Archives of Academic Emergency Medicine. 2026; 14(1): e19

OPEN ACCESS

REVIEW ARTICLE

Driver Drowsiness Detection using Machine Learning and Deep Learning Techniques: A Systematic Review

Saber Ghaffari fam[1,2], Ehsan Sarbazi[3], Senobar Naderian[4], Salman Khazaei[2,5], Mehmet Tatli[6], Hassan Soleimanpour[7]*

1. Department of Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
2. Student Research Committee, Hamadan University of Medical Sciences, Hamadan, Iran
3. Road Traffic Injury Research Center, Tabriz University of Medical Sciences, Tabriz, Iran
4. Student Research Committee, Tabriz University of Medical Sciences, Tabriz, Iran
5. Research Center for Health Sciences, Hamadan University of Medical Sciences, Hamadan, Iran
6. Emergency Medicine Department, Van Training and Research Hospital, University of Health Sciences , Van, Turkiye
7. Medical Philosophy and History Research Center, Imam Reza General Hospital, Tabriz University of Medical Sciences, Tabriz, Iran

Received: March 2026; Accepted: May 2026; Published online: 1 June 2026

Abstract: **Introduction:** Behavioral indicators have been increasingly utilized in machine learning (ML) and deep learning (DL) frameworks to enable automated driver drowsiness detection (DDD). This study aimed to investigate the available evidence on the modeling frameworks, datasets, input modalities, and performance metrics used in DDD systems. **Methods:** Studies were identified through systematic searches of PubMed, Scopus, Web of Science, and IEEE Xplore for English-language publications up to 31 August 2025. Eligible studies were original research that applied ML or DL techniques to image- or video-based behavioral features for automatic DDD. Studies focusing primarily on vehicle telemetry, physiological signals without behavioral imaging, reviews, editorials, and gray literature were excluded. We extracted the dataset, input data modality, driving context, inference mode, ML/DL methods, and performance metrics including accuracy, precision, recall, and F1-score. Risk of bias was evaluated using the PROBA-AI tool. **Results:** A total of 69 studies met the inclusion criteria. DL models outperformed ML, achieving higher median accuracy (94.48% vs. 91.80%) and significantly better F1-scores (93.15% vs. 84.00%). Median recall was comparable between DL and ML models (93.76% vs. 94.12%), and precision remained similarly high across both approaches (93.18% vs. 94.6%). Most DL methods employed Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), or hybrid CNN– Long Short-Term Memory (LSTM) architectures, whereas ML studies primarily relied on classical classifiers such as Support Vector Machine (SVM) and Random Forest (RF) supported by handcrafted behavioral features. PROBA-AI assessment revealed considerable methodological heterogeneity, with 27 studies rated as high risk, 25 as low risk, and 17 as moderate. **Conclusion:** DL models demonstrate clear performance advantages over ML approaches, particularly in accuracy and F1-score. However, substantial methodological variability-reflected in inconsistent dataset design, annotation practices, and validation strategies-continues to limit comparability across studies.

Keywords: Accidents; Machine learning; Deep learning; Fatigue; Drowsiness; Driver assistance

Cite this article as: Ghaffari fam S, Sarbazi E, Naderian S, et al. Driver Drowsiness Detection using Machine Learning and Deep Learning Techniques: A Systematic Review. Arch Acad Emerg Med. 2026; 14(1): e19. https://doi.org/10.22037/aaem.v14i1.2932.

1. Introduction

Driver drowsiness (DD) represents a significant threat to road safety. Drowsiness, commonly defined as a state of reduced alertness, manifests through diminished executive functioning, decreased cognitive effort, and involuntary muscular inhibition (1). Drowsiness compromises attentional control, slows reaction times, and impairs decision-making capacity, thereby elevating both the likelihood of accidents and the severity of associated injuries (2). According to the German Insurance Society, drowsy driving is implicated in approximately 25% of all road crashes in Germany (3). In France, driver drowsiness is associated with 14.9% of crash-related injuries and 20.6% of fatal collisions (4). Drowsiness can be triggered by multiple factors, including reduced arousal, elevated workload, and sleep-related disturbances. For example, long, monotonous rural roads with minimal traffic often create optimal conditions for the onset of this form of drowsiness (1). To mitigate this risk, researchers have increasingly turned to driver monitoring systems designed to

* **Corresponding Author:** Hassan Soleimanpour; Medical Philosophy and History Research Center, Imam Reza General Hospital, Tabriz University of Medical Sciences, Tabriz, Iran. Email: h.soleimanpour@gmail.com or soleimanpourh@tbzmed.ac.ir. Tel: +98 914116414, ORCID: https://orcid.org/0000-0002-1311-4096.

detect early indicators of drowsiness. Driver drowsiness detection (DDD) technologies evaluate a DD level and issue timely warnings about potential hazards that might otherwise remain unnoticed (5). Machine learning (ML), a central discipline within artificial intelligence, facilitates automated pattern recognition and underpins the ML and deep learning (DL) techniques widely applied in behavioral driver drowsiness detection (6, 7). Using large sets of training samples, DL algorithms efficiently learn and optimize the parameters necessary for effective DDD (8). DDD techniques are generally categorized into three groups based on the type of signals they assess: biological, vehicle-based, and behavioral approaches (9, 10). Behavioral metrics rely on observable indicators of drowsiness derived from the driver's head and facial characteristics. By monitoring head movements and facial cues-including those associated with the eyes, mouth, eyebrows, and breathing-these systems can identify the onset of driver drowsiness (11, 12).

Accordingly, real-time monitoring of the driver's state is crucial for enabling timely interventions, such as issuing alerts or activating automated driving functions. As a result, analysis based on images and videos can detect drowsiness-related indicators in real time and assess operator performance to ensure that tasks are executed safely and correctly (13).

Several previous reviews have examined driver drowsiness detection from various perspectives. Earlier surveys have offered broad overviews of physiological, vehicle-based, and behavioral indicators (10), and have summarized recent advances in sensor technologies and multimodal detection approaches (9). More recent scoping and systematic reviews have underscored persistent challenges, including inconsistent ground-truth labeling, limited real-world validation, and substantial variability in evaluation protocols (14). However, these reviews have not specifically addressed ML- and DL-based behavioral approaches, nor have they provided comparative evaluations of their performance across different datasets and driving contexts. This gap highlights the need for an updated synthesis focused exclusively on behavioral measures analyzed through ML and DL techniques.

Despite growing interest and notable advances in the field, the existing literature remains disjointed. Although numerous ML and DL architectures have been developed for driver drowsiness detection, few studies offer rigorous comparative evaluations, and no consensus has emerged regarding optimal model design. This study aimed to investigate the available evidence on the modeling frameworks, datasets, input modalities, and performance metrics used in DDD systems.

2. Methods

2.1. Study design and setting

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (15). Figure 1 displays the search strategy based on the PRISMA flow diagram. To find studies that were missed by the database search method, reference lists of the included articles were examined up to August 2025.

The review is structured around the following research questions: i: Which ML and DL architectures have been employed to analyze behavioral indicators of driver drowsiness, and what design patterns characterize these modeling approaches? ii: To what extent do these models demonstrate robust and generalizable performance in detecting drowsiness, as reflected in standardized metrics (accuracy, precision, recall, F1-score)? iii. Which datasets-characterized by their dataset names, Input Data Modality (images or videos), driving contexts (simulated or real-road), and inference modes (real-time or offline)-have been used to train and validate ML- and DL-based behavioral models for DDD, and how do these dataset characteristics affect methodological comparability and model performance? iv: What methodological strengths and weaknesses are evident across the included studies, as determined by PROBA-AI evaluations of participant selection, predictor specification, outcome labeling, and analytical rigor?

2.2. PICOS framework

The Population consisted of human driver behaviors; the Intervention referred to behavioral drowsiness-detection systems using ML or DL; the Comparison involved ground-truth labels derived from reference measures such as annotated video or image from facial landmarks; Outcomes included performance metrics (accuracy, precision, recall, and F1-score).

2.3. Protocol

Before 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 Reviews (15) under the reference CRD420251139663 (16). Researchers have used a variety of methods i.e., DL and ML to measure DDD, which can be used to carry out the detecting process. Figure 2 shows a brief categorization of DDD.

2.4. Search strategy

This systematic review involved searching four databases of PubMed, Scopus, Web of Science, and IEEE Xplore Digital Library. Furthermore, any relevant material discovered during the initial search was reviewed. The search was conducted until August 2025. To eliminate duplicates and ensure the chosen articles were pertinent to this review, we examined the titles and abstracts of the research works. The following terms were applied for searching the databases: ("Artificial Intelligence" OR "Machine learning" OR "Artificial neural network" OR "Behavioral measurement" OR "Adaptive learning" OR "Driver monitoring" OR "Deep learning" OR "Com-

puter vision" OR "neural nets" OR "Mathematical model"), AND ("Sleepiness" OR "Drowsiness" OR "Fatigue" OR "Feature extraction" OR "Distracted Driving" OR "Face" OR "Image" OR "Video" OR "Eye movements" OR "Visual attention model" OR "Feature extraction" OR "Driving style"), AND ("Crash" OR "Accidents").

A full electronic search strategy was developed and executed in (PubMed), ensuring reproducibility: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "computer vision" OR "adaptive learning" OR "driver monitoring" OR "mathematical model" AND ("sleepiness" OR "drowsiness" OR "fatigue" OR "eye movements" OR "visual attention" OR "face" OR "image" OR "video") AND ("driving" OR "crash" OR "traffic accident" OR "driver behavior")).

2.5. Eligibility criteria

The process involved identifying relevant studies through database searches, screening titles and abstracts, conducting full-text assessments to determine eligibility, and including studies that met all predefined criteria. Studies included in this review were required to meet all of the following criteria: i: the research had to evaluate an in-vehicle system specifically designed to mitigate DDD; ii: drowsiness had to be automatically detected by an external system using a predefined rule-based or algorithmic method that subsequently triggered an intervention; iii: the study had to be conducted in a realistic driving context, including passenger vehicle or commercial freight scenarios; iv: the analysis had to be based on original data collected from human participants; and v: only original, peer-reviewed research articles published in English with full-text with online availability were considered.

2.6. Exclusion criteria

Studies were excluded based on the following criteria: i: Data source: studies that did not use data from human participants, including those relying solely on computer simulations; ii: Publication type: review articles, commentaries, editorials, and gray literature were not considered; iii: Intervention focus: studies examining safety interventions not directly targeting DD were excluded; iv: Technology scope: interventions that were not in-vehicle technologies or did not incorporate a driver state monitoring system; and v: Type of impairment: research focusing on impairments caused by alcohol or drug use was excluded to maintain a clear focus on drowsiness.

Studies lacking direct reporting of performance metrics were excluded. studies other than those that returned results of "specificity," "sleepiness," "accidents," "deep learning" and "machine learning" (together with their synonyms) were excluded.

Following these calibration sessions, the researchers independently completed the title and abstract screening. To ensure greater consistency, one researcher (S.Gh) screened all studies, while the others (E.S and S.N) split the studies randomly. Initially, all paper titles and abstracts were reviewed to see if they fit the present review.

2.7. Data collection and extraction

The variables listed in the information record are first author, publication year, country, accuracy, recall (sensitivity), F1-score, precision (positive predictive value; PPV), modeling approach (ML, DL architecture), dataset type, driving context, and inference mode.

When multiple models were reported, the best-performing model was extracted for comparison. Following the collection of every article with abstracts, additional screening was carried out using the EndNote 8.

For data extraction, a structured spreadsheet was created in Microsoft Excel (Microsoft Corporation) to systematically capture key information from each included study.

2.8. PROBA-AI rating procedure

The risk of bias for each included study was assessed using the PROBA-AI tool, which evaluates methodological quality across five core domains: participants, predictors, outcome, analysis, and overall risk of bias. Each domain contains a series of signaling questions designed to identify potential sources of bias specific to artificial intelligence-based diagnostic or detection models. Responses to these questions are used to assign a domain-level rating of low risk, high risk, or unclear risk. A rating of low risk is given when the study provides sufficient methodological detail and meets all criteria without concerns; high risk indicates substantial methodological limitations or clear threats to validity; and unclear risk is applied when reporting is insufficient to make a definitive judgment. The overall risk of bias is determined by integrating all domain ratings, with any domain judged as high risk typically resulting in an overall high-risk classification.

2.9. Data synthesis

Given the substantial heterogeneity across the included studies, particularly in modeling objectives, input modalities, evaluation settings, and outcome measures, a quantitative meta-analysis was not feasible. Data synthesis was conducted using a structured narrative approach aligned with the aims guiding this review.

Several indicators have been used to evaluate a system's ability to identify participants who are fatigued or drowsy. They consist of accuracy, recall (sensitivity), F1-Score, and precision (17).

Accordingly, the methods for calculating all four indicators are provided below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$\text{Precision (PPV)} = \frac{TP}{TP + FP} \times 100$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100$$

$$\text{F1-Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100$$

TP = The participants were genuinely drowsy and were accurately classified as drowsy.
TN = The participants were truly in a non-drowsy condition and they were accurately classified as non-drowsy.
FP= is the quantity of subjects that were mislabeled as drowsy when, in fact, they were not.
FN= is the number of patients who were genuinely drowsy, but were incorrectly classified as non-drowsy by the system/model.

3. Results

3.1. Overview of included studies

This section introduces the main features of the 69 studies selected for this systematic review (figure 1). 43 studies utilized DL models, 22 studies utilized ML models, and 4 studies were hybrid (DL+ML).
In terms of geographical distribution, the included studies originated from a wide range of countries (Figure 3). As illustrated in the figure, India represents the most substantial contribution, with 20 studies accounting for 28.9% of the total. China follows with 9 studies (13%), and South Korea ranks third with 5 studies (7%). A group of five countries-the United States, the United Arab Emirates (UAE), Iran, Malaysia, and Bangladesh-each contributed 3 studies (4%). The remaining contributions are dispersed across several other nations, each representing a comparatively smaller share of the overall authorship.

3.2. Accuracy, precision, recall, F1-score benchmarks

This section provides a quantitative synthesis of model performance across the included studies, using median values and interquartile ranges (25th to 75th percentile) for the four primary evaluation metrics: accuracy, recall, F1-score, and precision (PPV). Table 1 summarizes the distributions separately for DL and traditional ML approaches, enabling a rigorous comparative assessment.

3.2.1. Accuracy

DL models demonstrate a clear overall performance advantage, with a median accuracy of 94.48% (interquartile range (IQR): 89.72% to 96.42%) compared to 91.80% for ML models (IQR: 84.50% to 96.20%). Beyond higher central tendency, the lower quartile of DL models notably exceeds that of ML models.

3.2.2. Recall (Sensitivity)

Recall values for DL and ML models are relatively close at the median level-93.76% for DL versus 94.12% for ML. However, DL exhibits greater stability and robustness across studies. Specifically, the DL lower quartile (89.90%) is considerably higher than that of ML (87.43%), reflecting more dependable detection sensitivity among DL approaches, even under varying conditions.

3.2.3. F1-score

The largest performance gap appears in the F1-score. DL models achieve a median of 93.15% (IQR: 86.03% to 95.65%), substantially outperforming ML models, which reach a median of only 84.00% (IQR: 79.80% to 86.80%). This demonstrates that DL architectures provide a far better balance between sensitivity and precision-an essential property for minimizing both missed detections and false alarms in DDD.

3.2.4. Precision (PPV)

DL models obtain a median PPV of 93.18% (IQR: 90.75% to 96.23%), whereas ML models lag behind with a median of 94.60% but with a much wider and lower quartile (81.70%). This suggests that DL systems are markedly more consistent in avoiding false positives, with three-quarters of DL models surpassing 90% precision-a level attained only by the top quartile of ML approaches.

3.3. Performance overview

Reported accuracies range widely, from as low as 67.0% to a perfect 100%, with several models achieving performances above 98%. The highest accuracy was reported by Macalisang using YOLOv3 for eye closure detection (18). Other top-performing models include Convolutional Neural Networks (CNNs), often enhanced with transfer learning, ensemble methods, or temporal modeling techniques like Long Short-Term Memory (LSTM) architectures. CNNs and their variants (e.g., DCNN, MobileNet, ResNet, VGG) are the most frequently used and often yield the highest accuracy, especially when combined with temporal models like LSTMs. Support Vector Machine (SVM) and Random Forest (RF) models, particularly when combined with feature engineering (e.g., EAR, PERCLOS), still achieve high performance (e.g., 99.45% by Al-Anizy with SVM) (19). Models combining CNNs with LSTMs or attention mechanisms (e.g., Xiao, 2019 (20); Xie, 2018 (21)) show strong results, capturing both spatial and temporal features.
Many high-performing models are deployed in real-time or near-real-time settings, indicating practical applicability. Performance varies significantly with dataset type (public vs. custom), data modality (video, images), and driving context (simulated vs. real-road).

3.4. Input data modality

Most DL studies (49 of 54) relied on raw video or image sequences, enabling end-to-end learning of facial and ocular cues without manual feature engineering. In contrast, ML studies showed greater variability: while ten used visual data, five depended primarily on handcrafted or non-visual signals. Overall, the field remains dominated by visual modalities, with DL models leveraging pixel-level inputs and ML models relying more heavily on engineered descriptors.

3.5. Characteristics of driving contexts

Most investigations were conducted in simulated or controlled environments (73.4%), which allow systematic manipulation of lighting, posture, and task load to obtain clean, well-annotated behavioral signals (supplementary Table A). Real-world or in-vehicle recordings accounted for a smaller but essential portion of the literature (20.3%), providing ecological validity by capturing natural variations in illumination, motion, and traffic conditions. A few studies (6.0%) used mixed contexts. DL models were particularly dominant in real-world settings, reflecting their ability to learn robust representations from noisy, unconstrained visual data, whereas ML approaches were more common in controlled environments where handcrafted features remain reliable.

3.6. Summary of datasets used

An analysis of dataset provenance across the 69 included studies reveals a predominant reliance on bespoke data collection: 46 studies (63.7%) employed custom datasets specifically recorded to satisfy study-specific requirements (e.g., controlled illumination, annotated video sequences, or simulator scenarios). Public benchmarks were used in 18 studies (29.9%); the most frequently cited repositories include NTHU-DDD, YawDD, the MRL eye dataset, and the ZJU Eye blink database. A minority of investigations (5 studies, 6%) adopted a hybrid strategy, augmenting public benchmarks with proprietary recordings to increase subject diversity and environmental variability. This distribution highlights two complementary trends: i: investigators continue to create targeted datasets to capture fine-grained behavioral cues under bespoke experimental protocols, and ii: the growing but still limited adoption of standardized benchmarks supports cross-study comparability [see supplementary Table B].

3.7. DL and ML models for DDD

Supplementary table A summarizes the ML and DL architectures applied for behavioral DDD; ML entries denote handcrafted-feature pipelines (e.g., EAR, PERCLOS) with classical classifiers (primarily SVM and RF), whereas DL entries denote end-to-end convolutional and spatio-temporal networks (e.g., CNN variants, 3D-CNN, and CNN+LSTM hybrids).

3.8. Types and frequencies of architectures

3.8.1. DL methods

DL models constituted the majority of the reviewed approaches (62.7%).

Most studies used CNNs or their enhanced variants, including VGG16/VGG19, ResNet families, MobileNet, YOLO-based detectors, DCNN, 3D-CNN, MTCNN, and ensemble CNNs. CNN-based models accounted for over 70% of all DL studies. Hybrid CNN + LSTM architectures represented a significant subgroup (12% of DL studies), reflecting the importance of modeling temporal dynamics such as blink duration and yawning progression. These results confirm that DL architectures demonstrate stronger generalization capacity-particularly in varying illumination and real-road conditions-though they require larger annotated datasets and greater computational resources.

3.8.2. ML methods

ML methods were used in 31.3%, most of which relied on handcrafted behavioral features such as EAR, PERCLOS, MAR, blink rate, and head-pose indicators. SVM were by far the most frequently used ML classifier, appearing in over 60% of ML-based studies. RF models accounted for roughly 15 to 20% of ML applications. Other ML techniques including Naïve Bayes, bayesian classifiers, FLDA, MLP, and AdaBoost were used infrequently, each appearing in fewer than 10% of the total ML papers.

3.8.3. Hybrid approaches

A smaller subset of studies (6.0%) implemented hybrid methods, typically combining: i: CNNs for spatial feature extraction; ii: LSTM/Recurrent Neural Network (RNN) units for sequential modeling; iii: ML classifiers (e.g., SVM or RF) applied on top of deep features. These hybrid systems reported competitive performance, often lying between pure DL and pure ML results, and offered advantages in real-time applications where interpretability and computational efficiency are important.

3.9. Real-time vs. offline detection

The findings indicate a pronounced dominance of real-time inference within both DL- and ML-based frameworks, with 92.1% of the included studies operating under real-time conditions. Notably, the vast majority of DL models (50 out of 54) and nearly all ML models (10 out of 11) were explicitly engineered for real-time deployment, underscoring the field's strong emphasis on immediate and continuous driver-state assessment. In contrast, only a small fraction of studies (n= 4) adopted offline inference, typically reflecting preliminary model development or evaluation phases that relied on pre-recorded datasets prior to real-world application. Collectively, these results highlight a clear methodological trend toward real-time operational capability, which remains essential for the practical integration of DD systems into in-vehicle driver-monitoring technologies [supplementary Table B].

3.10. Risk of bias evaluation

Evaluation of methodological quality using the PROBA-AI tool revealed a heterogeneous risk-of-bias profile across the 69 included studies. Overall, 27 studies were rated as high risk, reflecting substantial concerns related to dataset construction, labeling procedures, insufficient reporting of ground-truth generation, and inadequate validation strategies. Twenty-three studies demonstrated low risk, typically characterized by clear dataset provenance, transparent annotation methods, and the use of appropriate training–validation splitting and performance evaluation protocols. The remaining 17 studies were categorized as moderate risk, often due to partial reporting, reliance on small

or homogeneous datasets, or limited justification of modeling choices. Collectively, these findings highlight persistent methodological inconsistencies within the ML/DL-based behavioral drowsiness detection literature and underscore the need for standardized dataset curation, labeling frameworks, and evaluation practices to enhance reproducibility and generalizability [supplementary Table C].

4. Discussion

The findings of this systematic review highlight several important trends in the development of behavioral DDD systems. DL approaches consistently outperformed ML methods, most notably in terms of accuracy and F1-score. CNN- and CNN-LSTM-based architectures captured complex spatiotemporal cues underlying drowsy behavior more effectively than handcrafted feature pipelines.

Our analysis shows a clear methodological shift toward DL, with CNN-based architectures representing the majority of behavioral DD models and consistently outperforming ML approaches. CNN variants-together with hybrid CNN–LSTM designs-demonstrated strong capacity to capture spatial and temporal behavioral cues, yielding markedly higher median accuracy and F1-scores than ML pipelines. In contrast, classical ML methods, dominated by SVM and RF classifiers, remained dependent on handcrafted features such as EAR, PERCLOS, and MAR, making them more susceptible to noise and environmental variability. CNN-based architectures and CNN–LSTM hybrids demonstrated superior performance across various illumination and driving conditions (Yu et al., 2024; Adhithyaa, 2023) (21, 27). These findings confirm that DL architectures currently provide a more reliable framework for behavioral DD than feature-engineered ML algorithm.

The findings of this review clearly indicate that DL architectures outperform ML methods in behavioral DDD. DL models achieved notably higher median accuracy (94.48% vs. 91.80%) and substantially stronger F1-scores (93.15% vs. 84.00%), highlighting their superior balance of sensitivity and precision. Although recall and precision were broadly comparable between the two paradigms, DL approaches demonstrated greater stability, reflected in consistently higher lower-quartile values across studies. This superiority aligns with evidence from recent DL-focused systematic analyses, which emphasize that CNN, spatio-temporal, and hybrid CNN–LSTM models extract richer behavioral representations and offer stronger generalization in unconstrained driving environments (22).

The strong predominance of real-time inference observed across the included studies is consistent with prior reports emphasizing the necessity of responsive, in-vehicle monitoring systems for mitigating drowsiness-related risks. Earlier works have demonstrated that real-time DD supports continuous behavioral assessment and enables timely hazard prevention (23, 24). Studies employing simulated environments similarly highlight the value of real-time processing for replicating driver workload and fatigue progression under controlled conditions. This emphasis aligns with broader findings that stress the importance of moment-to-moment evaluation of ocular and facial indicators, such as blink patterns, gaze direction, and head movements, for accurate estimation of driver state (24, 25). Although a minority of studies still rely on offline analysis-primarily for algorithm development or dataset-based benchmarking-prior literature underscores that real-time operation remains essential for practical deployment in advanced driver-assistance systems and safety-critical contexts (26). Collectively, these findings suggest that real-time capability extends beyond a methodological preference and constitutes a fundamental requirement influencing system architecture and evaluation strategies, consistent with established frameworks in which timeliness represents a key component of quality of care (27).

The distribution of datasets across the reviewed studies highlights long-standing structural limitations in DD research. The predominance of custom datasets reflects the same issues previously emphasized in the literature, where heterogeneity in data collection protocols and limited transparency impede reproducibility and cross-study comparability (14, 28). Overall, consistent with prior systematic reviews on DDD (14), the evidence underscores a persistent trade-off: custom datasets enhance task-specific signal fidelity but restrict external validity, whereas benchmark datasets support reproducibility but lack contextual richness.

The partial reliance on public benchmarks such as NTHU-DDD, YawDD, the MRL Eye dataset, and the ZJU Eye blink database aligns with earlier findings suggesting that existing open-source repositories remain insufficiently comprehensive for robust model generalization (29, 30). These datasets typically offer limited variation in driving context, subject behavior, or environmental conditions, thereby limiting the trustworthiness of the models for real-world deployment readiness (31). Hybrid datasets used in a smaller subset of studies provide incremental improvements but still inherit the limitations of both proprietary and benchmark sources.

4.1. Trends and future work for DDD systems

The environment, available resources, and real-time application constraints are some of the factors that should be taken into consideration when determining which of the many methods for DD have been reviewed.

For DD in real time in controlled settings (such as lab or simulated conditions): CNNs can recognize facial traits including eye states and mouth opening with excellent accuracy in controlled situations with controlled lighting and other variables. Data should be recorded in well-lit situations at a high frame rate (30 fps or higher) for optimal outcomes. Images should be normalized using preprocessing approaches like Face alignment and histogram equalization to enhance feature extraction performance in a variety of scenarios. SVM and CNNs can be used together to improve classification, particularly when the task calls for binary classification and

the dataset is limited (32, 33). For use in real-world settings with varying lighting and ambient: It is advised to use hybrid models that combine CNNs and LSTMs in more dynamic and real-world situations when environmental variables like lighting, weather, and driver behavior change dramatically. These models can monitor how drowsiness changes over time because LSTMs are excellent at identifying temporal dependencies in sequential data. This is particularly helpful when examining signs of tiredness such as head motions and blinking. To manage ambient noise and enhance model accuracy in difficult situations, preprocessing techniques like noise reduction and geometric transformation (e.g., cropping, scaling) are crucial (33, 34).

5. Limitations

This review is subject to several methodological constraints. First, substantial heterogeneity across studies-particularly in dataset sources, labeling protocols, driving environments, and evaluation procedures-limits the comparability of reported performance metrics. Second, the predominance of simulator-based and controlled laboratory experiments may overestimate model performance relative to real-world driving conditions. Third, incomplete reporting in several studies, including missing details on sample characteristics, annotation methods, and validation schemes restricts the reliability of cross-study synthesis. Finally, aggregated metrics were derived from published results rather than raw outputs, preventing standardized recalculation and formal statistical comparison between DL and ML approaches.

6. Conclusions

This review examined behavioral approaches to DDD and identified a consistent performance advantage for DL architectures over ML methods. DL models, dominated by CNN and CNN–LSTM frameworks, achieved higher median accuracy, and F1-scores and exhibited reduced performance variability across studies. In contrast, ML pipelines relying on handcrafted features demonstrated lower robustness to illumination changes, subject variability, and environmental noise. The evidence remains limited due to heterogeneous datasets, non-standardized evaluation protocols, and a predominance of simulator-based experiments, which constrain cross-study comparability. Overall, DL architectures represent the most reliable computational direction for advancing behavioral DD systems, although further progress requires standardized benchmarks and more extensive real-world validation.

7. Declarations

7.1. Acknowledgments

We sincerely thank Dr. Homayoun Sadeghi-Bazargani (Tabriz University of Medical Sciences) for his crucial conceptual input and Dr. Jalal Poorolajal (Hamadan University of Medical Sciences) for his contributions to article screening, data extraction, and analysis. Despite his significant help, Dr. Poorolajal declined authorship due to limited domain expertise in this area. We would like to appreciate the cooperation of Clinical Research Development Unit, Imam Reza General Hospital, Tabriz, Iran in conducting this research.

7.2. Authors' contributions

Conceptualization: HS, SGF, ES, SN, SKh and MT. Methodology: HS and ES. Formal analysis: SGF and ES. Investigation: ES, MT and HS. Data curation: SN and SGF Writing—original draft preparation: ES, SN, MT and SGF. Writing—review and editing: ES, HS, SKh and SN. Visualization: HS and ES. Supervision: HS. Submission: HS. All authors have read and agreed to the last version of this manuscript.

7.3. Funding/Support

Not applicable.

7.4. Conflict of interest

None declared.

7.5. Data Availability

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

7.6. Using artificial intelligence chatbots

Artificial intelligence chatbots (including ChatGPT) were used solely to facilitate topic based searching within the literature and to assist with the extraction and organization of relevant information from published articles. These tools were not used for data analysis, result interpretation, or scientific decision making.

7.7. Abbreviations

AI: Artificial Intelligence
ANN: Artificial Neural Network
BiLSTM: Bidirectional Long Short-Term Memory
CHT: Circular Hough Transform
CNN: Convolutional Neural Network
CV: Computer Vision
DCNN: Deep Convolutional Neural Network
DDD: Driver Drowsiness Detection
DL: Deep Learning
EAR: Eye Aspect Ratio
ECNN: Ensemble Convolutional Neural Network
ECR: Eye Closure Ratio
EEG: Electroencephalography
EM-CNN: Eye and Mouth Convolutional Neural Network
FN: False Negative
FP: False Positive fps: Frames Per Second
HoG: Histogram of Oriented Gradients
KNN: K-Nearest Neighbors
LSTM: Long Short-Term Memory
MAR: Mouth Aspect Ratio

ML: Machine Learning
MLP: Multi-Layer Perceptron
MOR: Mouth Opening Ratio
MTCNN: Multi-Task Cascaded Convolutional Neural Network
NB: Naïve Bayes
NLR: Nose Length Ratio
NR: Not Reported
PERCLOS: Percentage of Eyelid Closure over Time
POM: Percentage of Mouth Opening
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
RF: Random Forest
RLDD: Real-Life Drowsiness Dataset
RNN: Recurrent Neural Network
ROI: Region of Interest
RPM: Revolutions Per Minute
SNN: Spiking Neural Network
SVM: Support Vector Machine
TLD: Track-Learning-Detection
TN: True Negative
TP: True Positive
FPS: Frames Per Second

7.8. Ethical considerations

This research was approved by the Ethics Committee of Hamedan University of Medical Sciences (Ethics Code: IR.UMSHA.REC.1404.819)

References

1. Ayas S, Donmez B, Tang X. Drowsiness mitigation through driver state monitoring systems: a scoping review. Human factors. 2024;66(9):2218-43.
2. Williamson A, Lombardi DA, Folkard S, Stutts J, Courtney TK, Connor JL. The link between fatigue and safety. Accident Analysis Prevention. 2011;43(2):498-515.
3. Zhang H, Ni D, Ding N, Sun Y, Zhang Q, Li X. Structural analysis of driver fatigue behavior: A systematic review. TRIP. 2023;21:100865.
4. Li D-H, Liu Q, Yuan W, Liu H-X. Relationship between fatigue driving and traffic accident. Journal of traffic and transportation engineering (Xi'an, Shaanxi). 2010;10(2):104-9.
5. Lenné MG, Jacobs EE. Predicting drowsiness-related driving events: a review of recent research methods and future opportunities. TIES. 2016;17(5-6):533-53.
6. Solanki P, Baldaniya D, Jogani D, Chaudhary B, Shah M, Kshirsagar A. Artificial intelligence: New age of transformation in petroleum upstream. Petroleum Research. 2022;7(1):106-14.
7. Masoudi N, Sarbazi E. Artificial intelligence in older adults' health. Int J Aging. 2024;2(1):e23.
8. Liu F, Chen D, Zhou J, Xu F. A review of driver fatigue detection and its advances on the use of RGB-D camera and deep learning. Engineering Applications of Artificial Intelligence. 2022;116:105399.
9. Albadawi Y, Takruri M, Awad M. A review of recent developments in driver drowsiness detection systems. Sensors. 2022;22(5):2069.
10. Ramzan M, Khan HU, Awan SM, Ismail A, Ilyas M, Mahmood A. A survey on state-of-the-art drowsiness detection techniques. IEEE Access. 2019;7:61904-19.
11. Bamidele AA, Kamardin K, Abd Aziz NSN, Sam SM, Ahmed IS, Azizan A, et al. Non-intrusive driver drowsiness detection based on face and eye tracking. IIJACSA. 2019;10(7).
12. Kiashari SEH, Nahvi A, Bakhoda H, Homayounfard A, Tashakori M. Evaluation of driver drowsiness using respiration analysis by thermal imaging on a driving simulator. Multimedia Tools and Applications. 2020;79:17793-815.
13. Vural U, Akgul YS. Eye-gaze based real-time surveillance video synopsis. Pattern Recognition Letters. 2009;30(12):1151-9.
14. El-Nabi SA, El-Shafai W, El-Rabaie E-SM, Ramadan KF, Abd El-Samie FE, Mohsen S. Machine learning and deep learning techniques for driver fatigue and drowsiness detection: a review. Multimedia Tools and Applications. 2024;83(3):9441-77.
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj. 2021;372: n71.
16. Ghaffari fam S SE, Naderian S, Soleimanpour H. Drivers' Drowsiness Detection Using Behavioral Measures by Machine Learning: A Systematic Review. Available online: https://www.crd.york.ac.uk/PROSPERO/view/CRD420251139663.
17. El-Nabi SA, El-Shafai W, El-Rabaie E-SM, Ramadan KF, Abd El-Samie FE, Mohsen S. Machine learning and deep learning techniques for driver fatigue and drowsiness detection: a review. Multimedia Tools and Applications. 2023:83(3): 9441-9477.
18. Macalisang JR, Alon AS, Jardiniano MF, Evangelista DCP, Castro JC, Tria ML. Drive-Awake: a YOLOv3 machine vision inference approach of eyes closure for drowsy driving detection. In: Proceedings of the 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET); 2021. Piscataway (NJ): IEEE.
19. Al-Anizy GJ, Nordin MJ, Razooq MM. Automatic driver drowsiness detection using haar algorithm and support vector machine techniques. Asian J Appl Sci. 2015;8(2):149.
20. Xiao Z, Hu Z, Geng L, Zhang F, Wu J, Li Y. Fatigue driving recognition network: fatigue driving recognition via convolutional neural network and long short-term memory

units. IET Intelligent Transport Systems. 2019;13(9):1410-6.
21. Xie Y, Chen K, Murphey YL. Real-time and robust driver yawning detection with deep neural networks. In: Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI); 2018. Piscataway (NJ): IEEE.
22. Adhithyaa N, Tamilarasi A, Sivabalaselvamani D, Rahunathan L. Face positioned driver drowsiness detection using multistage adaptive 3D convolutional neural network. Information Technology and Control. 2023;52(3):713-30.
23. 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 Transactions on Intelligent Transportation Systems. 2024; 25(11):16614-16624.
24. Priyanka S, Shanthi S, Kumar AS, Praveen V. Data fusion for driver drowsiness recognition: A multimodal perspective. Egyptian Informatics Journal. 2024;27:100529.
25. Benmohamed A, Zarzour H. A Deep Learning-Based System for Driver Fatigue Detection. Ingenierie des Systemes d'Information. 2024;29(5):1779.
26. Sedik A, Marey M, Mostafa H. An adaptive fatigue detection system based on 3d cnns and ensemble models. Symmetry. 2023;15(6):1274.
27. Sarbazi E, Sadeghi-Bazargani H, farahbakhsh M, Ala A, Pouraghaei A, Soleimanpour H. Development and psychometric testing of the quality of care for trauma patients scale using exploratory and confirmatory factor analysis. Front Emerg Med. 2026;9(4):e30.
28. Shaik ME. A systematic review on detection and prediction of driver drowsiness. Transportation research interdisciplinary perspectives. 2023;21:100864.
29. 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(21): e39479.
30. 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. Biomedical Signal Processing and Control. 2025;102:107262.
31. Sarbazi E, Sadeghi-Bazargani H, Sheikhalipour Z, Farahbakhsh M, Ala A, Soleimanpour H. Trust in medicine: a scoping review of the instruments designed to measure trust in medical care studies. Journal of Caring Sciences. 2024;13(2):116.
32. Chaabene S, Bouaziz B, Boudaya A, Hökelmann A, Ammar A, Chaari L. Convolutional Neural Network for Drowsiness Detection Using EEG Signals. Sensors. 2021;21(5):1734.
33. Salem D, Waleed M. Drowsiness detection in real-time via convolutional neural networks and transfer learning. Journal of Engineering and Applied Science. 2024;71(1):122.
34. Mahmud T, Saha B, Islam D, Aziz MT, Datta N, Barua K, et al., editors. Deep Learning Approach for Driver Drowsiness Detection in Real Time. Innovations in Cybersecurity and Data Science; 2024 2024//; Singapore: Springer Nature Singapore.
35. Mehta S, Dadhich S, Gumber S, Jadhav Bhatt A. Real-time driver drowsiness detection system using eye aspect ratio and eye closure ratio. In: Proceedings of the International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM); 2019; Jaipur, India.
36. Al Redhaei A, Albadawi Y, Mohamed S, Alnoman A, editors. Realtime Driver Drowsiness Detection Using Machine Learning. 2022 Advances in Science and Engineering Technology International Conferences (ASET); 2022: IEEE.
37. Dey S, Chowdhury SA, Sultana S, Hossain MA, Dey M, Das SK, editors. Real time driver fatigue detection based on facial behaviour along with machine learning approaches. 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON); 2019: IEEE.
38. Kumar A, Patra R, editors. Driver drowsiness monitoring system using visual behaviour and machine learning. 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE); 2018: IEEE.
39. Chinthalachervu R, Teja I, Kumar MA, Harshith NS, Kumar TS, editors. Driver Drowsiness Detection and Monitoring System using Machine Learning. Journal of Physics: Conference Series; 2022: IOP Publishing.
40. Saradadevi M, Bajaj P. Driver fatigue detection using mouth and yawning analysis. International journal of Computer science and network security. 2008;8(6):183-8.
41. Rehman HU, Naeem M, Khan M, Sikander G, Anwar S, editors. Eye Tracking based Real-Time Non-Interfering Driver Fatigue Detection System. 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI); 2018: IEEE.
42. Maior CBS, das Chagas Moura MJ, Santana JMM, Lins ID. Real-time classification for autonomous drowsiness detection using eye aspect ratio. Expert Systems with Applications. 2020;158:113505.
43. Alioua N, Amine A, Rziza M. Driver's fatigue detection based on yawning extraction. Int J Veh Technol. 2014;2014:1-7. 4
44. Vural E, Çetin M, Erçil A, Littlewort G, Bartlett M, Movellan J. Machine learning systems for detecting driver drowsiness. In: Takeda K, Ercil A, Hansen JHL, Abut H, eds. In-Vehicle Corpus and Signal Processing for Driver Behavior. Springer; 2009:97-110.
45. Tabrizi PR, Zoroofi RA, editors. Open/closed eye analysis for drowsiness detection. 2008 first workshops on image processing theory, tools and applications; 2008: IEEE.
46. Ed-Doughmi Y, Idrissi N, editors. Driver fatigue detection using recurrent neural networks. Proceedings of the 2nd international conference on networking, information systems & security; 2019.

47. Hashemi M, Mirrashid A, Beheshti Shirazi A. Driver safety development: Real-time driver drowsiness detection system based on convolutional neural network. SN Computer Science. 2020;1(5):1-10.
48. Padamata BK, Singothu JR. A machine learning approach for driver drowsiness detection. International Journal of Engineering Research and Applications. 2020;10(11):58-65.
49. Cheamanunkul S, Chawla S, editors. Drowsiness detection using facial emotions and eye aspect ratios. 2020 24th International Computer Science and Engineering Conference (ICSEC); 2020: IEEE.
50. Thakur R, Raj S, Pandey S. Driver Drowsiness Detection System Using Machine Learning. Advanced Production and Industrial Engineering: IOS Press; 2022. p. 31-8.
51. Phan A-C, Nguyen N-H-Q, Trieu T-N, Phan T-C. An efficient approach for detecting driver drowsiness based on deep learning. Applied Sciences. 2021;11(18):8441.
52. Zhao Z, Zhou N, Zhang L, Yan H, Xu Y, Zhang Z. Driver fatigue detection based on convolutional neural networks using EM-CNN. Computational intelligence and neuroscience. 2020;2020(1): 7251280.
53. Nguyen D-L, Putro MD, Jo K-H, editors. Eyes status detector based on light-weight convolutional neural networks supporting for drowsiness detection system. IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society; 2020: IEEE.
54. Salman RM, Rashid M, Roy R, Ahsan MM, Siddique Z. Driver drowsiness detection using ensemble convolutional neural networks on YawDD. arXiv preprint arXiv:211210298. 2021.
55. Chirra VRR, Uyyala SR, Kolli VKK. Deep CNN: A Machine Learning Approach for Driver Drowsiness Detection Based on Eye State. Rev d'Intelligence Artif. 2019;33(6):461-6.
56. Ghazal M, Haeyeh YA, Abed A, Ghazal S, editors. Embedded fatigue detection using convolutional neural networks with mobile integration. 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW); 2018: IEEE.
57. Jahan I, Uddin KA, Murad SA, Miah MSU, Khan TZ, Masud M, et al. 4D: a real-time driver drowsiness detector using deep learning. Electronics. 2023;12(1):235.
58. Arsić A, Ilić V, Pavković B, Samardžija D, editors. System for detecting driver's drowsiness, fatigue and inattention. 2021 29th Telecommunications Forum (TELFOR); 2021: IEEE.

Table 1: Summary statistics of model performance metrics across included studies

Metric	Deep learning			Machine learning		
	Median	25th percentile	75th percentile	Median	25th percentile	75th percentile
Accuracy	94.48	89.72	96.42	91.8	84.5	96.2
Recall	93.76	89.90	97.0	94.12	87.43	95.9
F1-Score	93.15	86.03	95.65	84.0	79.8	86.8
PPV	93.18	90.75	96.23	94.6	81.7	96.86

PPV: positive predictive value.

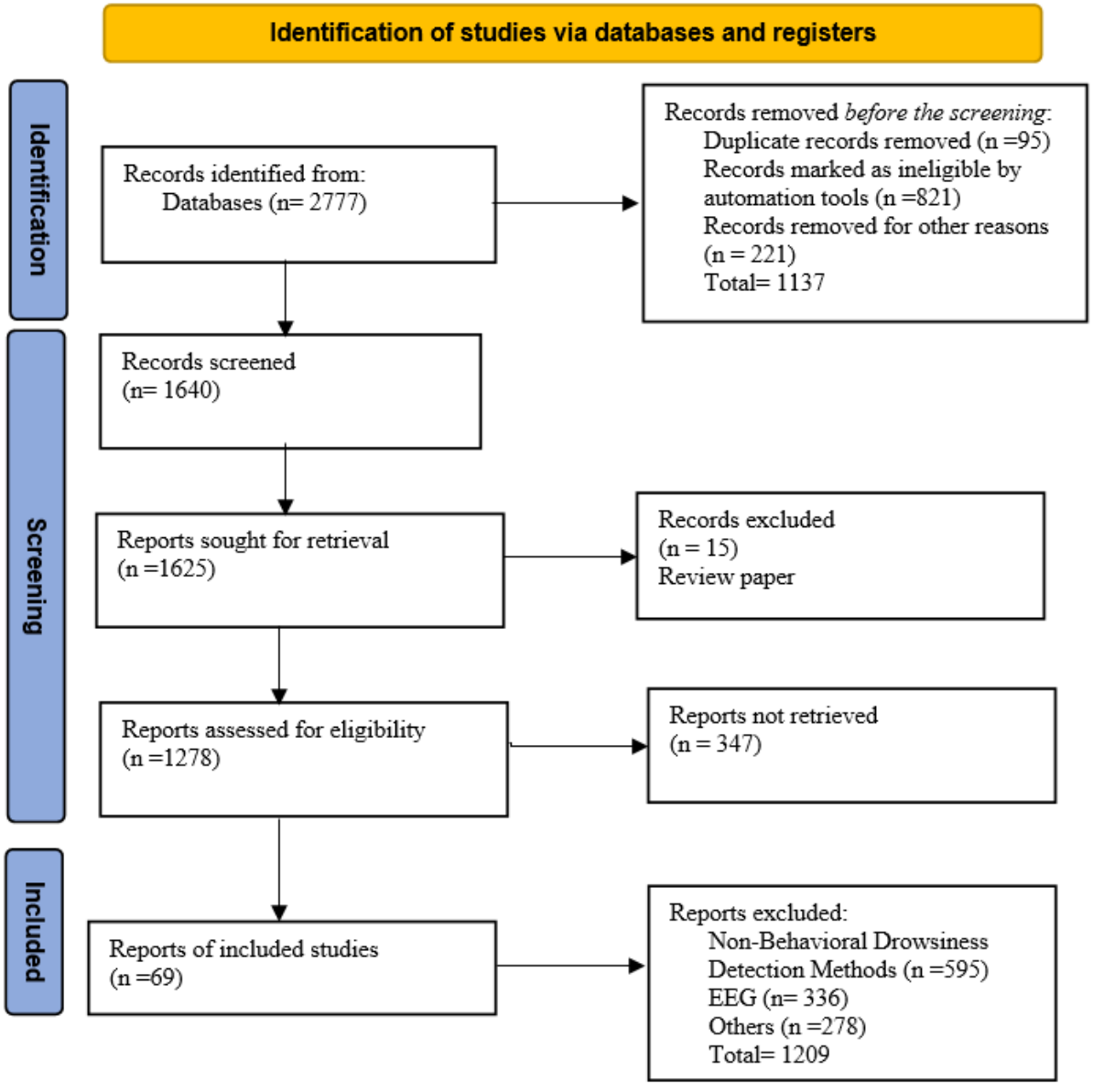

Figure 1: Flow diagram of the study selection for the review process. EEG: electroencephalogram.

Supplementary Table A 1: Summary of model performance metrics

Row	Author, year	ML/DL Method Used	Accuracy	Recall (Sensitivity)	F1-Score	Precision (PPV)
1	Al-Anizy 2015 (19)	SVM	99.45	NR	NR	NR
2	Mehta 2019 (35)	RF	84.0	84.0	84.0	84.0
		SVM	80.0	80.0	79.8	80.1
		Naive Bayes	80.0	80.0	79.8	80.7
3	Redhaei 2022 (36)	SVM	95.0	97.0	NR	95.0
		RF	99.0	96.0	NR	98.0
		Sequential NN	97.0	96.0	NR	97.0
4	Dey 2019 (37)	SVM	96.4	94.6	NR	NR
		FLDA	93.3	89.9	NR	NR
		Bayesian	85.7	96.9	NR	NR
5	Kumar 2020 (38)	Bayesian Classifier	85.4	97.3	NR	NR
		FLDA	92.6	89.6	NR	NR
		SVM	95.8	95.58	NR	NR
6	Chinthalachervu 2022 (39)	SVM	NR	95.58	NR	NR
7	Saradadevi 2008 (40)	SVM	85.0	84.0	NR	76.0
8	Rehman 2018 (41)	SVM	87.0	NR	NR	NR
9	Maior 2020 (42)	SVM	94.9	NR	NR	NR
		SVM + Personal Feedback	99.0	NR	NR	NR
		MLP	94.9	NR	NR	NR
		RF	91.8	NR	NR	NR
10	Alioua, 2014 (43)	SVM	98.0	NR	NR	NR
11	Vural, 2007 (44)	AdaBoost + MLR with FACS	96.0	NR	NR	NR
12	Tabrizi, 2008 (45)	Chromatic-based Eye Analysis + PERCLOS	98.8	NR	NR	NR
13	Ed-Doughmi, 2019 (46)	LSTM	92.71	NA	NA	NA
14	Hashemi 2020 (47)	FD-NN	98.15	86.7	NA	99.8
		TL-VGG16	95.45	NA	NA	NA
		TL-VGG19	94.96	NR	NA	NA
15	Xiao 2020 (20)	CNN + LSTM	98.96	95.83	95.83	95.83
16	Padamata 2020 (48)	CNN	96.42	NR	95.47	NA
17	Cheamanunkul 2019 (49)	SVM with EmoPy + EAR + PCA	81.7	NR	NR	NR
		RF with EmoPy + EAR + PCA	78.8	NR	NR	NR
		MLP with EmoPy + EAR + PCA	77.3	NR	NR	NR
		SVM with EmoPy + EAR (no PCA)	73.5	NR	NR	NR
18	Thakur 2022 (50)	CNN	95.78	NR/NA	95.43	NA
19	Phan 2021 (51)	MobileNet-V2	96.0	97.0	95.0	93.0
		ResNet	97.0	97.0	97.0	96.0
20	Zhao 2020 (52)	EM-CNN	93.62	93.64	NR	NR
21	Nguyen 2020 (53)	CNN	98.7	NR	NR	NR
22	Salman 2021 (54)	ECNN	NA	90.3	NA	97.0
23	Chirra 2019 (55)	CNN	96.42	NR	NR	NR
24	Ghazal 2018 (56)	CNN	95.0	NR	NR	NR
25	Jahan 2023 (57)	VGG16 + Transfer Learning	95.93	93.87	93.51	93.15
		VGG19 + Transfer Learning	95.03	95.47	95.14	94.82
		4D Model (Custom CNN)	97.53	97.06	97.20	97.35
26	Arsic 2021 (58)	FDN	90.79	NR	NR	NR
		REN	86.68	NR	NR	NR
		SEN	72.65	NR	NR	NR
27	Kulhandjian 2022 (59)	DCNN	95.0	NR	NR	NR
28	Saif 2020 (60)	DCNN and RF	98.97	NR	NR	NR
29	Park 2017 (61)	Supervised Deep Learning (AlexNet + VGG-FaceNet + FlowImageNet)	73.06	NR	NR	NR
30	Huynh, 2016 (62)	CNN	87.46	87.97	87.97	NR
31	Jia 2021 (63)	MTCNN	97.5	NR	NR	NR
32	Xie, 2018 (21)	CNN and LSTM	98.4	100	NR	NR
33	Tamanani 2021 (64)	CNN	91.8	92.0	92.0	92.8
34	Valsan 2021 (65)	CNN	92.5	NR	NR	NR
35	Jabbar 2018 (66)	CNN	83.33	NR	NR	NR
36	Pachouly 2020 (67)	CNN	94.0	0.95 (Closed) 0.93 (Open)	0.95 (Closed) 0.93 (Open)	0.95 (Closed) 0.93 (Open)

Supplementary Table A 1: Summary of model performance metrics

Row	Author, year	ML/DL Method Used	Accuracy	Recall (Sensitivity)	F1-Score	Precision (PPV)
37	Xiaofeng 2021 (68)	CNN	89.55	NR	NR	NR
38	Geng 2018 (69)	CNN	NR	89.58	92.47	95.56
39	Chand 2022 (70)	CNN	93.0	90.0 to 92.0	91.0 to 94.0	91.0 to 93.0
40	Panwar 2022 (71)	CNN	99.95	95.0		
41	Victoria 2022 (72)	CNN	67.0	1.00	75.0	60.0
42	Dipu 2021 (73)	CNN(SSD_MobileNet)	98.0	NR	NR	Eyes Close: 96.3 Eyes Open: 97.7 No Yawn: 99.8 Yawn: 98.0
43	Walizad 2022 (74)	CNN	95.0	Closed Eyes: 98.0	Closed Eyes: 95.0	Closed Eyes: 92.0
				Open Eyes: 92.0	Open Eyes: 94.0	Open Eyes: 98.0
44	George, 2016 (75)	CNN	89.61	NR	NR	NR
45	Celona, 2018 (76)	ERT + CNN	83.44	NR	NR	NR
46	Zhuang 2020 (77)	SESDM	NR	93.63	95.16	96.72
47	Zhang 2015 (78)	CNN	92.0	NR	NR	NR
48	Macalisang, 2021 (18)	YOLOv3	100	NR	NR	NR
49	Yu, 2017 (79)	3 D-DCNN	72.6	NR	NR	NR
50	Lyu, 2018 (80)	LSTM	90.05	NR	NR	NR
51	Wang, 2020 (81)	MTCNN	95.0	NR	NR	NR
52	Pattarapongsin, 2020 (82)	(SSD + ResNet-10 backbone) and Custom 6-layer CNN	EAR: 93.0	EAR: 92.8	NR	EAR: 93.2
			MAR: 94.9	MAR: 95.3	NR	MAR: 94.6
53	Dua, 2021 (83)	CNNs (AlexNet, VGG, FaceNet, FlowImageNet, ResNet)	85.0	82.0	84.09	86.3
54	Yarlagadda, 2020 (84)	RNN and LSTM	97.25	NR	NR	NR
55	Rajamohana, 2021 (85)	CNN and BiLSTM	94.0	96.0	NR	90.67
56	Shahverdy, 2020 (86)	CNN	99.98	100	99.9	100
57	Guo, 2019 (87)	CNN and TSC-LSTM	84.85	NR	NR	NR
58	Kim, 2018 (88)	CNN	95.75	NR	NR	NR
59	Deng, 2019 (89)	CNN	92.0	NR	NR	NR
60	Aydemir, 2021 (90)	CNN and LSTM	80.0	85.55	78.04	81.63
61	Mahmoud, 2021 (91)	CNN	86.0	NR	NR	NR
62	Shiau, 2019 (92)	Deep Neural Network with Time-Domain Convolution	95.86	NR	NR	NR
63	Yu 2024 (29)	LSTM	97.0	97.7	97.8	97.8
64	Adhithyaa, 2023(22)	3D-CNN	77.4	NR	78.1	NR
65	Almazroi, 2023 (93)	SVM	87.0	91.2	86.8	82.8
66	Benmohamed, 2024 (25)	Hybrid model (Structural Features + CNN + LSTM)	90.12	NA	NA	NA
67	Huang, 2020 (94)	RF-DCM + LSTM	89.42	NA	89.42	NA
68	Narayana, 2025 (95)	SVM	92.3	90.8	91.7	91.2
69	Kharat, 2025 (96)	YOLO based vision system	92.0	NA	NA	NA

Table presents a consolidated overview of the performance metrics reported across the included studies. Key indicators-such as accuracy, precision, recall, and F1-score-are listed for each ML and DL model. This table facilitates direct cross-study comparison and highlights variability in algorithmic performance under different modeling strategies. NA: not available; NR: not reported; PPV: positive predictive value; ML: machine learning; DL: deep learing.

Supplementary Table B 1: Dataset characteristics and implementation details

Authors	Country	Dataset Name	Input Data Modality	Driving Context	Inference Mode
Al-Anizy (19)	Malaysia	Custom Dataset	Video	Simulated	Real-time
Mehta (35)	India	Custom Dataset	Video	Simulated	Real-time
Redhaei (36)	UAE	Custom Dataset	Video	Simulated	Real-time
Dey (37)	Bangladesh	INVEDRIFAC	Video, Images	Simulated	Real-time
Kumaran (38)	India	Custom Dataset and INVEDRIFAC dataset	Video	Simulated	Real-time
Chinthalachervu (39)	India	INVEDRIFAC	Video and Images	In-vehicle (implied, as the system is designed for DDD)	Real-time, and Offline
Rehman (41)	Pakistan	Custom Dataset	Video	Real driving	Real-time
Maior (42)	Brazil	Multimodality Drowsiness Database	Video	Simulated	Real-time
Saradadevi (40)	India	Custom Dataset	Video and Images	In-vehicle (implied, as the system is designed for DDD)	Real-time
Alioua (43)	Morocco	Custom Dataset	Video	Real driving	Real-time
Vural (44)	Turkey	Custom Dataset	Video	Simulated	Real-time
Tabrizi (45)	Iran	IMM Database and Custom Dataset	Video and Images	Simulated	Real-time
Ed-Doughmi (46)	Morocco	National Tsing Hua University Computer Vision Lab	Video	Simulated	Real-time
Hashemi (47)	Iran	ZJU Eyeblink Database and Extended Dataset	Images	Simulated	Real-time
Xiao (20)	China	Custom Dataset and TJPU-FDD	Video	Simulated	Real-time
Padamata (48)	India	Custom Dataset	Images	Real driving	Real-time
Cheamanunkul (49)	Thailand	Custom Dataset	Video	Simulated	Real-time
Thakur (50)	India	MRL Eye Dataset and Custom Dataset	Images	In-vehicle (implied, as the system is designed for DDD)	Real-time
Phan (51)	Vietnam	API, Kaggle, RMFD, iStock	Video and Images	Simulated	Real-time
Zhao (52)	China	Custom Dataset	Images	Real driving	Real-time
Nguyen (53)	South Korea	Custom Dataset	Images	Stimulated	Real-time
Salman (54)	Malaysia	YawDD	Videos	Simulated	Offline
Chirra (55)	India	Custom Dataset	Images	In-vehicle (implied, as the system is designed for DDD)	Real-time
Ghazal (56)	UAE	CEW and LFW	Images	Simulated	Real-time
Jahan (57)	Bangladesh	MRL Eye Dataset	Images	Simulated	Real-time
Arsic (58)	Serbia	Custom Dataset	Video	Simulated	Real-Time
Kulhandjian (59)	USA	Custom Datasets	Video and Images	Real driving	Real-Time
Saif (60)	Malaysia	iBUG and Real-time monocular video	Video	Simulated	Real-Time
Park (61)	South Korea	NTHU	Video and Images	Simulated	Real-Time
Huynh (62)	South Korea	NTHU	Video	Simulated	Sequence-based classification
Jia (63)	China	WIDER FACE, and MTFL	Video and Images	Simulated	Real-Time
Xie (21)	USA	YawDD, NTHU-DDD, and a custom dataset	Video	Simulated	Real-time
Tamanani (64)	Canada	Custom Dataset	Video	Simulated	Real-Time
Valsan (65)	India	Custom Dataset	Video	Real driving	Real-Time
Jabbar (66)	Taiwan	Custom Dataset	Video	Simulated	Real-time
Pachouly (67)	India	MRL Eye Dataset	Video and Images	Simulated and real driving	Real-time
Xiaofeng (68)	China	Custom Dataset	Video and Images	Simulated and real driving	Real-time
Geng (69)	China	TJPU-FDD, ZJU, Brain4Cars	Video	Simulated	Real-time
Chand (70)	India	DDD Dataset, Extended Cohn-Kanade Dataset	Images	Real driving	Real-time
Panwar (71)	India	Custom Dataset	Video	Real driving	Real-time
Victoria (72)	India	Custom Dataset	Video	NA	Real-time
Dipu (73)	Bangladesh	Custom Dataset	Video and Images	Simulated	Real-time
Walizad (74)	India	MRL Eye Dataset	Images	NA	Real-time
George (75)	India	Eye Chimera database	Video and Images	NA	Real-time
Celona (76)	Italy	NTHU	Video and Images	Simulated	Real-time
Zhuang (77)	China	NTHU	Videos	Simulated	Real-time

Supplementary Table B 1: Dataset characteristics and implementation details

Authors	Country	Dataset Name	Input Data Modality	Driving Context	Inference Mode
Zhang (78)	USA	Custom Dataset	Videos	Real driving	Real-time
Macalisan (18)	Philippines	Google Images and Kaggle	Video and Images	Simulated	Real-time
Yu (79)	South Korea	NTHU	Video	Simulated	Real-time
Lyu (80)	China	NTHU and FI-DDD	Video	Real driving	Real-time
Wang (81)	China	Custom Dataset + multiple public face datasets	Video and Images	Simulated	Real-time
Pattarapongsin (82)	Thailand	FDDB, WFLW, and custom Dataset	Video and Images	Real driving	Real-time
Dua (83)	India	NTHU and Custom Dataset	videos	Simulated	Real-time
Yarlagadda (84)	India	Custom Dataset	Video	Real driving	Real-time
Rajamohana (85)	India	Public Dataset (mrl.cs.vsb.cz/eye)	Video and Images	Simulated	Real-time
Shahverdy (86)	Iran	Custom Dataset	NA	Real driving	Offline
Guo (87)	Taiwan	ACCV	Visual	Simulated	Real-time
Kim (88)	South Korea	Custom Dataset	Images	Simulated	Real-time
Deng (89)	China	Custom Dataset and Public Datasets	Video	Simulated	Real-time
Aydemir (90)	Turkey	Custom Dataset	Video	Simulated	Real-time
Mahmoud (91)	UAE	NTHU	Video	Simulated	Real-time
Shiau (92)	Japan	Custom Dataset	Video	Real driving	Real-time
Yu (29)	China	Custom Dataset	Multimodal (PPG, Facial, Head Pose)	Real-road	Offline
Adhithyaa (22)	India	KEC+ NTHU	Video	Real-road	Offline
Almazroi (93)	Hungary	NA	Video	Real-road	Offline
Benmohamed (25)	Taiwan	NTHU	Video	Real-road	Offline
Huang (94)	China	NTHU	Video	Simulated	Offline training + real-time inference
Narayana (95)	India	Custom dataset (facial video recordings)	Video	Simulated	Real-time
Kharat (96)	India	Custom dataset	Video	Real driving	Real-time

Table summarizes the dataset properties and implementation conditions of all reviewed studies. It outlines the dataset name, input data modality (public or custom), data modality, driving context (simulated or real-world), and inference mode. This table provides essential context for interpreting model performance by clarifying the environments in which data were collected and systems were evaluated. DDD: driver drowsiness detection; NA: not available.

Supplementary Table C 1: Risk of bias ratings for included studies based on the PROBA-AI checklist (PROBA-AI Risk of Bias Evaluation)

Authors	Participants	Predictors	Outcome	Analysis	Overall Risk of Bias
Al-Anizy (97)	High	Low	High	High	High
Mehta (35)	High	Low	High	High	High
Redhaei (36)	High	Low	High	High	High
Dey (37)	High	Low	High	High	High
Kumaran (38)	High	Low	High	High	High
Chinthalachervu (39)	High	Low	High	High	High
Rehman (41)	High	High	High	High	High
Maior (42)	Low	Low	Low	Low	Low
Saradadevi (40)	High	High	High	High	High
Alioua (43)	High	Low	Moderate	Moderate	Moderate
Vural (44)	Moderate	Low	Low	Low	Low
Tabrizi (45)	Moderate	Low	Moderate	Moderate	Moderate
Ed-Doughmi (46)	Moderate	Moderate	High	High	High
Hashemi (47)	Moderate	Low	Low	Low	Low
Xiao (20)	Moderate	Low	Moderate	Low	Moderate
Padamata (48)	Low	Low	Low	Moderate	Low
Cheamanunkul (49)	High	Moderate	High	Moderate	High
Thakur (50)	High	Moderate	High	Moderate	High
Phan (51)	Moderate	Low	Low	Moderate	Moderate
Zhao (98)	Moderate	Low	Low	Moderate	Moderate
Nguyen (53)	Moderate	Low	Low	Moderate	Moderate
Salman (54)	Moderate	Low	Low	Moderate	Moderate
Chirra (55)	High	Low	Low	High	High
Ghazal (56)	Low	Low	Low	Moderate	Low
Jahan (57)	Low	Low	Low	Low	Low
Arsic (58)	Low	Low	Low	Low	Low
Kulhandjian (59)	Low	Low	Low	High	Low
Saif (60)	Moderate	Low	Low	Moderate	Moderate
Park (61)	Low	Low	Low	High	Low
Huynh (62)	Low	Low	High	High	High
Jia (63)	High	Low	High	High	High
Xie (21)	Low	Low	Low	Moderate	Low
Tamanani (64)	Low	Low	Moderate	High	Moderate
Valsan (65)	High	Low	High	High	High
Jabbar (66)	Moderate	Low	Low	Low	Low
Pachouly (67)	High	Moderate	Moderate	High	High
Xiaofeng (68)	Low	Low	Low	Low	Low
Geng (69)	Moderate	Low	Low	Moderate	Moderate
Chand (70)	High	High	High	High	High
Panwar (71)	High	High	High	High	High
Victoria (72)	Moderate	Low	Low	Moderate	Moderate
Dipu (73)	High	Low	High	High	High
Walizad (74)	Moderate	Low	Low	Moderate	Moderate
George (75)	Low	Low	Low	Low	Low
Celona (76)	Low	Low	Low	Low	Low
Zhuang (77)	Low	Low	Low	Low	Low
Zhang (78)	Low	Low	Low	Low	Low
Macalisan (18)	High	High	High	High	High
Yu (79)	High	High	High	High	High
Lyu (80)	Low	Low	Low	Low	Low
Wang (81)	Moderate	Low	Low	High	Moderate
Pattarapongsin (82)	High	Low	Low	Moderate	Moderate
Dua (83)	Low	Low	Low	Low	Low
Yarlagadda (84)	Low	Low	Low	Low	Low
Rajamohana (85)	High	Moderate	Low	Moderate	High
Shahverdy (86)	Low	Low	Low	Moderate	Low
Guo (98)	Low	Low	Low	Low	Low
Kim (88)	High	Low	Moderate	High	High
Deng (89)	Moderate	Low	Low	Moderate	Moderate
Aydemir (90)	Low	Low	Low	Moderate	Low
Mahmoud (91)	Moderate	Low	Low	Moderate	Moderate

Supplementary Table C 1: Risk of bias ratings for included studies based on the PROBA-AI checklist (PROBA-AI Risk of Bias Evaluation)

Authors	Participants	Predictors	Outcome	Analysis	Overall Risk of Bias
Shiau (92)	Low	Low	Low	Low	Low
Yu (29)	High	High	High	High	High
Adhithyaa (22)	High	Low	High	High	High
Almazroi (93)	High	Low	High	High	High
Benmohamed (25)	High	Low	High	High	High
Huang (94)	Moderate	Low	Low	Moderate	Moderate
Narayana (95)	Low	Low	Low	Low	Low
Kharat (96)	Moderate	Low	Moderate	Moderate	High

Table reports the risk of bias assessment for each study using the PROBA-AI framework. Domain-level judgments and overall ratings are provided to reflect methodological rigor and potential sources of bias. This table supports a transparent appraisal of evidence quality and enhances the interpretability of findings presented in the review.

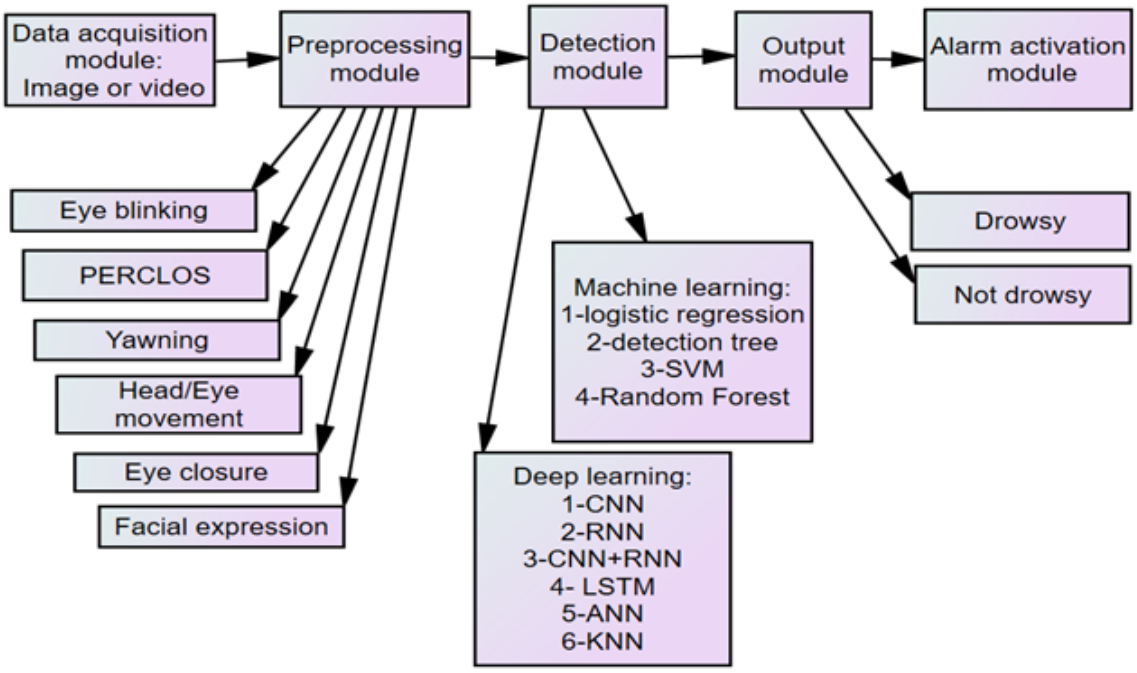

Figure 2: A typical schematic and flow of information in drowsiness detection systems that rely on images or videos (17). PERCLOS: Percentage of Eyelid Closure over Time; SVM: Support Vector Machine; CNN: Convolutional Neural Networks; RNN: Recurrent Neural Network; LSTM: Long Short-Term Memory; ANN: Artificial Neural Network; KNN: K-Nearest Neighbors.

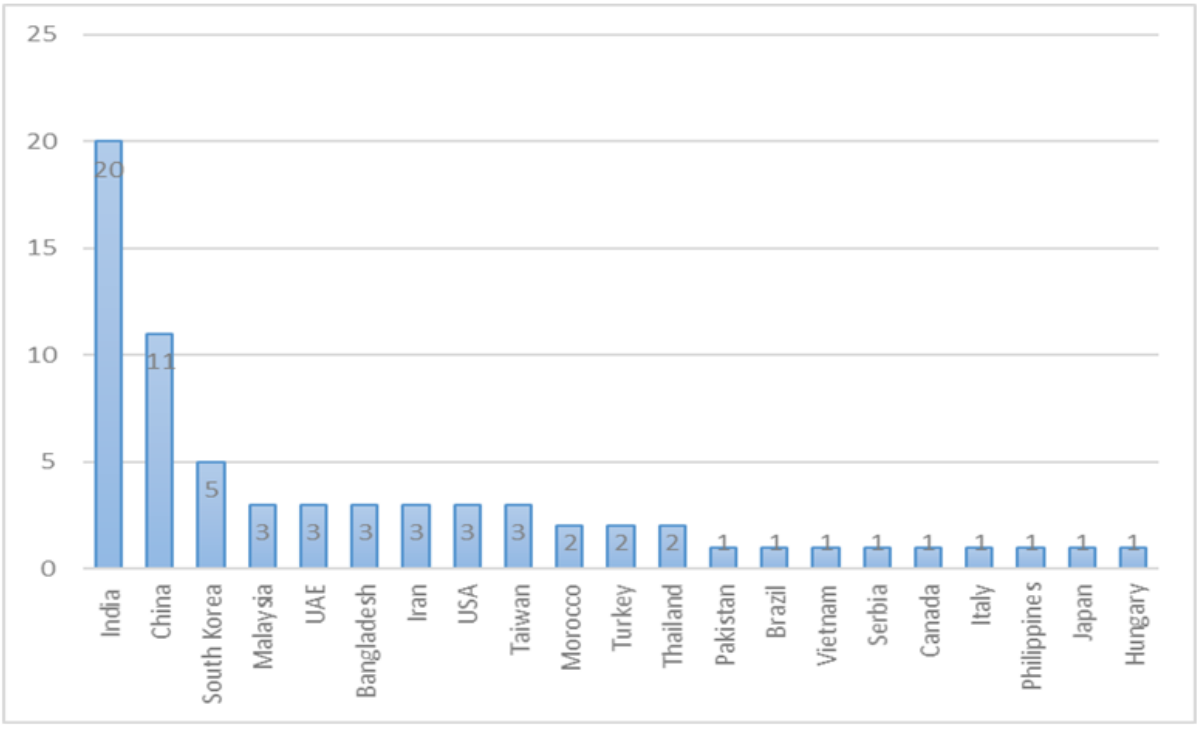

Figure 3: Country-wise distribution of selected studies.