TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 22, Số 11 (2025): 1980-1992
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 22, No. 11 (2025): 1980-1992
https://doi.org/10.54607/hcmue.js.22.11.5160(2025)

Bài báo nghiên cứu

FUSED-A: MÔ HÌNH ĐA LUỒNG DỰA TRÊN CƠ CHẾ CHÚ Ý ĐỂ PHÁT HIỆN SỚM BẠO LỰC HỌC ĐƯỜNG

Nguyễn Viết Hưng
Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam
**Tác giả liên hệ: Nguyễn Viết Hưng – Email: hungnv@hcmue.edu.vn*
Ngày nhận bài: 30-7-2025; Ngày nhận bài sửa: 30-10-2025; Ngày duyệt đăng: 15-11-2025

TÓM TẮT

Bạo lực học đường là một vấn đề nghiêm trọng, ảnh hưởng đến sức khỏe học sinh và chất lượng môi trường giáo dục. Tuy nhiên, hầu hết nghiên cứu hiện tại tập trung vào bạo lực trong bối cảnh công cộng hoặc điện ảnh, vốn khác biệt đáng kể với hành vi bạo lực học đường – thường tinh vi và khó quan sát. Bên cạnh đó, sự thiếu hụt dữ liệu chuyên biệt cũng là rào cản lớn trong việc phát triển hệ thống giám sát hiệu quả. Để khắc phục những hạn chế này, nghiên cứu đề xuất FUSED-A, một kiến trúc học sâu đa luồng tích hợp đặc trưng không gian–thời gian từ chuỗi ảnh RGB và dữ liệu khung xương 2D thông qua cơ chế Guided Dot-Product Attention. Mô hình cho phép học tương quan giữa chuyển động cơ thể và ngữ cảnh hình ảnh, giúp tăng cường độ chính xác trong nhận diện hành vi. Đồng thời, bộ dữ liệu EduSafe-Early được xây dựng với 10 lớp hành động nhằm phục vụ phát hiện sớm hành vi bất thường. Các thực nghiệm cho thấy FUSED-A vượt trội so với nhiều phương pháp hiện đại, mở ra hướng tiếp cận hiệu quả cho các hệ thống giám sát bạo lực học đường thông minh và ứng dụng thực tiễn cao.

Từ khóa: bạo lực học đường; nhận dạng hành vi; phát hiện hành vi bất thường; thị giác máy tính; tiền bạo lực; YOLO

1. Giới thiệu

Bạo lực học đường là hành vi hành hạ, ngược đãi, đánh đập; xâm hại thân thể, sức khỏe; lăng mạ, xúc phạm danh dự, nhân phẩm; cô lập, xua đuổi và các hành vi cố ý khác gây tổn hại về thể chất, tinh thần của người học xảy ra trong cơ sở giáo dục hoặc lớp độc lập (Government of Vietnam, 2017). Bạo lực học đường gây ra những hệ lụy sâu rộng, tác động trực tiếp và gián tiếp đến nhiều khía cạnh của cá nhân, gia đình, nhà trường và xã hội. Về mặt thể chất, bạo lực học đường gây ra những tổn thương nghiêm trọng dẫn đến các vấn đề sức khỏe lâu dài, cản trở khả năng tham gia học tập và hoạt động hàng ngày của học sinh (Nguyen et al., 2025). Về mặt tâm lí, nạn nhân phải đối mặt với những tổn thương nghiêm trọng như rối loạn lo âu, trầm cảm (Tran et al., 2025a; Ta et al., 2025), từ đó làm

Cite this article as: Nguyen, V. H. (2025). Fused-a: A multi-stream attention-based model for early detection of school violence. *Ho Chi Minh City University of Education Journal of Science, 22*(11), 1980-1992. https://doi.org/10.54607/hcmue.js.22.11.5160(2025)

giảm hiệu suất học tập và cản trở sự hình thành nhân cách. Bạo lực học đường còn góp phần gia tăng bất ổn cộng đồng, làm tăng gánh nặng cho hệ thống y tế công cộng và chi phí kinh tế liên quan đến việc khắc phục hậu quả, đồng thời làm suy giảm chất lượng nguồn nhân lực tương lai, gây ra những tác động lâu dài đối với sự phát triển bền vững của xã hội.

Trên toàn cầu, có hơn 150 triệu học sinh đã từng bị bạo lực học đường, khoảng 50% em học sinh trong độ tuổi 13-15 từng bị bắt nạt (UNICEF Vietnam, 2018). Còn ở Việt Nam, trong giai đoạn 5 năm từ 2017 đến 2022, đã xảy ra tổng cộng 2624 vụ bạo lực học đường, với 7209 đối tượng có liên quan (Nguyen, 2023). Trong bối cảnh đó, việc xây dựng một mô hình phát hiện sớm bạo lực học đường không chỉ mang ý nghĩa khoa học mà còn có giá trị thiết thực đối với công tác giáo dục và bảo vệ học sinh. Phát hiện sớm là yếu tố then chốt để can thiệp kịp thời, ngăn chặn sự leo thang của hành vi bạo lực và hạn chế hậu quả tâm lí kéo dài đối với nạn nhân.

Hiện nay, các công trình nghiên cứu về bạo lực thường tập trung vào việc phát hiện hành vi bạo lực xã hội (Tang et al., 2024; Mishra et al., 2025; Andrade et al., 2025) thường không hiệu quả trong môi trường thực tế như trường học, nơi các hành vi bạo lực có thể diễn ra tinh vi, nhanh chóng và thường bị che giấu dưới những hình thức tưởng chừng vô hại. Một trong những rào cản lớn nhất chính là sự khan hiếm của các bộ dữ liệu chuyên biệt cho bối cảnh học đường. Hầu hết các bộ dữ liệu hiện có tập trung vào bạo lực trong môi trường công cộng hoặc trong phim ảnh, với đặc điểm hành vi và ngữ cảnh rất khác biệt so với học đường. Việc thiếu dữ liệu không chỉ cản trở khả năng huấn luyện các mô hình học sâu có độ chính xác cao mà còn hạn chế tính khả dụng của các hệ thống giám sát thông minh trong môi trường giáo dục.

Để giải quyết những thách thức này, nghiên cứu này đề xuất FUSED-A, một kiến trúc học sâu đa luồng tích hợp đặc trưng không gian–thời gian từ chuỗi ảnh và dữ liệu khung xương, sử dụng cơ chế chú ý (attention) để tăng cường hiệu quả phát hiện hành vi bạo lực trong môi trường học đường. Những đóng góp chính của nghiên cứu này gồm:

- Mô hình FUSED-A được đề xuất, một kiến trúc đa luồng tích hợp đặc trưng từ chuỗi ảnh RGB và khung xương 2D thông qua cơ chế chú ý, nhằm nâng cao hiệu quả nhận diện hành động bạo lực học đường trong môi trường video.
- Bộ dữ liệu chuyên biệt về hành vi bạo lực học đường EduSafe-Early được đề xuất, bao gồm 10 lớp hành động bạo lực và không bạo lực.
- Thực nghiệm và phân tích hệ thống các mô hình học sâu như ResNet50, MobileNetV3 kết hợp với LSTM, BiLSTM, GRU và BiGRU để so sánh hiệu suất trên từng loại dữ liệu và phương pháp hợp nhất đặc trưng.
- Mô hình đề xuất đạt hiệu suất vượt trội so với các phương pháp SOTA trên bộ dữ liệu đề xuất EduSafe-Early.

2. Đối tượng và phương pháp nghiên cứu

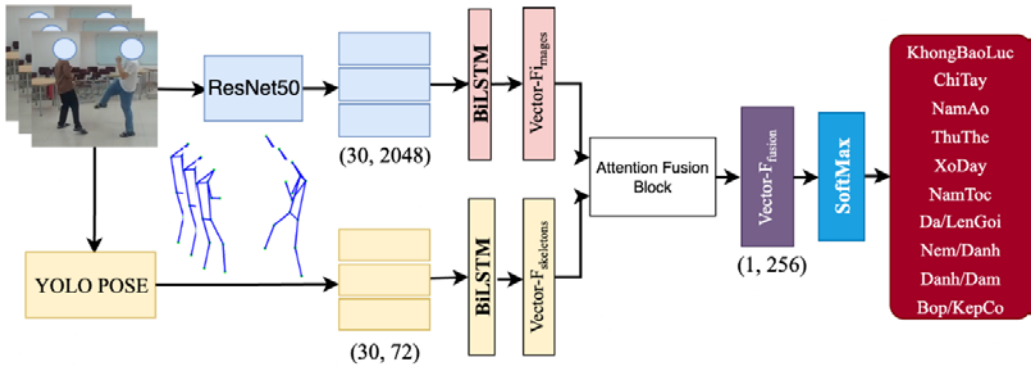
2.1. Các công trình liên quan

Các nghiên cứu gần đây đã khai thác tiềm năng của học sâu trong việc phát hiện hành vi bạo lực (Tran et al., 2024). (Omarov et al., 2022) đề xuất một phương pháp phát hiện hành vi bạo lực dựa trên khung xương người, không yêu cầu phần cứng tính toán cao, phù hợp với các hệ thống giám sát trong môi trường học đường. Phương pháp này gồm hai giai đoạn: trích xuất đặc trưng tư thế người từ chuỗi ảnh và phân loại hành động bằng mạng nơ-ron để xác định các hành vi bạo lực. Phương đề xuất đạt độ chính xác lên tới 97% trong nhận diện hành vi gây hấn trên bộ dữ liệu tự xây dựng. (Ye et al., 2020) phát triển hệ thống dựa trên video sử dụng KNN, đặc trưng optical flow và bộ phân loại hai tầng DT-SVM, đạt độ chính xác 97,6% và cho thấy khả năng phân biệt tốt giữa hành vi bạo lực và hoạt động thường nhật. Bên cạnh đó, (Haque et al., 2024) giới thiệu mô hình BrutNet kết hợp CNN và GRU, không sử dụng optical flow nhưng vẫn đạt độ chính xác 90% trên tập dữ AVDC. Gần đây, (Perseghin & Foresti, 2023) công bố hệ thống dựa trên CNN 2D được huấn luyện trên bộ dữ liệu Daily School Break (DSB), đạt độ chính xác 95% và có thể triển khai với chi phí thấp trong môi trường học đường thực tế. Gần đây, nghiên cứu của (Tran et al., 2025b) mở rộng hướng tiếp cận bằng cách ứng dụng mạng tích chập đồ thị (GCN) để khai thác đặc trưng chuyển động từ khung xương. Mô hình SKE-A3TGCN được đề xuất sử dụng cơ chế chú ý không gian–thời gian nhằm tập trung vào các khớp quan trọng trong hành vi bạo lực. Kết quả thực nghiệm trên nhiều tập dữ liệu như HockeyFight, RWF-2000 và Movies cho thấy mô hình đạt hiệu suất vượt trội, đặc biệt trong bối cảnh độ phân giải thấp và môi trường biến động.

Tuy đạt kết quả cao, phần lớn các mô hình trên vẫn dựa vào các bộ dữ liệu không chuyên biệt cho bạo lực học đường như Hockey Fight, Movies hoặc Violent Flow, vốn được thu thập từ phim ảnh, YouTube hoặc môi trường xã hội nói chung. Các tập dữ liệu này không phản ánh đúng tính chất tinh vi và kín đáo của bạo lực học đường, đồng thời thường có chất lượng hình ảnh thấp và thiếu đa dạng về bối cảnh. Ngay cả các tập dữ liệu chuyên biệt hơn như Daily School Break (DSB) (Perseghin & Foresti, 2023) hay video tự xây dựng của (Ye et al., 2020) cũng còn nhiều hạn chế, chủ yếu về quy mô (chỉ khoảng 100 video ngắn) và số lượng lớp hành động chưa phong phú. Điều này làm hạn chế khả năng tổng quát hóa của mô hình và triển khai thực tế trong trường học gặp nhiều khó khăn. Do đó, cần có một hướng tiếp cận phù hợp hơn với thực tế trường học, cả về mô hình lẫn dữ liệu.

2.2. Phương pháp phát hiện sớm bạo lực trong học đường

Nghiên cứu này mô hình FUSED-A nhằm phát hiện bạo lực học đường từ video giám sát bao gồm hai nhánh xử lí song song: nhánh RGB để khai thác đặc trưng không gian – thời gian từ dữ liệu hình ảnh và nhánh Skeleton để khai thác đặc trưng hành vi từ khung xương người. Cả hai nhánh đều sử dụng kiến trúc BiLSTM để mô hình hóa mối quan hệ theo chuỗi thời gian giữa các khung hình, sau đó được tích hợp thông qua cơ chế chú ý. Cách tiếp cận này cho phép mô hình tận dụng hiệu quả cả hai nguồn thông tin bổ trợ - thông tin ngữ cảnh từ hình ảnh RGB và thông tin hành vi từ chuyển động khung xương. Chi tiết về kiến trúc của FUSED-A được trình bày như trong Hình 1.



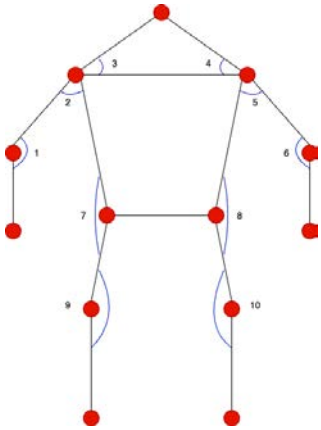
Hình 1. *Kiến trúc FUSED-A tích hợp đặc trưng không gian–thời gian từ chuỗi ảnh và dữ liệu khung xương*

2.2.1. Trích xuất đặc trưng không gian – thời gian

Mỗi video đầu vào được chuẩn hóa thành 30 khung hình có kích thước (224×224×3). Các khung hình này được đưa vào mạng ResNet50 đã tiền huấn luyện trên bộ dữ liệu ImageNet để trích xuất đặc trưng không gian. Mỗi khung hình được ánh xạ thành một vector đặc trưng 2048 chiều, kết quả là chuỗi đặc trưng có kích thước (30×2048) đại diện cho toàn bộ video. Chuỗi đặc trưng này sau đó được đưa vào hai lớp BiLSTM liên tiếp để mô hình hóa quan hệ thời gian giữa các khung hình. Lớp BiLSTM đầu tiên gồm 128 đơn vị ẩn và trả về toàn bộ chuỗi đầu ra nhằm bảo toàn thông tin theo từng thời điểm, trong khi lớp thứ hai gồm 64 đơn vị ẩn chỉ giữ lại hidden state (trạng thái ẩn) cuối cùng, nén toàn bộ chuỗi vào một biểu diễn đặc trưng cố định giàu thông tin. Để hạn chế hiện tượng quá khớp (overfitting), Dropout với tỉ lệ 0.2 được áp dụng sau mỗi lớp BiLSTM.

2.2.2. Trích xuất đặc trưng khung xương

Đối với nhánh Skeleton, mỗi khung hình được xử lí bằng YOLO11 Pose để phát hiện các điểm khung xương. Cụ thể, mỗi đối tượng người được biểu diễn bằng 36 đặc trưng, bao gồm tọa độ (x, y) của 13 điểm khớp chính và 10 góc tạo bởi các cặp khớp (như được thể hiện trong Hình 2). Trong nghiên cứu này, chúng tôi tập trung vào hai đối tượng chính trong mỗi khung hình, dẫn đến việc trích xuất 72 đặc trưng khung xương.



Hình 2. *Các đặc trưng điểm và góc khung xương được sử dụng*

Các góc đặc trưng phản ánh tư thế tại các khớp chính, được tính toán bằng công thức cosin giữa hai vector tạo thành bởi ba điểm liên tiếp A, B và C (với B là điểm trung gian). Góc tại điểm B được tính theo công thức:

$$\cos\theta = \frac{(x_1 - x_2)(x_3 - x_2) + (y_1 - y_2)(y_3 - y_2)}{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \cdot \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2}} \quad (1)$$

trong đó:

- $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ là tọa độ của của ba khớp A, B và C.
- θ là góc tạo bởi hai vector $\overrightarrow{BA}$ và $\overrightarrow{AB}$

Tương tự như nhánh RGB, các đặc trưng khung xương được sắp xếp thành chuỗi có kích thước (30×72) và đưa qua hai lớp BiLSTM với kiến trúc tương tự: lớp đầu có 128 đơn vị ẩn trả về chuỗi đầy đủ, lớp sau có 64 đơn vị ẩn và chỉ giữ lại đầu ra cuối cùng.

2.2.3. Kết hợp đặc trưng thông qua cơ chế chú ý tích vô hướng có hướng dẫn

Để kết hợp đặc trưng từ hai nhánh RGB và Skeleton, cơ chế **chú ý tích vô hướng có hướng dẫn** (Guided Dot-Product Attention – GDPA) được đề xuất để tối ưu hóa quá trình kết hợp thông tin. GDPA giúp mô hình tập trung vào những tín hiệu quan trọng hơn bằng cách tính toán điểm tương quan giữa hai nguồn dữ liệu. Cụ thể, thay vì áp dụng trọng số lên cả hai nhánh, chỉ có đặc trưng khung xương được điều chỉnh, trong khi đặc trưng hình ảnh được giữ nguyên. Điều này được thiết kế dựa trên việc đặc trưng RGB cung cấp thông tin ngữ cảnh phong phú và ổn định hơn, có thể đóng vai trò làm nguồn tham chiếu để điều chỉnh đặc trưng khung xương vốn thường bị ảnh hưởng bởi lỗi phát hiện điểm hoặc dao động chuyển động nhỏ. Việc giữ nguyên đặc trưng RGB trong khi điều chỉnh đặc trưng khung xương nhằm hướng dẫn nhánh động học học cách thích nghi tốt hơn với ngữ cảnh không gian – thời gian của toàn bộ cảnh quay.

Đầu tiên, ma trận chú ý được tính bằng các nhân hai vector đặc trưng từ hai nhánh là $F_{RGB}, F_{skeletons} \in \boldsymbol{R}^{128\times1}$, theo công thức:

$$M = F_{RGB} \times F_{skeletons}^{T} \quad (2)$$

Trong đó, $M \in \boldsymbol{R}^{128\times128}$ là ma trận chú ý, phản ánh mức độ liên quan giữa các đặc trưng hình ảnh và khung xương, còn $F_{skeletons}^{T} \in \boldsymbol{R}^{1\times128}$ là ma trận chuyển vị của đặc trưng khung xương sau khi đi qua các lớp BiLSTM. Sau đó, các điểm này được chuẩn hóa bằng hàm softmax để tạo ra ma trận trọng số chú ý α:

$$\alpha = softmax(M) \quad (3)$$

Ma trận này sau đó được sử dụng để điều chỉnh đặc trưng khung xương bằng phép nhân ma trận:

$$F'_{skeletons} = \alpha F_{skeletons} \quad (4)$$

Trong đó $F'_{skeletons} \in \boldsymbol{R}^{128\times1}$ là đặc trưng khung xương sau khi đã được điều chỉnh. Cuối cùng, đặc trưng hình ảnh ban đầu được ghép nối (concatenate) với đặc trưng khung xương đã được điều chỉnh để tạo ra một biểu diễn thống nhất:

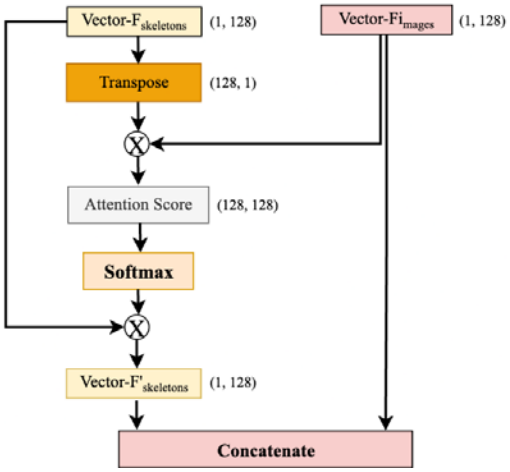
$$F_{fusion} = F_{RGB} || F'_{skeletons} \quad (5)$$

Việc chỉ điều chỉnh đặc trưng khung xương giúp mô hình tận dụng được thông tin từ cả hai nguồn dữ liệu một cách hiệu quả. Đặc trưng hình ảnh đóng vai trò như một nguồn tham chiếu cố định, trong khi đặc trưng khung xương được học cách thích nghi để phù hợp hơn với thông tin không gian và ngữ cảnh mà hình ảnh cung cấp. Kết quả là một mô hình nhận diện hành động mạnh mẽ, có khả năng khai thác đồng thời thông tin không gian, thời gian và động học để đưa ra dự đoán chính xác hơn. Hình 3 cung cấp góc nhìn trực quan về cơ chế chú ý GDPA được đề xuất.

Sau khi đặc trưng tổng hợp từ hai nhánh được hình thành thông qua cơ chế attention fusion, biểu diễn hợp nhất F_{fusion} sẽ được đưa vào một tầng Dense với số lượng nơ-ron bằng số lớp hành động. Tại đây, hàm kích hoạt softmax được sử dụng để chuẩn hóa đầu ra thành phân phối xác suất trên không gian nhãn, từ đó thu được vector dự đoán:

$$\hat{y} = Softmax(Dense(F_{fusion}) \quad (6)$$

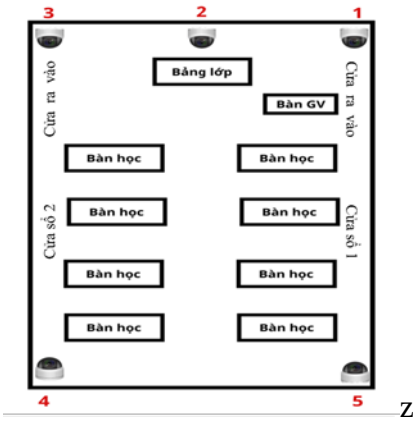
Trong đó, $\hat{y} \in \boldsymbol{R}^{10\times1}$ là vector xác suất của các loại hành động và mỗi phần tử trong $\hat{y}_i$ đại diện cho xác suất mô hình dự đoán video thuộc lớp hành động thứ i. Nhãn đầu ra cuối cùng sẽ được xác định bằng cách chọn phần tử có xác suất cao nhất trong $\hat{y}$.



Hình 3. *Cơ chế chú ý GDPA*

2.2.4. Bộ dữ liệu

Để giải quyết hạn chế về sự thiếu sót các bộ dữ liệu chuyên biệt cho môi trường học đường, chúng tôi xây dựng EduSafe-Early (Educational Safety Dataset for Early Violence Detection) - một tập dữ liệu được thiết kế dành riêng cho nhiệm vụ phát hiện sớm hành vi bạo lực trong bối cảnh giáo dục. Dựa trên mức độ thương tích gây ra cho nạn nhân, bạo lực học đường được phân thành ba cấp độ: (1) Đe dọa bạo lực, gồm các hành vi mang tính cảnh báo như chỉ tay, thủ thế, xô đẩy, nắm cổ áo nhưng chưa gây thương tích; (2) Bạo lực gây thương tích, bao gồm các hành vi như đánh, đấm, đá, lên gối, bóp cổ hoặc dùng vật ném gây tổn thương thể chất mức nhẹ đến trung bình; (3) Bạo lực nghiêm trọng, là những hành vi có thể gây chấn thương nghiêm trọng hoặc liên quan đến vũ khí. Tuy nhiên, mức độ thứ ba thường ít xuất hiện trong môi trường học đường.



Hình 4. *Mô tả năm góc quay thu thập bộ dữ liệu*

Dữ liệu được thu thập với sự tham gia của 12 sinh viên (8 nam, 4 nữ) từ Khoa Công nghệ Thông tin, Trường Đại học Sư phạm Thành phố Hồ Chí Minh, trong độ tuổi từ 19-21. Mỗi tình huống được thực hiện bởi một cặp sinh viên đóng vai nạn nhân và người thực hiện hành vi bạo lực, với tổ hợp giới tính đa dạng (nam–nam, nữ–nữ, nam–nữ). Video được ghi hình ở độ phân giải 720p@30fps, với camera cố định ở độ cao trên 2m, mô phỏng góc nhìn của hệ thống giám sát trường học. Việc quay phim được thực hiện tại 5 vị trí khác nhau, như minh họa ở Hình 4, nhằm đảm bảo sự đa dạng về góc quay và bối cảnh. Trong quá trình rà soát dữ liệu, chúng tôi nhận thấy một số video có tư thế hành động tương đồng được ghi lại từ cùng một góc quay, dẫn đến khả năng dư thừa thông tin. Để đảm bảo tính đa dạng và chất lượng của tập dữ liệu, các video trùng lặp đã được loại bỏ, chỉ giữ lại những mẫu có sự khác biệt rõ rệt về tư thế hoặc bối cảnh không gian.

Tập dữ liệu EduSafe-Early bao gồm tổng cộng 1000 video, được chia đều cho 10 lớp hành động, mỗi lớp gồm 100 video. Các lớp hành động được thiết kế để phản ánh phổ hành vi từ không bạo lực đến các mức độ bạo lực khác nhau, bao gồm: Chỉ tay, Nắm cổ áo, Xô đẩy, Thủ thế, Bóp/kẹp cổ, Nắm tóc, Đá/lên gối, Ném/đánh bằng vật, Đánh/đấm và Không bạo lực. Việc phân bổ cân bằng giữa các lớp không chỉ giúp mô hình học sâu phân biệt chính xác các hành vi bạo lực học đường, mà còn tăng khả năng nhận diện trong những tình huống có hành động tương đồng nhưng không mang tính bạo lực, từ đó giảm thiểu nhầm lẫn trong quá trình huấn luyện và suy luận mô hình. Hình 5 trình bày một số mẫu dữ liệu trong bộ dữ liệu EduSafe-Early.

Hình 5. *Một số mẫu dữ liệu trong bộ dữ liệu EduSafe-Early*

2.2.5. Phương pháp đánh giá

Hiệu suất của mô hình được đánh giá thông qua bốn chỉ số chính là Accuracy, Precision, Recall và F1-score, được tính dựa trên các công thức (7), (8), (9) và (10). Các chỉ số này phản ánh khả năng phân loại chính xác hành vi bạo lực và không bạo lực của mô hình. Trong đó, TP (True Positive) là số trường hợp bạo lực được phát hiện đúng, FP (False Positive) là số trường hợp không bạo lực nhưng bị phân loại nhầm thành bạo lực, TN (True Negative) là số trường hợp không bạo lực được nhận diện chính xác, và FN (False Negative) là số trường hợp bạo lực mà mô hình không phát hiện.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

3. Kết quả và thảo luận

3.1. So sánh với các phương pháp SOTA

Bảng 1 trình bày kết quả so sánh giữa mô hình đề xuất và một số phương pháp SOTA trên một bộ dữ liệu EduSafe-Early. Nhìn chung, mô hình được đề xuất đạt kết quả vượt trội trên cả bốn chỉ số, với Precision = 0.95, Recall = 0.94, F1-score = 0.94 và Accuracy = 0.94. So với (Dündar et al., 2024) đạt F1-score là 0.87 – mô hình FUSED-A cho thấy mức cải thiện đáng kể lên đến 7%, đặc biệt trong bối cảnh các hành vi bạo lực thường ngắn, tinh vi và dễ bị nhầm lẫn. Phương pháp (Haque et al., 2024) đạt kết quả cao với F1-score là 0.92, tuy nhiên vẫn thấp hơn so với mô hình đề xuất. Điều này cho thấy hiệu quả của chiến lược kết hợp đặc trưng có hướng dẫn (GDPA), cũng như tính phù hợp của kiến trúc đa luồng trong việc tích hợp thông tin thị giác và động học. Trong khi đó, (Islam et al., 2021) chỉ đạt F1-score là 0.80, cho thấy khoảng cách đáng kể về mặt hiệu suất, phản ánh sự tiến bộ của các mô hình học sâu trong những năm gần đây. Những kết quả này xác thực rằng kiến trúc FUSED-A, cùng với cơ chế chú ý tích vô hướng có hướng dẫn, mang lại lợi thế rõ rệt trong nhận diện hành vi bạo lực học đường, đặc biệt là trong giai đoạn đầu của hành vi – nơi các mô hình truyền thống thường gặp khó khăn.

Bảng 1. *So sánh hiệu suất của mô hình FUSED-A với các phương pháp trước đây trên bộ dữ liệu EduSafe-Early*

Nghiên cứu	Precision	Recall	F1-score	Accuracy
Islam et al., 2021	0.81	0.80	0.80	0.80
Dündar et al., 2024	0.88	0.87	0.87	0.87
Haque et al., 2024	0.93	0.92	0.92	0.92
Skeleton-BiLSTM (đề xuất)	0.66	0.65	0.65	0.65
ResNet50-BiLSTM (đề xuất)	0.92	0.92	0.91	0.92
FUSED-A (đề xuất)	**0.95**	**0.94**	**0.94**	**0.94**

Đáng chú ý, khi so sánh với hai nhánh đơn được tách ra từ FUSED-A, mô hình ResNet50-BiLSTM trên chuỗi ảnh RGB đạt F1-score = 0.91 – cao hơn hẳn so với nhánh Skeleton-BiLSTM (F1-score = 0.65), phản ánh vai trò nổi trội của đặc trưng hình ảnh trong nhận diện hành vi. Tuy nhiên, khi tích hợp hai nguồn dữ liệu qua GDPA, hiệu suất mô hình được cải thiện thêm 3%, cho thấy sự bổ sung giá trị từ đặc trưng khung xương khi được dẫn hướng bởi đặc trưng RGB. Điều này khẳng định rằng FUSED-A không chỉ đơn thuần là sự kết hợp hai nhánh mà còn là sự phối hợp có định hướng giúp tăng cường khả năng phân biệt hành vi tinh vi trong môi trường học đường.

3.2. Thực nghiệm cắt bỏ

3.2.1. Thực nghiệm trên đặc trưng RGB

Chúng tôi đã tiến hành thử nghiệm hai kiến trúc dựa trên CNN phổ biến là MobiNetV3 và ResNet50, kết hợp với bốn kiến trúc RNN phổ biến là LSTM, BiLSTM, GRU và BiGRU có khả năng xử lí dữ liệu chuỗi hiệu quả, nhằm đánh giá khả năng nhận diện hành động dựa trên chuỗi ảnh RGB. Các kết quả được trình bày trong Bảng 2. Kết quả thực nghiệm cho thấy mô hình sử dụng MobileNetV3 kết hợp BiLSTM đạt hiệu suất cao với Accuracy và F1-score đạt 0.91, khẳng định ưu thế của việc khai thác quan hệ thời gian theo cả hai chiều. Nhóm mô hình sử dụng ResNet50 đạt hiệu suất cao hơn so với MobileNetV3, cho thấy khả năng trích xuất đặc trưng không gian ưu việt của ResNet50. Đặc biệt, mô hình ResNet50 kết hợp BiLSTM đạt độ chính xác 0.92 và F1-score 0.91, khẳng định hiệu quả của BiLSTM trong việc khai thác mối quan hệ không gian - thời gian. Nhìn chung, các mô hình sử dụng BiLSTM đều đạt hiệu suất cao nhất trong cả hai nhóm, nhờ khả năng khai thác thông tin theo hai chiều thời gian. Bên cạnh đó, ResNet50 cũng cho thấy hiệu quả nhỉnh hơn MobileNetV3, đặc biệt khi kết hợp với BiLSTM.

Bảng 2. *Kết quả thực nghiệm dựa trên đặc trưng RGB.*

Model	**Model**	**Precision**	**Recall**	**F1-score**	**Accuracy**
MobileNetV3	LSTM	0.89	0.89	0.89	0.89
	BiLSTM	***0.92***	***0.91***	***0.91***	***0.91***
	GRU	0.88	0.86	0.87	0.87
	BiGRU	0.91	0.90	0.90	0.90
ResNet50	LSTM	0.91	0.90	0.89	0.90
	BiLSTM	**0.92**	**0.92**	**0.91**	**0.92**
	GRU	0.91	0.90	0.90	0.91
	BiGRU	0.92	0.91	0.91	0.91

3.2.2. Thực nghiệm trên đặc trưng khung xương

Ba mô hình LSTM, BiLSTM và BiGRU được thử nghiệm để học tập các đặc trưng khung xương. Các kết quả được trình bày trong Bảng 3. Tương tự như ở nhánh RGB, mô hình BiLSTM tiếp tục đạt hiệu suất cao nhất với độ chính xác 0.65 và F1-score 0.65, cho thấy khả năng phân loại ổn định và hiệu quả khi xử lí dữ liệu khung xương. Trong khi đó, mô hình LSTM có kết quả thấp hơn một chút, với Accuracy và F1-score đạt 0.63, còn BiGRU có hiệu suất thấp nhất khi chỉ đạt 0.60 Accurcy và 0.59 F1-score

Bảng 3. *Kết quả thực nghiệm dựa trên đặc trưng khung xương*

Model	Precision	Recall	F1-score	Accuracy
BiGRU	0.61	0.69	0.59	0.60
LSTM	0.65	0.63	0.63	0.63
BiLSTM	**0.66**	**0.65**	**0.65**	**0.65**

3.2.3. Phương pháp kết hợp đặc trưng

Dựa trên kết quả thực nghiệm sơ bộ với hai loại đặc trưng RGB và khung xương, chúng tôi lựa chọn ResNet50 làm mạng trích xuất đặc trưng không gian cho nhánh RGB và sử dụng BiLSTM làm mô hình học chuỗi thời gian cho cả hai nhánh nhằm khai thác hiệu quả quan hệ theo thời gian giữa các khung hình. Bảng 4 trình bày kết quả so sánh giữa ba chiến lược hợp nhất đặc trưng được triển khai trong mô hình FUSED-A, bao gồm Guided Dot-Product Attention (GDPA) – phương pháp được đề xuất, Concatenation và Element-wise Addition.

Bảng 4. *So sánh hiệu suất giữa các phương pháp hợp nhất đặc trưng*

Phương pháp kết hợp	Precision	Recall	F1-Score	Accuracy
Element-wise Addition	0.93	0.92	0.92	0.92
Concatenation	0.93	0.93	0.93	0.93
GDPA	**0.95**	**0.94**	**0.94**	**0.94**

Kết quả thực nghiệm cho thấy **Guided Dot-Product Attention (GDPA)** là phương pháp hợp nhất đặc trưng hiệu quả nhất trong ba chiến lược được so sánh. Điểm độc đáo của GDPA nằm ở cơ chế hợp nhất không đối xứng, trong đó đặc trưng khung xương – vốn nhạy cảm với chuyển động – được điều chỉnh theo đặc trưng RGB đóng vai trò làm nguồn tham chiếu giàu ngữ cảnh. Cách tiếp cận này cho phép mô hình học được các tương quan động học trong mối liên hệ với bố cục cảnh quan và các yếu tố không gian, từ đó nâng cao khả năng phân biệt hành vi bạo lực trong các tình huống phức tạp và tinh vi – vốn thường gặp trong môi trường học đường. So với các phương pháp ghép nối trực tiếp hoặc cộng phần tử, GDPA mang lại biểu diễn đặc trưng hài hòa và có tính định hướng rõ ràng hơn, tạo nền tảng cho những cải tiến tiếp theo trong các hệ thống giám sát thông minh.

3.3. Thảo luận

Kết quả cho thấy mô hình đề xuất đạt hiệu suất vượt trội trong phát hiện sớm hành vi bạo lực học đường, với Accuracy và F1-score đều đạt 94% khi sử dụng kiến trúc ResNet50–BiLSTM cho chuỗi ảnh RGB, kết hợp với BiLSTM cho dữ liệu khung xương thông qua cơ chế hợp nhất GDPA. Điều này khẳng định hiệu quả của chiến lược kết hợp đặc trưng có điều hướng, cho phép mô hình tập trung vào những tín hiệu chuyển động quan trọng trong ngữ cảnh không gian – thời gian. Trong khi các phương pháp hợp nhất thông thường như concatenation hoặc element-wise addition chỉ đơn thuần ghép nối thông tin, GDPA cho phép điều chỉnh động học của chuỗi hành vi theo các đặc trưng thị giác ổn định hơn, từ đó cải thiện độ chính xác trong các tình huống phức tạp và dễ nhầm lẫn.

BiLSTM cho thấy khả năng mô hình hóa hiệu quả các chuỗi thời gian hai chiều, đặc biệt phù hợp với bài toán phân tích video. Mặc dù mô hình đơn thuần dựa trên dữ liệu skeleton cho kết quả chưa thực sự nổi bật, nhưng khi được tích hợp với đặc trưng RGB và xử lí qua cơ chế attention, hiệu suất được cải thiện rõ rệt. Điều này làm nổi bật giá trị của việc khai thác dữ liệu đa nguồn trong bài toán nhận diện hành vi bạo lực.

Một đóng góp khác là việc xây dựng EduSafe-Early, bộ dữ liệu chuyên biệt cho môi trường học đường, bao gồm 10 lớp hành động từ phi bạo lực đến bạo lực gây thương tích. Đây là cơ sở quan trọng giúp mô hình học tốt hơn và có khả năng tổng quát hóa trong môi trường thực tế. Cuối cùng, nghiên cứu này đề xuất một hướng tiếp cận khả thi cho các hệ thống giám sát thông minh trong giáo dục, với khả năng nhận diện sớm hành vi bất thường và hỗ trợ giáo viên can thiệp kịp thời. Tuy nhiên, nghiên cứu vẫn còn nhiều thách thức, đặc biệt về việc sử dụng các kiến trúc nặng như Resnet50 làm công cụ trích xuất đặc trưng. Việc tối ưu hóa mô hình để đảm bảo tính ổn định và chính xác trong các điều kiện ánh sáng, góc quay và bối cảnh khác nhau sẽ là hướng phát triển tiếp theo trong tương lai.

4. Kết luận và kiến nghị

Nghiên cứu này đề xuất FUSED-A, một mô hình học sâu đa luồng kết hợp thông tin từ chuỗi ảnh RGB và dữ liệu khung xương 2D thông qua cơ chế chú ý tích vô hướng có hướng dẫn (GDPA), nhằm nâng cao hiệu quả phát hiện sớm hành vi bạo lực học đường. Kết quả thực nghiệm cho thấy FUSED-A đạt hiệu suất vượt trội với Accuracy và F1-score đều đạt 94%, cao hơn so với các phương pháp học sâu đơn luồng và các chiến lược hợp nhất truyền thống. Một đóng góp quan trọng khác là việc xây dựng EduSafe-Early, bộ dữ liệu chuyên biệt đầu tiên tại Việt Nam dành cho bài toán phát hiện sớm bạo lực học đường, với 10 lớp hành động được gán nhãn cẩn thận và phản ánh trung thực đặc điểm hành vi trong môi trường giáo dục. Điều này giúp tăng cường tính thực tiễn và khả năng tổng quát hóa của mô hình trong các ứng dụng giám sát thực tế.

Từ những kết quả đạt được, nghiên cứu kiến nghị rằng các hệ thống giám sát an ninh trong trường học nên tích hợp các mô hình học sâu đa phương thức như FUSED-A để phát hiện sớm các hành vi nguy cơ, từ đó hỗ trợ giáo viên và cán bộ quản lí can thiệp kịp thời. Trong tương lai, các hướng nghiên cứu mở có thể bao gồm: (1) tích hợp thêm dữ liệu âm thanh tăng cường hiệu suất; (2) phát triển mô hình nhẹ phù hợp với thiết bị giám sát có cấu hình thấp; và (3) triển khai thử nghiệm thực tế tại các trường học để đánh giá hiệu quả hoạt động và mức độ chấp nhận của người dùng.

❖ ***Tuyên bố về quyền lợi:*** *Tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Andrade, J. P. F., Si, T., Cavalcanti, A. P., Nascimento, A. C., & Miranda, P. B. (2025). SUSAN: A deep learning-based architecture for violence detection against women in surveillance videos. *Expert Systems with Applications, 280*, 127337. https://doi.org/10.1016/j.eswa.2025.127337

Government of Vietnam. (2017). *Decree No. 80/2017/ND-CP dated July 17, 2017 on a safe, healthy and friendly education environment which prevents and stops school violence*. Government Portal of Vietnam.

Dündar, N., Keçeli, A. S., Kaya, A., & Sever, H. (2024). A shallow 3D convolutional neural network for violence detection in videos. *Egyptian Informatics Journal, 26*, 100455. https://doi.org/10.1016/j.eij.2024.100455

Haque, M., Nyeem, H., & Afsha, S. (2024). BrutNet: A novel approach for violence detection and classification using DCNN with GRU. *The Journal of Engineering, 2024*(4), e12375. https://doi.org/10.1049/tje2.12375

Islam, Z., Rukonuzzaman, M., Ahmed, R., Kabir, M. H., & Farazi, M. (2021). Efficient two-stream network for violence detection using separable convolutional LSTM. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533633

Nguyen, L. (2023). Trên .600 vụ bạo lực học đường có tính chất phức tạp, chuyên gia đề xuất giải pháp [Over 2,600 Violent School Incidents of a Complex Nature, Experts Propose Solutions]**.** *People's Representative Newspaper*. https://daibieunhandan.vn/giao-duc--y-te1/tren-2-600-vu-bao-luc-hoc-duong-co-tinh-chat-phuc-tap-chuyen-gia-de-xuat-giai-phap-i331004/

Mishra, S., Jain, V., Saraf, Y. A., Kandasamy, I., & WB, V. (2025). Deep neuro-fuzzy system for violence detection. *Neurocomputing, 619*, 129007. https://doi.org/10.1016/j.neucom.2024.129007

Nguyen, V. H., Ta, C. P., Le, T. L., Ngo, Q. K., & Tran, T. N. (2025). C-ViDNet: a model for supporting violence detection in schools. *Ho Chi Minh City University of Education Journal of Science, 22*(5), 801-813. https://doi.org/10.54607/hcmue.js.22.5.4699(2025)

Omarov, B., Narynov, S., Zhumanov, Z., Gumar, A., & Khassanova, M. (2022). A skeleton-based approach for campus violence detection. *Computers, Materials & Continua, 72*(1). https://doi.org/10.32604/cmc.2022.024566

Perseghin, E., & Foresti, G. L. (2023). A shallow system prototype for violent action detection in Italian public schools. *Information, 14*(4), 240. https://doi.org/10.3390/info14040240

Ta, P., Tran, N., Nguyen, H., & Nguyen, H. D. (2025). Detecting signs of depression on social media: A machine learning analysis and evaluation. *Sustainable Futures, 100827*. https://doi.org/10.1016/j.sftr.2025.100827

Tang, Y., Chen, Y., Sharifuzzaman, S. A., & Li, T. (2024). An automatic fine-grained violence detection system for animation based on modified faster R-CNN. *Expert Systems with Applications, 237*, 121691. https://doi.org/10.1016/j.eswa.2023.121691

Tran, N., Nguyen, H., Ly, D., & Nguyen, H. D. (2024). Violence detection using skeleton data with graph convolutional networks. In *International Conference on Intelligent Systems and Data Science* (pp. 86–97). Springer. https://doi.org/10.1007/978-981-97-9616-8_7

Tran, N., Nguyen, H., Ly, D., Ngo, K., & Nguyen, H. D. (2025b). Advancing violence detection with graph-based skeleton motion analysis. *SN Computer Science, 6*(6), 1-18. Springer. https://doi.org/10.1007/s42979-025-04118-7

Tran, N., Ta, P., Nguyen, H., Nguyen, H. D., & Le, A.-C. (2025a). Hybrid contextual and sentiment-based machine learning model for identifying depression risk in social media. *Expert Systems with Applications, 291*, 128505. https://doi.org/10.1016/j.eswa.2025.128505

UNICEF Vietnam. (2018, September 6). More than 150 million adolescents worldwide are subjected to school violence [Press release]. *UNICEF*. Retrieved July 22, 2025, from https://www.unicef.org/vietnam/vi/thông-cáo-báo-chí/hơn-150-triệu-thanh-thiếu-niên-trên-thế-giới-bị-bạo-lực-học-đường

Ye, L., Wang, L., Ferdinando, H., Seppänen, T., & Alasaarela, E. (2020). A video-based DT–SVM school violence detecting algorithm. *Sensors, 20*(7), 2018. https://doi.org/10.3390/s20072018

FUSED-A: A MULTI-STREAM ATTENTION-BASED MODEL FOR EARLY DETECTION OF SCHOOL VIOLENCE

Nguyen Viet Hung

Ho Chi Minh City University of Education, Vietnam

Corresponding author: Nguyen Viet Hung – Email: hungnv@hcmue.edu.vn

Received: July 30, 2025; Revised: October 30, 2025; Accepted: November 15, 2025

.

ABSTRACT

School violence is a serious issue that affects students' well-being and the overall quality of the educational environment. However, most research focuses on violence in public or cinematic contexts, which significantly differ from school-based violence, often subtle and difficult to detect. Moreover, the lack of specialized datasets remains a major barrier to developing effective surveillance systems. To address these limitations, this study proposes FUSED-A, a multi-stream deep learning architecture that integrates spatio-temporal features from RGB image sequences and 2D skeleton data through a Guided Dot-Product Attention (GDPA) mechanism. The model enables learning correlations between body motion and visual context, thereby enhancing the accuracy of behavior recognition. Additionally, the EduSafe-Early dataset is introduced, comprising 10 action classes specifically designed for early detection of abnormal behaviors. Experimental results demonstrate that FUSED-A outperforms several state-of-the-art methods, offering a promising and practical approach for intelligent school violence surveillance systems.

Keywords: abnormal behavior detection; action recognition; computer vision; pre-violence; school violence; YOLO