**XÂY DƯNG MÔ HÌNH AI NHẬN DIỆN VÀ PHÂN LOẠI LỖI KỸ THUẬT TRONG ĐỘNG TÁC SQUAT SỬ DỤNG THỊ GIÁC MÁY TÍNH VÀ MÔ HÌNH HỌC SÂU**

**Lê Mậu Hải Đăng, Đỗ Hải Đăng, Lưu Trọng Hiếu, Nguyễn Đức Anh, Đặng Hoàng Long**

Học viện Công nghệ Bưu chính Viễn thông

[*DangLMH.B22AT083@stu.ptit.edu.vn*](mailto:DangLMH.B22AT083@stu.ptit.edu.vn)*,*[*DangDH.B22AT082@stu.ptit.edu.vn*](mailto:DangDH.B22AT082@stu.ptit.edu.vn)*,* [*AnhND.B22AT016@stu.ptit.edu.vn*](mailto:AnhND.B22AT016@stu.ptit.edu.vn)*,* [*HieuLT.B22AT119@stu.ptit.edu.vn*](mailto:HieuLT.B22AT119@stu.ptit.edu.vn)*,*

***Corresponding author: longdh@ptit.edu.vn***

***TÓM TẮT*** *— Nghiên cứu này trình bày phương pháp đơn giản và hiệu quả dựa trên ứng dụng của thị giác máy tính kết hợp với các mô hình học máy và mô hình học sâu nhằm nhận diện, phát hiện các lỗi phổ biến khi thực hiện động tác squat. Mô hình được huấn luyện trên bộ dữ liệu từ video tự quay và video trực tuyến, với mỗi video được chia thành các đoạn tập hoàn chỉnh và phân loại theo bốn nhóm lỗi chính. Phương pháp của chúng tôi là sử dụng pose estimation để trích xuất các khung xương và tính toán các góc khớp như góc gối, góc thân người và góc lưng sau đó đưa qua mô hình GRU (Gated Recurrent Unit) kết hợp với attention giúp nhận diện và phân loại các lỗi kỹ thuật. Kết quả thực nghiệm cho thấy mô hình có khả năng phát hiện các lỗi sai với độ chính xác cao, hỗ trợ người tập gym cải thiện kỹ thuật và giảm thiểu nguy cơ chấn thương trong quá trình tập luyện.*

***ABSTRACT —*** *This study presents a simple yet effective approach that leverages computer vision in combination with machine learning and deep learning models to detect and classify common mistakes during squat exercises. The model is trained on a dataset composed of both self-recorded and online videos, with each video segmented into complete repetitions and categorized into four main error types. Our method utilizes pose estimation to extract skeletal keypoints and compute joint angles such as knee angle, torso angle, and back angle, which are then fed into a GRU (Gated Recurrent Unit) model enhanced with attention mechanisms to identify and classify technical errors. Experimental results demonstrate that the model achieves high accuracy in detecting incorrect movements, thereby assisting gym-goers in improving their technique and reducing the risk of injury during workouts.*

**Từ khóa** *— Squat, thị giác máy tính, học máy, học sâu, khung xương, pose estimation, GRU, attention*

1. **GIỚI THIỆU**

Trong lĩnh vực thể dục thể thao, Squat là một bài tập điển hình, đặc biệt trong các bài tập phát triển cơ bắp chân và thân dưới. Tuy nhiên, việc thực hiện squat sai kỹ thuật có thể dẫn đến những chấn thương nghiêm trọng, ảnh hưởng đến quá trình tập luyện và sức khỏe người tập. Để tối ưu hóa hiệu quả bài tập và giảm thiểu rủi ro, việc phát hiện và sửa chữa các lỗi sai trong quá trình thực hiện squat là vô cùng quan trọng. Các lỗi sai phổ biến trong squat bao gồm: lỗi nghiêng người quá mức (excessive lean), độ sâu không đủ (shallow depth), và gù lưng (roundback). Việc phát hiện chính xác những lỗi này giúp người tập điều chỉnh tư thế, cải thiện kỹ thuật và giảm thiểu nguy cơ chấn thương.

Hiện nay, công nghệ nhận diện hình ảnh và các mô hình học máy, học sâu đã được áp dụng rộng rãi trong nhiều lĩnh vực, trong đó có thể thao. Các nghiên cứu trước đây đã chỉ ra rằng, việc sử dụng các công nghệ này có thể giúp phát hiện các lỗi kỹ thuật trong các bài tập thể thao dựa trên phân tích các khớp và tư thế của người tập. Việc áp dụng các mô hình học máy như Decision Tree, mô hình học sâu như CNN (Convolutional Neural Network), RNN (Recurrent Neural Network)[6] và mạng nơ-ron hồi tiếp dài hạn LSTM (Long Short-Term Memory) đã mang lại kết quả khả quan trong việc nhận diện các vấn đề kỹ thuật trong tập luyện. Một nghiên cứu trước đây đã sử dụng mô hình LSTM để nhận diện hoạt động con người từ dữ liệu cảm biến [1]. Mô hình này sau khi huấn luyện có độ chính xác lên tới 94%; tuy nhiên, quá trình huấn luyện diễn ra khá lâu với 500 epochs. Do đó, bài báo này đã đề xuất một giải pháp mới cho vấn đề nhận diện, phân tích hành động của con người nhằm phát hiện các lỗi sai kỹ thuật trong quá trình hoạt động thể chất. Ngoài ra, việc sử dụng các công nghệ như pose estimation còn giúp tách biệt rõ các yếu tố hình học trong động tác, tạo điều kiện cho mô hình học sâu khai thác các đặc trưng theo thời gian một cách hiệu quả hơn. Điều này đặc biệt hữu ích trong các bài tập có tính tuần hoàn như squat, nơi mà sự thay đổi góc và biên độ chuyển động là yếu tố then chốt để phân biệt tư thế đúng và sai.

Nghiên cứu này nhằm xây dựng một mô hình sử dụng mô hình GRU kết hợp với attention[10,11] để phát hiện các lỗi sai trong bài tập squat sau khi trích xuất được thông tin về các góc khớp bằng phương pháp pose estimation. Mô hình này sử dụng các chỉ số như góc gối, góc thân người và góc lưng để phân loại các lỗi như excessive lean, shallow depth, và roundback. Bộ dữ liệu sẽ được xây dựng từ các video quay lại các động tác squat, kết hợp với các kỹ thuật tăng cường dữ liệu (data augmentation) như xoay, lật video và điều chỉnh tốc độ để làm phong phú thêm bộ dữ liệu. Mặc dù có nhiều mô hình hiện đại như LSTM (Long Short-Term Memory), TCN (Temporal Convolutional Network), chúng tôi chọn sử dụng GRU[9,12] là vì một số lý do sau. Đầu tiên, số lượng dữ liệu là một vấn đề lớn của chúng tôi. LSTM yêu cầu một lượng tài nguyên lớn về dữ liệu; trong khi đó GRU không đòi hỏi quá nhiều dữ liệu. Thứ hai, GRU có cấu trúc đơn giản, ít tham số hơn LSTM[7], giúp tăng tốc độ huấn luyện mà vẫn giữ được khả năng học các mối quan hệ dài hạn. Cuối cùng, mô hình TCN có thể yêu cầu số lượng tham số lớn và phức tạp trong việc điều chỉnh các lớp chập để đạt hiệu quả tốt nhất. Điều này là một nhược điểm của mô hình này so với GRU. Vậy nên chúng tôi chọn GRU kết hợp với Attention, một cơ chế có thể xác định các bước quan trọng hơn và chỉ ra mô hình cần tập trung vào các bước đó. Qua đó, phương pháp của chúng tôi có thể đạt độ chính xác cao trong quá trình huấn luyện mô hình.

Phần còn lại của bài báo có cấu trúc như sau: trước khi đưa ra giải pháp cụ thể ở Phần III sẽ là một số công trình nghiên cứu có liên quan ở Phần II. Sau đó, chúng tôi sẽ trình bày kết quả thực nghiệm ở Phần IV, và cuối cùng đưa ra kết luận cho nghiên cứu ở Phần V.

1. **CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN**

***A. Các phương pháp phát hiện tư thế***

Hiện nay, nhiều hệ thống phân tích tư thế người dùng sử dụng mạng nơ-ron tích chập (CNN)[2,13] kết hợp với mô hình nhận diện điểm mốc như BlazePose từ Mediapipe hoặc OpenPose[8,14]. Các mô hình này áp dụng kỹ thuật ước lượng keypoint với 33 điểm giải phẫu học quan trọng (vai, đầu gối, khuỷu tay, v.v.) nhằm hỗ trợ đếm số lần tập và đánh giá động tác.Một số hệ thống web-based còn tích hợp các kiến trúc nhẹ như BlazeFace, BlazePalm để nhận diện khuôn mặt và bàn tay, đồng thời sử dụng heatmap và hồi quy để dự đoán chính xác toạ độ. Dù các mô hình CNN này cho phép xử lý hơn 20 FPS và hoạt động gần thời gian thực, chúng vẫn dựa trên xử lý theo từng khung hình riêng lẻ, thiếu thông tin chuỗi thời gian, dẫn đến việc đánh giá động tác đôi khi chưa ổn định khi người dùng chuyển động nhanh hoặc bị khuất hình.

Bên cạnh đó, các mô hình học máy truyền thống như Random Forest[3, 4] cũng từng được ứng dụng trong bài toán phân loại tư thế nhờ khả năng huấn luyện nhanh và dễ triển khai. Tuy nhiên, bản chất của Random Forest là dựa vào các đặc trưng tĩnh tại từng thời điểm, do đó không tận dụng được mối quan hệ theo chuỗi trong chuyển động cơ thể. Mặt khác, mô hình này hoạt động kém hiệu quả trên dữ liệu chiều cao hoặc thưa (như dữ liệu skeleton), dễ bị ảnh hưởng bởi các đặc trưng phân loại có nhiều cấp độ, và thiếu tính thích nghi trong môi trường thay đổi liên tục. Những hạn chế này khiến Random Forest khó có thể đảm bảo độ chính xác và độ mượt cần thiết trong các bài toán theo dõi tư thế người dùng thực tế.

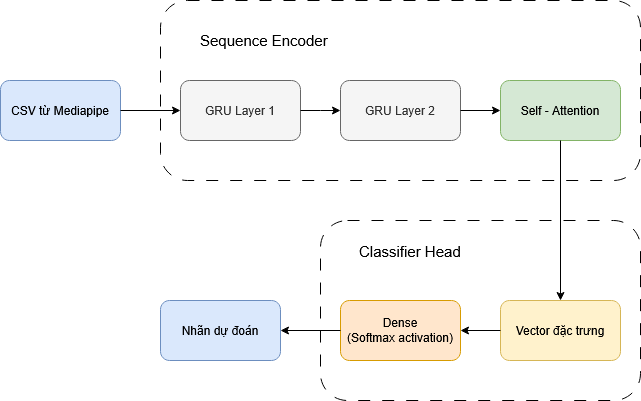
Mạng nơ-ron hồi tiếp LSTM (Long Short-Term Memory)[5] là một biến thể cải tiến của RNN, được thiết kế nhằm giải quyết các bài toán chuỗi có mối liên hệ dài hạn. Trong bài toán phân loại chuyển động, LSTM có khả năng ghi nhớ thông tin từ nhiều khung hình liên tiếp, giúp mô hình học được đặc trưng động học thay vì chỉ dựa vào tư thế tĩnh tại một thời điểm. Khi áp dụng vào dữ liệu xương (skeleton data) thu thập từ cảm biến Kinect, LSTM có thể học được sự thay đổi không gian–thời gian của các khớp trên cơ thể, từ đó tăng độ chính xác trong việc nhận dạng hành động. Tuy nhiên, một số nhược điểm vẫn tồn tại: LSTM có thời gian huấn luyện tương đối dài, đòi hỏi lượng dữ liệu lớn để tránh overfitting, đồng thời gặp khó khăn khi xử lý các chuỗi có độ dài không đồng nhất nếu không có bước chuẩn hóa đầu vào như padding. Ngoài ra, mô hình cũng khá nhạy cảm với việc lựa chọn siêu tham số (số tầng, số đơn vị ẩn, learning rate...), và hiệu suất có thể giảm đáng kể nếu không được tối ưu đúng cách. Những yếu tố này khiến LSTM tuy hiệu quả nhưng vẫn cần đánh đổi về mặt tính toán và tài nguyên.

***B. Ứng dụng của AI trong hỗ trợ thể thao***

Trí tuệ nhân tạo (AI) đang ngày càng được ứng dụng rộng rãi trong lĩnh vực thể thao nhằm cải thiện hiệu suất tập luyện, giảm chấn thương và cung cấp phản hồi cá nhân hóa cho người dùng. Các hệ thống phân tích chuyển động dựa trên thị giác máy tính giúp đánh giá tư thế của vận động viên, từ đó đưa ra gợi ý điều chỉnh kỹ thuật. Một số nghiên cứu gần đây đã tích hợp AI vào thiết bị đeo hoặc camera thông minh, cho phép ghi lại chuyển động và đánh giá theo thời gian thực. Không chỉ dừng lại ở phân tích tư thế, AI còn hỗ trợ xây dựng giáo án tập luyện cá nhân hoá, nhận diện mệt mỏi qua biểu cảm hoặc tư thế để từ đó đưa ra các quyết định giống như một huấn luyện viên thực thụ. Một trong những hướng ứng dụng nổi bật khác là sử dụng AI để phát hiện lỗi kỹ thuật ngay trong thời gian thực, từ đó phát tín hiệu cảnh báo hoặc hướng dẫn sửa tư thế. Điều này giúp người tập điều chỉnh kịp thời mà không cần huấn luyện viên trực tiếp, đặc biệt hữu ích trong bối cảnh tập luyện cá nhân tại nhà ngày càng phổ biến sau đại dịch COVID-19.

1. **PHƯƠNG PHÁP**

Chúng tôi đề xuất một phương pháp phân loại tư thế squat dựa trên mô hình học sâu kết hợp giữa GRU và cơ chế attention. Mô hình được thiết kế để xử lý chuỗi dữ liệu pose trích xuất từ video, với mục tiêu phân loại động tác squat thành tư thế đúng hoặc một trong ba dạng sai khác nhau. Cụ thể, mô hình sử dụng lớp GRU để học các đặc trưng theo thời gian từ chuỗi pose đầu vào. Sau đó, cơ chế attention được áp dụng nhằm tập trung vào các bước thời gian quan trọng trong chuỗi, giúp mô hình làm nổi bật các khung hình có ý nghĩa trong việc nhận diện sai sót về tư thế. Việc kết hợp GRU với attention cho phép mô hình vừa nắm bắt được động học của chuyển động, vừa học được các mối quan hệ không tuyến tính giữa các bước thời gian, từ đó nâng cao hiệu quả phân loại.



***Mô hình phân loại tư thế squat sử dụng GRU + attention***

Phần này trình bày chi tiết quy trình huấn luyện cơ bản của GRU + attention để phân loại các tư thế squat

1. *Tiền xử lí dữ liệu*

Đầu vào của mô hình là chuỗi thời gian đại diện cho một lần thực hiện động tác squat, được trích xuất từ video góc ngang (bên trái hoặc phải). Mỗi video được chuẩn hóa về độ dài cố định 100 frame bằng kỹ thuật co giãn thời gian. Tại mỗi frame, ba góc chính liên quan đến khớp gối, hông và vai được tính toán từ tọa độ khớp thu được bằng thư viện MediaPipe. Các góc này phản ánh đặc trưng hình thái quan trọng của tư thế squat và được chuẩn hóa về khoảng [0, 1] bằng phương pháp min-max scaling. Với các chuỗi ngắn hơn 100 frame, kỹ thuật padding bằng giá trị 0 được áp dụng để đảm bảo tính đồng nhất đầu vào. Dữ liệu đầu ra sau xử lý có dạng tensor 2 chiều với kích thước (100, 3), trong đó mỗi hàng tương ứng với một frame và ba cột là ba góc chuyển động chính – đóng vai trò đầu vào trực tiếp cho mô hình GRU kết hợp attention.

1. *Kiến trúc mô hình*

Mô hình đề xuất sử dụng kiến trúc mạng nơ-ron hồi tiếp có cổng (Gated Recurrent Unit – GRU) kết hợp với cơ chế attention, nhằm khai thác hiệu quả các đặc trưng theo thời gian từ chuỗi góc chuyển động được trích xuất từ video động tác squat. Đầu vào của mô hình là chuỗi thời gian có kích thước cố định (100, 3), trong đó 100 là số lượng frame, và 3 là ba góc đặc trưng được tính toán tại mỗi frame. Trước khi đưa vào GRU, dữ liệu được xử lý qua lớp Masking để loại bỏ ảnh hưởng của các giá trị padding (0.0), đảm bảo mô hình chỉ học từ các phần tử thực. Tiếp theo, hai lớp GRU liên tiếp, mỗi lớp gồm 64 đơn vị ẩn, được sử dụng để trích xuất đặc trưng động theo chuỗi thời gian. Đầu ra tại tất cả các thời điểm từ lớp GRU cuối được giữ lại, cho phép cơ chế attention hoạt động trên toàn bộ chuỗi thay vì chỉ dựa vào đầu ra tại thời điểm cuối.

Cơ chế Attention được áp dụng nhằm học phân phối trọng số thích hợp trên các frame, từ đó làm nổi bật các thời điểm chứa thông tin đặc trưng quan trọng. Đầu ra của attention được kết hợp với đầu ra GRU bằng phép cộng phần tử (residual), tạo nên một biểu diễn ngữ cảnh giàu thông tin hơn. Toàn bộ chuỗi đầu ra sau đó được rút gọn thành một vector duy nhất thông qua lớp Global Average Pooling. Vector này được đưa qua hai lớp fully-connected, trong đó lớp cuối sử dụng hàm kích hoạt softmax để phân loại tư thế thành một trong bốn lớp: perfect, not deep enough, rounded back và excessive lean. Nhờ vào kiến trúc đơn giản và có khả năng khai thác đặc trưng động học mạnh mẽ, mô hình GRU + Attention không chỉ đạt hiệu quả trong việc phân loại mà còn đảm bảo tốc độ huấn luyện nhanh, phù hợp với các hệ thống có giới hạn về tài nguyên tính toán. Điều này đặc biệt hữu ích trong các ứng dụng triển khai thực tế, nơi mà thời gian phản hồi và độ nhẹ của mô hình là yếu tố quan trọng.

1. *Ưu điểm so với các phương pháp đi trước*

Mô hình GRU kết hợp Attention được lựa chọn nhằm khắc phục hạn chế của các phương pháp như Random Forest, CNN hay LSTM thuần. GRU có cấu trúc gọn nhẹ, giúp mô hình học được sự thay đổi liên tục của tư thế theo thời gian mà không yêu cầu tài nguyên tính toán cao. Việc tích hợp Attention cho phép mô hình tập trung vào các khung hình quan trọng nhất trong một rep squat, từ đó nâng cao độ chính xác phân loại. Nhờ đó, GRU + Attention trở thành giải pháp hiệu quả, cân bằng giữa độ chính xác, tốc độ xử lý và khả năng ứng dụng thực tế trong nhận diện tư thế.

1. **THỰC NGHIỆM**
2. **Dữ liệu**
3. *Thu thập dữ liệu*

Dữ liệu được thu thập từ nhiều nền tảng trực tuyến phổ biến như YouTube, TikTok, Instagram và Reddit, nhằm đảm bảo sự đa dạng về nguồn gốc, điều kiện ánh sáng, tốc độ cũng như cách thể hiện động tác squat. Tuy nhiên, do dữ liệu từ Internet thường có chất lượng không đồng đều hoặc thiếu nhãn rõ ràng, nhóm cũng tiến hành tự quay thêm các video thực hiện động tác squat trong điều kiện kiểm soát. Khối lượng dữ liệu tự quay chiếm tỷ lệ lớn hơn và đóng vai trò chính trong quá trình huấn luyện, giúp đảm bảo chất lượng và độ tin cậy của bộ dữ liệu.

Để phục vụ cho việc đánh giá kỹ thuật động tác squat một cách chính xác, chỉ các video được quay từ góc nhìn bên trái hoặc bên phải của người tập (góc ngang) mới được lựa chọn. Các góc quay chính diện hoặc từ phía sau bị loại bỏ do không cung cấp đủ thông tin hình học cần thiết để quan sát rõ các yếu tố như độ nghiêng thân trên, độ cong cột sống hay độ sâu của động tác. Ngoài ra, các video bị rung lắc, thiếu sáng hoặc có vật cản che khuất cơ thể người tập cũng được loại bỏ trong giai đoạn tiền xử lý nhằm đảm bảo đầu vào phù hợp cho các bước trích xuất pose.

Tất cả các video sau khi được thu thập sẽ được xử lý thủ công để cắt ngắn lại, sao cho mỗi video chỉ chứa duy nhất một lần thực hiện động tác squat (tương ứng với một rep). Việc rút gọn này nhằm mục tiêu tạo ra một bộ dữ liệu rõ ràng, trong đó mỗi đơn vị dữ liệu tương ứng với một hành động squat riêng biệt, phục vụ cho việc huấn luyện và đánh giá mô hình một cách chính xác hơn. Ngoài việc tách rep, nhóm cũng đồng thời kiểm tra xem chuỗi chuyển động trong mỗi đoạn có đầy đủ giai đoạn bắt đầu, hạ người, và đứng dậy hay không trước khi đưa vào pipeline xử lý tiếp theo.

1. *Gán nhãn dữ liệu*

Để phân loại tư thế squat trong bài toán này, chúng tôi sử dụng bốn nhãn chính: not deep enough, rounded back, excessive lean, và perfect. Việc gán nhãn này dựa trên các góc được tính toán từ các điểm đặc trưng của cơ thể người tập, được trích xuất từ video sử dụng Mediapipe. Các góc được xác định giữa các khớp chính như đầu gối, hông, vai, và cổ để đánh giá độ chính xác của tư thế squat. Not deep enough được gán khi góc gập của đầu gối quá lớn (lớn hơn 120°), cho thấy người tập chưa thực hiện động tác squat đủ sâu. Tư thế này làm giảm hiệu quả của bài tập và không đảm bảo mức độ phát triển cơ bắp tối đa. Rounded back là nhãn dành cho các trường hợp người tập có lưng cong trong khi thực hiện squat, được phát hiện qua các góc liên quan đến các điểm tại hông, vai và mũi. Tư thế này rất nguy hiểm, dễ dẫn đến chấn thương cột sống nếu không được điều chỉnh kịp thời. Excessive lean được gán khi góc giữa thân người (góc giữa vai, hông và đầu gối) quá nhỏ, thể hiện rằng người tập có xu hướng nghiêng người quá mức về phía trước trong khi squat. Điều này có thể làm mất đi tính hiệu quả của động tác và gây căng thẳng không cần thiết cho lưng dưới.

Cuối cùng, perfect là nhãn dành cho các tư thế squat đạt yêu cầu, khi góc gập của các khớp đạt mức lý tưởng (thường từ 90° đến 120° đối với góc gối), và lưng giữ thẳng, không có hiện tượng nghiêng quá hoặc cong lưng. Đây là tư thế squat an toàn và hiệu quả. Ngoài các góc cơ thể, chúng tôi còn sử dụng tọa độ z của các khớp để xác định góc quay của người tập. Cụ thể, nếu tọa độ z của khớp vai phải nhỏ hơn tọa độ z của vai trái, điều này chỉ ra rằng camera đã được đặt ở bên phải người tập. Việc này giúp đảm bảo độ chính xác trong việc xác định vị trí và hướng của người tập trong video, hỗ trợ quá trình gán nhãn chính xác hơn.

1. *Tăng cường dữ liệu*

Để mở rộng quy mô và tăng độ đa dạng cho tập dữ liệu huấn luyện, chúng tôi áp dụng một số kỹ thuật tăng cường dữ liệu (data augmentation) lên toàn bộ video sau khi gán nhãn. Cụ thể, hai phép biến đổi đầu tiên bao gồm điều chỉnh màu sắc (color jitter) và thêm nhiễu Gaussian nhẹ (noise injection), giúp mô phỏng các điều kiện ánh sáng và chất lượng video khác nhau. Sau đó, toàn bộ video đã được biến đổi tiếp tục được áp dụng phép lật ngang (horizontal flip), nhằm tạo ra các mẫu phản chiếu hình học hợp lệ.

Nhờ quy trình tăng cường này, tổng số mẫu video sau xử lý đạt 3036, với phân bố cụ thể theo từng nhãn như sau: 906 mẫu perfect, 762 mẫu excessive lean, 756 mẫu not deep enough và 612 mẫu rounded back. Các nhãn này được giữ nguyên từ video gốc, đảm bảo tính nhất quán trong gán nhãn sau tăng cường. Việc đa dạng hóa dữ liệu bằng các phép biến đổi hình ảnh không chỉ giúp tăng kích thước tập huấn luyện mà còn cải thiện khả năng khái quát của mô hình khi gặp phải các điều kiện ánh sáng, góc quay hay hình thể người tập khác nhau trong môi trường thực tế. Điều này đóng vai trò quan trọng trong việc nâng cao độ bền và hiệu suất tổng thể của hệ thống nhận diện.

1. **Chi tiết Thực hiện**
2. *Chi tiết quá trình huấn luyện*

Trong quá trình huấn luyện, chúng tôi sử dụng 900 mẫu nhãn perfect và 500 mẫu cho mỗi nhãn còn lại, được trích từ tập đã tăng cường. Tập dữ liệu này được chia thành 80% để huấn luyện và 20% để đánh giá, đồng thời giữ cân bằng phân bố nhãn giữa các tập con. Ngoài ra, chúng tôi cũng áp dụng cơ chế early stopping dựa trên độ chính xác tập validation để ngăn mô hình huấn luyện quá mức, đồng thời sử dụng dropout nhằm giảm thiểu nguy cơ overfitting. Các siêu tham số như learning rate, batch size và số epoch được lựa chọn thông qua các thử nghiệm nhỏ, đảm bảo quá trình huấn luyện ổn định và hiệu quả.

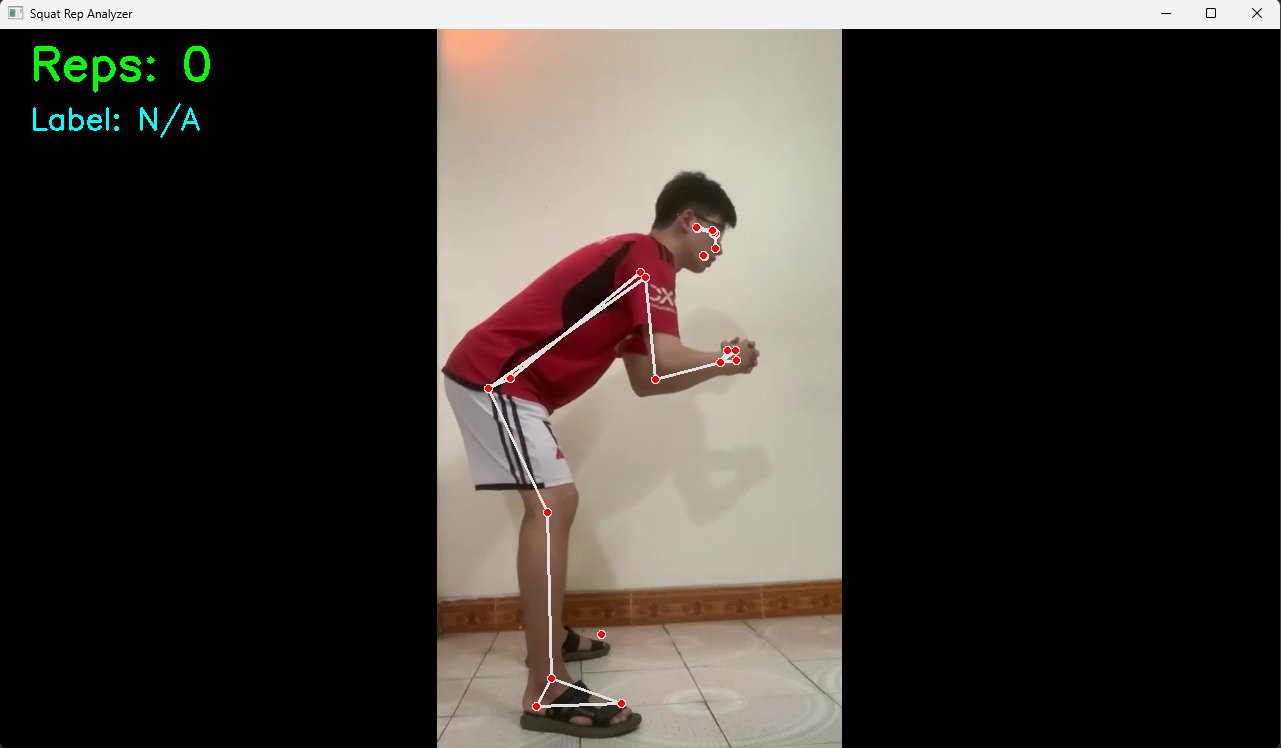
Mô hình được đề xuất có cấu trúc tuần tự, kết hợp giữa mạng hồi tiếp có cổng (Gated Recurrent Unit – GRU) và cơ chế attention, nhằm khai thác hiệu quả đặc trưng động từ chuỗi góc chuyển động trong các đoạn video thực hiện động tác squat. Chuỗi đầu vào gồm ba đặc trưng góc học được tại mỗi khung hình, được đưa qua hai tầng GRU liên tiếp, cho phép mô hình học được quan hệ phụ thuộc theo thời gian ở cả mức thấp và mức khái quát cao hơn. Cơ chế attention sau đó được tích hợp để giúp mô hình tập trung vào những thời điểm có giá trị phân biệt cao trong chuỗi, từ đó tăng cường khả năng nhận diện tư thế. Biểu diễn chuỗi thu được sau attention được rút gọn thành một vector toàn cục, rồi đưa qua các tầng kết nối đầy đủ (fully-connected) để thực hiện phân loại. Tầng cuối sử dụng hàm kích hoạt softmax để xác định tư thế đầu vào thuộc một trong bốn lớp: perfect, not deep enough, rounded back, hoặc excessive lean. Kiến trúc được thiết kế nhằm đạt được sự cân bằng giữa hiệu quả biểu diễn, độ chính xác và khả năng triển khai trên các hệ thống tính toán với tài nguyên hạn chế.

1. **Chỉ số đánh giá**

Số lượng tham số là một chỉ số quan trọng trong việc đánh giá độ phức tạp của mô hình học máy. Số lượng tham số đại diện cho tổng số trọng số và độ dịch chuyển mà mô hình cần học trong quá trình huấn luyện. Một mô hình với số lượng tham số cao có thể học được nhiều đặc trưng phức tạp hơn từ dữ liệu, nhưng cũng đồng nghĩa với việc tăng chi phí tính toán và khả năng bị overfitting (quá khớp) nếu không được điều chỉnh đúng cách. Do đó, việc tối ưu hóa số lượng tham số là rất quan trọng trong việc phát triển mô hình hiệu quả.

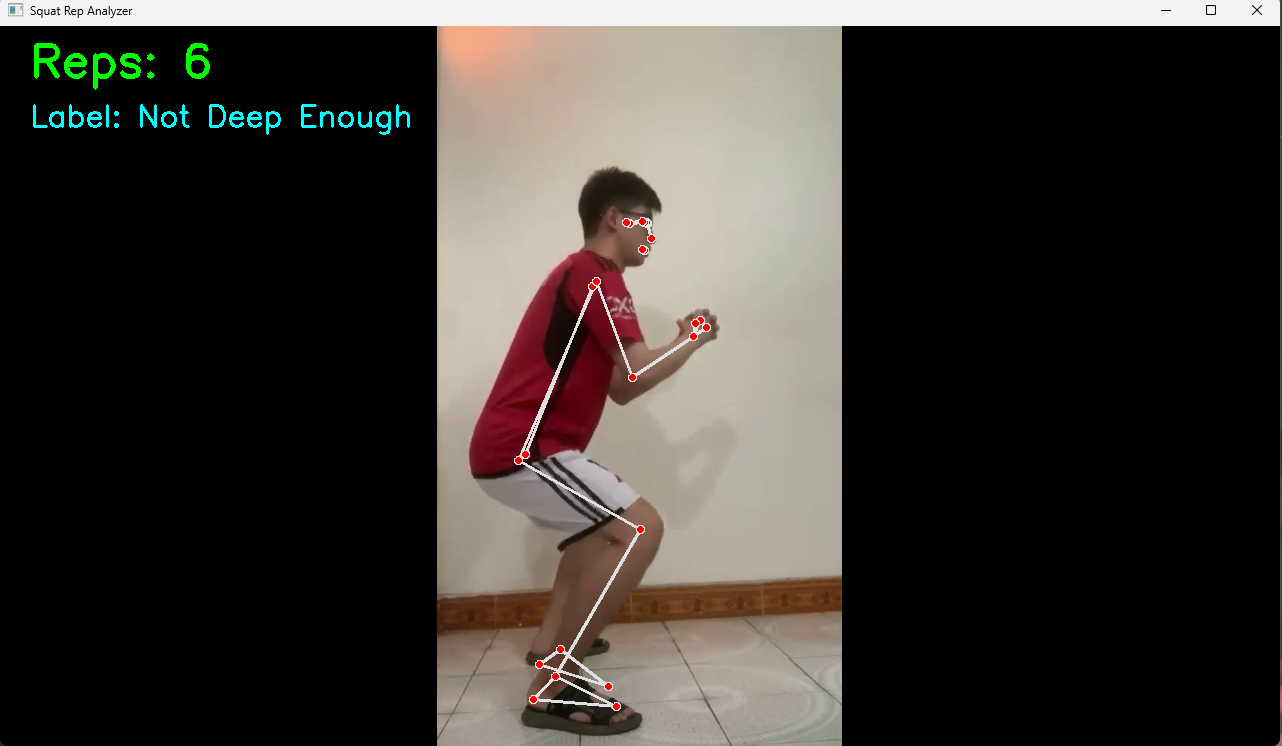
Accuracy là một trong những chỉ số đánh giá phổ biến nhất để đo lường hiệu suất của mô hình phân loại. Accuracy tính toán tỷ lệ dự đoán chính xác của mô hình, tức là tỷ lệ số dự đoán đúng so với tổng số dự đoán. Chỉ số này phản ánh khả năng của mô hình trong việc phân loại đúng các đối tượng trong tập dữ liệu kiểm tra. Một mô hình có độ chính xác cao cho thấy khả năng phân biệt các lớp đối tượng tốt, trong khi một mô hình có độ chính xác thấp có thể gặp vấn đề với việc phân loại sai hoặc không đủ khả năng nhận diện đúng đối tượng.

1. **Ứng dụng**



**Hình 1**

Hình 1 minh họa giao diện hệ thống huấn luyện viên ảo ở trạng thái ban đầu khi người dùng chưa bắt đầu thực hiện động tác squat. Lúc này, bộ đếm số rep hiển thị giá trị bằng 0 và nhãn chất lượng được gán là “N/A” – biểu thị rằng hệ thống chưa xác định được chất lượng do chưa phát hiện chu trình squat hoàn chỉnh. Dữ liệu khung xương được trích xuất liên tục qua mỗi khung hình, tuy nhiên hệ thống chỉ kích hoạt đánh giá khi người dùng bắt đầu thực hiện chuyển động rõ ràng với các điều kiện về biên độ và trình tự động tác được đảm bảo. Trạng thái “N/A” không chỉ đóng vai trò làm giá trị mặc định ban đầu, mà còn giúp đảm bảo hệ thống không sinh ra nhãn sai trong giai đoạn người dùng đang chuẩn bị hoặc điều chỉnh tư thế. Chỉ khi hệ thống nhận diện được một chuỗi chuyển động đầy đủ với đủ biên độ và tốc độ đặc trưng, quy trình đánh giá mới được kích hoạt. Điều này giúp giảm thiểu các cảnh báo nhiễu và tăng độ tin cậy của phản hồi trong môi trường sử dụng thực tế.



**Hình 2**

Hình 2 là ví dụ về quá trình demo của hệ thống khi người dùng thực hiện động tác squat. Tại thời điểm này, hệ thống đã ghi nhận được 6 rep hợp lệ và gán nhãn “Không đủ sâu” (Not Deep Enough) cho rep hiện tại. Mỗi lần squat được xác định dựa trên một chu kỳ chuyển động đầy đủ và được gán nhãn tương ứng thông qua mô hình học máy đã huấn luyện. Khung xương hiển thị trên video giúp người dùng dễ dàng nhận diện sai sót tư thế, trong khi thông tin rep và nhãn được trình bày trực quan ở phía trái màn hình, hỗ trợ người tập điều chỉnh kỹ thuật trong thời gian thực. Về lâu dài, hệ thống có thể được tích hợp vào các thiết bị thông minh như gương tập gym, camera AI hoặc ứng dụng di động nhằm cung cấp phản hồi tư thế theo thời gian thực. Khả năng mở rộng sang các bài tập khác như deadlift hoặc overhead press cũng hoàn toàn khả thi nếu bổ sung thêm dữ liệu và tái huấn luyện mô hình. Điều này mở ra hướng phát triển thành một huấn luyện viên ảo đa năng, hỗ trợ người dùng luyện tập an toàn và hiệu quả ngay tại nhà.

1. **Phân tích và đánh giá**

Chúng tôi tiến hành so sánh hiệu suất của mô hình GRU kết hợp attention được đề xuất với một tập hợp các kiến trúc học sâu phổ biến, bao gồm cả các mô hình baseline đơn giản như MLP, GRU và LSTM, cùng với các kiến trúc nâng cao hơn như BiLSTM, CNN + LSTM và LSTM + attention. Kết quả chi tiết được trình bày trong Bảng 1.

Mô hình GRU + attention đạt độ chính xác cao nhất với 88.17%, cho thấy khả năng khai thác thông tin theo thời gian hiệu quả hơn so với các mô hình còn lại. So với các baseline đơn giản, mô hình này vượt trội rõ rệt: cao hơn GRU thuần gần 10% (78.63%), LSTM hơn 7.1% (81.06%) và MLP đến gần 14% (74.33%). Điều này khẳng định rằng việc tích hợp cơ chế attention giúp mô hình tập trung vào các frame quan trọng, nâng cao khả năng phân biệt giữa các tư thế squat.

Đối với các mô hình nâng cao, GRU + attention vẫn duy trì hiệu suất dẫn đầu. Cụ thể, LSTM + attention đạt 83.82%, BiLSTM đạt 80.29%, và CNN + LSTM đạt 80.57%. Dù các mô hình này có khả năng biểu diễn mạnh và sâu hơn, nhưng vẫn chưa đạt đến mức hiệu quả như GRU + attention. Điều này phần nào cho thấy rằng trong các bài toán nhận diện chuỗi ngắn như tư thế squat (100 frame), mô hình GRU — với kiến trúc gọn nhẹ hơn — vừa đủ để nắm bắt đặc trưng chuỗi mà vẫn giữ được tốc độ hội tụ nhanh và tính tổng quát cao.

Tóm lại, GRU kết hợp attention là một kiến trúc hiệu quả, cân bằng tốt giữa độ chính xác, chi phí huấn luyện và khả năng triển khai thực tế. Mô hình này thể hiện ưu thế rõ rệt so với cả các baseline cơ bản và các kiến trúc nâng cao, là lựa chọn phù hợp trong bối cảnh nhận diện động tác ngắn dựa trên dữ liệu thời gian. Ngoài độ chính xác cao, kiến trúc này còn mang lại lợi thế về tốc độ hội tụ trong huấn luyện, ít yêu cầu tuning phức tạp và dễ tích hợp vào các pipeline xử lý thời gian thực. Do vậy, GRU + Attention không chỉ phù hợp về mặt kỹ thuật mà còn mang tính ứng dụng cao trong các hệ thống giám sát tư thế tập luyện ngoài đời thực.

**Bảng 1.** Kết quả thực nghiệm trên tập test

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| MLP | 74.33% |
| GRU only | 78.63% |
| LSTM only | 81.06% |
| CNN + LSTM | 80.57% |
| BiLSTM | 80.29% |
| LSTM + Attention | 83.82% |
| **GRU + Attention** | **88.17%** |

1. **KẾT LUẬN**

Trong bài báo này, chúng tôi trình bày một phương pháp mới nhằm cải thiện hiệu suất của mô hình phát hiện lỗi sai khi thực hiện động tác squat bằng cách kết hợp tri thức từ các mô hình học sâu chuyên dụng trong nhận dạng chuyển động và hình ảnh. Bằng cách sử dụng thị giác máy tính và mô hình học sâu các vùng quan sát chuyển động và vùng dữ liệu chuẩn, phương pháp của chúng tôi đã đạt được kết quả khả quan trong việc phát hiện các lỗi sai khi squat. Đặc biệt, mô hình của chúng tôi cho thấy độ chính xác vượt trội so với các phương pháp hiện tại trong việc nhận diện các sai sót tư thế. Hướng phát triển trong tương lai của nghiên cứu này là mở rộng mô hình bằng cách sử dụng các mạng học sâu và các cơ sở dữ liệu lớn hơn để nâng cao độ chính xác và khả năng phát hiện đa dạng các lỗi sai khi thực hiện động tác squat. Kết quả này khẳng định tiềm năng ứng dụng của mô hình trong việc hỗ trợ huấn luyện viên ảo, đặc biệt trong các phòng tập gym hiện đại. Với chi phí triển khai thấp và khả năng thích ứng cao, hệ thống hoàn toàn có thể tích hợp vào các thiết bị di động hoặc nền tảng web nhằm phục vụ người dùng cuối.

1. **REFERENCES**

[1] Schalk Wilhelm Pienaar, Reza Malekian, Human Activity Recognition Using LSTM-RNN Deep Neural Network Architecture https://arxiv.org/abs/1905.00599.language knowledge distillation. International Conference on Learning Representations, 2022

[2]Mrs. P. Swathi, R. Pavani Pranathi, K. S. G. Sai Rohith, M. Cathey Sagar, S. Jayavardhan, AI-BASED FITNESS TRAINER APPLICATION

[3] Breiman (2001), "Random Forests", Machine Learning 45(1), trang 5-32.

[4]Mrs. K. Prathyusha, T.Lakshmi Narayana, P.Vijay Kumar, R.Venkatesh, PERSONAL AI TRAINER

[5]Yi Wang, Xiaowen Zhu and Chengzhang Qu, Fitness Movements Recognition and Evaluation Based on LSTM

[6] Kang, J., Kim, S., & Choi, C. (2018). Human activity recognition using recurrent neural networks. Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), 1826-1832

[7] Shi, Y., & Wu, Y. (2018). Human Action Recognition using LSTM and GRU. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8]Cao et al. (2017), "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields"

[9] Mohsen, S. (2023). Human Activity Recognition Using GRU Deep Learning Algorithm

[10] Haque, M. N., Tonmoy, M. T. H., Mahmud, S., Ali, A. A., Khan, M. A. H., & Shoyaib, M. (2019). GRU-based attention mechanism for human activity recognition. In Proceedings of the 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT 2019)

[11] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2017). An End‑to‑End Spatio‑Temporal Attention Model for Human Action Recognition from Skeleton Data. Proceedings of the AAAI Conference on Artificial Intelligence.

[12] Wang, J., Zhu, X., & Qu, C. (2018). Fitness Movements Recognition and Evaluation Based on GRU.

[13] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale Video Classification with Convolutional Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1725–1732

[14] Bazarevsky, V., Kartynnik, Y., Vakunov, A., Tkachenka, A., & Grundmann, M. (2020). BlazePose: On-device Real-time Body Pose Tracking.