Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3704~3718
e-ISSN: 2548-6861

Image Classification using DenseNet-121 Based on MediaPipe Face Mesh for Real-Time Drowsiness Detection

Raihan Ramadhan Hamzah [1], Christy Atika Sari [2]*, Eko Hari Rachmawanto [3], Rabei Raad Ali [4]
[1,2,3]Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
[4]Department of Computer Engineering Technology, Northern Technical University (NTU), Iraq
111202315095@mhs.dinus.ac.id [1], christy.atika.sari@dsn.dinus.ac.id [2], eko.hari@dsn.dinus.ac.id [3]
rabei@ntu.edu.iq [4]

Article Info

Article history:

Received 2026-07-06
Revised 2026-08-08
Accepted 2026-08-10

Keyword:

Drowsiness Detection,
DenseNet-121,
MediaPipe Face Mesh,
Accuracy,
Region of Interest.

ABSTRACT

The high rate of traffic accidents caused by driver drowsiness and microsleep highlights the urgent need for reliable driver monitoring systems. However, conventional Eye Aspect Ratio (EAR) methods often fail due to their high sensitivity to changes in head poses and ambient lighting conditions, while standard Convolutional Neural Network (CNN) models impose heavy computational loads on hardware. This study aims to implement and evaluate a real-time drowsiness detection system by integrating the DenseNet-121 architecture with MediaPipe Face Mesh. The proposed method utilizes MediaPipe Face Mesh to isolate the left and right eye Regions of Interest (ROI) independently, using a proportional padding of 35%, which are then classified using a DenseNet-121 transfer learning model fine-tuned in two stages across its last 30 layers. Evaluation was conducted using a custom dataset of 2,000 source images from five subjects, yielding 3,926 eye-region samples after extraction and quality filtering, assessed using a Subject-Independent Leave-One-Subject-Out (LOSO) cross-validation protocol. Across five folds, the model achieved a mean accuracy of 83.22% (standard deviation 13.22 percentage points) and a mean AUC of 0.879 (standard deviation 0.131), with performance variation across subjects found to correlate with inter-subject differences in eye-closure expressiveness, where the two lowest performing subjects also exhibited the lowest AUC values (0.707 and 0.769). The system achieved an average total latency of 198.00 ms per frame, equivalent to 5.1 FPS. These findings indicate that the integration of MediaPipe Face Mesh and DenseNet-121 shows meaningful potential for real-time drowsiness monitoring, while also highlighting the importance of subject-independent evaluation and cross-domain generalization for reliable real-world deployment.

I. INTRODUCTION

The number of traffic accidents in Indonesia remains fairly high every year. According to data from the Indonesian National Police Traffic Corps (Korlantas Polri), the number of accidents reached 150,279 cases in 2024 and increased further in 2025. According to [1], driver drowsiness is one of the main causes of traffic accidents. A drowsy condition can reduce a driver's level of concentration and alertness, thereby slowing reaction time in responding to road situations. In addition, drowsiness can also lead to microsleep. Research by [2] state that microsleep is a brief sleep state lasting 1 to 30 seconds, during which the affected person fails to respond to motor sensory input and becomes unconscious. This condition has the potential to endanger driving safety, so a system capable of detecting drowsiness at an early stage is needed to minimize the risk of traffic accidents. The implementation of an intelligent monitoring system has been shown to reduce fatality rates by up to 20% [3].

Various methods have been developed to automatically detect drowsiness, one of which is through a computer vision-based approach. A systematic review of various deep learning models for driver drowsiness detection shows that CNN-based approaches and facial landmark analysis remain the most widely adopted methods due to their balance between accuracy and computational efficiency [4]. Recent research trends also show a shift toward lighter and more efficient CNN architectures to support real-time drowsiness detection in vehicle systems [5]. This approach enables real-time analysis of facial conditions and is non-invasive, so it does not disturb the user [2]. Several previous studies have used Convolutional Neural Network (CNN)-based methods to detect drowsiness through facial images. This method is able to achieve a high level of accuracy in identifying drowsy conditions [1]. CNNs can directly learn various head positions, face orientations, and even the use of accessories (such as glasses) and head tilt, by extracting features that are resistant to pose changes, thereby significantly improving detection capability [6]. Although effective, CNN performance is highly dependent on data quality and the preprocessing process used to determine the Region of Interest (ROI). To improve system efficiency, the use of facial landmarks through MediaPipe Face Mesh has become a crucial approach. No longer used merely to compute mathematical values such as the Eye Aspect Ratio, which tends to be sensitive to changes in facial angle, landmark points now serve as a precise feature extractor and location aligner for the ROI, mapping the eye area in real time to anticipate errors caused by changes in face position [7], [8], [9].

The integration of MediaPipe with more advanced Deep Learning architectures, such as DenseNet-121, offers a solution to overcome the computational load limitations of conventional CNN models. The DenseNet-121 architecture has the advantage of feature reuse, in which each layer receives additional input from all previous layers, making the model lighter while still retaining the ability to extract highly detailed visual patterns of the eyes. Research by [10] shows that the use of DenseNet-121 can achieve up to 98% accuracy in identifying eye conditions. Another study by [11] emphasizes that, in addition to other behavioral indicators such as mouth movement and changes in facial expression, eye activity is also a crucial component in the early identification of driver fatigue. By utilizing MediaPipe to localize the eye area and DenseNet-121 to classify its condition, the system can work more accurately and efficiently for real-time implementation compared to relying solely on traditional EAR calculations. The EAR system is highly dependent on the accuracy of facial landmark detection and the setting of a static fixed threshold to determine eye closure. However, the use of a single threshold is often less effective due to variations in eye shape between individuals, with previous studies showing differences in optimal values ranging from 0,18 [12] to 0,25 [13].

Several previous studies have examined various methods for detecting driver drowsiness. Research by [1] shows that CNN-based methods can detect drowsiness with a high level of accuracy through facial image analysis. The application of combined architectures such as CNN-LSTM has also been developed to improve detection robustness, as it can learn unique facial features and behavioral patterns associated with sustained drowsiness [14]. Another study of [15] shows that facial landmark-based approaches can efficiently detect various indicators of drowsiness, such as eye blinking and head movement. In addition, [2] also asserts that the computer vision approach has the advantage of being non-invasive and applicable in real time.

Although various methods have been developed, most previous studies still rely on manual calculation of geometric eye features, which are highly sensitive to changes in lighting. The dependence of the EAR method on the accuracy of facial landmark detection makes it prone to performance degradation under low-light conditions or when exposed to glare, leading to inaccurate eye-shape calculations. According to a study by [16], real-world problems such as night driving and light glare effects significantly disrupt the accuracy of the standard EAR threshold, causing large performance differences between bright and dark environments. The use of standard CNN models also entails a heavy computational load, making them less than optimal for devices with limited specifications. The DenseNet-121 architecture presents itself as a superior solution because it has a feature-reuse mechanism that allows for deeper and more accurate extraction of the eye's visual patterns while remaining computationally efficient [17]. Therefore, this study fills the gap left by previous research by analyzing the implementation of DenseNet-121 based on MediaPipe Face Mesh. In this system, MediaPipe is used specifically to precisely localize and extract the eye area, which is then processed by DenseNet-121 as the main classification model. This approach is expected to produce a more efficient drowsiness detection system capable of operating in real time.

While the combination of MediaPipe Face Mesh with CNN-based classifiers has been explored in prior work [7], [18], as well as more recently by Ersoy et al. [19], who combined MediaPipe-based landmark extraction with separate CNN classifiers for eye and mouth state achieving 99.1% eye-state accuracy, these studies primarily focus on demonstrating feasibility and reporting aggregate accuracy, without addressing two aspects that are critical for real-world deployment. First, neither study evaluates model performance using a subject-independent testing protocol, leaving open the question of how well such systems generalize to individuals not seen during training. Second, none of these studies investigates the specific architectural choices involved in the MediaPipe-to-CNN pipeline, such as the extent of ROI padding or the number of CNN layers fine-tuned, both of which directly affect the quality of the extracted eye region and the resulting classification performance. The present study addresses these gaps by (1) adopting a Subject-

Independent Leave-One-Subject-Out (LOSO) cross-validation protocol to obtain a leakage-free estimate of generalization to unseen individuals, and (2) applying DenseNet-121 with a proportional ROI padding equal to 35% of the bounding box dimensions and a two-stage fine-tuning strategy across its last 30 layers, with the resulting performance and its variation across subjects analyzed in detail in Section III.

Based on the background described above, the problem statements of this study are: how to implement a MediaPipe Face Mesh-based DenseNet-121 architecture to detect drowsiness in real time; how capable the DenseNet-121 model is at classifying images of open and closed eye conditions as indicators of drowsiness; and how the overall system performs in detecting drowsiness in terms of accuracy level and computation time. Based on these problem statements, the objectives of this study are to implement a MediaPipe Face Mesh-based DenseNet-121 architecture to detect drowsiness in real time, to analyze the capability of the DenseNet-121 model in classifying images of open and closed eye conditions as indicators of drowsiness, and to evaluate the overall system performance in detecting drowsiness based on accuracy level and computation time efficiency.

II. METHODOLOGY

A. Previous Research

Various previous studies have examined the implementation of computer vision-based methods for automatically detecting driver drowsiness. The implementation of a real-world vehicle monitoring system has been shown to reduce the traffic accident rate by up to 20% [3]. Various automatic methods have been developed by researchers, which can broadly be grouped into three main approaches: a global vision approach based on convolutional neural networks, a facial geometric feature approach, and an advanced deep architecture approach.

The first group of studies focuses on the use of pure CNNs to directly classify the driver's facial status [1], [7]. CNNs are considered highly effective because they have a natural ability to learn characteristics directly from image pixels without manual modification. Research by Gosai and Barad [6] shows that CNN models are able to extract features that can learn various changes in pose, head tilt, and the use of accessories. This characteristic is reinforced by Florez et al. [7], who implemented real-time eye-status identification by first using MediaPipe Face Mesh to localize and extract the eye region, followed by CNN-based classification (InceptionV3, VGG16, and ResNet50V2) to distinguish between open and closed eyes. A related study by the same research group extended this MediaPipe-plus-CNN pipeline by incorporating Mouth Aspect Ratio analysis for a combined eye-and-mouth drowsiness detection system on embedded hardware [18]. While both studies demonstrate the general viability of combining MediaPipe-based facial landmark extraction with CNN classification, neither systematically evaluates model generalization across previously unseen individuals, nor investigates the influence of ROI padding or the number of fine-tuned layers on classification performance, aspects that are directly addressed in the present study However, the main weakness of this group of standard CNN-based studies lies in the heavy computational load and large resource requirements, which trigger bottlenecks when run on embedded hardware with limited specifications [17].

The second group of studies shifted to a geometric feature approach, specifically using the Eye Aspect Ratio (EAR) method and facial landmarks [9], [15]. This approach analyzes drowsiness through measurable behavioral indicators such as eye blinking, mouth movement, and the driver's head tilt level. The main advantage of EAR-based methods is that the process is non-invasive and runs very quickly in real time [2]. However, a fundamental weakness of this method is its heavy reliance on a static fixed threshold value. Variations in eye shape between individuals cause the optimal threshold value to become inconsistent, as shown by a study of Dewi et al. [12] set an optimal threshold value of 0.18, whereas a study by Fathima and Girisha [13] set a different threshold value of 0,25. Kai Yuen et al. [16] also explained that this purely geometric method tends to experience a drastic drop in performance when tested under challenging environmental conditions, such as night driving with minimal lighting and exposure to glare effects from other vehicles' headlights, which disrupt the eye landmark coordinates.

Research from the third group attempts to overcome the obstacles found in the two groups above by integrating a combination of modern landmarks with more complex deep learning architectures [8], [14]. To address the issue of facial-angle sensitivity in geometric calculations, FaceMesh landmark technology has begun to be implemented, as it has been shown to reduce the effect of face orientation and maintain precise tracking of eyelid position. In terms of classification modeling, Chen et al. proposed a combined CNN-LSTM architecture that leverages bilinear features capturing both spatial and temporal (time-series) aspects of sustained drowsy behavior. Meanwhile, the use of architectures such as DenseNet-121 offers a solution for reducing computational load through a feature-reuse mechanism. Through this mechanism, each layer receives additional input from all previous layers, which makes the extraction of the eyelid's visual patterns far more detailed without increasing the number of model parameters [17]. Research by Phan et al. [10] explains that the use of DenseNet-121 in combination with an intelligent warning system can produce a very high eye-classification accuracy rate of up to 98%. A comparative mapping and the distinguishing positions of each of these previous studies are systematically summarized in Table I.

TABLE I
PREVIOUS RESEARCH

Researcher	Main Approach	Parameters / Indicators	Method Advantages	Method Weakness / Limitations
Gosai & Barad	Standard CNN	Full facial image, head position, accessories	Resistant to pose changes and facial accessories	Heavy computational load, less optimal for limited devices
Florez et al.	CNN Classification	Eye status identification (open/closed)	High accuracy in distinguishing eyelid conditions	Highly dependent on initial preprocessing data quality
Dewi et al.	Geometric EAR	Eye-opening ratio with a *fixed threshold* of 0.18	Simple algorithm and lightweight computation	Static threshold is not flexible for individual eye-shape variation
Fathima & Girisha	EAR & Landmark	Eye-opening ratio with a *fixed threshold* of 0.25	Responsive detection to changes in blinking	Produces a different optimal value, indicating threshold inconsistency
Kai Yuen et al.	*Computer Vision* EAR	Eye monitoring under varying light intensity	Able to recognize early signs of drowsiness (*microsleep*)	Accuracy drops due to night-lighting disturbances and light glare
Nomura et al.	*Facemesh* & Thermal	Facial landmark coordinates and thermal facial imagery	Able to reduce distortion effects caused by face orientation/angle	Requires additional sensor hardware, which is more expensive
Chen et al.	*Hybrid* CNN-LSTM	Fusion of spatial features and sequential temporal patterns	Robust in recognizing sustained, recurring drowsiness	Sequential architecture requires high working memory
Phan et al.	DenseNet-121 & IoT	Eye-image extraction integrated with a smart device	Very high eye-condition classification accuracy, reaching 98%	Has not yet integrated dynamic ROI alignment for a moving face

Based on the overall analysis of the advantages and limitations in Table I, this study is designed to close the existing research gap. This study proposes an integrated system that utilizes MediaPipe Face Mesh not for sensitive mathematical EAR calculations, but purely as a dynamic and precise ROI location extractor for the eye area, in order to anticipate changes in the driver's facial angle. The resulting cropped eye images are then classified using the DenseNet-121 architecture. Through this combination, the weakness of the static EAR threshold is successfully eliminated, the issue of light sensitivity can be mitigated through the deep learning model, and the computational load can be minimized as much as possible due to the efficiency of DenseNet-121's *feature reuse* mechanism.

B. Research Stages

To provide an overview of the developed drowsiness detection system, the workflow is illustrated in Figure 1. The system integrates the MediaPipe Face Mesh framework for face localization with the DenseNet-121 deep learning model for real-time drowsiness classification. A webcam captures a real-time video stream, which is divided into image frames in BGR format and converted to RGB for facial landmark detection. MediaPipe Face Mesh generates normalized 3D facial landmarks, enabling extraction of the left and right eye Regions of Interest (ROIs). Each eye is cropped separately using landmark-based bounding boxes with an additional 35% padding to preserve contextual information. An ablation study comparing single-eye extraction and combined-eye cropping, presented in Section III, demonstrates that the single-eye approach provides superior classification performance.

The extracted eye ROIs are resized to 224 × 224 pixels and normalized using the DenseNet-121 preprocessing function based on ImageNet statistics. After batch expansion, the images are fed into the DenseNet-121 model, which extracts discriminative eyelid features through dense feature reuse. The resulting features are processed by a Global Average Pooling layer and a fully connected layer, while the final Sigmoid activation outputs a probability between 0 and 1. Using a threshold of 0.5, the system classifies the driver's state as Normal (probability > 0.5) or Drowsy (probability ≤ 0.5), with the result displayed in real time. All processing stages operate continuously and efficiently, while detailed mathematical formulations and parameter configurations are described in the following subsection.

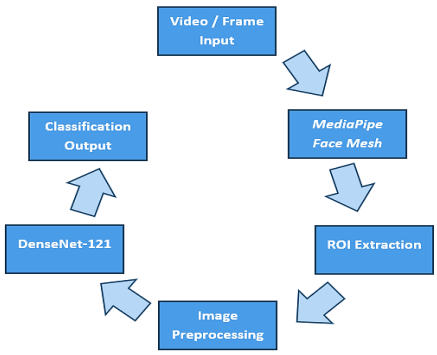

Figure 1. Research Stages

C. Drowsiness Image Dataset

Drowsiness is a state of reduced consciousness and alertness that may lead to microsleep, a brief loss of awareness lasting 1–30 seconds, during which drivers fail to

respond to visual and motor stimuli, significantly increasing accident risk [2]. Among various behavioral indicators, eye activity has the strongest correlation with early drowsiness detection [11]. In computer vision, drowsiness is identified through changes in eyelid geometry and eye texture, where open eyes indicate alertness, while partially or fully closed eyes reflect increasing fatigue. Deep learning models exploit these visual patterns without requiring manual eyelid boundary detection.

To evaluate the proposed approach, a custom dataset was collected using Google's Teachable Machine with a standard webcam, where all images were automatically resized to 224 × 224 pixels. The dataset consisted of 2,000 images (1,000 alert and 1,000 drowsy) captured from five subjects under consistent indoor lighting. Images were manually labeled into Normal and Drowsy classes, with the latter including partially and fully closed eyes representing early microsleep. Using MediaPipe Face Mesh, the left and right eye regions were extracted independently, producing up to two samples per image. After removing duplicate entries and filtering invalid eye detections using the Eye Aspect Ratio (EAR) range of 0.0–0.45, the final dataset contained 3,926 eye-region samples. To ensure robust and leakage-free evaluation despite the limited number of subjects, the model was assessed using Subject-Independent Leave-One-Subject-Out (LOSO) cross-validation, where one subject was reserved for testing while the remaining four were used for training in each of the five folds.

D. Object Detection and Region of Interest (ROI) Extraction

Object detection is one of the main pillars in the field of computer vision, combining two computational tasks simultaneously: image classification to identify an object's category, and spatial localization to determine the object's position within the image using a bounding box. In the context of a driving monitoring system, object detection is specifically applied as a face detection task. This process functions to recognize and localize the presence of a human face area amid a dynamic and complex background environment. Reliable face detection is a highly crucial foundational stage before the system can carry out a more in-depth analysis of the more specific facial components.

After the face has been successfully identified within the image space, the next stage that determines the system's efficiency is Region of Interest (ROI) extraction. The concept of ROI extraction is implemented by isolating and cropping a specific spatial segment from the full facial image, in this case focused on the driver's eye area, to be used as the main input for the classification model [9]. The use of facial landmarks for region-of-interest isolation has been proven effective in various facial expression and condition recognition tasks, including under partially occluded facial conditions [20]. A non-invasive, camera-based visual approach is also a top choice in real-time drowsiness detection systems because it does not burden the driver with additional devices [21]. There re several theoretical reasons why eye-area ROI extraction is a highly crucial approach in the architecture of a drowsiness detection system. First, specific cropping is useful for eliminating irrelevant background information, such as variations in hair shape, cabin shadows, or the movement of other facial organs that could obscure the main model's learning focus. Second, reducing the input matrix dimensions from the full facial image down to just the eyelid area makes the computational load of the neural network much lighter when processing pixels. This can minimize latency values and optimize the system's response speed to meet real-time processing requirements [2]. Third, feature extraction focused exclusively on changes in the eyelid area will minimize the risk of overfitting, allowing the model to become more sensitive in recognizing subtle visual patterns of both open and closed eye conditions [7].

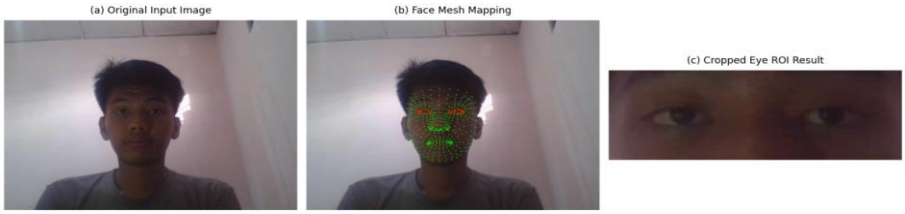

Figure 2. Visualization stages of input image processing: (a) original input image from the webcam, (b) mapping of the facial landmark coordinate mesh matrix via MediaPipe Face Mesh, and (c) the resulting isolated ROI eye-area cropping segmentation with padding space expansion

To perform adaptive localization of the eyelid's coordinate points, facial-landmark technology based on MediaPipe Face Mesh is used as the extraction instrument. MediaPipe Face Mesh is a modern machine learning architecture capable of estimating human facial geometry through 468 three-dimensional (3D) landmark points quickly and efficiently [8], [9]. This algorithm works through a two-stage pipeline approach, beginning with the use of a lightweight face detector that localizes the outermost boundary of the face, followed by a landmark regression model that maps the entire surface coordinates of the facial topology in detail. The main advantage of MediaPipe Face Mesh over traditional geometric feature methods is its high resistance to distortion effects caused by changes in the driver's head orientation and tilt angle [8], [15]. By utilizing specific landmark coordinate indices on the edges of the upper and lower eyelids as well as the internal and external eye corners, the system can determine the eye ROI cropping boundaries dynamically and precisely. This mechanism ensures that the ROI location alignment process remains consistent in anticipating errors

caused by changes in the driver's facial position while driving. The visual implementation of the face acquisition stage, the facial landmark point mesh mapping, up to the final result of isolating the driver's eye ROI image, is illustrated in Figure 2.

E. Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is one of the architectures within the deep learning family specifically designed to process, analyze, and recognize spatial patterns in two-dimensional matrix-shaped data such as digital images. CNNs have a key advantage in automatically learning a hierarchy of visual features, starting from low-level patterns such as edges and lines up to more complex high-level patterns. In drowsiness detection implementations, this method is capable of achieving a high level of accuracy in identifying drowsy conditions through direct facial image analysis [1]. This is because the CNN architecture can directly learn various head positions, face orientations, head tilt levels, and even the use of accessories such as glasses, by extracting pose-invariant features, thereby significantly improving detection capability.

The first and most fundamental layer in a CNN is the Convolution Layer, which in this study is applied to extract local visual features from the driver's face. Through the periodic sliding of the filter matrix, the network can automatically capture important characteristics such as the shape of the eyelid edges, the curvature of the under-eye area, and the eyebrow lines. The two-dimensional convolution operation between the input image and this filter matrix is systematically defined in Equation (1). The variables in Equation (1) are described as follows: *S(i,j)* is the resulting pixel value of the convolution at coordinates (i, j), I represents the input image matrix, K represents the filter matrix, while m and n denote the spatial dimensions of the filter used. After local feature values are extracted through the convolution operation, the resulting feature map is passed through an activation function to introduce non-linearity into the neural network. Without an activation function, the neural network would only act as an ordinary linear regression model incapable of learning the complex visual patterns of the eyelids. The activation function commonly implemented in CNN hidden layers is the Rectified Linear Unit (ReLU), which is applied after each convolution operation to eliminate negative pixel values resulting from face extraction, setting them to zero. This is important for accelerating the model training process so that the neural network can quickly recognize the eyelid's visual patterns without experiencing vanishing gradients. The ReLU activation function is systematically expressed in Equation (2). Based on Equation (2), the variable x is the input pixel value from the convolution result, while f(x) is the output value after passing through the ReLU activation, where negative values are eliminated to zero and positive values are retained linearly.

The next step after non-linear activation is spatial-dimension reduction using a Pooling Layer. This process is a crucial part of controlling the number of computational parameters within the network, saving working memory, and reducing the risk of overfitting. The most commonly applied pooling technique is Max Pooling, used to reduce the spatial dimensions of the driver's eye-area feature map. This reduction in matrix size is carried out to save the system's computational memory while still retaining essential information, such as the degree of eye-opening width, by taking the highest pixel intensity. The local Max Pooling operation on sub-matrix R is formulated in Equation (3). In Equation (3), $P_{i,j}$ represents the reduced pixel value at the pooling layer, while $X_{m,n}$represents the original pixel value in the feature map that falls within the local region R. At the end of the network architecture, the reduced spatial feature map goes through a flattening process into a one-dimensional vector to be passed on to a Fully Connected Layer or Dense Layer. This layer has a conventional artificial-neural-network structure in which every neuron is fully connected to all neurons in the previous layer, mapping high-level spatial features into a classification decision. Since drowsiness classification in this study is binary drowsiness and wakefulness, the neuron in the final *Dense* layer is passed through a Sigmoid activation function to produce a final decision probability within a range of 0 to 1. The Sigmoid activation function is mathematically expressed in Equation (4). In Equation (4), $\sigma(z)$represents the output probability value of the binary classification, *z* represents the total input value received by the neuron in the output layer, and *e* is Euler's constant, equal to 2.718. A probability value close to 1 indicates that the model predicts the image belongs to the drowsy-eye class, while a value close to 0 indicates a normal or awake eye condition [1], [7].

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I\,(i-m, j-n)\cdot K(m,n) \quad (1)$$

$$f(x) = \max(0, x) \quad (2)$$

$$P_{i,j} = \max_{(m,n)\in R} X_{m,n} \quad (3)$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (4)$$

F. DenseNet-121 Architecture

Densely Connected Convolutional Networks, or DenseNet-121, is one of the modern deep neural network architectures that optimizes the flow of information and gradients between layers through direct, feed-forward connections from each layer to the next. The integration of this architecture serves as a superior solution for overcoming the computational load limitations of conventional CNN models. The competitive advantage of DenseNet-121 lies in its feature-reuse mechanism, in which each layer receives additional input from all previous layers. Unlike residual architectures, which combine feature maps through a mathematical summation operation, DenseNet-121 combines features using a matrix concatenation operation. This mechanism minimizes the need to relearn redundant features, making the model far lighter and more parameter-efficient, while still retaining a very deep and detailed capability for extracting the eyelid's visual

patterns [17]. The feature-reuse mechanism in DenseNet enables better parameter efficiency compared to conventional CNN architectures, since each layer receives input from all previous layers [22].

The main structural components of DenseNet-121 are divided into two important elements: the *Dense Block* and the *Transition Layer*. Within a *Dense Block*, all layers are densely connected to one another. Systematically, if a *Dense Block* has a total of *l* layers, then the *l*-th layer will receive feature maps from all previous layers as a single, integrated input. This dense connectivity relationship is formulated in detail in Equation (5). Based on Equation (5), the notation $[x_0, x_1, x_2, \dots, x_{l-1}]$ represents the concatenation operation of all feature maps produced from layer 0 up to the layer before *l*. Meanwhile, $H_l(\cdot)$denotes a nonlinear composite function consisting of a sequence of three processing operations: *Batch Normalization* (BN), followed by the *Rectified Linear Unit* (ReLU) activation function, and ending with a convolution operation. Each composite function $H_l(\cdot)$is configured to produce *k* new feature maps, where the variable *k* is defined as the architecture's growth rate. The existence of this constant growth-rate parameter ensures that the addition of channel dimensions to the input matrix remains linearly controlled, thereby preventing data-dimension bloat as the network grows deeper.

The second important component is the *Transition Layer*, which is specifically placed between two adjacent *Dense Blocks*. Since the concatenation operation in Equation (5) requires matrix spatial dimensions to match, the *Transition Layer* plays an important role in performing downsampling through a combination of a 1x1 convolution operation and 2x2 *Average Pooling*. In addition to reducing the image's spatial resolution, this layer also functions to compress the number of feature channels in order to maintain system memory efficiency. If a *Dense Block* outputs *m* feature map channels, the *Transition Layer* will reduce this number of channels to a new output value determined by a compression factor (θ), as expressed in Equation (6). In Equation (6), m_{out}represents the total number of feature map channels after compression, m_{in}denotes the total number of input feature channels before passing through the transition layer, while θrepresents the compression factor with a value range of $0 < \theta \leq 1$. In the standard default DenseNet-121 architecture, the compression factor value is set to $\theta = 0.5$, meaning the transition layer is able to cut the feature channel depth dimension in half to save on image computation processing load.

$$x_l = H_l([x_0, x_1, x_2, \dots, x_{l-1}]) \quad (5)$$

$$m_{out} = \lfloor \theta \cdot m_{in} \rfloor \quad (6)$$

In this drowsiness detection system, the parameter-efficiency characteristics of the Dense Block and Transition Layer structures are leveraged through a Transfer Learning approach using model weights pre-trained on the ImageNet dataset [17]. The fine-tuning strategy is implemented in two sequential stages. In the first stage, the entire DenseNet-121 backbone is kept frozen, and only the newly added classification head is trained, allowing the randomly initialized head weights to stabilize before any adjustment is made to the pretrained backbone. In the second stage, the final 30 layers of DenseNet-121 are unfrozen for dynamic retraining, using a substantially reduced learning rate to preserve the previously learned ImageNet representations while allowing adaptation to the specific characteristics of human eyelid images. This two stage strategy aims to prevent the large gradient updates from an untrained classification head from degrading the pretrained feature representations, a risk that is particularly relevant given the limited size of the available training data. An architectural ablation comparing this two stage fine-tuning strategy against a single stage strategy that unfreezes the final 50 layers from the beginning of training, following an alternative configuration suggested during the research advisory process, is presented in Section III, with both strategies found to yield statistically comparable performance. This modification aims to allow the model to adaptively tailor its high-level visual feature extraction capability to variations in the characteristics of human eyelid images, resulting in sharp classification performance that remains responsive when running on in-vehicle computing hardware in real time [10], [17]. A fine-tuning strategy that unfreezes a number of DenseNet-121's final layers has been shown to significantly improve classification accuracy across various medical imaging domains [23]. The use of pretrained ImageNet weights through a transfer learning approach has also been applied in similar architectures for open- and closed-eye classification [24].

G. MediaPipe Face Mesh

The method proposed in this study is designed as an integrated framework that combines the strength of high-level facial geometry tracking from *MediaPipe Face Mesh* with the sharp spatial classification capability of the *DenseNet-121* architecture. The reliability of MediaPipe in extracting body and facial landmarks has also been validated outside the context of drowsiness detection, for instance in predicting aggressive behavior in individuals with dementia, demonstrating the generalizability of this framework across various cases [25]. This approach is theoretically proposed to address the fundamental weakness of the conventional *Eye Aspect Ratio* (EAR) method. In the traditional method, eye status is determined by calculating the geometric distance ratio of the eyelids using the standard formula stated in Equation (7). Based on Equation (7), p_1 through p_6 are the two-dimensional (2D) spatial coordinates of the eyelid landmark points. Although computationally lightweight, the EAR method has a high theoretical vulnerability because it is heavily dependent on a static fixed threshold value. Variations in biological eye shape between individuals cause the optimal threshold value to become inconsistent, such as the threshold differences between 0.18 [12] and 0.25 [13] found in previous studies. Because its calculation is based on linear distance,

changes in the driver's facial angle orientation (*head pose*) and reduced light intensity at night can easily distort the accuracy of landmark-coordinate detection, ultimately leading to failures in the ratio-value calculation [16].

$$\text{EAR} = \frac{|p_2 - p_6| + |p_3 - p_5|}{2|p_1 - p_4|} \quad (7)$$

To eliminate the dependence on a static threshold and the aforementioned geometric distortion, the proposed method shifts the decision-making burden away from mathematical eyelid-distance calculations toward deep visual pixel-texture analysis using a neural network. The concept flow of this system, illustrated previously in Figure 1, follows this same block diagram pattern: starting from raw video frame capture (input frame), facial landmark extraction, Region of Interest cropping, image normalization, and finally determination of the driver's drowsiness status probability.

Mathematically, the first stage of the proposed method begins by extracting the entire facial surface topology from the input image (I_{input}) using the landmark regression function from MediaPipe Face Mesh to generate a set of 468 three-dimensional (3D) coordinate points, as formulated in Equation (8). In Equation (8), L represents the set of all facial landmark points, where each element $l_i$contains spatial coordinate information (*x, y, z*). Through these 3D coordinates, the effect of changes in the driver's facial angle can be reduced, since the landmarks are resistant to spatial rotation [8]. The second stage is to perform dynamic isolation of the eye area. The system takes a sub-set of landmark coordinates that specifically map the edges of the left and right eyelids ($L_{\text{mata}} \subset L$). From these points, the system determines the extreme bounding box boundaries for cropping, followed by dimension alignment and pixel-value scaling. The process of forming this normalized input image (I_{norm}) is formulated in Equation (9). The operation in Equation (9) produces two isolated eye-area image matrices, one per eye, each with a uniform resolution of 224 by 224 pixels in RGB color channels, with pixel intensity values normalized according to the DenseNet-121 specific preprocessing function. This ROI preprocessing stage theoretically functions as an external-environment noise reducer while also reducing the data dimension load.

The final stage involves mapping high-level spatial features using the binary classification function of the *Transfer Learning DenseNet-121* model (f_{DenseNet}). The matrix $I_{\text{norm}}$is passed into the neural network, which utilizes the *feature-reuse* mechanism to extract subtle visual characteristics of the eyelid, such as skin wrinkles, the degree of eye-opening width, and the drooping shadow of drowsy eyes, regardless of an individual's biological eye size [17]. The final decision, in the form of a drowsiness probability value (*y*), is calculated based on Equation (10).

$$L = f_{\text{MediaPipe}}(I_{\text{input}}), \quad L = \{l_1, l_2, \dots, l_{468}\} \quad (8)$$

$$I_{\text{norm}} = \text{Preprocess}\left(\text{Resize}\left(\text{Crop}(I_{\text{input}}, L_{\text{mata}}), (224,224)\right)\right) \quad (9)$$

$$y = f_{\text{DenseNet}}(I_{\text{norm}}), \quad y \in [0,1] \quad (10)$$

The resulting probability value *y* is then evaluated using a binary decision threshold of $r = 0.5$. If $y \geq 0.5$, the system automatically classifies the driver into the drowsiness class, indicating symptoms of fatigue or microsleep, whereas a value of $y < 0.5$ classifies the driver as being in an alert or *wakefulness* condition [2]. Through this theoretical integration, the proposed method is designed to deliver a drowsiness detection system that is adaptive to variations in each driver's physical characteristics and computationally efficient for real-time implementation [10], [17]. Robustness to varying ambient lighting conditions was not explicitly evaluated in this study and is therefore not claimed; live deployment testing did, however, reveal a related sensitivity to rapid head movement, discussed further in Section III.I, which was partially mitigated through a temporal smoothing mechanism.

H. Confusion Matrix

A *Confusion Matrix* is a quantitative evaluation component used to measure the performance of a classification model by mapping the neural network's prediction results against the actual ground-truth class labels. This evaluation is presented as a two-dimensional *N* x *N* matrix table, where *N* represents the number of categories or classification target classes being tested. Through this measurement tool, the model's classification performance is not only assessed based on the overall accuracy value, but also analyzed in depth to detect the types of prediction errors made by the model during the testing stage. In the binary drowsiness classification scenario, this matrix table is constructed based on the cross-combination of four fundamental parameters. In the context of the proposed system, the *True Positive* (TP) parameter describes the network's success rate in correctly identifying an actual drowsy eye image into the drowsy class. Conversely, True Negative (TN) describes the situation in which an actual driver eye image in a normal or alert condition is correctly classified into the not-drowsy class. Meanwhile, False Positive (FP) indicates a false-alarm prediction error, i.e., when a driver's eye image that is actually normal or not drowsy is incorrectly classified by the system into the drowsy class. Conversely, False Negative (FN) is the most critical standard to evaluate in terms of driving safety, because this condition occurs when an actual driver eye image indicating severe drowsiness or microsleep is incorrectly detected by the model as a normal or not-drowsy condition.

Based on the collected values of the four parameters above, the reliability of the classification model is assessed in detail using four main performance metrics: accuracy, precision, recall (sensitivity), and F1-Score. The first metric is accuracy, which describes the overall ratio of correct predictions to the total number of test data samples. The accuracy value calculation is defined in Equation (11). Where TP (*True Positive*) is the number of drowsy eye images correctly

classified as drowsy, TN (*True Negative*) is the number of not-drowsy eye images correctly classified as not drowsy, FP (*False Positive*) is the number of not-drowsy eye images incorrectly classified as drowsy, and FN (*False Negative*) is the number of drowsy eye images incorrectly classified as not drowsy.

The second metric is precision, which measures the level of accuracy between the number of samples correctly predicted as positive and the total number of samples declared positive by the model. This metric is very important for minimizing false alarms that could disturb the driver's concentration. The precision value is calculated using Equation (12). The third metric is sensitivity (*recall*), which measures the model's ability to retrieve and identify all samples belonging to the actual positive class category. This sensitivity value is formulated in Equation (13). The fourth metric is the *F1-Score*, which represents the harmonic mean of the precision and sensitivity components. This metric is a highly reliable indicator for testing the balance of a classification model's performance, especially when facing an imbalanced data distribution. The formula for calculating the *F1-Score* is stated in Equation (14). Through the use of the evaluation formulas above, the effectiveness of the network architecture in determining the driver's eyelid image can be objectively assessed, so that the validity of the classification system's test results can be scientifically justified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (13)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (14)$$

III. Results And Discussion

A. Model Training and Validation Results

Given the adoption of the Subject-Independent LOSO protocol described in Section II.C, five separate DenseNet-121 models were independently trained and validated, one for each fold. In every fold, the model was initialized from ImageNet pre-trained weights and trained in two sequential stages as described in Section II.F: first with the entire backbone frozen and only the classification head trained (Adam optimizer, learning rate = 1×10^{-3}, up to 15 epochs), followed by fine-tuning of the last 30 layers of the backbone (Adam optimizer, learning rate = 1×10^{-5}, up to 20 epochs). An early stopping mechanism (monitoring training loss, patience of 5 epochs, restoring the best-performing weights) was applied at each stage to prevent overfitting on the training subjects. Across all five folds, training consistently converged before reaching the maximum epoch limit, indicating stable learning behavior under the fine-tuning configuration. The detailed per-fold performance resulting from this training protocol is presented in Section III.

B. Classification Performance Evaluation

Following the Subject-Independent LOSO protocol, model performance was evaluated separately on each of the five held-out test subjects, none of whom contributed any images to the corresponding fold's training or validation data. Table II summarizes the resulting accuracy, macro-averaged precision, recall, F1-score, and AUC for each fold, along with the mean and standard deviation across all five folds.

TABLE II
LOSO Cross-Validation Performance Per Fold

Fold (Test Subject)	n (test)	*Accuracy*	Precision	Recall	F1-Score	AUC
1 (Subject 0)	762	91.86%	86.23%	98.62%	92.01%	0.950
2 (Subject 1)	799	69.59%	75.16%	58.40%	65.73%	0.707
3 (Subject 2)	800	85.25%	92.73%	76.50%	83.84%	0.967
4 (Subject 3)	778	70.05%	72.75%	66.75%	69.62%	0.769
5 (Subject 4)	787	99.36%	98.77%	100.00%	99.38%	0.999
Mean (SD)	3,926	83.22% (13.22)	85.13% (11.37)	80.05% (18.72)	82.11% (14.35)	0.879 (0.131)

Across the five folds, the model achieved a mean accuracy of 83.22%, with a standard deviation of 13.22 percentage points, and a mean AUC of 0.879, with a standard deviation of 0.131. These values are substantially lower and more variable than the near-perfect accuracy of 97.21% obtained when the identical modeling approach was evaluated using a random, subject-mixed data split, as detailed in Section III.E. This performance drop reflects a genuine measure of the model's ability to generalize to previously unseen individuals, rather than an artifact of memorized subject-specific visual features. Notably, Subjects 1 and 3 consistently exhibited the lowest performance across all evaluation metrics, including the lowest AUC values of 0.707 and 0.769 respectively, a pattern further examined in the following paragraph through direct comparison of their Eye Aspect Ratio distributions and visual inspection of misclassified samples. To verify that this aggregate estimate was not an artifact of a single stochastic training run, the LOSO procedure was independently repeated. The resulting mean accuracy of 84.20% was closely comparable to the primary run reported in Table II, despite per-fold variation attributable to random weight initialization, indicating that the aggregate five-fold estimate is reasonably stable and reproducible.

Substantial variation in per-fold performance was observed, ranging from 69.59% (Subject 1) to 99.36% (Subject 4) in accuracy. This variation was further investigated by computing the Eye Aspect Ratio (EAR) separately for each subject's drowsy and alert images,

followed by direct visual inspection of misclassified samples. Subjects 1 and 3, whose folds yielded the two lowest accuracies (69.59% and 70.05%, respectively), exhibited the smallest mean EAR differences between drowsy and alert conditions among all five subjects, with overlapping distributions of 0.0684 and 0.0141 respectively, compared to no measurable overlap for Subjects 0, 2, and 4. This indicates that these two subjects' drowsy and alert expressions were physically the least distinguishable in terms of eye closure, plausibly due to a partial rather than complete eyelid closure when simulating the drowsiness condition. Visual inspection of false negative and false positive samples for these two subjects further revealed that many misclassified crops lacked clearly discernible eyelid structure, consistent with the constrained native resolution of the extracted eye regions discussed in Section II.C. In contrast to occasional threshold miscalibration patterns reported elsewhere in the literature, Subject 4 in this study achieved both the highest AUC (0.999) and the highest accuracy (99.36%), indicating consistent alignment between ranking capability and threshold based classification for this subject. Taken together, these findings suggest that the observed performance variance is attributable to a combination of genuine inter-subject differences in eye-closure expressiveness and the constrained spatial resolution of the source imagery, rather than to limitations of the classification model or threshold calibration.

For context, prior studies applying deep learning and transformer-based architectures to eye-state or drowsiness classification have reported accuracies in the 96--100% range. For example, Hassan et al. reported open-eye and closed-eye classification accuracies of 99.15% and 99.03% using Vision Transformer and Swin Transformer architectures, respectively, and 98.65% using a DenseNet169 variant [26]. The considerably lower accuracy obtained in the present study (83.22%) should be interpreted in light of the evaluation protocol rather than as a direct indicator of inferior model capability. Many existing studies, including the aforementioned works, evaluate performance using data splitting strategies that do not explicitly enforce subject independence between training and testing sets, a practice that risks subject-level information leakage similar to that identified in this study's preliminary, non-subject-independent evaluation. The Leave-One-Subject-Out protocol adopted here provides a stricter and arguably more realistic estimate of generalization to entirely unseen individuals. This suggests that reported accuracies in the broader drowsiness detection literature may, in some cases, be optimistic relative to real-world deployment conditions, and highlights subject-independent evaluation as an important methodological consideration for future work in this field.

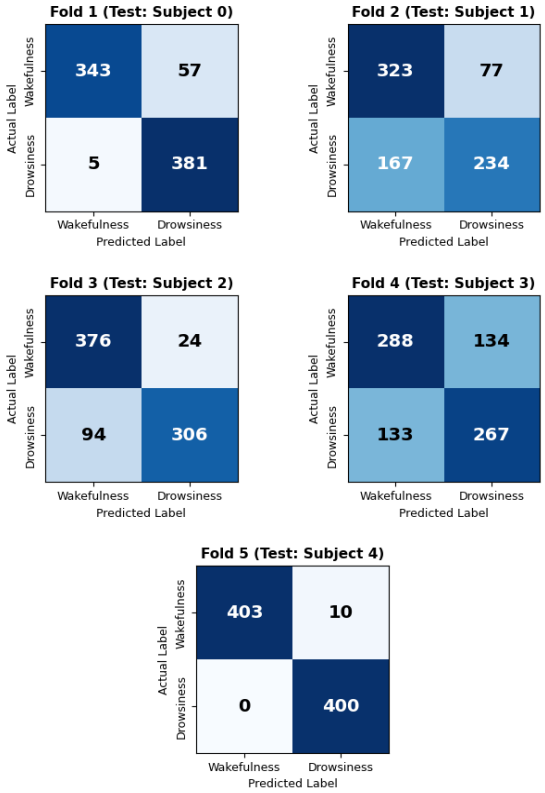

Figure 4. Confusion matrices per LOSO fold

C. ROC Curve and AUC Value Analysis

Figure 5 presents the AUC obtained for each of the five LOSO folds. The AUC values ranged from 0.707 (Subject 1) to 0.999 (Subject 4), with a mean of 0.879 and a standard deviation of 0.131 across folds. This indicates that, on average, the model retains a strong ability to rank drowsy and alert samples correctly, but that this ranking ability varies considerably depending on the individual subject held out for testing. Three of the five folds achieved an AUC above 0.90, suggesting that the model generalizes well to most previously unseen individuals; the two exceptions, Subject 1 (AUC = 0.707) and Subject 3 (AUC = 0.769), performed moderately above the chance level (AUC = 0.5), consistent with the EAR distribution overlap identified for these subjects in Section III.B, which indicates that these subjects' drowsy and alert expressions were inherently more difficult to distinguish, even independent of the classification model used.

Unlike patterns occasionally reported in the literature where AUC and accuracy diverge for a given subject due to threshold miscalibration, the present results exhibit consistent alignment between the two metrics across all five folds: subjects with lower AUC values (Subjects 1 and 3) also exhibited the lowest classification accuracy, while the subject with the highest AUC (Subject 4, AUC = 0.999) also achieved the highest classification accuracy (99.36%). This consistency suggests that, for the present dataset and model configuration, the default 0.5 decision threshold was reasonably well calibrated across subjects, and that the

observed performance variation is more plausibly attributable to genuine inter-subject differences in eye-closure expressiveness, as discussed in Section III.B, rather than to threshold-related miscalibration.

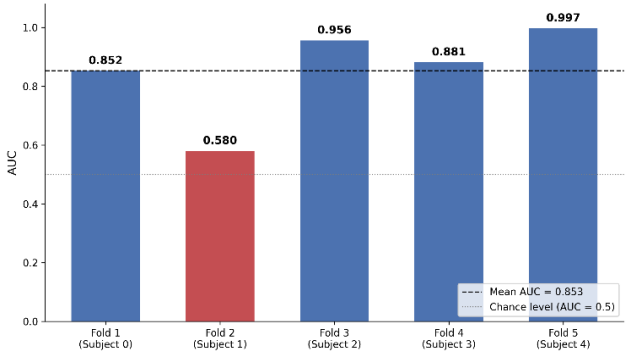

Figure 5. AUC per LOSO fold

D. Real-Time Computational Detection Speed Analysis

The computational speed of the proposed system was benchmarked on 100 randomly sampled images from the dataset, using a Google Colab environment equipped with an NVIDIA T4 GPU. The evaluation metrics were divided into three main processing stages to identify computational bottlenecks: the image preprocessing stage, the Region of Interest (ROI) extraction of the eye area using MediaPipe Face Mesh, and the status classification inference stage using the DenseNet-121 model. The total overall latency per frame is calculated based on the accumulated time of these three stages via Equation (15).

$$T_{\text{total}} = T_{\text{preprocess}} + T_{\text{mediapipe}} + T_{\text{inference}} \quad (15)$$

As detailed in Table III, the system achieved an average total latency of 198.00 ms per frame, corresponding to an average processing speed of 5.1 FPS, measured on a CPU-only consumer laptop (Intel Core i7-7820HQ, 4 cores/8 threads, 2.90 GHz, 32 GB RAM, no dedicated GPU). Among the three processing stages, DenseNet-121 classification accounted for the overwhelming majority of the total latency (98.02%), while MediaPipe-based ROI extraction and image preprocessing contributed comparatively negligible portions (0.94% and 0.89%, respectively).

The dominant contribution of the classification stage to overall latency is primarily attributable to the absence of dedicated GPU acceleration on the evaluation device, combined with the per-image invocation pattern used in this implementation, in which each frame is processed individually through the model's prediction function rather than in batches. This is a known characteristic of CPU-bound deep learning inference, where the lack of parallel hardware acceleration substantially increases per-call computation time relative to GPU-based execution. Other real-time embedded drowsiness detection systems, such as the CNN-based eye and mouth analysis approach combined with Mouth Aspect Ratio reported by Florez et al., have demonstrated that substantially lower latency is achievable on resource-constrained devices through dedicated embedded optimization [18]. This suggests that, while a processing speed of 5.1 FPS remains marginally adequate for the intended drowsiness monitoring application given that microsleep episodes are typically sustained over 1 to 30 seconds, further optimization, such as batched or asynchronous inference, model quantization, GPU acceleration, or conversion to a lightweight inference format (e.g., TensorFlow Lite), represents a promising and increasingly necessary direction to bring the system's responsiveness closer to real-time requirements on consumer-grade hardware.

TABLE III
COMPUTATIONAL LATENCY AND SYSTEM FRAME RATE

Image Processing Stage	Average Computation Time (ms)	Latency Contribution (%)
Image Preprocessing (Resize & Normalize)	1.86 ms	0.94%
MediaPipe Face Mesh & ROI Extraction	2.06 ms	1.04%
Status Classification Inference (DenseNet-121)	194.08 ms	98.02%
Total Latency per Frame	198.00 ms	100.00%
Average Processing Speed (FPS)	5.1 FPS	

Figure 6. Sample outputs from the real-time drowsiness detection system, correctly classifying (top) an open-eye condition as "Wakefulness" with a smoothed probability of 0.18, and (bottom) a closed-eye condition as "Drowsiness" with a smoothed probability of 0.71.

Figure 6 presents two representative frames captured from the live system, illustrating correct classification of both eye conditions, with on-screen frame rates of 4.3 FPS for the wakefulness example and 3.9 FPS for the drowsiness example, broadly consistent with the aggregate benchmark of 5.1 FPS reported in Table III. The displayed classification for each frame reflects the smoothed prediction described in Section III.I, rather than a single-frame raw prediction. It should be noted that this qualitative demonstration illustrates correct system operation on two representative instances and does not, by itself, constitute a generalization claim; the quantitative generalization performance of the system across unseen individuals is separately established through the Subject-Independent LOSO evaluation reported in Section III.B.

E. Empirical Validation of Subject-Level Leakage

To empirically substantiate the risk of data leakage associated with random image-level splitting, as raised during the peer review process, the identical DenseNet-121 architecture and training protocol, adapted with a single-stage fine-tuning strategy on the last 50 layers following an alternative configuration suggested during the research advisory process, was evaluated under two splitting configurations. Under a random 70/20/10 split without subject-level separation, the model achieved a test accuracy of 97.21%. Under the subject-independent LOSO protocol, the same architecture achieved a substantially lower average accuracy of 84.01% (standard deviation 12.58 percentage points), a gap of approximately 13.2 percentage points. This result provides direct empirical confirmation that random splitting allows the model to exploit subject-specific visual characteristics rather than learning generalizable drowsiness indicators, corroborating the concerns raised during peer review regarding potential data leakage.

F. Ablation Study: Single-Eye versus Combined Dual-Eye ROI Extraction

An architectural ablation was conducted to evaluate an alternative ROI extraction strategy in which the left and right eye regions were merged into a single combined crop, rather than extracted and classified independently as described in Section II.B. Under this combined-crop configuration, the identical LOSO protocol yielded a mean accuracy of 72.30% (standard deviation 16.00 percentage points), a substantial decrease from the 83.22% achieved using independent single-eye extraction. This performance degradation is attributable to three compounding factors: a reduction in the effective number of training samples, as each source image contributes one combined sample rather than two independent samples; the inclusion of visually non-informative regions, such as the nasal bridge, within the combined bounding box; and geometric distortion introduced when resizing the resulting wide, non-square crop to the model's square input dimensions. These findings support the retention of independent single-eye extraction as the preferred configuration for this dataset.

G. Cross-Dataset Generalization Evaluation

To assess the model's generalization beyond the visual characteristics of its training domain, the five LOSO-trained models were evaluated on an independent subset of the MRL Eye Dataset, a publicly available infrared-based eye-state dataset comprising 37 subjects. A stratified subset of 1,405 images, balanced across subjects and eye states, was used for evaluation. Across all five folds, the models achieved a mean accuracy of 47.80% (standard deviation 0.06 percentage points), with the predicted probability distribution collapsing to a near-constant value close to zero (mean = 0.0028, standard deviation = 0.0303) regardless of the true label, a pattern equivalent to majority-class prediction rather than meaningful discrimination. This severe degradation is attributable to the substantial domain shift between the RGB, visible-light training data and the infrared-based evaluation data, which differs considerably in color distribution, contrast characteristics, and illumination properties. This result indicates that the model's learned representations remain highly specific to the visual domain represented in the training data and do not transfer to a structurally different sensor modality, underscoring the importance of evaluating generalization across visually diverse conditions rather than relying solely on within-domain subject-independent evaluation.

H. Comparison with Prior Studies

Table IV compares the subject-independent accuracy achieved in this study against results reported in prior drowsiness and eye-state detection literature that similarly employed evaluation protocols emphasizing generalization to unseen individuals. While direct comparison across studies should be interpreted cautiously given differences in modality, dataset composition, and task formulation, particularly for the EEG-based comparison, the present study's accuracy compares favorably against these subject-independent or evaluation-rigorous baselines, suggesting that the proposed approach is competitive within this stricter evaluation paradigm.

TABLE IV
COMPARISON WITH PRIOR STUDIES

Study	Method	Dataset	Accuracy
Park et al. (2016)	AlexNet + VGG-FaceNet + FlowImageNet ensemble	NTHU-DDD	73.06%
Dwivedi et al. (2014)	CNN representation learning	Custom dataset	78.00%
Subject-Independent CNN-LSTM (2021)	CNN-LSTM, LOSO (EEG-based)	Public EEG dataset	72.97%
This study	MediaPipe + DenseNet-121, LOSO	Custom dataset	83.22%

I. Real-Time Deployment Observations and Temporal Smoothing

During live deployment testing, an additional practical limitation was observed that had not been apparent from the static image evaluation reported in Table II. Specifically, single-frame predictions exhibited noticeable instability during rapid head movement, occasionally producing incorrect momentary classifications despite the subject maintaining a consistent wakefulness or drowsiness state. This behavior is consistent with the model having been trained exclusively on static, unblurred images, and suggests that motion blur and transient MediaPipe landmark detection lag represent a meaningful real-world challenge not captured by the LOSO evaluation protocol. A temporal smoothing mechanism, in which the displayed classification is derived

from a rolling average of the raw probability output across the seven most recent frames, was found to substantially mitigate this instability during informal testing, as illustrated in Figure 6. A more rigorous quantitative evaluation of this mitigation, for instance by systematically comparing raw versus smoothed prediction accuracy on video recordings involving controlled head movement, is recommended as future work.

J. Comparison with EAR-Based Baseline and a Lightweight Architecture

To directly address the comparison requested during peer review, an EAR-based classifier was evaluated under the identical LOSO protocol, with the decision threshold re-optimized independently for each fold using only the four training subjects, ensuring the comparison remained fully subject-independent for both methods. Under this evaluation, the EAR baseline achieved a mean accuracy of 84.37% (standard deviation 15.11 percentage points) and a mean AUC of 0.933 (standard deviation 0.087), marginally exceeding the DenseNet-121 model's accuracy of 83.22% and AUC of 0.879. This result should be interpreted carefully rather than as a refutation of the motivation for a deep learning approach. First, the EAR calculation operates directly on MediaPipe's sub-pixel landmark coordinates and is therefore unaffected by the constrained native pixel resolution of the extracted eye-region images discussed in Section II.C, a limitation that plausibly constrains DenseNet-121's ability to learn fine-grained textural features. Second, the present dataset was collected under a single, controlled indoor lighting condition, meaning the lighting- and pose-sensitivity limitations of EAR motivated in Section I were not actively exercised during this evaluation. The optimal EAR threshold also varied noticeably across folds (0.25 to 0.27), consistent with the inter-individual threshold inconsistency reported in prior literature, even though this variation did not, in this particular dataset, translate into a decisive accuracy disadvantage relative to the deep learning approach.

To additionally address the suitability of the proposed system for resource-constrained edge devices, DenseNet-121 was compared against MobileNetV3-S, a substantially more compact architecture purpose-built for mobile and embedded deployment. Under the identical LOSO protocol, preprocessing pipeline, and two-stage fine-tuning strategy, MobileNetV3-S achieved a mean accuracy of 77.25% (standard deviation 14.68 percentage points) and a mean AUC of 0.873 (standard deviation 0.136). While DenseNet-121 maintained a consistent accuracy advantage of approximately 6 percentage points, the two architectures achieved nearly identical AUC values, a difference of only 0.006, within the range of run-to-run variance observed elsewhere in this study, suggesting comparable underlying discriminative capability despite the accuracy gap. In terms of model size, MobileNetV3-S contains approximately 1.1 million parameters (4.2 MB), compared to DenseNet-121's 7.3 million parameters (27.9 MB), a 6.6-fold reduction, which may be particularly relevant for deployment on embedded automotive hardware with limited memory and storage. Table V summarizes these comparisons. These findings suggest that the comparative advantage of a deep-learning-based approach over EAR may be more pronounced under more visually diverse and challenging conditions not represented in the present dataset, and that a full latency benchmark of MobileNetV3-S on the deployment hardware described in Section III.D remains a valuable direction for future work.

TABLE V
SUMMARY COMPARISON OF CLASSIFICATION APPROACHES

Method	Accuracy (Mean ± SD)	AUC (Mean ± SD)	Notes
EAR Baseline	84.37% ± 15.11%	0.933 ± 0.087	Threshold re-optimized per fold, range 0.25-0.27; landmark-coordinate-based, unaffected by image resolution
MobileNetV3-S	77.25% ± 14.68%	0.873 ± 0.136	Identical protocol to DenseNet-121; 1.1M parameters (4.2 MB), 6.6x smaller than DenseNet-121
DenseNet-121 (Proposed)	83.22% ± 13.22%	0.879 ± 0.131	Two-stage fine-tuning, 224x224 RGB input, 7.3M parameters (27.9 MB)

K. Practical Implementation Considerations for In-Vehicle Deployment

Beyond the technical performance metrics reported in the preceding sections, several practical considerations merit discussion regarding the integration of the proposed system into an operational in-vehicle environment. First, regarding camera integration, the present system was evaluated using a standard, unmounted webcam positioned to approximate a driver-facing view; a production deployment would instead require a fixed, dashboard- or steering-column-mounted camera module with a consistent viewing angle relative to the driver's seating position, potentially incorporating infrared illumination to maintain MediaPipe Face Mesh detection reliability during nighttime driving, a condition not evaluated in this study. Second, concerning driver alert mechanisms, the binary drowsiness classification produced by the system would need to be integrated with an escalating alert strategy rather than a single-threshold warning, for instance progressing from a subtle visual or auditory cue upon initial sustained drowsiness detection to a more assertive alert, such as seat vibration or an audible alarm, if the smoothed drowsiness prediction persists beyond a clinically informed duration threshold, consistent with the 1 to 30 second microsleep window discussed in Section I. The temporal smoothing mechanism introduced in Section III.I provides a natural foundation for such duration-based escalation logic, as the rolling prediction window could be extended to track

sustained drowsiness state over several seconds rather than merely stabilizing frame-level noise. Third, with respect to long-term operational evaluation, the present study's assessment was limited to short, controlled testing sessions; a genuine deployment would require extended evaluation across full driving sessions and diverse drivers to characterize the system's false alarm rate under naturalistic conditions, such as prolonged blinking, conversation, or glancing at mirrors, which may superficially resemble drowsiness-related eye closure but do not indicate an unsafe driving state. Establishing an acceptable false alarm rate under such naturalistic conditions, rather than accuracy alone, is likely to be the primary determinant of driver acceptance and sustained use of the system in practice.

IV. CONCLUSION

Based on the results of this study, which were obtained using a Subject-Independent Leave-One-Subject-Out (LOSO) cross-validation protocol, the integration of MediaPipe Face Mesh and the DenseNet-121 architecture demonstrates the ability to classify driver drowsiness from eye-region images without relying on a static, manually defined EAR threshold. Across five folds, each evaluated on a subject entirely unseen during training, the proposed system achieved a mean accuracy of 83.22% (standard deviation 13.22 percentage points) and a mean AUC of 0.879 (standard deviation 0.131), with per-fold accuracy ranging from 69.59% to 99.36%. This considerable variation across subjects indicates that, while the model performs very well for most individuals, its generalization to entirely new users remains inconsistent, and further analysis attributed this variation to a combination of genuine inter-subject differences in eye-closure expressiveness and the constrained native spatial resolution of the source imagery. A supplementary evaluation using a random, subject-mixed data split demonstrated that this stricter subject-independent estimate is substantially more conservative than commonly reported alternatives, with accuracy inflating to 97.21% when subject independence was not enforced, underscoring the importance of the evaluation protocol adopted in this study. Cross-dataset evaluation on an independent, infrared-based eye-state dataset further revealed a substantial degradation in generalization performance, indicating that the model's learned representations remain sensitive to shifts in visual domain. The system was measured to operate at an average total latency of 198.00 ms per frame, corresponding to a processing speed of 5.1 FPS on a CPU-only consumer laptop without dedicated GPU acceleration, indicating that further optimization or hardware acceleration would be needed to achieve conventional real-time video frame rates, though this speed remains plausible for the intended drowsiness monitoring use case given the typically sustained duration of microsleep episodes.

These findings suggest that the proposed approach shows meaningful potential for driver drowsiness monitoring applications, but that its current performance and evaluation scope do not yet support strong claims of readiness for real-world in-vehicle deployment. Several limitations should be acknowledged. First, the dataset used in this study was collected from only five subjects under a single, controlled indoor lighting condition, which limits the diversity of head poses, eyewear usage, and illumination scenarios represented in training and evaluation; this limitation is further compounded by the constrained native pixel resolution of the extracted eye regions, a consequence of the data collection tool used, which was found to correlate with reduced classification performance for a subset of subjects. Second, this study did not include a direct empirical comparison against the traditional EAR method under matched testing conditions, so the extent of improvement over EAR-based approaches remains to be quantitatively established. Third, cross-dataset evaluation revealed that the model's learned representations do not transfer well to a visually distinct domain, indicating that claims of generalization should, for now, be limited to visual conditions similar to those represented in the training data. Fourth, live deployment testing revealed a sensitivity to rapid head movement not captured by the static image evaluation protocol, which was partially, though not exhaustively, mitigated through a temporal smoothing mechanism applied to the prediction output.

Future research should therefore prioritize expanding the dataset to include a larger and more diverse pool of subjects, incorporating variations such as eyeglasses usage, nighttime or low-light conditions, and a wider range of head poses, captured at a higher native image resolution to preserve fine-grained eyelid detail. A direct comparative evaluation against EAR-based and other existing baseline methods under identical subject-independent testing conditions is also recommended to substantiate claims of improvement. Further work is also needed to improve cross-domain generalization, potentially through domain adaptation techniques or the inclusion of visually diverse training data spanning multiple sensor types. Finally, a more rigorous quantitative evaluation of temporal smoothing strategies under controlled head movement conditions is recommended, alongside a full latency and power-consumption benchmark of the MobileNetV3-S alternative evaluated in Section III.J, to more comprehensively establish the accuracy-efficiency trade-off for edge deployment scenarios. The practical deployment considerations discussed in Section III.K, particularly regarding naturalistic false alarm rates and duration-based alert escalation, should also be empirically validated through extended real-world driving trials prior to production deployment.

REFERENCES

[1] E. L. Istikhomah Puspita Sari and I. K. Agung Enriko, “Sistem Peringatan Tersemat untuk Pengemudi Mengantuk,” *Jurnal Riset Rekayasa Elektro*, vol. 5, no. 1, p. 75, Jun. 2023, doi: 10.30595/jrre.v5i1.17922.

[2] R. P. F. A. Nadapdap, J. M. Jabbar, and M. Pamungkas, "Faktor Risiko Microsleep Pada Driver Ojek Online Di Antapani Tahun 2023," *STIKES Dharma Husada*, 2023.

[3] J.-D. Wu and C.-H. Chang, "Driver Drowsiness Detection and Alert System Development Using Object Detection," *Traitement du Signal*, vol. 39, no. 2, pp. 493–499, Apr. 2022, doi: 10.18280/ts.390211.

[4] T. Fonseca and S. Ferreira, "Drowsiness Detection in Drivers: A Systematic Review of Deep Learning-Based Models," *Applied Sciences*, vol. 15, no. 16, p. 9018, Aug. 2025, doi: 10.3390/app15169018.

[5] R. I. Rabbi, P. P. Em, and Md. J. Hossen, "Smart driving with AI: A review of CNN approaches to drowsiness detection," *Array*, vol. 29, p. 100675, Mar. 2026, doi: 10.1016/j.array.2025.100675.

[6] P. Gosai and U. Barad, "Real time Driver's Drowsiness Detection by Convolution Neural Network (CNN) of Deep Learning Approach," *International Journal of Research in Advent Technology*, vol. 11, no. 1, pp. 7–16, Mar. 2023, doi: 10.32622/ijrat.111202307.

[7] R. Florez, F. Palomino-Quispe, R. J. Coaquira-Castillo, J. C. Herrera-Levano, T. Paixão, and A. B. Alvarez, "A CNN-Based Approach for Driver Drowsiness Detection by Real-Time Eye State Identification," *Applied Sciences*, vol. 13, no. 7849, pp. 1–18, Jul. 2023, doi: 10.3390/app13137849.

[8] A. Nomura, A. Yoshida, K. Nagumo, and A. Nozawa, "Reducing the effect of face orientation using FaceMesh landmarks in drowsiness estimation based on facial thermal images," *Artif. Life Robot.*, vol. 30, no. 2, pp. 317–324, May 2025, doi: 10.1007/s10015-024-01001-1.

[9] D. Kim, H. Park, T. Kim, W. Kim, and J. Paik, "Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition.," *Sci. Rep.*, vol. 13, no. 18264, pp. 1–14, Oct. 2023, doi: 10.1038/s41598-023-44955-1.

[10] A.-C. Phan, T.-N. Trieu, and T.-C. Phan, "Driver drowsiness detection and smart alerting using deep learning and IoT," *Internet of Things*, vol. 22, p. 100705, Jul. 2023, doi: 10.1016/j.iot.2023.100705.

[11] A. Altameem, A. Kumar, R. C. Poonia, S. Kumar, and A. K. J. Saudagar, "Early Identification and Detection of Driver Drowsiness by Hybrid Machine Learning," *IEEE Access*, vol. 9, pp. 162805–162819, Nov. 2021, doi: 10.1109/ACCESS.2021.3131601.

[12] C. Dewi, R.-C. Chen, C.-W. Chang, S.-H. Wu, X. Jiang, and H. Yu, "Eye Aspect Ratio for Real-Time Drowsiness Detection to Improve Driver Safety," *Electronics (Basel).*, vol. 11, no. 19, p. 3183, Oct. 2022, doi: 10.3390/electronics11193183.

[13] Taufiya Fathima and Dr. H Girisha, "Real-Time Driver Drowsiness Detection Using Eye Aspect Ratio and Facial Landmark Analysis," *International Research Journal on Advanced Engineering Hub (IRJAEH)*, vol. 3, no. 09, pp. 3432–3438, Sep. 2025, doi: 10.47392/IRJAEH.2025.0503.

[14] S. Chen, Z. Wang, and W. Chen, "Driver Drowsiness Estimation Based on Factorized Bilinear Feature Fusion and a Long-Short-Term Recurrent Convolutional Network," *Information*, vol. 12, no. 3, pp. 1–15, Dec. 2020, doi: 10.3390/info12010003.

[15] S. Sugeng and T. N. Nizar, "Deteksi Aktivitas Mata, Mulut Dan Kemiringan Kepala Sebagai Fitur Untuk Deteksi Kantuk Pada Pengendara Mobil," *Komputika : Jurnal Sistem Komputer*, vol. 12, no. 1, pp. 83–91, May 2023, doi: 10.34010/komputika.v12i1.9688.

[16] S. C. Kai Yuen, N. H. B. Zakaria, G. Eg Su, R. Hassan, S. Kasim, and T. Sutikno, "Real-time smart driver sleepiness detection by eye aspect ratio using computer vision," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 677–686, Apr. 2024, doi: 10.11591/ijeecs.v34.i1.pp677-686.

[17] A. Soultana, F. Benabbou, N. Sael, and S. Bohsissin, "A new deep learning model based on convolutional neural network and residual blocks for driver drowsiness detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 3, p. 1692, Mar. 2025, doi: 10.11591/ijeecs.v37.i3.pp1692-1701.

[18] N. Datta *et al.*, "Convolutional neural network-based real-time drowsy driver detection for accident prevention," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 23, no. 3, pp. 682–693, Jun. 2025, doi: 10.12928/telkomnika.v23i3.26059.

[19] G. Ersoy, M. A. Tatar, E. Tonbul, and S. Kırbız, "Improving Driver Drowsiness Detection via Personalized EAR/MAR Thresholds and CNN-Based Classification," Apr. 2026, [Online]. Available: http://arxiv.org/abs/2604.22479

[20] M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People," *Sensors*, vol. 23, no. 3, p. 1080, Jan. 2023, doi: 10.3390/s23031080.

[21] Y. Albadawi, A. AlRedhaei, and M. Takruri, "Real-Time Machine Learning-Based Driver Drowsiness Detection Using Visual Features," *J. Imaging*, vol. 9, no. 5, p. 91, Apr. 2023, doi: 10.3390/jimaging9050091.

[22] T. Chauhan, H. Palivela, and S. Tiwari, "Optimization and fine-tuning of DenseNet model for classification of COVID-19 cases in medical imaging," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100020, Nov. 2021, doi: 10.1016/J.JJIMEI.2021.100020.

[23] A. Bello, S.-C. Ng, and M.-F. Leung, "Skin Cancer Classification Using Fine-Tuned Transfer Learning of DENSENET-121," *Applied Sciences*, vol. 14, no. 17, p. 7707, Aug. 2024, doi: 10.3390/app14177707.

[24] I. Kayadibi, G. E. Güraksın, U. Ergün, and N. Özmen Süzme, "An Eye State Recognition System Using Transfer Learning: AlexNet-Based Deep Convolutional Neural Network," *International Journal of Computational Intelligence Systems*, vol. 15, no. 49, pp. 1–19, Jul. 2022, doi: 10.1007/s44196-022-00108-2.

[25] I. Galanakis, R. F. Soldatos, N. Karanikolas, A. Voulodimos, I. Voyiatzis, and M. Samarakou, "A MediaPipe Holistic Behavior Classification Model as a Potential Model for Predicting Aggressive Behavior in Individuals with Dementia," *Applied Sciences*, vol. 14, no. 22, p. 10266, Nov. 2024, doi: 10.3390/app142210266.

[26] O. F. Hassan, A. F. Ibrahim, A. Gomaa, M. A. Makhlouf, and B. Hafiz, "Real-time driver drowsiness detection using transformer architectures: a novel deep learning approach," *Sci. Rep.*, vol. 15, no. 1, p. 17493, May 2025, doi: 10.1038/s41598-025-02111-x.