

HỆ THỐNG PHÁT HIỆN BUỒN NGỦ KHI LÁI XE THEO THỜI GIAN THỰC DỰA TRÊN PHÂN TÍCH HÀNH VI KHUÔN MẶT THEO THỜI GIAN

Lương Văn Linh^{1*}, Nguyễn Thị Hương Lý²

¹ Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế

² Khoa Công nghệ Thông tin, Trường Đại học Nha Trang

* Email: 250401013601@husc.edu.vn

Ngày nhận bài: 02/6/2026; ngày hoàn thành phản biện: 28/6/2026; ngày duyệt đăng: 30/6/2026

TÓM TẮT

Phát hiện buồn ngủ khi lái xe là một thành phần quan trọng trong các hệ thống hỗ trợ lái xe an toàn. Tuy nhiên, nhiều phương pháp hiện nay vẫn dựa trên các ngưỡng cố định hoặc xử lý độc lập từng khung hình, do đó chưa khai thác hiệu quả đặc tính liên tục theo thời gian của các hành vi như nhắm mắt kéo dài hoặc ngáp. Bài báo đề xuất một hệ thống phát hiện buồn ngủ dựa trên phân tích chuỗi đặc trưng khuôn mặt theo thời gian. Hệ thống sử dụng MediaPipe Face Mesh để trích xuất các đặc trưng EAR, MAR và tư thế đầu. Thông tin về tư thế đầu được sử dụng như một tín hiệu hỗ trợ nhằm tăng độ tin cậy của quá trình phát hiện buồn ngủ. Các đặc trưng EAR và MAR được mô hình hóa bằng Bi-LSTM để khai thác sự thay đổi theo thời gian, trong khi tư thế đầu được phân loại bằng mạng DNN nhẹ. Để thích ứng với sự khác biệt giữa những người dùng, hệ thống áp dụng cơ chế hiệu chuẩn thích ứng và chiến lược bầu chọn đa số nhằm giảm các cảnh báo sai. Kết quả thực nghiệm cho thấy hệ thống đạt độ chính xác 98,14% trên tập dữ liệu Internal và 76,41% trong kịch bản đánh giá chéo Internal → NTHU-DDD, đồng thời vẫn đáp ứng yêu cầu xử lý thời gian thực với tốc độ khoảng 25–30 FPS trên CPU phổ thông mà không cần GPU chuyên dụng.

Từ khóa: Bi-LSTM; hỗ trợ lái xe an toàn; phát hiện buồn ngủ; phân tích hành vi khuôn mặt; thời gian thực.

1. MỞ ĐẦU

Buồn ngủ và mất tập trung khi lái xe là những nguyên nhân quan trọng dẫn đến tai nạn giao thông nghiêm trọng trên toàn thế giới. Theo thống kê của National Highway Traffic Safety Administration (NHTSA), riêng tại Hoa Kỳ trong năm 2017 đã ghi nhận

khoảng 91.000 vụ tai nạn liên quan đến tình trạng người lái thiếu tỉnh táo, làm gần 50.000 người bị thương và hơn 800 người thiệt mạng (National Highway Traffic Safety Administration, 2023). Điều này cho thấy nhu cầu phát triển các hệ thống giám sát có khả năng phát hiện sớm trạng thái buồn ngủ và hỗ trợ cảnh báo người lái trong thời gian thực là hết sức cần thiết.

Các phương pháp phát hiện buồn ngủ hiện nay chủ yếu được chia thành ba nhóm gồm: (i) phương pháp dựa trên tín hiệu sinh lý như EEG hoặc EOG, (ii) phương pháp dựa trên hành vi điều khiển phương tiện và (iii) phương pháp thị giác máy tính dựa trên phân tích khuôn mặt người lái (Albadawi, Takruri, & Awad, 2022). Trong đó, các phương pháp thị giác máy tính ngày càng được quan tâm nhờ ưu điểm không xâm lấn, chi phí triển khai thấp và khả năng tích hợp trên các nền tảng phần cứng phổ thông (Albadawi, Takruri, & Awad, 2022; Alajlan & Ibrahim, 2023). Nhiều nghiên cứu gần đây đã áp dụng các mô hình học sâu để nhận diện trạng thái buồn ngủ từ hình ảnh khuôn mặt và đạt được kết quả khả quan trong các ứng dụng thời gian thực (Ly, Nguyen, & Kiet, 2025).

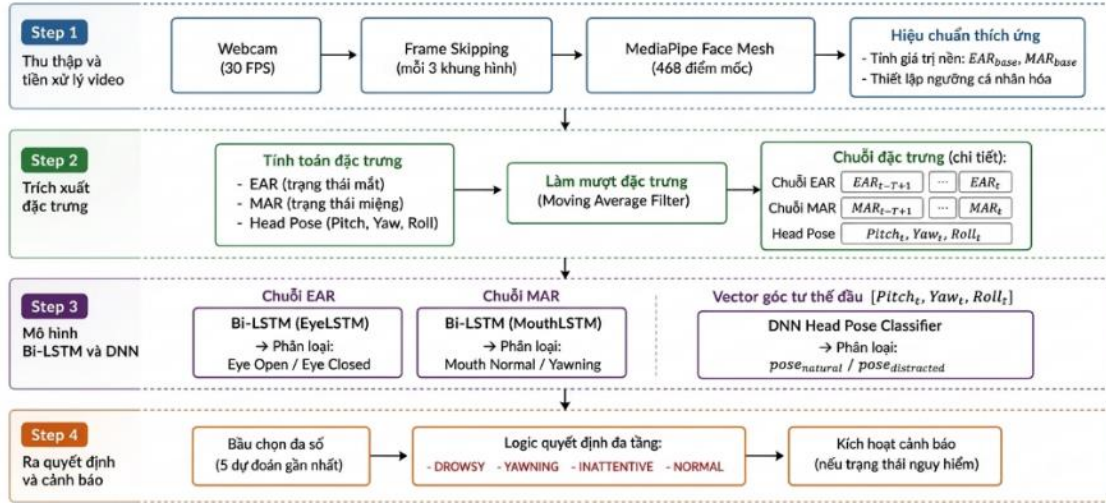
Tuy nhiên, các hệ thống hiện có vẫn còn một số hạn chế. Các phương pháp dựa trên ngưỡng cố định của Eye Aspect Ratio (EAR) hoặc Mouth Aspect Ratio (MAR) thường khó thích nghi với sự khác biệt về đặc điểm khuôn mặt giữa các cá nhân cũng như điều kiện quan sát khác nhau (Soukupová & Čech, 2016; Abtahi, Hariri, & Shirmohammadi, 2011). Bên cạnh đó, việc xử lý độc lập từng khung hình chưa khai thác hiệu quả tính liên tục theo thời gian của các hành vi như nhắm mắt kéo dài hoặc ngáp liên tiếp (Kim et al., 2023). Ngoài ra, một số kiến trúc học sâu hiện đại mặc dù đạt độ chính xác cao nhưng đòi hỏi tài nguyên tính toán lớn, gây khó khăn cho việc triển khai trên các hệ thống nhúng hoặc thiết bị chỉ sử dụng CPU (Jarndal et al., 2025).

Xuất phát từ các hạn chế trên, bài báo đề xuất một hệ thống phát hiện buồn ngủ theo thời gian thực dựa trên phân tích chuỗi hành vi khuôn mặt. Hệ thống sử dụng MediaPipe Face Mesh để trích xuất các đặc trưng EAR, MAR và tư thế đầu, đồng thời áp dụng Bi-LSTM nhằm mô hình hóa sự biến đổi theo thời gian của trạng thái mắt và hành vi ngáp. Trong nghiên cứu này, thông tin về tư thế đầu và trạng thái mắt tập trung chỉ được sử dụng như tín hiệu hỗ trợ nhằm nâng cao độ tin cậy của quá trình phát hiện buồn ngủ, không phải là mục tiêu phát hiện chính của hệ thống. Bên cạnh đó, cơ chế hiệu chuẩn thích ứng được tích hợp để cá nhân hóa các ngưỡng phát hiện cho từng người dùng, kết hợp với chiến lược bầu chọn đa số nhằm giảm cảnh báo sai và nâng cao độ ổn định của hệ thống. So với nghiên cứu trước của nhóm tác giả dựa trên YOLO và CNN (Ly, Nguyen, & Kiet, 2025), phương pháp đề xuất chuyển từ phân tích từng khung hình sang khai thác thông tin chuỗi thời gian của hành vi khuôn mặt, đồng thời vẫn đảm bảo khả năng xử lý thời gian thực trên CPU phổ thông mà không cần GPU chuyên dụng.

2. PHƯƠNG PHÁP ĐỀ XUẤT

2.1. Kiến trúc tổng quan hệ thống

Hệ thống đề xuất gồm bốn thành phần chính: (i) thu nhận video, (ii) trích xuất đặc trưng khuôn mặt, (iii) phân tích trạng thái người lái và (iv) ra quyết định và kích hoạt cảnh báo (Hình 1).



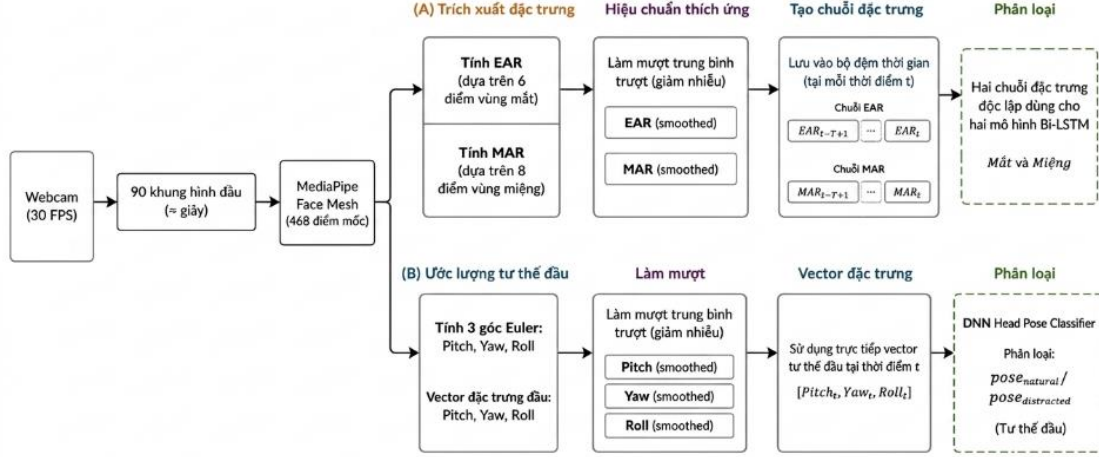
Hình 1. Quy trình xử lý của hệ thống phát hiện buồn ngủ đề xuất

Video đầu vào được thu nhận từ webcam với tốc độ khoảng 30 FPS và xử lý bằng MediaPipe Face Mesh để xác định 468 điểm mốc khuôn mặt theo thời gian thực. Từ các điểm mốc này, hệ thống tính toán các đặc trưng hình học gồm EAR, MAR và các góc tư thế đầu.

Trong giai đoạn khởi tạo, cơ chế hiệu chuẩn thích ứng được áp dụng để thiết lập ngưỡng EAR và MAR cho từng người dùng. Sau đó, chuỗi EAR và MAR được đưa vào hai mô hình Bi-LSTM nhằm phân loại trạng thái mắt và hành vi ngáp, trong khi thông tin tư thế đầu được sử dụng để phát hiện trạng thái mất tập trung. Cuối cùng, các kết quả dự đoán được tổng hợp thông qua cơ chế bầu chọn đa số và chiến lược ra quyết định đa tầng trước khi phát cảnh báo.

2.2. Hiệu chuẩn thích ứng và trích xuất đặc trưng

Các phương pháp dựa trên ngưỡng cố định thường gặp khó khăn khi áp dụng cho những người dùng có đặc điểm khuôn mặt khác nhau như mắt nhỏ bẩm sinh, đeo kính hoặc thay đổi góc quan sát (Soukupová & Čech, 2016; Abtahi, Hariri, & Shirmohammadi, 2011). Để giảm ảnh hưởng của sự khác biệt giữa các cá nhân, nghiên cứu đề xuất cơ chế hiệu chuẩn thích ứng nhằm thiết lập ngưỡng phù hợp cho từng người dùng ngay khi hệ thống khởi động (Hình 2).



Hình 2. Quy trình hiệu chuẩn thích ứng EAR/MAR và trích xuất đặc trưng

Trong khoảng 3 giây đầu tiên, hệ thống giả định người lái đang ở trạng thái bình thường với mắt mở, miệng đóng và tư thế đầu hướng về phía trước. Chỉ các khung hình có khuôn mặt được phát hiện ổn định mới được sử dụng để tính các giá trị EAR_{base} và MAR_{base} , nhằm hạn chế ảnh hưởng của nhiễu trong giai đoạn hiệu chuẩn. Giá trị nền của từng cá nhân được xác định như sau:

$$EAR_{base} = \frac{1}{N} \sum_{i=1}^N EAR_i \quad (1)$$

$$MAR_{base} = \frac{1}{N} \sum_{i=1}^N MAR_i \quad (2)$$

Trong các công thức (1) và (2), N là số khung hình được sử dụng trong giai đoạn hiệu chuẩn; i là chỉ số của từng khung hình; EAR_i và MAR_i lần lượt là giá trị EAR và MAR tại khung hình thứ i ; EAR_{base} và MAR_{base} là giá trị trung bình của EAR và MAR trong giai đoạn hiệu chuẩn.

Các ngưỡng phát hiện sau đó được cá nhân hóa theo các biểu thức:

$$T_{close} = 0.6 \times EAR_{base} \quad (3)$$

$$T_{open} = 0.8 \times EAR_{base} \quad (4)$$

$$T_{yawn} = \max(1.4 \times MAR_{base}, 0.4) \quad (5)$$

Trong đó, T_{close} là ngưỡng xác định trạng thái nhắm mắt, T_{open} là ngưỡng xác định trạng thái mắt mở trở lại và T_{yawn} là ngưỡng phát hiện hành vi ngáp. Các hệ số trong (3)–(5) được lựa chọn thông qua các thử nghiệm trên tập dữ liệu nội bộ. Nhóm tác giả khảo sát nhiều giá trị và lựa chọn bộ tham số cho kết quả cân bằng giữa khả năng phát hiện và tỷ lệ cảnh báo nhầm, sau đó giữ cố định trong toàn bộ các thí nghiệm. Do các ngưỡng được tính trực tiếp từ EAR_{base} và MAR_{base} của từng người dùng, hệ thống có thể thích ứng với sự khác biệt về đặc điểm khuôn mặt giữa các cá nhân.

Sau giai đoạn hiệu chuẩn, MediaPipe Face Mesh tiếp tục cung cấp 468 điểm mốc khuôn mặt ba chiều theo thời gian thực (Zhang, Zhu, & Ye, 2020). Từ các điểm mốc này, hệ thống tính toán các đặc trưng EAR (Soukupová & Čech, 2016), MAR (Abtahi, Hariri, & Shirmohammadi, 2011) và ba góc Euler gồm Pitch, Yaw và Roll (Lepetit, Moreno-Noguer, & Fua, 2009). Trong đó, EAR và MAR phản ánh trạng thái mở hoặc đóng của mắt và miệng, còn các góc Euler mô tả chuyển động của đầu và hỗ trợ phát hiện trạng thái mất tập trung.

Để hạn chế nhiễu tức thời, các đặc trưng được làm mượt bằng trung bình trượt trước khi đưa vào giai đoạn phân loại. Chuỗi EAR và MAR được lưu trong bộ đệm thời gian để xây dựng đầu vào cho các mô hình Bi-LSTM (Hochreiter & Schmidhuber, 1997; Graves & Schmidhuber, 2005; Wijnands et al., 2020; Kim et al., 2023), trong khi vector đặc trưng $[Pitch, Yaw, Roll]$ được sử dụng cho bộ phân loại tư thế đầu.

2.3. Mô hình hóa chuỗi thời gian bằng Bi-LSTM

Các hành vi liên quan đến buồn ngủ như nhắm mắt kéo dài hoặc ngáp thường diễn ra liên tục theo thời gian thay vì xuất hiện tại một khung hình đơn lẻ. Vì vậy, nghiên cứu sử dụng Bi-LSTM để khai thác sự thay đổi của các đặc trưng EAR và MAR theo thời gian (Hochreiter & Schmidhuber, 1997; Graves & Schmidhuber, 2005).

Đối với mỗi bài toán, hệ thống áp dụng cơ chế cửa sổ trượt với kích thước $T = 12$ để xây dựng chuỗi đầu vào:

$$X_t = \{x_{t-T+1}, x_{t-T+2}, \dots, x_t\} \quad (6)$$

trong đó X_t là chuỗi đặc trưng đầu vào tại thời điểm t ; x_t là giá trị EAR hoặc MAR tại thời điểm t ; T là kích thước cửa sổ thời gian và được thiết lập bằng 12 khung hình. Mỗi mẫu đầu vào được biểu diễn dưới dạng tensor kích thước (12,1).

Trong quá trình suy luận, hệ thống thực hiện dự đoán sau mỗi ba khung hình nhằm giảm chi phí tính toán. Tuy nhiên, các đặc trưng EAR và MAR vẫn được cập nhật liên tục vào bộ đệm thời gian để duy trì tính liên tục của chuỗi đầu vào. Với tốc độ thu nhận khoảng 30 FPS, cửa sổ thời gian $T = 12$ tương ứng khoảng 0,4 giây dữ liệu liên tiếp của hành vi người lái. Chi tiết kiến trúc mô hình và cấu hình huấn luyện được trình bày trong Mục 3.4.

2.4. Phân loại trạng thái tư thế đầu

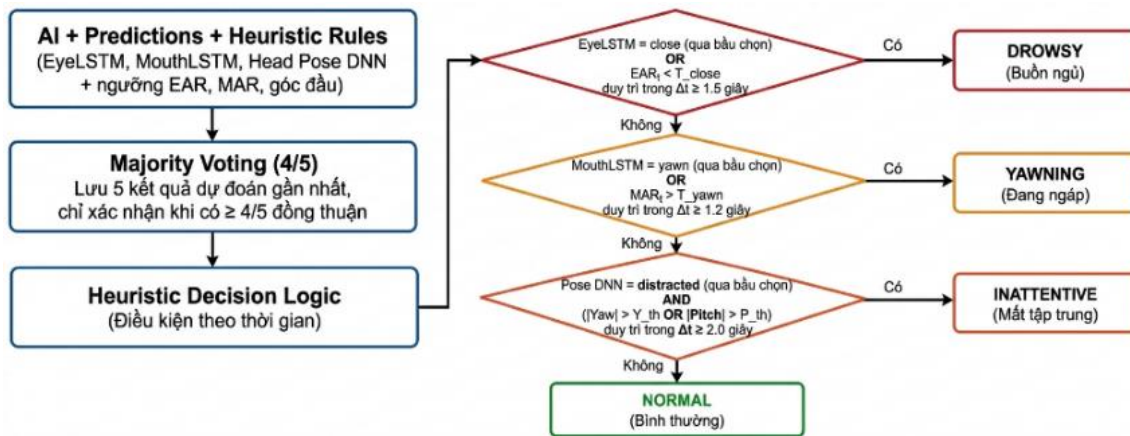
Bên cạnh các đặc trưng từ mắt và miệng, hệ thống sử dụng thông tin tư thế đầu để hỗ trợ phát hiện trạng thái mất tập trung của người lái. Tại mỗi khung hình, ba góc Euler gồm Pitch, Yaw và Roll được ước lượng từ các điểm mốc khuôn mặt thông qua bài toán Perspective-n-Point (PnP) (Lepetit, Moreno-Noguer, & Fua, 2009).

Một mạng DNN gọn nhẹ được sử dụng để phân loại trạng thái tư thế đầu dựa trên vector đặc trưng gồm ba góc Euler đã được chuẩn hóa. Mô hình được huấn luyện

trên tập dữ liệu AFLW2000-3D với các nhãn được xây dựng theo quy trình trình bày tại Mục 3.3. Kết quả phân loại được kết hợp với các luật heuristic trong giai đoạn ra quyết định của hệ thống. Chi tiết kiến trúc mô hình và cấu hình huấn luyện được trình bày trong Mục 3.4.

2.5. Chiến lược ra quyết định và cảnh báo

Để tăng độ ổn định trong môi trường thực tế, hệ thống sử dụng cơ chế ra quyết định đa tầng kết hợp giữa mô hình học sâu và các luật heuristic thay vì phụ thuộc hoàn toàn vào kết quả dự đoán tức thời của từng khung hình. Trong cơ chế này, kết quả phân loại tư thế đầu được sử dụng như một điều kiện hỗ trợ để tăng độ tin cậy của quá trình xác nhận trạng thái buồn ngủ, thay vì được sử dụng như một tiêu chí phát hiện độc lập.



Hình 3. Cơ chế ra quyết định đa tầng và kích hoạt cảnh báo của hệ thống

Mỗi lần suy luận của mô hình Bi-LSTM sử dụng cửa sổ thời gian $T = 12$ (khoảng 0,4 giây). Kết quả của các cửa sổ liên tiếp được lưu trong bộ đệm quyết định và cập nhật theo thời gian. Nhờ đó, hệ thống theo dõi được các trạng thái kéo dài thay vì chỉ dựa trên kết quả của một cửa sổ đơn lẻ. Để giảm ảnh hưởng của các dự đoán nhiễu, hệ thống áp dụng cơ chế bầu chọn đa số trên năm kết quả phân loại gần nhất. Một trạng thái chỉ được xác nhận khi có ít nhất bốn trong năm dự đoán đồng thuận (Hình 3).

Dựa trên kết quả phân loại và các đặc trưng hình học, hệ thống xác định bốn trạng thái gồm buồn ngủ, ngáp, mất tập trung và bình thường. Trạng thái buồn ngủ được kích hoạt khi các dự đoán nhắm mắt của mô hình Bi-LSTM duy trì liên tiếp trong thời gian trên 1,5 giây. Trạng thái ngáp được xác định khi giá trị MAR vượt ngưỡng trong một khoảng thời gian nhất định hoặc được phát hiện bởi mô hình Bi-LSTM. Trạng thái mất tập trung được kích hoạt khi tư thế đầu vượt ngưỡng cho phép và đồng thời được xác nhận bởi bộ phân loại DNN. Khi phát hiện các trạng thái nguy hiểm, hệ thống sẽ kích hoạt cảnh báo âm thanh để hỗ trợ người lái.

3. DỮ LIỆU VÀ THIẾT LẬP THÍ NGHIỆM

3.1. Tập dữ liệu tự thu thập

Dữ liệu nội bộ được thu thập từ 10 người tham gia trong môi trường mô phỏng lái xe bằng webcam với tốc độ khoảng 30 FPS. Các mẫu bao gồm trạng thái mắt mở, mắt nhắm, ngáp và thay đổi tư thế đầu; toàn bộ video được xử lý bằng MediaPipe Face Mesh để tạo các chuỗi EAR và MAR phục vụ huấn luyện Bi-LSTM.

Bảng 1. Thông tin tổng quan tập dữ liệu tự thu thập

Thuộc tính	Giá trị
Số người tham gia	10
Nam/Nữ	6/4
Người đeo kính	3
Tốc độ ghi hình	~30 FPS
Môi trường thu thập	Trong nhà
Hành vi ghi nhận	Mắt mở, mắt nhắm, ngáp, tư thế đầu
Tổng số mẫu chuỗi	> 9.700

Các chuỗi EAR và MAR được xây dựng bằng cửa sổ trượt với kích thước $T=12$. Dữ liệu được chia theo người dùng với 8 người cho huấn luyện và 2 người cho kiểm tra nhằm tránh rò rỉ dữ liệu giữa các cửa sổ liên tiếp.

3.2. Tập dữ liệu NTHU-DDD

National Tsing Hua University Driver Drowsiness Detection (NTHU-DDD) (Ouyang, Yang, & Yang, 2016) là tập dữ liệu công khai được sử dụng rộng rãi trong nghiên cứu phát hiện buồn ngủ của người lái xe. Tập dữ liệu bao gồm 36 người tham gia với nhiều điều kiện ghi hình khác nhau như ban ngày, ban đêm, đeo kính, kính râm và thay đổi góc quan sát. Ngoài trạng thái lái xe bình thường, tập dữ liệu còn bao gồm các hành vi như nhắm mắt kéo dài, ngáp, quay đầu và mất tập trung, giúp đánh giá khả năng tổng quát hóa của hệ thống trong nhiều điều kiện thực tế khác nhau.

Bảng 2. Thông tin tổng quan tập dữ liệu NTHU-DDD

Thuộc tính	Giá trị
Số người tham gia	36
Điều kiện ánh sáng	Ban ngày, ban đêm
Người đeo kính và kính râm	Có
Góc nhìn đầu	Đa dạng

Thuộc tính	Giá trị
Hành vi ghi nhận	Tỉnh táo, buồn ngủ, ngáp, nhắm mắt
Tốc độ ghi hình	~30 FPS

Trong nghiên cứu này, tập dữ liệu NTHU-DDD được sử dụng trong các thí nghiệm nhằm đánh giá khả năng tổng quát hóa của mô hình. Tùy theo từng kịch bản thực nghiệm, tập dữ liệu này được sử dụng làm tập huấn luyện, tập kiểm tra hoặc kết hợp với dữ liệu nội bộ. Các giao thức đánh giá cụ thể được trình bày trong Mục 4.2.

3.3. Tập dữ liệu AFLW2000-3D

Đối với bài toán phân loại tư thế đầu, nghiên cứu sử dụng tập dữ liệu công khai AFLW2000-3D (Zhu et al., 2016), gồm khoảng 2.000 ảnh khuôn mặt cùng các tham số tư thế đầu, từ đó suy ra ba góc Euler gồm Pitch, Yaw và Roll.

Do AFLW2000-3D không cung cấp sẵn nhãn cho bài toán phân loại trạng thái tập trung hay mất tập trung, nhóm tác giả xây dựng nhãn nhị phân dựa trên các góc Euler. Việc sử dụng hai góc Yaw và Pitch làm cơ sở gán nhãn được tham khảo từ các nghiên cứu về giám sát người lái (Doshi & Trivedi, 2012; Zhao et al., 2020). Các mẫu có $|Yaw| \leq 20^\circ$ và $|Pitch| \leq 15^\circ$ được gán nhãn *pose_natural*, trong khi các mẫu còn lại được gán nhãn *pose_distracted*. Góc Roll không tham gia vào quá trình gán nhãn nhưng vẫn được giữ lại trong vector đặc trưng đầu vào của mô hình DNN.

Sau khi gán nhãn, dữ liệu được cân bằng giữa hai lớp và chia theo tỷ lệ 80% cho huấn luyện và 20% cho kiểm tra. Thống kê dữ liệu được trình bày trong Bảng 3.

Bảng 3. Thống kê dữ liệu cho bài toán phân loại tư thế đầu

Lớp dữ liệu	Huấn luyện	Kiểm tra
<i>pose_natural</i>	800	200
<i>pose_distracted</i>	800	200

3.4. Cấu hình huấn luyện và giao thức đánh giá

Hai mô hình Bi-LSTM được huấn luyện độc lập cho hai bài toán phân loại trạng thái mất và trạng thái ngáp. Mỗi mô hình gồm một lớp Bidirectional LSTM với 64 đơn vị ẩn ở mỗi chiều, tiếp theo là lớp Dense 32 nút sử dụng hàm kích hoạt ReLU và lớp đầu ra Softmax hai nhãn. Đối với bài toán phân loại tư thế đầu, nghiên cứu sử dụng một mạng DNN gồm hai lớp Dense với 32 và 16 nút, kết hợp lớp Dropout trước lớp Softmax.

Cả hai mô hình đều sử dụng bộ tối ưu Adam và hàm mất mát Sparse Categorical Crossentropy. Bi-LSTM được huấn luyện với learning rate 0,001, batch size 64, tối đa 60 epoch và cơ chế EarlyStopping (patience = 10). Đối với DNN, learning rate là 0,001, batch size là 32 và số epoch tối đa là 50.

Các thí nghiệm trong nghiên cứu gồm bốn nhóm: (i) đánh giá trên tập dữ liệu nội bộ với chiến lược chia theo người dùng; (ii) đánh giá khả năng thích nghi trên tập dữ liệu công khai NTHU-DDD; (iii) phân tích ảnh hưởng của kích thước cửa sổ thời gian đối với mô hình Bi-LSTM; và (iv) đánh giá ảnh hưởng của hiệu chuẩn thích ứng, Bi-LSTM và cơ chế bầu chọn đa số đối với hiệu năng của hệ thống.

4. KẾT QUẢ VÀ THẢO LUẬN

4.1. Đánh giá dữ liệu nội bộ

Bảng 4 trình bày kết quả đánh giá các mô hình thành phần trên tập dữ liệu nội bộ với chiến lược phân chia theo người dùng nhằm đảm bảo các đối tượng xuất hiện trong tập kiểm tra không xuất hiện trong quá trình huấn luyện.

Bảng 4. Kết quả đánh giá các mô hình thành phần trên tập dữ liệu nội bộ

Model	Accuracy	F1-Score
EyeLSTM	99,5%	99,5%
MouthLSTM	93,5%	93,5%
PoseDense	89,1%	89,0%

Mô hình EyeLSTM đạt kết quả cao nhất với độ chính xác 99,5%, cho thấy chuỗi EAR chứa thông tin phân biệt rõ ràng giữa trạng thái mắt mở và mắt nhắm. Trong khi đó, bài toán phát hiện ngáp và phân loại tư thế đầu có độ khó cao hơn do chịu ảnh hưởng của biểu cảm khuôn mặt, góc quan sát và sự đa dạng trong hành vi người dùng, dẫn tới hiệu năng thấp hơn ở MouthLSTM và PoseDense.

4.2. Đánh giá khả năng tổng quát hóa

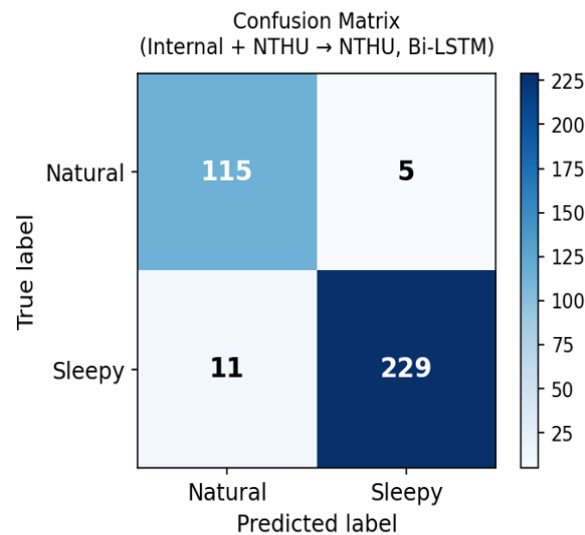
Để đánh giá khả năng thích nghi của hệ thống đối với các điều kiện chưa xuất hiện trong quá trình huấn luyện, nghiên cứu thực hiện các thí nghiệm đánh giá chéo giữa tập dữ liệu nội bộ và tập dữ liệu công khai NTHU-DDD.

Bảng 5 trình bày bốn kịch bản đánh giá gồm Internal $\rightarrow$ Internal, Internal $\rightarrow$ NTHU-DDD, NTHU-DDD $\rightarrow$ Internal và Internal + NTHU-DDD $\rightarrow$ NTHU-DDD. Trong kịch bản Internal $\rightarrow$ NTHU-DDD, tập NTHU-DDD đóng vai trò tập kiểm thử độc lập. Ở các kịch bản còn lại, tập dữ liệu này được sử dụng để huấn luyện hoặc kết hợp với dữ liệu nội bộ theo từng giao thức thực nghiệm. Đối với tất cả các kịch bản, tập huấn luyện và tập kiểm tra được tách biệt hoàn toàn; không có mẫu dữ liệu nào xuất hiện đồng thời trong cả hai tập, do đó không xảy ra hiện tượng rò rỉ dữ liệu.

Bảng 5. Kết quả đánh giá khả năng tổng quát hóa của LSTM và Bi-LSTM

Training	Testing	LSTM Acc.	Bi-LSTM Acc.	Improvement
Internal	Internal	97,72%	98,62%	+0,90
Internal	NTHU-DDD	74,52%	76,41%	+1,89
NTHU-DDD	Internal	83,28%	85,14%	+1,86
Internal + NTHU-DDD	NTHU-DDD	92,57%	95,63%	+3,06

Kết quả cho thấy Bi-LSTM luôn đạt độ chính xác cao hơn LSTM trong tất cả các kịch bản thực nghiệm. Khoảng cách này trở nên rõ rệt hơn trong các thí nghiệm đánh giá chéo miền dữ liệu, cho thấy việc khai thác thông tin theo cả hai hướng thời gian giúp mô hình thích nghi tốt hơn với các biến thiên về ánh sáng, góc nhìn và đặc điểm khuôn mặt. Ngoài ra, việc kết hợp dữ liệu từ nhiều nguồn huấn luyện giúp cải thiện đáng kể khả năng tổng quát hóa trên tập NTHU-DDD.



Hình 4. Confusion Matrix của mô hình Adaptive + Bi-LSTM + Voting trên tập NTHU-DDD

Hình 4 trình bày ma trận nhầm lẫn của mô hình cuối cùng trên tập dữ liệu NTHU-DDD. Kết quả cho thấy phần lớn sai số xuất hiện giữa các trạng thái có đặc điểm hành vi tương đồng như nhắm mắt ngắn hạn và buồn ngủ mức độ nhẹ.

4.3. Đánh giá đóng góp của các thành phần

Để phân tích vai trò của từng thành phần trong hệ thống, nghiên cứu thực hiện các thí nghiệm loại bỏ thành phần như trình bày trong Bảng 6. Trong cấu hình Fixed Threshold, hệ thống sử dụng các ngưỡng EAR = 0,25 và MAR = 0,40 cho tất cả người dùng mà không thực hiện hiệu chuẩn theo từng cá nhân. Hai giá trị này được lựa chọn dựa trên các ngưỡng thường được sử dụng trong các nghiên cứu phát hiện buồn ngủ

dựa trên đặc trưng EAR và MAR (Soukupová & Čech, 2016; Abtahi et al., 2011). Cấu hình này được sử dụng làm đường cơ sở để so sánh với cơ chế hiệu chuẩn thích ứng và đánh giá mức cải thiện của phương pháp đề xuất.

Bảng 6. Đóng góp của các thành phần trong hệ thống

Configuration	Accuracy
Fixed Threshold	48,05%
Adaptive Threshold	60,75%
Adaptive + LSTM	97,53%
Adaptive + Bi-LSTM	97,88%
Adaptive + Bi-LSTM + Voting	98,14%

So với cấu hình Fixed Threshold, cơ chế hiệu chuẩn thích ứng giúp độ chính xác tăng từ 48,05% lên 60,75%. Khi bổ sung mô hình LSTM và Bi-LSTM, độ chính xác tăng lần lượt lên 97,53% và 97,88%, cao hơn đáng kể so với hai cấu hình chỉ sử dụng ngưỡng. Sau khi áp dụng cơ chế bầu chọn đa số, độ chính xác tiếp tục tăng lên 98,14%. Cơ chế này giúp hạn chế ảnh hưởng của các dự đoán sai xuất hiện ở một số khung hình riêng lẻ.

4.4. Ảnh hưởng của kích thước cửa sổ thời gian

Bảng 7 trình bày ảnh hưởng của kích thước cửa sổ thời gian đối với hiệu năng và độ trễ của hệ thống.

Bảng 7. Ảnh hưởng của kích thước cửa sổ thời gian

T	Accuracy	Latency (ms)	FPS
12	98,13%	25,47	39,3
30	99,20%	52,20	19,2
60	99,49%	99,32	10,1

Việc tăng kích thước cửa sổ thời gian giúp mô hình khai thác thêm thông tin ngữ cảnh và cải thiện độ chính xác. Tuy nhiên, độ trễ suy luận cũng tăng tương ứng, làm giảm tốc độ xử lý của hệ thống. Giá trị 39,3 FPS trong Bảng 7 phản ánh tốc độ suy luận của mô hình Bi-LSTM với cửa sổ thời gian $T = 12$. Khi tích hợp toàn bộ hệ thống, bao gồm MediaPipe Face Mesh, các mô hình Bi-LSTM, DNN và cơ chế bầu chọn đa số, tốc độ xử lý thực tế đạt khoảng 25–30 FPS trên CPU phổ thông. Với $T = 12$, hệ thống vẫn duy trì độ chính xác trên 98% đồng thời đáp ứng yêu cầu xử lý thời gian thực.

4.5. Hiệu năng thời gian thực

Kết quả cho thấy phần lớn chi phí tính toán tập trung ở bước phát hiện điểm mốc khuôn mặt bằng MediaPipe Face Mesh. MediaPipe Face Mesh chiếm khoảng 15 ms trong tổng độ trễ suy luận 18–20 ms (Bảng 8), tương đương khoảng 75% thời gian xử lý của toàn bộ hệ thống, trong khi thời gian suy luận của các mô hình Bi-LSTM và DNN chỉ khoảng 3 ms. Điều này cho thấy chi phí tính toán chủ yếu đến từ quá trình phát hiện điểm mốc khuôn mặt và trích xuất đặc trưng, thay vì từ các mô hình học sâu.

Do đó, tốc độ xử lý của toàn bộ hệ thống không chỉ phụ thuộc vào thời gian suy luận của các mô hình Bi-LSTM và DNN mà còn bao gồm thời gian phát hiện điểm mốc khuôn mặt và trích xuất đặc trưng. Mặc dù tổng độ trễ suy luận của hệ thống chỉ khoảng 18–20 ms, tốc độ xử lý thực tế còn chịu ảnh hưởng của quá trình thu nhận khung hình từ camera, hiển thị giao diện và các tác vụ xử lý đồng thời của hệ điều hành. Vì vậy, hệ thống đạt tốc độ khoảng 25–30 FPS khi vận hành trên CPU phổ thông mà không yêu cầu GPU chuyên dụng.

Bảng 8. Thời gian xử lý của các thành phần

Module	Processing time
MediaPipe Face Mesh	~15 ms
Tính toán EAR, MAR, Head Pose	~1 ms
Suy luận Bi-LSTM / Dense DNN	~3 ms
Tổng Inference Latency	~18–20 ms

4.6. So sánh với các nghiên cứu liên quan

Bảng 9 trình bày kết quả của phương pháp đề xuất cùng với một số nghiên cứu tiêu biểu về phát hiện buồn ngủ của người lái. Do các nghiên cứu sử dụng bộ dữ liệu, giao thức đánh giá và mục tiêu thực nghiệm khác nhau nên các kết quả chỉ mang tính tham khảo và không được xem là phép so sánh trực tiếp.

Bảng 9. So sánh với các nghiên cứu liên quan

Nghiên cứu	Đặc trưng	Phương pháp	Dữ liệu	Accuracy
Abtahi et al. (2011)	MAR	Rule-based	YawDD	81,0%
Wijnands et al. (2020)	Facial landmarks	3D CNN	NTHU-DDD	77,6%
Kim et al. (2023)	Eye landmarks + Head Pose	solvePnP + Landmark-based	IR-Dataset	> 99%

<i>Nghiên cứu</i>	<i>Đặc trưng</i>	<i>Phương pháp</i>	<i>Dữ liệu</i>	<i>Accuracy</i>
Nghiên cứu đề xuất	EAR + MAR + Head Pose	Bi-LSTM + DNN	Internal	98,14%
Nghiên cứu đề xuất	EAR + MAR + Head Pose	Bi-LSTM + DNN	Internal → NTHU-DDD	76,41%

Có thể thấy các nghiên cứu gần đây khai thác nhiều loại đặc trưng khuôn mặt để phát hiện trạng thái buồn ngủ của người lái. Abtahi et al. sử dụng đặc trưng ngập kết hợp phương pháp dựa trên luật; Wijnands et al. áp dụng mạng tích chập 3D nhằm khai thác thông tin không gian-thời gian trên thiết bị di động; trong khi Kim et al. kết hợp phát hiện trạng thái mắt và ước lượng tư thế đầu từ các điểm mốc khuôn mặt. Trong nghiên cứu này, hệ thống kết hợp đồng thời các đặc trưng EAR, MAR và tư thế đầu với mô hình Bi-LSTM và DNN, đồng thời tích hợp cơ chế hiệu chuẩn thích ứng và chiến lược bầu chọn đa số. Trên tập dữ liệu nội bộ, phương pháp đạt độ chính xác 98,14%. Khi đánh giá chéo theo kịch bản Internal → NTHU-DDD, độ chính xác đạt 76,41%, phản ánh ảnh hưởng của sự khác biệt giữa hai miền dữ liệu nhưng vẫn cho thấy mô hình có khả năng hoạt động trên dữ liệu chưa được sử dụng trong quá trình huấn luyện.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đề xuất một hệ thống phát hiện buồn ngủ của người lái dựa trên phân tích chuỗi đặc trưng khuôn mặt theo thời gian, trong đó trạng thái mất tập trung được sử dụng như một tín hiệu hỗ trợ nhằm tăng độ tin cậy. Thay vì xử lý độc lập từng khung hình, phương pháp đề xuất khai thác sự thay đổi liên tục của các đặc trưng EAR và MAR bằng mô hình Bi-LSTM, đồng thời kết hợp thông tin tư thế đầu để hỗ trợ phát hiện trạng thái mất tập trung. Bên cạnh đó, cơ chế hiệu chuẩn thích ứng và chiến lược bầu chọn đa số giúp nâng cao khả năng cá nhân hóa và cải thiện độ ổn định của hệ thống.

Kết quả thực nghiệm cho thấy hệ thống đạt hiệu năng khả quan trên tập dữ liệu nội bộ và duy trì tốc độ xử lý khoảng 25–30 FPS trên CPU phổ thông mà không yêu cầu GPU chuyên dụng.

Mặc dù đạt được những kết quả tích cực, nghiên cứu vẫn còn một số hạn chế như quy mô dữ liệu nội bộ còn nhỏ và chủ yếu được thu thập trong môi trường mô phỏng, chưa phản ánh đầy đủ các điều kiện vận hành thực tế như thay đổi ánh sáng, rung lắc phương tiện hoặc hiện tượng che khuất khuôn mặt. Do đó, mặc dù mô hình đạt kết quả khả quan trên tập dữ liệu nội bộ và các thí nghiệm đánh giá chéo, hiệu năng khi triển khai trong các điều kiện lái xe thực tế vẫn cần được kiểm chứng trên các bộ dữ liệu đa dạng hơn. Trong tương lai, nghiên cứu sẽ tập trung mở rộng đánh giá trên các bộ dữ liệu đa dạng hơn, tối ưu hóa mô hình cho các thiết bị nhúng và tích hợp thêm các nguồn

dữ liệu đa phương thức nhằm nâng cao độ tin cậy của hệ thống trong các ứng dụng hỗ trợ lái xe an toàn.

TÀI LIỆU THAM KHẢO

- [1]. Abtahi, S., Hariri, B., & Shirmohammadi, S. (2011). Driver drowsiness monitoring based on yawning detection. In *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC)* (pp. 1–4). IEEE.
<https://doi.org/10.1109/IMTC.2011.5944101>
- [2]. Alajlan, N. N., & Ibrahim, D. M. (2023). DDD TinyML: A TinyML-based driver drowsiness detection model using deep learning. *Sensors*, 23(12), 5696.
<https://doi.org/10.3390/s23125696>
- [3]. Albadawi, Y., Takruri, M., & Awad, M. (2022). A review of recent developments in driver drowsiness detection systems. *Sensors*, 22(5), 2069. <https://doi.org/10.3390/s22052069>
- [4]. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610. <https://doi.org/10.1016/j.neunet.2005.06.042>
- [5]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [6]. Jarndal, A., Tawfik, H., Siam, A. I., Alsayouf, I., & Cheaitou, A. (2025). A real-time vision transformers-based system for enhanced driver drowsiness detection and vehicle safety. *IEEE Access*, 13, 1790–1803. <https://doi.org/10.1109/ACCESS.2024.3522111>
- [7]. Kim, D., Park, H., Kim, T., Kim, W., & Paik, J. (2023). Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition. *Scientific Reports*, 13, 18264. <https://doi.org/10.1038/s41598-023-45114-1>
- [8]. Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision*, 81(2), 155–166.
<https://doi.org/10.1007/s11263-008-0152-6>
- [9]. Ly, N. T. H., Nguyen, H. T. H., & Kiet, H. T. (2025). Xây dựng hệ thống phát hiện buồn ngủ khi lái xe bằng phương pháp kết hợp YOLO và mạng CNN. Trong *Kỷ yếu Hội thảo khoa học Công nghệ Thông tin và Truyền thông năm 2025* (tr. 355–359). Nhà xuất bản Khoa học Tự nhiên và Công nghệ.
- [10]. National Highway Traffic Safety Administration. (2023). *Drowsy driving: Countermeasures that work*. U.S. Department of Transportation. Truy cập từ
<https://www.nhtsa.gov/book/countermeasures-that-work/drowsy-driving>
- [11]. Ouyang, Y., Yang, J., & Yang, M.-H. (2016). *NTHU Driver Drowsiness Detection Dataset*. National Tsing Hua University. Truy cập ngày 30 tháng 6 năm 2026 từ
<http://cv.cs.nthu.edu.tw/php/callforpaper/datasets/DDD>
- [12]. Soukupová, T., & Čech, J. (2016). Real-time eye blink detection using facial landmarks. Trong *Proceedings of the 21st Computer Vision Winter Workshop (CVWW 2016)* (tr. 1–8). Rimske Toplice, Slovenia.

- [13]. Wijnands, J. S., Thompson, J., Nice, K. A., Aschwanden, G. D. P. A., & Stevenson, M. (2020). Real-time monitoring of driver drowsiness on mobile platforms using 3D neural networks. *Neural Computing and Applications*, 32, 9731–9743. <https://doi.org/10.1007/s00521-019-04506-0>
- [14]. Zhang, F., Zhu, X., & Ye, M. (2020). MediaPipe Face Mesh: Real-time face landmark detection and tracking. *arXiv preprint arXiv:2006.10959*. <https://arxiv.org/abs/2006.10959>
- [15]. Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Face alignment across large poses: A 3D solution. Trong *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (tr. 146–155). Las Vegas, NV, USA.
- [16]. Doshi, A., & Trivedi, M. M. (2012). On the roles of eye gaze and head pose in predicting driver's intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems*, 13(3), 1265–1276. <https://doi.org/10.1109/TITS.2012.2185793>
- [17]. Zhao, Z., Xia, S., Xu, X., Zhang, L., Yan, H., Xu, Y., & Zhang, Z. (2020). Driver distraction detection method based on continuous head pose estimation. *Computational Intelligence and Neuroscience*, 2020, Article 9606908. <https://doi.org/10.1155/2020/9606908>

A REAL-TIME DRIVER DROWSINESS DETECTION SYSTEM BASED ON TEMPORAL FACIAL BEHAVIOR ANALYSIS

Luong Van Linh^{1*}, Nguyen Thi Huong Ly²

¹ Faculty of Information Technology, University of Sciences, Hue University

² Faculty of Information Technology, Nha Trang University

*Email: 250401013601@husc.edu.vn

ABSTRACT

Driver drowsiness detection is an important component of advanced driver assistance systems. However, many existing approaches still rely on fixed thresholds or frame-by-frame analysis, limiting their ability to capture the temporal characteristics of prolonged eye closure and yawning. This study presents a real-time driver drowsiness detection system based on temporal analysis of facial feature sequences. The system employs MediaPipe Face Mesh to extract the Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and head pose information. Head pose information is incorporated only as a complementary cue to improve the reliability of drowsiness detection. EAR and MAR sequences are modeled using a Bidirectional Long Short-Term Memory (Bi-LSTM) network to model temporal dependencies, while head pose is classified using a lightweight Deep Neural Network (DNN). To account for inter-user facial variations, the system employs an adaptive calibration mechanism and a majority voting strategy to reduce false alarms. Experimental results show that the proposed method achieves 98.14% accuracy on the internal dataset and 76.41% under the Internal → NTHU-DDD cross-dataset evaluation protocol, while meeting real-time processing requirements at approximately 25–30 FPS on a commodity CPU without dedicated GPU hardware.

Keywords: advanced driver assistance systems (ADAS); Bi-LSTM; driver drowsiness detection; facial behavior analysis; real-time processing.