Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

Comprehensive Benchmark of Yolov11n, SSD MobileNet, CenterFace, Yunet, FastMtCnn, HaarCascade, and LBP for Face Detection in Video Based Driver Drowsiness

Agnestia Agustine Djoenaidi Go[1], Farrikh Alzami[2,*], Muhammad Naufal[1], Harun Al Azies[1], Sri Winarno[1], Ricardus Anggi Pramunendar[1], Rama Aria Megantara[1], Isa Iant Maulana[8], Mohammad Arif[2]

[1] Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
[2] Fakultas Ilmu Komputer, Program Studi Sistem Informasi, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1] 111202213996@mhs.dinus.ac.id, [2,*]alzami@dsn.dinus.ac.id, [3] m.naufal@dsn.dinus.ac.id, [4] harun.alazies@dsn.dinus.ac.id, [5] sri.winarno@dsn.dinus.ac.id, [6] ricardus.anggi@dsn.dinus.ac.id, [7] aria@dsn.dinus.ac.id, [8] 111202214416@mhs.dinus.ac.id, [9] 112202206932@mhs.dinus.ac.id
Correspondence Author Email: alzami@dsn.dinus.ac.id
Submitted: **10/11/2025**; Accepted: **16/12/2025**; Published**: 16/12/2025**

Abstract−Face detection is a critical foundation of video-based drowsiness monitoring systems because all downstream tasks such as eye-closure estimation, yawning detection, and head movement analysis depend entirely on correctly identifying the face region. Many previous studies rely on detector-generated outputs as ground truth, which can introduce bias and inflate model performance . To avoid this limitation, I manually constructed a ground truth dataset using 1,229 frames extracted from 129 yawning and microsleep videos in the NITYMED dataset. Ten representative frames were sampled from each video using a face-guided extraction script, and all frames were manually annotated in Roboflow following the COCO format to ensure accurate bounding box labeling under varying lighting, head poses, and facial deformation. Using this manually annotated dataset, I conducted a comprehensive benchmark of seven face-detection algorithms: YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP. The evaluation focused on localization quality using Intersection over Union (IoU ≥ 0.5) and Dice Similarity, allowing each algorithm's predicted bounding box to be directly compared against human defined ground truth. The results show that HaarCascade achieved the highest IoU and Dice scores, particularly in frontal and well-lit frames. FastMtCnn also produced strong alignment with a high number of correctly matched frames. CenterFace and SSD MobileNet demonstrated smooth bounding box fitting with competitive Dice scores, while YOLOv11n and YuNet delivered moderate but stable performance across most samples. LBP showed the weakest results, mainly due to its sensitivity to lighting variations and soft-texture regions. Overall, this benchmark provides an unbiased and comprehensive comparison of modern and classical face-detection algorithms for video-based driver-drowsiness applications.

Keywords: Face Detection; Drowsiness Monitoring; IoU Evaluation; Video-Based Analysis; Deep Learning

1. INTRODUCTION

Driver drowsiness is one of the main factors that reduces driving safety because it directly affects reaction time, concentration, and awareness on the road. Behaviors such as yawning, slow eyelid closure, and microsleep usually appear before a driver becomes fully impaired, which makes early detection extremely important. When these early signs are missed, the driver's reaction time slows down, their vehicle control decreases, and the risk of accidents increases. Because of this, many vision based drowsiness detection systems monitor facial cues in real time. However, these systems depend heavily on accurate face localization. If the face is not detected correctly, then the system cannot analyze eye closure, yawning, or head movement. In other words, the entire performance of the system starts from the face detector.

Several studies have shown how serious drowsiness-related accidents are. Widyastuti and Brilianti reported that in Yogyakarta, drowsy driving has a strong correlation with accident frequency, especially during late-night hours[1], [2] . They also found that drivers who continue driving for long periods without rest are significantly more prone to fatigue even when they believe they are still able to concentrate. Their findings emphasize that drowsiness is not a minor issue but a major safety threat that requires proper monitoring.

Other researchers have also discussed how dangerous drowsiness is during driving, and many of them highlight that detecting early facial cues becomes essential before any deeper drowsiness indicator can be analyzed. Cai et al.[3] found that sleep deprivation leads to serious on-road impairments, including increased lane departures, slower reaction time, and a higher risk of near-crash events, with younger drivers experiencing even stronger performance decline. Meanwhile, Essahraui et al.[4] showed that drowsy driving contributes to tens of thousands of accidents each year worldwide, and modern detection systems often fail because they react too late or rely on intrusive sensing methods. Their findings reinforce that drowsiness is not only a physiological issue but a direct road-safety threat, and that early, reliable facial analysis starting from accurate face localization is a critical foundation for preventing accidents. Meanwhile, Saleem [5] conducted a large systematic review showing that sleepiness and sleep deprivation consistently increase road-traffic accident risk across multiple countries, with several studies reporting that drowsy drivers can have up to twelve times higher crash probability. These findings reinforce that drowsiness is not only a physiological problem but a direct threat to driving safety, which makes early and reliable face localization extremely important before analyzing yawning, eyelid closure, or microsleep behaviors under real-world conditions. Onososen et al.[6] developed a vision-based drowsiness monitoring system using YOLOv8 and showed that construction workers often experience significant fatigue due to long working hours, heavy physical tasks, and unstable environmental conditions.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

Their findings highlight that drowsiness reduces reaction time, weakens situational awareness, and increases the likelihood of on-site accidents. They also emphasized that traditional manual observation methods are slow, subjective, and error-prone, making automated vision-based monitoring a more reliable solution for detecting early signs of fatigue. This supports the idea that early and accurate facial analysis—starting from robust face detection—is essential for preventing safety-critical failures in real-world environments.

Because this study focuses on face localization under yawning and microsleep conditions, the choice of algorithms becomes important. In this research, this study focus on seven widely used face-detection algorithms, representing different detector categories: YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP. Several studies have also evaluated traditional OpenCV based detectors, showing how classical methods such as HaarCascade and LBP remain widely used because of their simplicity and low computational cost. Hasan and Sallow [7] demonstrated that OpenCV provides a stable framework for implementing face detection and recognition, and their review highlights that HaarCascade and LBP continue to be effective for real-time applications despite their sensitivity to lighting and pose variations. Their findings support the use of classical detectors in lightweight systems, which aligns with this study's comparison between classical and deep-learning-based models.These models cover a broad spectrum from deep-learning single-shot detectors to lightweight, edge-optimized models and classical feature-based approaches. Tran et al. [8] carried out a comprehensive survey comparing several deep-learning-based face detectors such as SSD, MTCNN, RetinaNet, and YuNet. Their results showed that each model has different strengths depending on accuracy, processing speed, and the ability to detect small or partially occluded faces. They also noted that real-time applications like driver-monitoring systems depend heavily on fast and stable face localization, which aligns with the focus of this study on evaluating detector performance in yawning and microsleep conditions. moreover, Cai et al. [9] presented a comprehensive systematic review of YOLO-based object detection models and highlighted how recent YOLO versions continue to improve speed, accuracy, and small-object detection performance across various real-time applications. Their findings underline that YOLO architectures are designed to maintain stable localization even under challenging visual conditions, which makes them relevant for tasks that require fast and reliable face detection such as drowsiness monitoring. Several studies have also evaluated face detection models in real-time scenarios, especially those using deep-learning single-shot detectors. Liu et al. [10] demonstrated that an improved SSD-based detector combined with temporal tracking can significantly enhance face localization stability in video sequences, especially under challenging conditions such as occlusion, scale variation, and lighting changes. Their findings show that SSD architectures remain highly relevant for real-time face detection tasks because they balance detection speed and accuracy, which aligns with this study's focus on evaluating both lightweight classical models and deeper single-shot detectors under yawning and microsleep conditions. Recent studies also highlight how widely the YOLO family is used in real-time automotive applications because of its strong balance between speed and detection stability. Gheorghe et al. [11] showed through a large-scale bibliometric review that YOLO has become one of the most frequently adopted object-detection models in modern transportation systems, including traffic monitoring, autonomous-vehicle perception, and intelligent safety applications. Their findings indicate that YOLO continues to dominate real-time detection tasks due to its fast inference and consistent localization performance across various visual conditions, which reinforces the importance of evaluating YOLO based detectors in vision-based drowsiness monitoring systems. Pandey et al. [12] also reported that classical detectors such as HaarCascade remain highly reliable for frontal-face scenarios and still outperform many deep-learning models in specific conditions. Their study showed that HaarCascade provides stable and accurate face localization in real-time environments by relying on simple feature extraction and a lightweight cascade structure, which supports the relevance of evaluating classical detectors alongside modern deep-learning approaches in this study. Another study also demonstrated that classical detectors such as the Viola–Jones Haar Cascade remain effective for real-time facial analysis. Upendra et al.[13] implemented a real-time face-mask detection system using Haar Cascade for face localization and reported a stable and accurate performance, achieving 98% detection accuracy on live video streams. Their findings show that Haar-based detectors continue to be reliable for real-time applications, supporting the relevance of evaluating classical detectors alongside deep-learning models in this study.

To address the lack of evaluation specifically under yawning and microsleep behaviors, this study created a focused ground-truth dataset from the NITYMED collection. From 129 yawning and microsleep videos, this study extracted ten representative frames per video using a face-guided sampling script to ensure that each selected frame contains a visible face. All frames were manually annotated in Roboflow following the COCO format. Manual annotation ensures that the bounding boxes reflect the actual facial region even under challenging conditions such as low light, partial occlusion, wide yawns, and head tilts. This avoids the biased results that occur when detector-generated outputs are used as ground truth.

Using this manually annotated dataset, this study evaluated seven commonly used face-detection algorithms such as YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP. Because the primary goal of this study is to measure localization accuracy, the evaluation focuses on Intersection over Union (IoU ≥ 0.5) and Dice Similarity instead of classification metrics like precision or recall. The evaluation shows that HaarCascade provides the highest IoU and Dice scores on yawning and microsleep frames. YOLOv11n, YuNet, FastMtCnn, and SSD offer consistent but moderate localization performance, while CenterFace performs well but runs slower on CPU only hardware. Classical detectors like HaarCascade and LBP are known to be sensitive to lighting and non-frontal

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

poses, but HaarCascade still performed strongly in stable lighting conditions, which aligns with the results of this evaluation.

2. RESEARCH METHODOLOGY

This research was conducted in several systematic stages to ensure the results align with the research objectives. The designed stages are expected to achieve accurate, measurable, and comparable results across the selected face detection algorithms. The overall research workflow is illustrated in Figure 1.

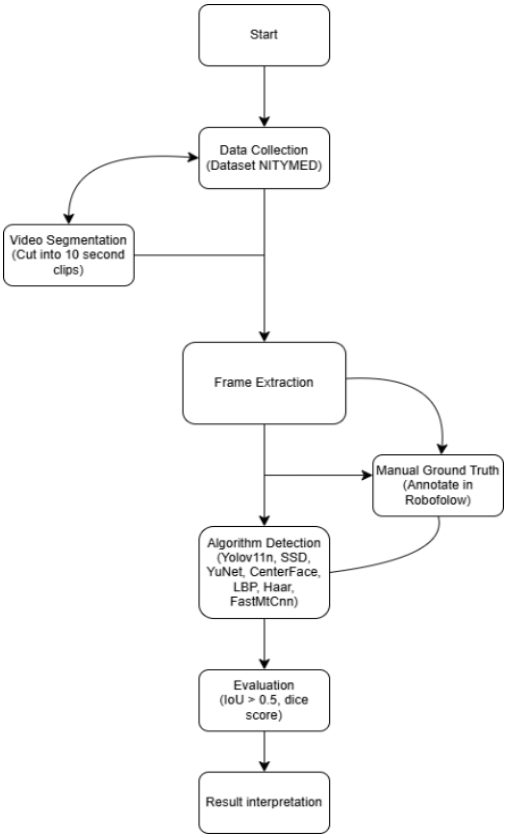


Figure 1. purpose method flowchart

Figure 1 illustrates the complete workflow used in this study, starting from the initial dataset preparation until the final interpretation of results. The purpose of this flowchart is to show that every stage in the experiment follows a clear sequence so that all algorithms are tested under equal and fair conditions. The first stage is Data Collection, where all yawning and microsleep videos from the NITYMED dataset are gathered. At this point, we only selected the clips that clearly show facial visibility because the next steps depend heavily on whether the subject's face is visible or not. The second stage is Preprocessing, which includes cutting the videos into 10-second segments and extracting frames using OpenCV. This step ensures that the important facial moments, such as yawning, blinking, or head dropping, are captured properly. Preprocessing also helps remove unnecessary frames that do not contain useful facial cues. Next is the Algorithm Comparison stage. In this part, all face detection algorithms (YOLOv11n, SSD MobileNet, CenterFace, FastMtCnn, YuNet, HaarCascade, and LBP) receive the exact same input frames. The purpose is to make sure the comparison is not biased. Each algorithm gives its own bounding box output, which will later be compared with ground truth. The Ground Truth Annotation stage is where the manually labeled bounding boxes from Roboflow are stored. These annotations serve as the reference for checking whether each algorithm's predictions match the actual facial region. This manual ground truth ensures that the evaluation is objective and not depending on any algorithm. Finally, the Testing and Evaluation stage calculates the IoU and Dice Similarity between the predicted bounding boxes and the ground truth. This is the stage that determines how accurate each algorithm is under yawning and microsleep conditions. The results from this stage are then interpreted to conclude which algorithm performs the most consistently.

2.1 Data Collection Stage

The dataset used in this study is the NITYMED drowsiness dataset that can be seen in table 1, which contains four subfolders:

Microsleep_Female, Microsleep_Male, Yawning_Female, and *Yawning_Male*. In this research, we used all four subsets, not only the female folders. The reason for including both male and female videos is to make the

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

evaluation more representative and to capture a wider variation of facial characteristics, lighting conditions, and head movements that naturally occur across different subjects.

A Python script was used to segment long videos into 10-second clips using MoviePy. Videos shorter than 10 seconds were automatically removed to avoid incomplete sequences. This segmentation ensures that each clip contains meaningful drowsiness indicators such as yawning, slow blinking, and head dropping moment. In several previous studies, the quality of the collected frames was also mentioned as one of the key factors that affects the accuracy of facial analysis tasks. For example, one study showed that unstable lighting, head movement, and low-quality facial regions tend to reduce the reliability of detection models, especially when the subject is already showing drowsiness symptoms such as slow blinking or head nodding. Saleem (2022) also pointed out that unstable lighting, head movement, and low-quality face regions can reduce the reliability of face-based detection models, especially when the subject is already showing drowsiness signs like slow blinking or head-nodding. This condition is very similar to the NITYMED dataset, where yawning and microsleep frames must be clear before they can be processed properly [5].

Table 1. Dataset Used

Dataset Name	Type	Number of Dataset	Description
Yawning_Female	Video	59	Videos of female subjects who repeatedly yawn and frequently blink as indicators of sleepiness.
Microsleep_Female	Video	11	Videos of female subjects experiencing microsleep episodes characterized by head dropping, slow blinking, and brief loss of alertness.
Microsleep_Male	Video	12	Videos of male subjects showing microsleep behaviors such as head nodding, reduced facial responsiveness, and momentary lapses in attention.
Yawning_Male	Video	47	Videos of male subjects who frequently yawn and blink repeatedly as signs of drowsiness.

Table 1 provides a detailed overview of the specific subsets from the NITYMED drowsiness dataset that were used in this study. Each subset represents a different type of drowsiness-related facial behavior, allowing the evaluation to cover variations such as yawning, blinking frequency, head movements, and microsleep episodes in both male and female subjects. The Yawning_Female subset contains 59 videos and is characterized by clear repetitive yawning behavior accompanied by frequent blinking, which typically appears in early stages of drowsiness. This subset helps the evaluation capture large mouth openings and rapid eyelid movements that often distort facial shape and challenge face detectors. The Microsleep_Female subset consists of 11 videos showing female subjects experiencing brief microsleep episodes. The facial cues in these clips usually include head dropping, slowed blinking, and temporary loss of facial responsiveness. These behaviors introduce non-frontal head poses, making them important for testing detector stability under movement. The Microsleep_Male subset includes 12 videos of male subjects with similar microsleep behaviors. Compared to the female subset, the male videos show slightly more pronounced head-nodding patterns and reduced responsiveness, creating additional challenges for bounding-box accuracy because the face may partially move out of frame or tilt at sharper angles. The Yawning_Male subset contains 47 videos featuring repetitive yawning and repeated blinking in male subjects. Since yawning causes strong facial deformation especially around the mouth and jawline this subset is essential for evaluating whether detectors can maintain accurate face localization despite large changes in facial geometry.

2.2 Preprocessing Stage

Each 10-second clip was converted into image frames using OpenCV at 10 fps. This step ensures that each video contributes several representative frames that capture head movement, yawning, or eyelid-closure moments. After the extraction, all frames were passed through all seven face-detection algorithms used in this study: YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP. The purpose of this stage is to standardize the detection output and ensure that every algorithm receives identical input frames under the same conditions. No fine-tuning or retraining was performed; all detectors were executed using their default pre-trained configurations. Several works focusing on real-time video analysis also highlight the importance of maintaining a consistent preprocessing pipeline. They explain that extracting frames at a fixed FPS helps keep the input stable, especially when the system needs to capture sudden facial changes such as yawning, blinking, or micro-nodding. Ali-Gombe et al. (2021) mentioned that using a fixed extraction rate helps stabilize the input for real-time face analysis and prevents models from missing fast facial transitions. Their preprocessing strategy is very similar to the approach applied in this study[14].

2.3 Algorithm Comparison Stage

All algorithms in this study were tested using their official implementations from the DeepFace GitHub repository, ensuring that each model runs under the same configuration without additional training or modification. The selected

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

detectors represent different categories of face detection approaches. YOLOv11n is a modern deep-learning single-shot detector designed for real-time performance. SSD MobileNet is a lightweight CNN-based model optimized for speed on limited hardware. CenterFace represents an anchor-free detection approach that focuses on efficient face localization with stable output. YuNet is designed specifically for edge devices, offering fast inference and low computational requirements. FastMtCnn is a pyramidal deep-learning model that uses multi-stage refinement, allowing it to produce stable face predictions even under varied head poses. Meanwhile, HaarCascade and LBP are classical feature-based methods that remain widely used due to their speed and simple architecture, although they typically struggle under challenging lighting or rapid head movements. To maintain consistency, all detection outputs from these seven algorithms were exported into JSON format for frame-by-frame comparison. Classical detectors such as HaarCascade also remain effective for frontal face scenarios and have been used in several recent studies. Ariffin et al.[15] demonstrated that HaarCascade still achieves high accuracy on multiple public datasets and can outperform several traditional machine-learning approaches, reinforcing the relevance of including classical detectors in comparative evaluations.

Several studies also highlight the strengths and limitations of MTCNN-based methods when dealing with challenging environments. Liu et al. (2021) [16] reported that the standard MTCNN architecture performs well on clean and frontal-face scenarios but tends to degrade under complex backgrounds, low-quality frames, and non-frontal head poses. Their study showed that MTCNN requires additional optimization modules to maintain stable detection performance in real-world conditions, which aligns with the behavior observed in the NITYMED dataset where yawning and microsleep movements often distort the facial shape or cause partial occlusion. This finding supports the inclusion of FastMtCnn in this evaluation since it inherits the multi-stage refinement pipeline of MTCNN but provides faster and more stable predictions under variable head poses.his ensures that the evaluation is objective, reproducible, and based on identical input data across all detectors. Other studies also compared deep-learning-based detectors and classical feature-based detectors under different conditions. They found that deep learning models like YOLO usually remain more stable when the face changes shape or orientation, while classical models such as HaarCascade or LBP tend to fail more often under uneven lighting or movement. A systematic review on YOLO based models also noted that deep-learning detectors such as YOLO are usually more stable when the face changes orientation or shape, while classical detectors like HaarCascade or LBP often fail under uneven lighting or fast movements. In addition, findings from Yakovleva et al. [17] show that CenterFace provides stable face detection performance across moderate rotation variations, but its accuracy remains lower compared to more advanced deep learning detectors such as RetinaFace, DSFD, and SCRFD. Their study also highlights that CenterFace is highly efficient and lightweight, making it suitable for real-time scenarios, although its detection quality tends to degrade when the face undergoes extreme pose changes or appears very small in the frame. These results support the inclusion of CenterFace in this evaluation because it represents a computationally efficient yet practically relevant detector for comparison alongside modern deep-learning model.

2.4 Ground Truth Reference

The ground truth in this study was manually generated using the Roboflow annotation tool. After assigning an initial bounding box and labeling each image with a face, every bounding box was manually corrected by the researchers. This ensures that the ground truth is fully validated by humans and remains completely independent from any of the seven face-detection algorithms used in this study (YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP). In total, 1,290 images (10 frames × 129 videos) were annotated manually following the standard COCO format, which includes the fields (x, y, width, height). This manual annotation process was essential to prevent bias and to make sure all algorithms were evaluated under the same objective conditions. The goal was to create bounding boxes that represent facial regions accurately, even in difficult situations such as dim lighting, wide yawns, strong facial deformation, tilted head posture, and partially obscured faces. These conditions are common in yawning and microsleep scenarios, so high quality ground truth is necessary to assess localization performance fairly. Similar annotation approaches are also used in previous research, where manually created bounding boxes are preferred because they eliminate the risk of using detector-generated labels that could favor one algorithm over another. Saleem (2022) also emphasized that manually validated ground-truth annotations are more reliable when the dataset contains challenging facial conditions such as tilted heads, partial occlusion, or rapid changes in expression. For this reason, manual annotation is considered the safest and most objective method for evaluating face-detection algorithms [5]. By creating a fully human-annotated dataset, all seven algorithms are compared against the same standard, ensuring that the evaluation of YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP is fair, unbiased, and reproducible.

2.5 Testing and Evaluation Stage

In this stage, all seven face detection algorithms YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP were evaluated by comparing their predicted bounding boxes against the manually annotated ground truth. The evaluation relied on two localization-focused metrics: Intersection over Union (IoU ≥ 0.5) and Dice Similarity, since the main objective of this study is to measure bounding-box accuracy rather than classification performance. IoU was used to quantify how much spatial overlap exists between the predicted box and the ground truth, while Dice Similarity was used to assess how consistent the predicted bounding-box region is with the actual

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

facial area. A prediction was considered correct when its IoU exceeded 50%. All evaluation scripts were executed under the exact same computational environment to ensure fairness among the seven algorithms. To respond to the reviewer's concern about reproducibility, the hardware specifications used during the experiments are explicitly documented. All tests were performed on a laptop equipped with an AMD Ryzen 9 5900HX processor (16 CPUs), 16 GB RAM, and Integrated Radeon Graphics, running Windows 11 Home 64-bit. Because the system does not include a dedicated GPU, all seven algorithms were evaluated using CPU-only processing. This setup ensures that the reported FPS values and processing times reflect consistent and comparable performance across all models.

Reporting hardware configuration is essential because inference speed especially FPS can vary significantly depending on the device used. Related work also highlights this issue. Several studies show that CPU-only execution generally results in lower FPS compared to GPU acceleration, but it still provides reliable performance for benchmarking when all algorithms are tested under the same hardware. Ali-Gombe et al. (2021) emphasized the importance of reporting precise hardware specifications, noting that differences between CPU and GPU environments can greatly influence runtime measurements. Their findings reinforce why documenting hardware details is crucial for reproducibility and accurate interpretation of FPS results in face-detection research[14]. By running YOLOv11n, SSD MobileNet, CenterFace, YuNet, FastMtCnn, HaarCascade, and LBP under identical conditions and comparing them against a fully human-verified ground truth, this evaluation provides an objective and fair comparison of localization accuracy and processing efficiency across all seven algorithms.

3. RESULT AND DISCUSSION

This experiment used the manually annotated ground truth that we created using Roboflow. In total, 1,229 frames were extracted from the selected drowsiness videos. Every frame was labeled manually following the COCO format (x, y, width, height). This manual labeling ensures that the comparison across algorithms is fair, because the evaluation is no longer dependent on any specific model. All detectors were compared directly to these human-made annotations.

3.1 Detection Accuracy and Localization Quality

In this evaluation, this study used IoU (with a ≥ 0.5 threshold) and Dice Similarity to measure how well each algorithm's predicted bounding box overlapped with the manually annotated ground truth. The full results for all detectors are shown in Table 2 and Table 3, and each metric gives a slightly different picture of how the algorithms behave on yawning and microsleep frames. OpenCV HaarCascade achieved the highest IoU value (0.5846) and also the highest Dice score (0.7284). From what we observed during testing, this makes sense because many of the frames still contain fairly clear frontal faces, and HaarCascade usually performs best when the head position is stable and the lighting is not too extreme. This finding is also consistent with prior research showing that Haar Cascade performs particularly well when the face is captured under stable illumination and moderate distances. Putra et al.[18] reported that Haar Cascade can reliably detect facial components such as the nose and lips at light intensities between 80–140 lux and within a 30–120 cm detection range, demonstrating that the method remains effective when the face structure is clearly visible and not heavily distorted. Their results also showed that the Haar Cascade classifier is capable of maintaining stable detection even in low-cost, real-time environments, reinforcing why this algorithm performs strongly on frontal and well-lit frames in my dataset. This observation is also supported by Yap et al.[19], who showed that Haar Cascade achieves higher accuracy when the face appears in a frontal position and under stable lighting conditions. Their benchmark results demonstrated that increasing illumination consistency and maintaining upright face orientation significantly improves detection rates while reducing false positives, which aligns directly with the behavior observed in this dataset. Several recent studies also confirm that HaarCascade remains effective under frontal and well-lit conditions because it is optimized for clear facial structures and controlled illumination. Ahmad et al.[20] demonstrated that HaarCascade can still achieve reliable face detection and recognition in real-time applications as long as the lighting is moderate and the face orientation stays within a limited angle range (approximately ±40°), which aligns closely with the behavior observed in this dataset . Even though HaarCascade is an older method, in this specific dataset it ended up aligning its bounding boxes the closest to the manually labeled ground truth. Matched-frame count was also the highest (869), although the runtime was the slowest among all detectors (605 seconds).

CenterFace also performed quite well, with an IoU of 0.5570 and a Dice score of 0.7080. Its Dice value is the second highest after HaarCascade, which means its bounding box shape often followed the ground truth nicely. However, its matched frame count (581 frames) is lower, and the runtime (138.2 seconds) is somewhere in the middle faster than HaarCascade but slower than YuNet and YOLOv11. SSD MobileNet achieved an IoU of 0.5494 and a Dice of 0.7033, which also puts it on the higher end of overlap quality. SSD seems to produce smoother bounding boxes, and even though its IoU is slightly lower than HaarCascade and CenterFace, the higher Dice score shows that the predicted box shape fits well around the face. Its matched frame count (581) is the same as CenterFace, and the runtime is relatively fast (108.2 seconds).

FastMtCnn delivered strong results too, especially in terms of matched frames (823), which is the second highest after HaarCascade. Its IoU was 0.5418 and Dice was 0.6947. So even though its overlap scores are slightly lower than SSD and CenterFace, it detects faces more consistently across the dataset. The runtime (211 seconds) is not the fastest, but still acceptable considering how stable its performance is. YOLOv11 and YuNet have very similar

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

IoU and Dice values, both sitting around the 0.52–0.53 range for IoU and 0.68 for Dice. YOLOv11 detected more frames (786), while YuNet is the fastest detector overall with only 37.1 seconds of runtime. In terms of accuracy, they are not the strongest, but they give a good balance between detection count and speed. YOLOv11 also tends to produce more false positives, which affects its IoU even though it detects many frames. YuNet is the fastest detector overall with only 37.1 seconds of runtime, which aligns with its design as a millisecond-level lightweight face detector optimized for edge devices[21]

LBP had the lowest performance, with an IoU of 0.5149, a Dice score of 0.6360, and the smallest matched frame count (550). This is expected because LBP relies heavily on texture patterns, and many frames in this dataset have soft lighting, shadows, or slight motion blur. Those conditions make LBP struggle more compared to the other detectors. When looking at the results as a whole, HaarCascade clearly gives the best overlap with the ground truth, especially in frontal and well-lit frames. However, it is not the fastest detector, and it is not the most stable when the head rotates or when the lighting changes suddenly. Detectors like FastMtCnn, YOLOv11, and YuNet are still competitive in terms of detection count and speed, so the best detector really depends on which trade off is more important: accuracy, speed, or the number of faces detected. In addition to these quantitative findings, the differences in algorithm behavior also highlight how each detector responds to the unique challenges present in yawning and microsleep conditions. Frames in this dataset often include rapid facial deformation, partial occlusion from head tilting, and sudden illumination shifts caused by movement, all of which influence how consistently the detectors can maintain accurate localization. Classical methods such as HaarCascade and LBP rely heavily on hand-crafted features, which makes them more sensitive to lighting and pose variations; however, HaarCascade still manages to outperform modern models in clear frontal scenarios because its feature filters are highly optimized for upright face structures. Deep-learning detectors like YOLOv11, FastMtCnn, and SSD tend to generalize better under non-frontal poses, but their bounding boxes sometimes deviate from the precise facial region when yawning causes major geometric distortion. Meanwhile, lightweight models such as YuNet show that efficiency does not always translate into high stability although extremely fast, YuNet occasionally underestimates the face area when the illumination becomes uneven. These variations across algorithms illustrate that face detection performance is not determined by accuracy alone, but also by how robustly the model adapts to real-world facial dynamics. Considering these factors, it becomes clear that each detector carries its own strengths and limitations, and their suitability ultimately depends on the specific environmental constraints and performance priorities of the intended drowsiness-monitoring system. A system that prioritizes precision in controlled environments might benefit from HaarCascade, while applications that require rapid response on low-power hardware may achieve better performance with models like YuNet or SSD. This complexity reinforces the importance of conducting multi-metric evaluations, as relying on a single indicator such as IoU or runtime cannot fully represent the practical usability of each face-detection algorithm in challenging real-world scenarios.

Table 2. Accuracy Results IoU

Algorithm	IoU	Dice	Matched Frames	Runtime (s)
HaarCascade	0.5846	0.7284	869	605.0
SSD	0.5494	0.7033	581	108.2
FastMtCnn	0.5418	0.6947	823	211.0
Yolov11	0.5265	0.6840	786	85.9
Yunet	0.5249	0.6834	656	37.1
LBP	0.5149	0.6360	550	385.9
CenterFace	0.5570	0.7080	581	138.2

3.2 Visual Result Samples

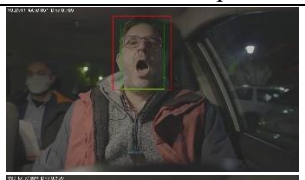
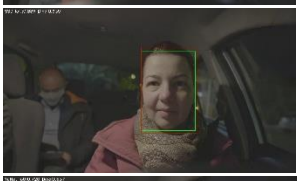
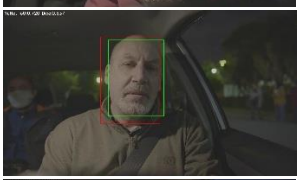
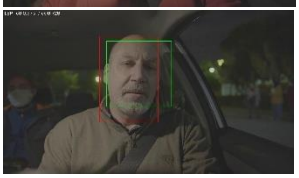
During the evaluation process, our script also generated several visual examples for each algorithm. These samples place the ground truth bounding box and the predicted bounding box on top of the same frame, making it much easier to see how well each detector actually aligns with the manually annotated face region. In these images, the red box represents the manual ground truth annotation that we created in Roboflow, while the green box represents the predicted bounding box produced by each face-detection model. Every visualization also includes the IoU and Dice scores at the top-left corner, so we can immediately understand how closely the model's prediction matches the ground truth.

From these visual samples in table 3, each algorithm shows its own pattern of behavior. FastMtCnn consistently produces one of the closest alignments, as the green box often fits neatly inside the red ground-truth box with very minimal shifting. YOLOv11, on the other hand, tends to predict a slightly larger bounding box that covers the entire head. While this helps capture the full face area, it also causes the green box to extend beyond the ground-truth region, which can lead to more false positives. SSD MobileNet generally follows the face shape quite well, but we noticed that in yawning frames the predicted box sometimes shifts slightly downward, especially when the mouth opens widely. Meanwhile, YuNet can miss parts of the face in dim lighting, producing a noticeably smaller box that sits inside the red GT box. LBP also struggles with darker frames; its predictions often shift upward or downward because the algorithm relies heavily on texture patterns. In contrast, OpenCV HaarCascade performs very strongly when the

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

face is frontal and the lighting is stable the green box almost completely overlaps the red box even though it may lose accuracy when the head starts rotating. For CenterFace, the predictions look tight and well centered in most frames. The green box usually aligns cleanly with the red GT box, although during wide-yawn conditions, it tends to slightly underestimate the chin area. Even so, its visual consistency makes CenterFace one of the more reliable detectors in terms of bounding-box shape and stability.

Table 3. Comparative Visual Results of Face-Detection Algorithms

Algorithm	Visual Sample	Result
YOLOv11		YOLOv11 predicts a larger bounding box, capturing the full head but sometimes exceeding the GT region.
SSD		SSD aligns fairly well but tends to shift slightly downward in yawning frames.
YuNet		YuNet sometimes detects only part of the face when the light is dim. Green box is noticeably smaller than the GT red box
HaarCascade		HaarCascade performs strongly in frontal faces; the green and red boxes overlap closely.
LBP		LBP detects the face but often shifts upward or downward. Works better when the face is fully lit.
FastMtCnn		FastMtCnn has one of the closest alignments. The green box fits right inside the red GT box.
CenterFace		CenterFace produces tight and well-centered boxes. Overlap with the red GT is strong, though it slightly underestimates the chin area during wide yawns.

3.3 Processing Time

The runtime results were obtained on a laptop equipped with an AMD Ryzen 9 5900HX (16-thread) processor, Radeon integrated GPU, 16 GB of RAM, and Windows 11 Home 64-bit. No external GPU acceleration was used, so the reported speeds represent pure CPU-based inference performance. Processing time is also an important aspect in this experiment because a face-detection system for drowsiness monitoring should ideally work in real time. Each algorithm was tested on the same set of 1,229 frames, and the total runtime was recorded. From these results, every detector shows different speed characteristics depending on its model structure and computational load. YuNet is the fastest among all algorithms, with a total runtime of 37.1 seconds, which makes it the most efficient model for real-time applications. YOLOv11 also performs relatively fast at 85.9 seconds, especially considering that it is a modern deep-learning detector. SSD MobileNet runs in 108.2 seconds, making it another efficient option, balancing both speed and accuracy. CenterFace shows a medium level runtime of 138.2 seconds, not as fast as YuNet or YOLOv11, but still faster than FastMtCnn. FastMtCnn, although quite strong in detection consistency, required 211.0 seconds, placing it among the slower deep-learning detectors. LBP took 385.9 seconds, which is surprisingly slow for a classical method, likely due to repeated multi-scale scanning. The slowest of all is OpenCV HaarCascade with 605.0 seconds,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

even though it achieved the highest IoU and Dice scores. HaarCascade's long runtime indicates that its strong accuracy in this dataset comes with a significant computational cost. The speed comparison shows that the fastest detectors are not necessarily the most accurate, and the most accurate detector (HaarCascade) is by far the slowest. This creates a clear trade-off between detection quality and processing efficiency. YuNet is the fastest detector among all algorithms (37.1 seconds total runtime) This result is consistent with prior research showing that YuNet was specifically engineered for high-efficiency CPU-based inference and can reach millisecond-level speed while maintaining competitive accuracy [21].

3.4 Discussion

Looking at the results across accuracy, localization quality, matched-frame count, and processing time, each detector shows different strengths depending on the evaluation criteria. OpenCV HaarCascade clearly provides the best overlap with the manually annotated ground truth, achieving the highest IoU (0.5846) and the highest Dice score (0.7284). This aligns well with the visual samples, where its green prediction box often overlaps almost perfectly with the red ground-truth box in frontal and well lit frames. However, this accuracy comes with the drawback of being the slowest algorithm in the entire experiment, taking 605 seconds to process all frames. So while HaarCascade is strong in localization, it is not ideal for real-time deployment unless optimized. CenterFace, SSD, and FastMtCnn form a middle group that balances accuracy and speed. CenterFace and SSD have Dice scores above 0.70, indicating that their bounding-box shapes match the ground truth quite smoothly. FastMtCnn, on the other hand, stands out in detection consistency it matched 823 frames, the second highest after HaarCascade. Even though its IoU and Dice are slightly lower, FastMtCnn often produces very tight alignments in the visual examples. SSD runs faster than CenterFace and produces smoother bounding boxes, while CenterFace maintains strong shape quality but is slightly slower. YOLOv11 and YuNet deliver similar overlap accuracy (IoU around 0.52–0.53 and Dice around 0.68), but they differ significantly in detection speed and behavior. YOLOv11 detects many frames (786) and runs relatively fast, but it tends to draw larger boxes that sometimes exceed the face region. YuNet, although less consistent, is extremely fast (37.1 seconds), making it the most suitable for lightweight real-time systems where speed is the main priority. LBP consistently shows the weakest performance overall. With the lowest IoU (0.5149) and Dice (0.6360), and the smallest matched frame count (550), it struggles particularly in low light or soft-texture scenes. It also runs unexpectedly slow for a classical model. The visual results also reflect this: LBP's green box often shifts up or down, especially when the lighting is uneven. When combining all observations accuracy metrics, runtime speed, and visual alignment the results show that no single detector is perfect for all conditions. HaarCascade gives the best accuracy but is the least efficient. YuNet is the fastest but sacrifices some overlap precision. FastMtCnn and SSD offer good middle-ground performance, while YOLOv11 provides strong detection coverage with moderate accuracy. CenterFace stands out for producing clean, tight bounding boxes but still misses some frames compared to YOLOv11 and FastMtCnn. Each algorithm's usability depends entirely on whether the system needs higher accuracy, faster speed, or a balance of both.

4. CONCLUSION

This study compared seven face-detection algorithms YOLOv11, YuNet, OpenCV HaarCascade, SSD MobileNet, CenterFace, FastMtCnn, and LBP using a manually annotated ground-truth dataset derived from 1,229 frames of yawning and microsleep scenes. By using fully manual COCO format annotations, the evaluation becomes more objective and no longer biased toward any specific detector. From the experimental results, each algorithm shows its own strengths and limitations. OpenCV HaarCascade achieved the highest localization accuracy, with the highest IoU (0.5846) and Dice score (0.7284). This matches the visual samples, where HaarCascade aligns very closely with the manually annotated bounding boxes, especially in frontal and well-lit conditions. However, it is also the slowest model, requiring 605 seconds to process all frames, which makes it less suitable for real-time use. CenterFace, SSD MobileNet, and FastMtCnn form a strong middle group. CenterFace provides tight and clean bounding boxes with a Dice score of 0.7080, while SSD produces smooth and stable predictions with competitive overlap scores. FastMtCnn stands out with one of the highest matched-frame counts (823), showing consistent detection across many scenes even though its IoU and Dice are slightly lower than SSD and CenterFace. YOLOv11 and YuNet achieve moderate localization accuracy (IoU around 0.52–0.53), but differ in their advantages. YOLOv11 detects many frames (786) and runs relatively fast, though it often predicts slightly larger bounding boxes. YuNet is by far the fastest model (37.1 seconds total runtime), making it ideal for lightweight real-time applications, even though its localization accuracy is slightly lower. LBP shows the weakest performance overall, with the lowest IoU (0.5149), lowest Dice score (0.6360), and smallest matched-frame count (550). The visual examples confirm that LBP is highly sensitive to lighting variations and often shifts its bounding box upward or downward. Overall, the results demonstrate no single detector is universally superior. HaarCascade produces the most accurate overlap but is extremely slow. YuNet is the fastest but sacrifices some precision. FastMtCnn and SSD provide a good balance between detection consistency and localization quality. YOLOv11 offers wide detection coverage with moderate overlap accuracy, while CenterFace consistently produces clean bounding boxes but misses more frames than YOLOv11 and FastMtCnn. These findings highlight that the choice of detector should depend on the application's prioritieswhether accuracy, speed, stability, or overall detection coverage is more important. Future work may include evaluating these algorithms on more diverse

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1775−1784
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8678

subjects, additional lighting conditions, and continuous video sequences. Combining fast models like YuNet with high-accuracy models like HaarCascade or CenterFace may also create a hybrid approach that balances speed and precision for real-time drowsiness monitoring system.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to IDSS research group from Universitas Dian Nuswantoro

REFERENCES

[1] Nur Rachmi Widyastuti and Dani Fitria Brilianti, "Impact of Drowsiness on Road Traffic Accidents in Yogyakarta," *Journal of Scientific Research, Education, and Technology (JSRET)*, vol. 3, no. 4, pp. 1651–1661, 2024, doi: 10.58526/jsret.v3i4.555.

[2] Farrikh Alzami, Muhammad Naufal, Harun Al Azies, Sri Winarno, and Moch Arief Soeleman, "Time Distributed MobileNetV2 with Auto-CLAHE for Eye Region Drowsiness Detection in Low Light Conditions," *(IJACSA) International Journal of Advanced Computer Science and Applications,* p. 13, 2024, doi: 10.14569/IJACSA.2024.0151146.

[3] Anna W. T. Cai, Jessica E. Manousakis, and Bikram Singh, "On-road driving impairment following sleep deprivation differs according to age," *Scientific Reports*, vol. 11, p. 21561, 2021, doi: 10.1038/s41598-021-99133-y.

[4] Siham Essahraui, Ismail Lamaakal, and Ikhlas El Hamly, "Real-Time Driver Drowsiness Detection Using Facial Analysis and Machine Learning Techniques," *Sensors*, vol. 25, no. 3, p. 812, 2025, doi: 10.3390/s25030812.

[5] Shehzad Saleem, "Risk Assessment of Road Traffic Accidents Related to Sleepiness During Driving: A Systematic Review," *East Mediterranean Health Journal*, vol. 28, no. 9, pp. 695–700, 2022, doi: 10.26719/emhj.22.055.

[6] Adetayo Olugbenga Onososen, Innocent Musonda, and Damilola Onatayo, "Drowsiness Detection of Construction Workers: Accident Prevention Leveraging YOLOv8 Deep Learning and Computer Vision Techniques," *Buildings*, vol. 15, no. 3, p. 500, 2025, doi: 10.3390/buildings15030500.

[7] Ramadan TH. Hasan and Amira Bibo Sallow, "Face Detection and Recognition Using OpenCV," *JOURNAL OF SOFT COMPUTING AND DATA MINING*, vol. 2, no. 2, p. 12, 2021, doi: https://doi.org/10.30880/jscdm.2021.02.02.008.

[8] Guodong Guo and Na Zhang, "A survey on deep learning based face recognition," *Computer Vision and Image Understanding*, vol. 189, p. 102805, 2019, doi: https://doi.org/10.1016/j.cviu.2019.102805.

[9] Z. Cai, K. Zhou, and Z. Liao, "A Systematic Review of YOLO-Based Object Detection in Medical Imaging: Advances, Challenges, and Future Directions," *Computers, Materials and Continua*, vol. 85, no. 2, pp. 2255–2303, Sep. 2025, doi: 10.32604/cmc.2025.067994.

[10] Yilin Liu, Ruian Liu, Shengxiong Wang, Da Yan, Bo Peng, and Tong Zhang, "Video Face Detection Based on Improved SSD Model and Target Tracking Algorithm," *Journal of Web Engineering*, vol. 21, no. 2, 2022, doi: https://doi.org/10.13052/jwe1540-9589.21225.

[11] C. Gheorghe, M. Duguleana, R. G. Boboc, and C. C. Postelnicu, "Analyzing Real-Time Object Detection with YOLO Algorithm in Automotive Applications: A Review," *CMES - Computer Modeling in Engineering and Sciences*, vol. 141, no. 3, pp. 1939–1981, Oct. 2024, doi: 10.32604/cmes.2024.054735.

[12] Anurag Pandey, Divyansh Choudhary, Ritik Agarwal, Tushar Shrivastava, and Kriti, "Face detection using Haar cascade classifier," *AECE 2022*, p. 599, 2022, doi: 10.2139/ssrn.4157631.

[13] Hruthik S. Upendra, Shruti Suman, Sai S. Vishnu, and Jaya Dharani, "Real-Time Face Mask Detection using OpenCV and Deep Learning," *CEUR-WS (Vol. 3085)*, p. 6, Sep. 2021, doi: 10.1109/AECE62803.2024.10911331.

[14] Adamu Ali-Gombe, Eyad Elyan, Carlos Francisco Moreno-García, and Johan Zwiegelaar, "Face Detection with YOLO on Edge," in Proceedings of the International Neural Networks Society (INNS, volume 3). Springer, Cham, 2021, pp. 284–292. doi: 10.1007/978-3-030-80946-1_25.

[15] Noor Afiza Binti Mohd Ariffin, Usman Abdul Gimba, and Ahmad Musa, "Face detection based on Haar Cascade and Convolution Neural Network (CNN)," Journal of Advanced Research in Computing and Applications, vol. 38, p. 11, 2025, doi: https://doi.org/10.37934/arca.38.1.111.

[16] Gaiping Liu, Jianmei Xiao, and Xihuai Wang, "Optimization of Face Detection Algorithm based on MTCNN," *Semantic*, [Online]. Available: https://www.semanticscholar.org/paper/Optimization-of-Face-Detection-Algorithm-based-on-Liu-Xiao/8e23ccf923cb9223accf282582919817c89b0706

[17] Olena Yakovleva, Andrii Kovtunenko, Valentyn Liubchenko, Vadym Honcharenko, and Oleg Kobylin, "Face Detection for Video Surveillance-based Security System", International Conference on Computational Linguistics and Intelligent Systems, April, 2023

[18] Radimas Putra M.D.L, Sirojul Hadi, and Parama Diptya Widayaka, "Low Cost System for Face Mask Detection Based Haar Cascade Classifier Method", Matrik, vol. 21, no. 1, 2021, doi: 10.30812/matrik.v21i1.1187.

[19] Yap Jia Hui and Lee Siaw Chong, "Face Detection withHaar Cascades Method," Enhanced Knowledge in Sciences and Technology, vol. 2, no.1, 2022, doi: https://doi.org/10.30880/ekst.2022.02.01.033.

[20] Adlan Hakim Ahmad *et al.*, "Real time face recognition of video surveillance system using haar cascade classifier" Indonesian Journal of Electrical Engineering and Computer Science (IJEECS) vol. 21, no. 3, 2021, doi: :10.11591/ijeecs.v21.i3.pp1389-1399.

[21] Wei Wu, Hanyang Peng, and Shiqi Yu, "YuNet: A Tiny Millisecond-level Face Detector," *Machine Intelligence Research, vol. 20, April*, 2023, doi: 10.1007/s11633-023-1423-y.