]