Driver Drowsiness Detection: Comparative Analysis of BiLSTM and CNN+BiLSTM Architectures

Luma T. L. de Souza[1], **Thiago M. Paixão**[1], **Richard J. M. G. Tello**[1]

[1]Instituto Federal do Espírito Santo (IFES) - Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, Serra - ES - Brasil

lumatavares501@gmail.com

{thiago.paixao, richard}@ifes.edu.br

Abstract. *Driver drowsiness can negatively affect a person's ability to stay alert, compromising not only their own safety but also the safety of others. In this work, we propose a comparison between a binary CNN+BiLSTM model and a BiLSTM model to predict whether an individual is alert or not. Using the UTA-RLLD dataset, which contains videos of individuals actually experiencing drowsiness, we process the positions of the eyes and mouth, as well as the distance between the chin and the nose, transforming these features into vectors that allow the model to capture the spatial information of each frame. A Bidirectional Long Short-Term Memory (BiLSTM) network is employed to capture the temporal dynamics across frames, including gradual changes in eye closure, yawning, and head movements. The experimental results show that the CNN+BiLSTM model achieves higher accuracy on the test dataset (77.59%) compared to the model using only BiLSTM layers (70.69%), demonstrating the advantage of integrating convolutional layers with BiLSTM.*

1. Introduction

Drowsiness is an intermediate state between wakefulness and sleep, during which an individual may experience reduced alertness and impaired performance, posing significant risks on the road [Ebrahim Shaik 2023]. Commonly, when a driver becomes drowsy, they may experience symptoms such as drooping eyelids and an increasing urge to sleep. Drivers who have b een on the road for extended periods may have a poor awareness of how fatigued they are and underestimate their need for rest [Ghoddoosian et al. 2019, Chen et al. 2022].

According to the United States National Highway Traffic Safety Administration (NHTSA), drowsiness was responsible for 3,662 injuries and 4,121 fatal crashes between 2011 and 2015 [National Center for Statistics and Analysis 2017]. In the U.S., drowsiness-related traffic incidents account for approximately 20% of all traffic accidents and over 40% of major traffic incidents [Chen et al. 2022]. Similarly, data from the Brazilian Association of Traffic Medicine (ABRAMET), based on records from the Brazilian Federal Highway Police, indicate that more than 20,000 traffic accidents were caused by drowsy drivers between January 2014 and July 2020 [ABRAMET 2020]. Together, these statistics highlight the critical need for continued research and the development of effective strategies to mitigate this issue.

Current approaches to drowsiness detection typically fall into three categories: physiological measurements, performance-based metrics, and facial feature-based methods:

- *Physiological measurements* rely on signals such as electroencephalograms (EEG), electrocardiograms (ECG), and electromyograms (EMG). Even though these methods offer high accuracy, they are intrusive due to the need for a large number of sensors to be attached to the individual's body.
- *Performance-based metrics* are based on indicators such as steering wheel movement, lane deviation, and driving speed.
- *Facial feature-based methods* rely on computer vision and machine learning models to extract facial features of the driver, offering a less intrusive and cheaper alternative to physiological measurements, as it can be implemented using a smartphone camera.

In this study, we explore a hybrid approach that combines 1-D Convolutional neural networks (CNNs) and long-short-term memory (LSTM) networks for facial feature-based drowsiness detection. Models use facial characteristics such as the aspect ratio of the eye (EAR), the aspect ratio of the lips (LAR), and a third feature that was shown to be relevant to enhance prediction: the distance between the chin and the nose. This distance increases when the individual is drowsy, due to yawning or lowering the head, and decreases when they are alert.

Since these features evolve over time, temporal context is essential for accurate prediction. To address this, the system analyzes a time window of 120 frames to generate each prediction, corresponding to two minutes of video at 1 frame per second (fps). The findings of this study highlight the differences between model architectures, their efficiency, and their applicability in real-world scenarios, contributing to the development of a robust system that could potentially help prevent accidents in the future.

2. Related Work

A literature review was conducted to gain a better understanding of the problem. Most of these studies adopt computer vision approaches to detect drowsiness, offering a low-cost and non-invasive solution to the problem.

In [Basit et al. 2022], a hybrid model called ConvLSTM is proposed, which combines LSTM and CNN architectures. The input to the model consists of images with the Region of Interest (ROI) cropped; in this case, the regions correspond to the eyes and face. When driver drowsiness is detected, an alarm is triggered.

In another approach [Bekhouche et al. 2022], the system initially identifies the driver's face and then employs transfer learning to extract facial features using a CNN previously trained on a face recognition dataset. A vector of temporal features is then created and fed into a binary classifier to determine whether the person is alert or not.

In [Agarkar et al. 2023], the landmarks of the eyes and mouth are used to calculate the Eye Aspect Ratio (EAR) and the distance between the upper and lower lips to determine whether a person is yawning. However, this system cannot detect a face when the person has their hand over their mouth. In such cases, a model is used to detect whether the person is yawning with their hand on their face or not. The study in [Desai et al. 2024]

proposed a hybrid CNN+LSTM model —unlike our approach, which employs a bidirectional LSTM— their model uses the EAR of both eyes, MAR, and head pose as inputs to predict drowsiness.

Moreover, as previously mentioned, drowsiness can be detected using different techniques, such as electroencephalography (EEG). In [Lee and An 2023], participants were induced into a drowsy state. The objective of the study was to detect and classify the transition between alert and sleepy states, as well as the individual's consciousness states: awake, sleep, and drowsiness. However, the process of acquiring EEG data can be invasive for the individual.

While these studies effectively tackle the problem, many demand significant computational power or involve techniques that may cause discomfort to individuals. As a result, their applicability in real-world scenarios or deployment on low-power and embedded systems is limited. To address this, we propose a less invasive and computationally efficient model.

3. Proposed Methodology

This section discusses the methodology elements of the conducted study, including dataset and data preprocessing, the investigated deep models, and the experimental procedure.

3.1. Dataset

The dataset used in this study is the UTA-RLDD (Real-Life Drowsiness Dataset), created by the University of Texas at Arlington [Ghoddoosian et al. 2019]. It contains 180 hours of video originally recorded at a frame rate of less than 30 fps and includes 60 different healthy individuals who are genuinely experiencing drowsiness. Each participant has three distinct videos, each showing them in an alert, low vigilant, and drowsy state.

The videos were recorded by the participants using their phones or webcams, which resulted in a wide variety of resolutions and video qualities. The labels were provided by the participants themselves, taking into account their own state at the time of recording. In this work, the scope is limited to detecting the drowsy and alert classes.

3.2. Data Preprocessing

As a preprocessing step, frames were extracted from videos at a rate of 1 fps. After extracting the frames, we used DeepFace [Serengil and Ozpinar 2024], a facial recognition framework, to crop the face regions. Once the face region is cropped, facial landmarks are extracted using the MediaPipe framework [Zhang et al. 2020], allowing the face to be represented by 468 points,only the facial landmarks representing the eyes, mouth, nose, and chin are utilized. From these points,a feature vector is generated containing the EAR (Eye Aspect Ratio) and LAR (Lip Aspect Ratio), as well as the distance between the chin and the nose, which helps to detect when the head is tilted forward, a common sign of drowsiness. Together, these features help characterize and detect drowsiness.

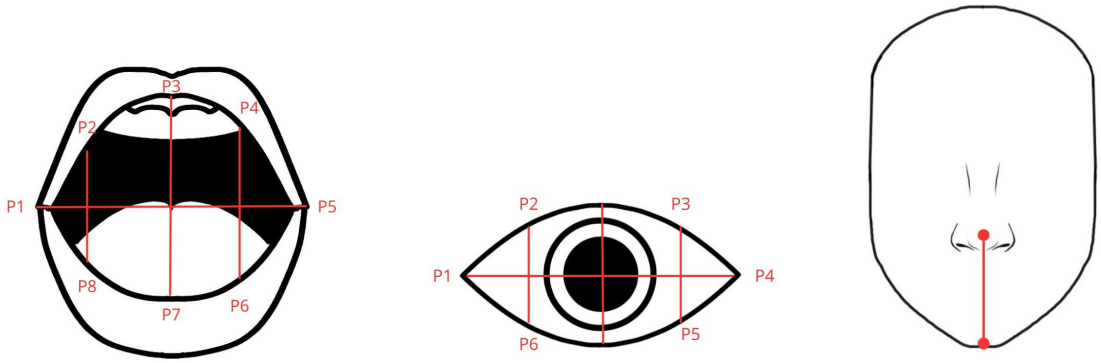


Figure 1. Landmarks around the (a) mouth, (b) eyes, and (c) head points.

LAR is defined as

$$LAR = \frac{|p_2 - p_8| + |p_3 - p_7| + |p_4 - p_6|}{2 \cdot |p_1 - p_5|}, \tag{1}$$

where $p_1, \ldots, p_8$ represent the mouth landmarks illustrated in Figure 1(a). Similarly, we have

$$EAR = \frac{|p_3 - p_5| + |p_2 - p_6|}{2 \cdot |p_1 - p_4|}, \tag{2}$$

where $p_1, \ldots, p_6$ correspond to the eye landmarks depicted in Figure 1b.

The distance between the nose and the chin is given by

$$ChinNoseDistance = |Nose - Chin|. \tag{3}$$

The equations for EAR and LAR represent, respectively, the eye closure ratio – which is the ratio between vertical eye distances and the horizontal eye width – and the ratio between the upper-lower lip distance and the distance between the mouth corners, [Agarkar et al. 2023].

3.3. Models

In this study, two architectures were compared: a BiLSTM, which relies solely on temporal features, and a hybrid CNN+BiLSTM model, which integrates both spatial and temporal information. The results indicate that the inclusion of CNN-derived spatial features contributes to improved performance.

The CNN+BiLSTM model combines convolutional layers for local feature extraction, a bidirectional Long Short-Term Memory (BiLSTM) network for temporal modeling, and fully connected layers for final classification.The input to the model does not consist of raw images but rather sequences of 120 feature vectors, each vector representing a video frame. Each vector contains four features: the Eye Aspect Ratio (EAR) for both eyes, the Lips Aspect Ratio (LAR), and the distance between the nose and the chin, which together capture information about eye closure, mouth opening, and head position. All of these features are numeric values representing their corresponding measurements.

In this architecture, the convolutional component is implemented as Conv1D layers that operate over the temporal sequences of feature vectors to extract local spatial–temporal patterns. It is composed of two Conv1D layers with 64 and 128 filters, respectively, both using a kernel size of 3 and ReLU activation. Each convolutional layer is followed by batch normalization and max pooling with a pool size of 2, which reduces sequence length and helps stabilize the learning process.

The resulting feature maps are then passed to a bidirectional LSTM layer, which captures temporal dependencies in both directions. This layer is regularized with L2 and followed by a dropout rate of 0.5 to mitigate overfitting. Finally, a dense layer and a sigmoid output layer produce the binary prediction.

To assess the impact of the convolutional component, a simplified BiLSTM model was also implemented. In this version, the convolutional layers were removed, retaining only the bidirectional LSTM followed by the dense and output layers. This configuration makes it possible to isolate the contribution of temporal modeling alone and provides a direct baseline for comparison with the hybrid CNN+BiLSTM.

3.4. Experimental Procedure

The dataset was split into 70% for training, 20% for validation, and 10% for testing, resulting in 42, 12, and 6 videos for each stage, respectively, in order to prevent data leakage. The models were trained using the Adam optimizer with a learning rate of 0.001, batch size 8, and binary cross-entropy as the loss function. To mitigate overfitting, L2 regularization (with $\lambda = 10^{-4}$) is applied. During training, an early stopping strategy was applied: if the validation loss did not improve for 10 consecutive epochs, the training process was stopped.

Furthermore, a preliminary study was conducted using different window sizes for the models. The window sizes tested were 30, 60, and 120 frames, and the resulting test accuracies were compared for both the CNN+BiLSTM and the BiLSTM-only models.

After training, the models were evaluated in the test set, and performance metrics were computed: accuracy, precision, recall, and validation loss.

Computer Setup: The experiments were conducted on a machine with the following specifications: an Intel(R) Core(TM) i9-9900KF CPU @ 3.60GHz, 32GB of RAM, a 1TB hard drive, a 260GB SSD, and an NVIDIA GeForce RTX 2060 GPU with 6GB of dedicated memory.The implementation was developed using the TensorFlow-Keras framework.

4. Results and Discussion

As an initial step, a preliminary study was conducted to compare the models' performance across different window sizes, with the results shown in Table1.

In both cases, using a window size of 120 frames resulted in the best test performance. However, for the BiLSTM model, the difference in accuracy between the 60- and 120-frame windows is approximately 2 percentage points, whereas for the CNN+BiLSTM model, the difference is around 7 percentage points. Although longer window sizes result in increased computation time, they provide more accurate results, which is particularly important in drowsiness detection scenarios.

Table 1. Test accuracy of CNN+BiLSTM and BiLSTM models for different window sizes.

windows size	CNN+BiLSTM	BiLSTM
30	70.54%	66.80%
60	70.59%	68.91%
120	77.59%	70.69%

Moreover, focusing on the models in the scenarios that achieved the best results—BiLSTM and CNN+BiLSTM with a window size of 120 frames—the following outcomes were obtained.

For CNN+BiLSTM, the best epoch was achieved at epoch 27 whereas for BiLSTM the best epoch was 12. The validation performance for both models is shown in Table 2. The results reveal a noticeable performance gap between the CNN+BiLSTM and the BiLSTM model. The CNN+BiLSTM model achieved an accuracy of 73.73%, whereas the BiLSTM model reached only 65.25% – a difference of over 8 percentage points.

Figures 2 and 3 show the accuracy and loss curves for the CNN+BiLSTM and BiLSTM models, with validation metrics summarized in Table 2. The CNN+BiLSTM model achieved the highest validation accuracy (73.73%) and recall (72.88%), reaching its best performance at epoch 27, with a precision of 72.88%. In contrast, the BiLSTM model reached its peak performance at epoch 12, with lower accuracy (65.25%) and considerably lower recall (57.63%).

These results indicate that the CNN+BiLSTM model consistently identifies drowsy instances with higher reliability. Moreover, the BiLSTM model's validation loss increased sharply after its peak, suggesting stronger overfitting. Overall, the CNN+BiLSTM model demonstrated better generalization and more robust performance compared to the BiLSTM-only model.

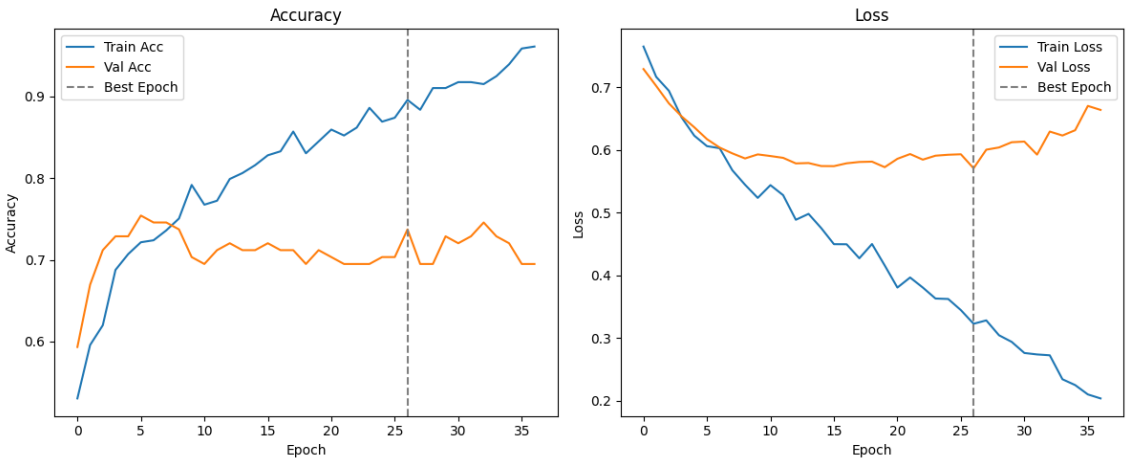


Figure 2. Accuracy and loss evolution during training for CNN+BiLSTM only model.

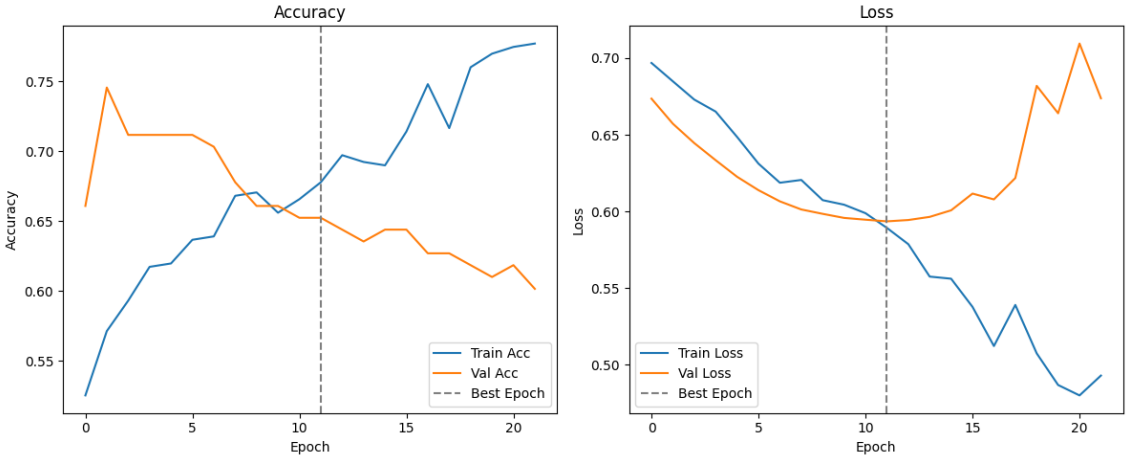
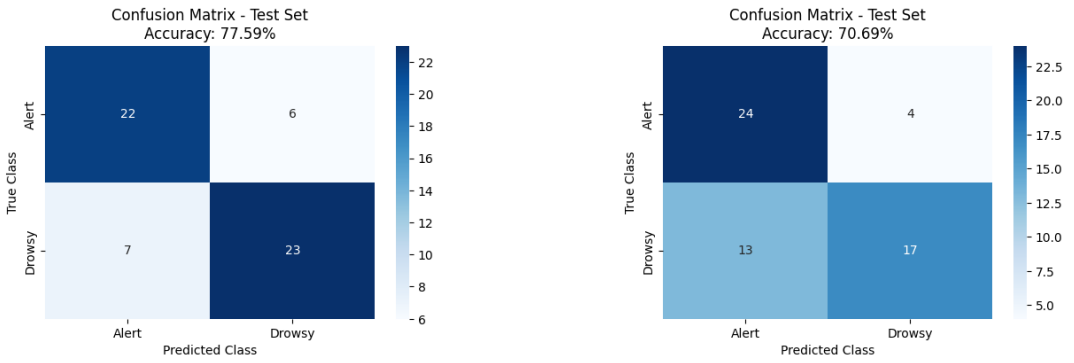


Figure 3. Accuracy and loss evolution during training for BiLSTM only model.

Table 2. Validation performance for CNN+BiLSTM (best epoch=27) and BiLSTM (best epoch=12).

Metric	CNN+BiLSTM	BiLSTM
Accuracy	73.73%	65.25%
Precision	74.14%	68.00%
Recall	72.88%	57.63%
Loss	0.5708	0.5935

In the test phase, the following results were obtained: the confusion matrices for both models are presented in Figure 4, along with the evaluation metrics on the test set, including precision, recall, and F1-score, are reported in Table 3.



(a) Hybrid CNN+BiLSTM model **(b) BiLSTM-only model**

Figure 4. Confusion matrix for the models with chin and nose features: (a) hybrid CNN+BiLSTM model and (b) BiLSTM-only model.

The classification reports for the test set (58 samples) highlight clear differences in performance between the BiLSTM-only and CNN+BiLSTM models. The BiLSTM-only model achieved an overall accuracy of 71.00%, with a macro-average precision of 0.73, recall of 0.71, and F1-score of 0.70. While it demonstrated high recall for the Alert class (0.86), it struggled with Drowsy instances (recall 0.57), indicating that many drowsy cases were misclassified as alert, as evidenced in Figure 4a. Its precision for Drowsy was

Table 3. Comparison of the classification reports for BiLSTM-only and CNN+BiLSTM models on the test set (58 samples).

Class	BiLSTM-only				CNN+BiLSTM			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Alert	0.65	0.86	0.74	28	0.76	0.79	0.77	28
Drowsy	0.81	0.57	0.67	30	0.79	0.77	0.78	30
Accuracy			70.69				77.59	
Macro avg	0.73	0.71	0.70	58	0.78	0.78	0.78	58
Weighted avg	0.73	0.71	0.70	58	0.78	0.78	0.78	58

higher (0.81), suggesting that when the model predicted drowsy, it was usually correct, but overall it failed to capture all true drowsy instances.

In contrast, the CNN+BiLSTM model improved all metrics, achieving an overall accuracy of 78.00%, with macro- and weighted-average precision, recall, and F1-score of 0.78. Both Alert and Drowsy classes showed balanced precision and recall (0.76–0.79 for Alert, 0.77–0.79 for Drowsy), indicating that the hybrid model consistently identifies both alert and drowsy states with higher reliability. These results demonstrate that the inclusion of CNN-derived spatial features enhances temporal modeling, leading to better generalization and more robust detection of drowsiness.

5. Conclusion

Throughout this research, the problem of driver drowsiness detection was addressed with a focus on developing a system that is less inconvenient for the individual to use and easier to implement. In addition, a comparative analysis between the CNN+BiLSTM and the standalone BiLSTM models was conducted to provide empirical evidence of the contribution of the CNN component to the BiLSTM architecture. The experimental results demonstrated that the integration of CNN with BiLSTM significantly improves classification accuracy, thereby increasing system reliability. The CNN+BiLSTM model achieved an accuracy of 77.59% on the test set, indicating that there is still room for improvement.

Although the model delivers satisfactory performance, its parameter count and computational requirements should be considered, as they directly impact deployment feasibility in real-time or resource-constrained environments. Future work should therefore not only focus on improving predictive performance but also on optimizing the model's size and inference efficiency, as well as exploring alternative strategies to achieve more robust and practical drowsiness detection.

References

ABRAMET (2020). Problemas na saúde de motoristas causaram mais de 280 mil acidentes nas rodovias desde 2014. Site da ABRAMET: https://abramet.com.br/noticias/problemas-na-saude-de-motoristas-causaram-mais-de-280-mil-acidentes-nas-rodovias-desde-2014-aponta-abramet/. Acesso em: 20 set. 2025.

Agarkar, A. S., Gandhiraj, R., and Panda, M. K. (2023). Driver drowsiness detection and warning using facial features and hand gestures. In *2023 2nd International Conference*

on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), pages 1–6.

Basit, M. S., Ahmad, U., Ahmad, J., Ijaz, K., and Ali, S. F. (2022). Driver drowsiness detection with region-of-interest selection based spatio-temporal deep convolutional-lstm. In *2022 16th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6.

Bekhouche, S. E., Ruichek, Y., and Dornaika, F. (2022). Driver drowsiness detection in video sequences using hybrid selection of deep features. *Knowledge-Based Systems*, 252:109436.

Chen, J., Fang, Z., Wang, J., Chen, J., and Yin, G. (2022). A multi-view driver drowsiness detection method using transfer learning and population-based sampling strategy. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 3386–3391.

Desai, M. M., Kathad, K., and Modi, N. (2024). Real-time driver drowsiness detection using hybrid cnn-lstm model with facial feature and behavioral analysis. In *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, pages 197–202.

Ebrahim Shaik, M. (2023). A systematic review on detection and prediction of driver drowsiness. *Transportation Research Interdisciplinary Perspectives*, 21:100864.

Ghoddoosian, R., Galib, M., and Athitsos, V. (2019). A realistic dataset and baseline temporal model for early drowsiness detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lee, C. and An, J. (2023). Lstm-cnn model of drowsiness detection from multiple consciousness states acquired by eeg. *Expert Systems with Applications*, 213:119032.

National Center for Statistics and Analysis (2017). Drowsy driving 2015 (crash·stats brief statistical summary). Technical Report DOT HS 812 446, National Highway Traffic Safety Administration, Washington, DC.

Serengil, S. and Ozpinar, A. (2024). A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107.

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., and Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.