TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 20, Số 10 (2023): 1775-1788
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 20, No. 10 (2023): 1775-1788
https://doi.org/10.54607/hcmue.js.20.10.3729(2023)

Bài báo nghiên cứu

GIẢI PHÁP XÁC THỰC NGƯỜI HỌC BẰNG PHƯƠNG PHÁP ĐA YẾU TỐ SINH TRẮC HỌC

Nguyễn Quốc Trung[1*], Nguyễn Võ Phi Long[1], Lê Đức Long[1], Nguyễn Đình Thúc[2]

[1]Trường Đại học Sư phạm Thành phố Hồ Chí Minh, Việt Nam

[2]Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

[*]Tác giả liên hệ: Nguyễn Quốc Trung – Email: trungnq@hcmue.edu.vn

Ngày nhận bài: 16-02-2023; ngày nhận bài sửa: 18-3-2023; ngày duyệt đăng: 20-3-2023

TÓM TẮT

Học tập trực tuyến, đặc biệt là các hệ kiểm tra đánh giá người học từ xa phải đối mặt với các vấn đề xác thực người học. Hiện tại, phương pháp xác thực người học trên các hệ thống này chủ yếu dựa trên cách tiếp cận là tài khoản-mật khẩu (username-password), một phương pháp phổ biến, dễ dùng. Nhưng để đảm bảo an toàn, người dùng cần sử dụng mật khẩu mạnh và phức tạp, các mật khẩu này thường rất khó nhớ. Bên cạnh đó, một số hệ thống học tập trực tuyến đã triển khai việc xác thực người học bằng các phương pháp sinh trắc học, theo đó người học có thể đăng nhập mà không cần mật khẩu. Tuy nhiên, việc chỉ sử dụng trực tiếp một yếu tố sinh trắc lại gặp phải nhiều vấn đề, nhất là tính ổn định của thông tin sinh trắc. Bài báo này đề xuất một giải pháp xác thực đa yếu tố sử dụng sinh trắc học khuôn mặt và giọng nói. Đề xuất đã được thử nghiệm trên tập dữ liệu VoxCeleb1, mang lại hiệu quả vượt trội với độ chính xác là 99,1%, so với 95,62% khi chỉ sử dụng khuôn mặt và 98.66% khi chỉ sử dụng giọng nói.

Từ khóa: xác thực; sinh trắc học; đa yếu tố sinh trắc học; đơn yếu tố sinh trắc học; xác thực người học; hệ thống học tập trực tuyến

1. Giới thiệu

Xác thực (Authentication) là quá trình xác định danh tính của người dùng khi truy vào hệ thống, nhằm trả lời câu hỏi "Ai đang truy cập ?". Hầu hết các hệ thống thông tin yêu cầu người dùng xác thực trước khi cho phép truy cập và tiến hành các thao tác, nhờ đó dữ liệu được bảo mật và toàn vẹn. Các phương pháp xác thực người dùng có thể được phân loại thành: (1) Mật khẩu (password aproach), (2) OTP (One-time password aproach), (3) Token (Token aproach), (4) Sinh trắc học (Biometric aproach), (5) Hỗn hợp (Hybrid aproach).

Cách tiếp cận (1) và (2) dựa trên sự ghi nhớ của người dùng nên dẫn đến nguy cơ quên hoặc bị đánh cắp. Đối với (3), xác thực phụ thuộc vào việc bảo quản keycard, smartcard, USB, hoặc tệp token. Do (1), (2), (3) sử dụng thông tin xác thực có thể chuyển giao từ cá nhân này sang cá nhân khác (transferable data) nên có tính bảo mật thấp hơn (4)

Cite this article as: Nguyen Quoc Trung, Nguyen Vo Phi Long, Le Duc Long, & Nguyen Dinh Thuc (2023). Learner authentication VIA biometric-based multi-factor. *Ho Chi Minh City University of Education Journal of Science*, *20*(10), 1775-1788.

(Sabhanayagam et al., 2018). Cụ thể, ưu điểm của sinh trắc học là tính có sẵn ở mỗi cá nhân, không thể chuyển giao và không thể bị lãng quên (Sharma, 2014). Thêm vào đó, các đặc điểm của con người rất khó sao chép và giả mạo, nên sinh trắc học ngày càng trở thành phương pháp xác thực phổ biến và hiệu quả (Dinca & Hancke, 2017).

Xác thực đơn yếu tố sinh trắc học là phương pháp xác thực chỉ dựa trên một đặc điểm sinh trắc học duy nhất. Xác thực đa yếu tố sinh trắc học (multi biomatric-factors authentication, hay còn gọi là multibiometrics) là sử dụng 2 hay nhiều yếu tố sinh trắc kết hợp với nhau (Drozdowski et al., 2020). Đa yếu tố sinh trắc học được chứng minh có khả năng khắc phục những nhược điểm của hệ thống xác thực sinh trắc học đơn yếu tố, đồng thời nâng cao độ an toàn, bảo mật cho hệ thống (Siddiqui et al., 2014). Mặc dù, mang tính hứa hẹn cao nhưng xác thực đa yếu tố sinh trắc học chưa được triển khai nhiều trong thực tế.

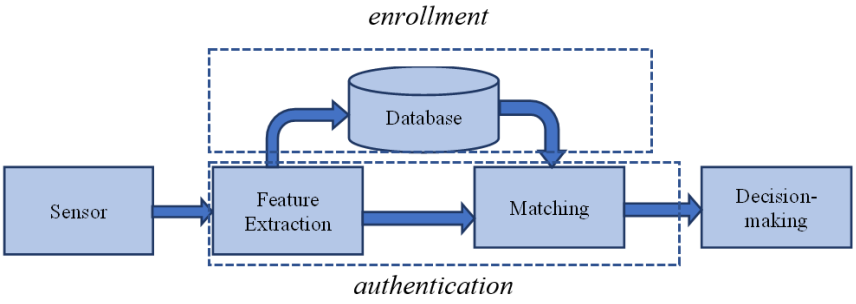
Nhận thấy sự cần thiết của việc triển khai xác thực dựa trên đa yếu tố sinh trắc học trên các ứng dụng, đặc biệt là trên các nền tảng học và kiểm tra trực tuyến, bài báo tập trung nghiên cứu và xây dựng một giải pháp xác thực đa yếu tố sử dụng sinh trắc học khuôn mặt và giọng nói.

2. Nội dung nghiên cứu

2.1. Tổng quan về sinh trắc học

Sinh trắc học là nghiên cứu phương pháp luận về đo lường và phân tích dữ liệu sinh học cho mục đích xác thực hoặc nhận dạng người dùng (Sabhanayagam et al., 2018). Sinh trắc học có thể được phân loại thành: sinh trắc học sinh lí (Physiological biometrics) – bao gồm các đặc điểm khác biệt và duy nhất của cơ thể vật lí như: khuôn mặt, võng mạc, mống mắt, vân tay, tay và tĩnh mạch tay, hình dạng bàn tay và ngón tay, dấu vân tay… và sinh trắc học hành vi (Behavioral biometrics) – bao gồm các đặc điểm về hành động để có thể phân biệt duy nhất một người như: thao tác gõ phím, chữ kí và dáng đi...

Một hệ thống sinh trắc học gồm nhiều module đảm nhận những vai trò khác nhau (Sabhanayagam et al., 2018): Cảm biến và thu dữ liệu (Sensor module), trích xuất đặc trưng (Feature extraction module), lưu trữ dữ liệu sinh trắc (Database module), tính độ tương đồng (Matching module), đưa ra quyết định (Decision-making module), được minh họa qua Hình 1.



Hình 1. *Sơ đồ hệ thống sinh trắc học*

Một số yếu tố sinh trắc học tiêu biểu

Mỗi yếu tố sinh trắc học có ý nghĩa và giá trị xác thực riêng (Gavrilova & Monwar, 2011). Lựa yếu tố sinh trắc học để áp dụng cho bài toán xác thực phụ thuộc vào nhiều yếu tố khác nhau, Bảng 1 trình bày sự so sánh về ưu, nhược điểm của các yếu tố sinh trắc học tiêu biểu.

Bảng 1. *So sánh các yếu tố sinh trắc học phổ biến và tiêu biểu*

Yếu tố	Ưu điểm	Nhược điểm	Độ phủ dân số
Vân tay Fingerprint	• Độ chính xác cao • Thời gian tính toán nhanh • Ổn định và không thay đổi theo lứa tuổi	• Vết cắt, sẹo làm ảnh hưởng lớn đến hiệu quả xác thực. Có thể giả mạo bằng các biện pháp tinh vi làm giả dấu vân tay, ngón tay • Chịu ảnh hưởng mạnh mẽ từ môi trường khi mưa, bụi bám vào bề mặt quét vân tay • Chưa phổ biến trên thiết bị phổ thông	• Là duy nhất với mỗi cá nhân • Khác nhau ở cặp song sinh
Khuôn mặt Face	• Được sử dụng rộng rãi • Có thể triển khai trên nhiều nền tảng. • Được nghiên cứu và phát triển mạnh. • Sử dụng thiết bị phổ thông: camera	• Khuôn mặt có thể thay đổi theo thời gian • Khó phân biệt được cặp song sinh • Phụ thuộc vào điều kiện môi trường: ánh sáng, sương mù… • Có khả năng bị giả mạo	• Độ hiếm cao, nhưng có thể xuất hiện cá nhân có khuôn mặt tương tự • Giống nhau ở cặp song sinh
Mống mắt Iris	• Độ chính xác cao • Thời gian tính toán nhanh • Ổn định và không đổi theo lứa tuổi • Không cần tiếp xúc vật lí với hệ thống	• Tốn chi phí do cần thiết bị chuyên dụng • Cần sự hợp tác của người dùng, có thể gây khó chịu • Phải đặt mắt gần thiết bị • Cần nền tảng và ứng dụng phù hợp	• Là duy nhất với mỗi cá nhân • Khác nhau ở cặp song sinh
Giọng nói Voice	• Độ chính xác cao • Được sử dụng rộng rãi • Có thể triển khai trên nhiều nền tảng • Sử dụng thiết bị phổ thông: microphone	• Dễ bị ảnh hưởng bởi chất lượng của micro và tiếng ồn • Một số vấn đề sức khỏe ở cổ họng có thể ảnh hưởng đến độ chính xác • Cần có chức năng chống giả mạo (anti-spoof) trước khi xác thực	• Độ hiếm cao, có thể xuất hiện cá nhân có giọng nói tương tự • Có sự khác nhau ở cặp song sinh

Việc lựa chọn yếu tố sinh trắc học được quyết định bởi tính khả dụng của cảm biến tương ứng. Hệ thống xác thực phải đảm bảo rằng người dùng có thể dễ dàng sử dụng cảm biến. Bài báo triển khai xác thực dựa trên đa yếu tố khuôn mặt và giọng nói, vì máy ảnh và microphone có sẵn trên các thiết bị mà người học sử dụng để truy cập vào hệ thống trực tuyến.

Bên cạnh các yếu tố sinh trắc học phổ biến, còn nhiều yếu tố sinh trắc khác ít phổ biến hơn được nghiên cứu và cho thấy độ chính xác cao, được mô tả ở Bảng 2. Các yếu tố sinh trắc học này ít hơn phổ biến do chúng cần các cảm biến chuyên dụng và khó lắp đặt.

Bảng 2. *Thống kê một số yếu tố sinh trắc học khác*

Yếu tố sinh trắc học	Mô tả
Đường chỉ tay (Palm print)	Sử dụng đặc trưng từ đường nét và hoa văn chỉ tay (Palm print), có độ chính xác cao tương tự như dùng vân tay (Fingerprint)
Dáng cơ thể khi di chuyển (Gait)	Sử dụng đặc trưng từ khoảng thời gian giữa các bước chân, các bước nghỉ, các bước chạy hoặc sử dụng hình dáng khi di chuyển. Có độ chính xác phụ thuộc nhiều vào môi trường và tình trạng sức khỏe của cá nhân
Hệ thống mạch máu (Hand Vein)	Sử dụng đặc trưng từ cấu trúc mạch máu ở bàn tay, hình ảnh mạch máu có thể được quét qua thông một thiết bị đặc biệt. Được nghiên cứu và vẫn đang trong thời gian phát triển, có độ chính xác cao
Hình ảnh nhiệt của cơ thể (Thermograms)	Sử dụng đặc trưng từ mức nhiệt độ của các bộ phận cơ thể, từ đó chuyển thành hình ảnh của cơ thể biểu diễn ở dạng nhiệt . Các thiết bị chuyên dụng có thể đo nhiệt độ với độ chính xác cao, có sự tương đồng với phương pháp xác thực dựa khuôn mặt (Face Recognition), thường được dùng trong y khoa và quân đội

2.2. Đa yếu tố sinh trắc học và đơn yếu tố sinh trắc học

2.2.1. Nhược điểm của xác thực dựa trên đơn yếu tố sinh trắc học

Một hệ thống xác thực đơn yếu tố thường phải đối mặt với các vấn đề sau (Dinca & Hancke, 2017):

- Dữ liệu bị nhiễu (Noise in sensed data): Cảm biến bị lỗi hoặc được bảo trì không đúng cách có thể thu về dữ liệu bị biến dạng và nhiễu.
- Giả mạo (Spoof attacks): Một số đặc điểm về thể chất và hành vi dễ bị tấn công giả mạo như giả giọng nói, ngón tay giả làm từ gelatin, mô hình khuôn mặt bằng silicon.
- Độ phủ dân số (Insufficient population coverage): Trên quy mô dân số lớn, đơn yếu tố sinh trắc học bị giảm hiệu quả khi các cá nhân có độ tương đồng cao về một yếu tố sinh trắc học nhất định (Asha, & Chellappan, 2008). Ví dụ: Hai người xa lạ có thể có gương mặt rất giống nhau dù chiều cao, cân nặng, giọng nói, của họ hoàn toàn khác nhau.

2.2.2. Ưu điểm của xác thực dựa trên đa yếu tố sinh trắc học

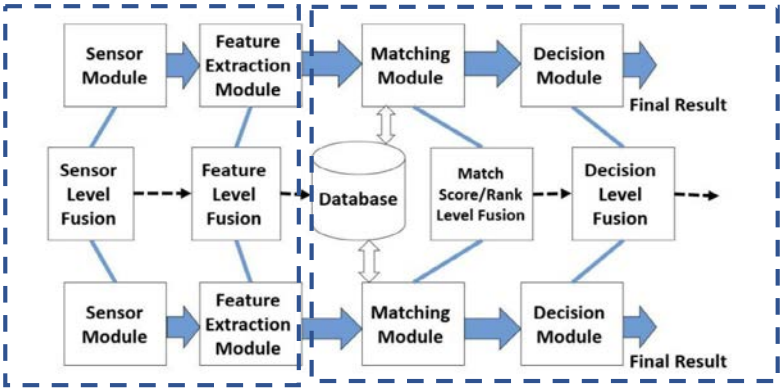
Đa yếu tố sinh trắc học đã được nghiên cứu hơn 2 thập kỉ (Siddiqui et al., 2014) và có nhiều ưu điểm so với đơn yếu tố sinh trắc học:

- Ít phổ biến và độ phủ dân số (Universality and large population coverage): Đa yếu tố sinh trắc học giải quyết vấn đề về độ phủ dân số hiệu quả hơn. Ví dụ: Hai anh em sinh đôi nhưng dấu vân tay sẽ khác nhau.
- Chống giả mạo tốt hơn (Resistance to Spoofing): Việc kẻ tấn công có thể tạo ra nhiều hơn hai sinh trắc học giả mạo phù hợp là rất ít khả thi. Thêm vào đó, hệ thống có thể yêu cầu xác thực tại thời điểm ngẫu nhiên với thời gian có hạn.
- Giảm nhiễu dữ liệu (Reduced noisy data): Hệ thống đa yếu tố sinh trắc học thích ứng tốt với điều kiện môi trường vì dựa trên kết quả từ nhiều yếu tố khác.

Sự hiệu quả của phương pháp đa yếu tố sinh trắc học phụ thuộc vào phương pháp kết hợp của chúng (Multi-Biometric Fusion). Phương pháp kết hợp phổ biến được triển khai có

thể được phân loại thành trước khi tính độ tương đồng (before matching) và sau khi tính độ tương đồng (after matching), (Drozdowski et al., 2020), được minh họa ở Hình 2.

- Kết hợp cảm biến (Sensor Level Fusion): Kết hợp dữ liệu từ nhiều cảm biến khác nhau. Phương pháp này khó triển khai và không hiệu quả khi các yếu tố sinh trắc có nhiều sự khác biệt.
- Kết hợp đặc trưng (Feature Level Fusion): Kết hợp các vector đặc trưng thu được của từng yếu tố lại với nhau. Gặp khó khăn khi muốn thay đổi, cái tiến thuật toán hoặc mô hình học sâu. Đây là phương pháp phức tạp và khó triển khai nhất.
- Kết hợp điểm (Score Level Fusion): Là phương pháp hiệu quả và linh hoạt nhất. Trong phương pháp này, độ tương đồng của từng yếu tố cần được chuẩn hóa như: minimum maximum (MM), hyperbolic tangent (HT), và z-score (ZS)… sau đó tổng hợp dựa trên các hệ số cụ thể và đưa ra điểm tổng hợp.
- Kết hợp quyết định (Decision level fusion) Là phương pháp phổ biến nhất dễ triển khai nhất trong các phương pháp. Mỗi yếu tố sinh trắc sẽ đưa ra kết quả của mình (thành công, không thành công). Quyết định sau cùng sẽ được dựa trên các phép toán logic AND hoặc OR để đưa ra kết quả sau cùng.



Hình 2. *Minh họa phân loại các phương pháp kết hợp đa yêu tố sinh trắc học dựa trên các giai đoạn kết hợp* (Gad et al., 2015)

2.3. Đề xuất giải pháp xác thực đa yếu tố sinh trắc học

Nghiên cứu đề xuất một hệ thống xác thực người học dựa trên đa yếu tố sinh trắc học gồm khuôn mặt và giọng nói, trong đó sử dụng phương pháp kết hợp điểm (score level fusion) vì dễ cải tiến và sự khác biệt lớn về tính chất của dữ liệu sinh trắc học của hai yếu tố khuôn mặt và giọng nói. Cụ thể bài báo tập trung vào:

- Xây dựng mô hình xác thực của từng yếu tố: giọng nói và khuôn mặt. Trong đó tập trung cải tiến mô hình xác thực giọng nói và kế thừa mô hình xác thực khuôn mặt từ các nghiên cứu trước.
- Đề xuất phương pháp kết hợp mới, nhanh, đơn giản, dễ mở rộng và hiệu quả dựa trên kết hợp điểm (score fusion) bằng phương pháp chuẩn hóa min-max.

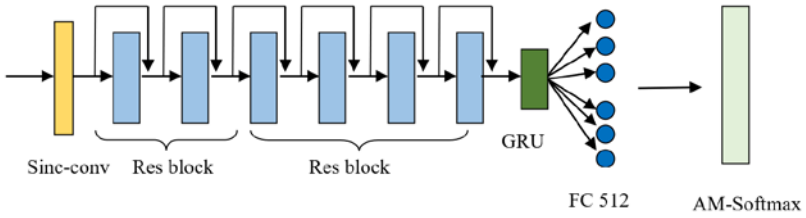
2.3.1. Mô hình nhận dạng người dùng dựa trên yếu tố giọng nói

Dựa trên nội dung lời nói, nhận dạng người dùng bằng giọng nói chia thành các phân loại:

- Phụ thuộc vào nội dung lời nói (Text-Dependent): Nội dung và số lượng các câu nói trong hệ thống là cố định và có giới hạn. Nội dung lời nói khi xác thực và dữ liệu đối sánh trong cơ sở dữ liệu phải giống nhau ở mỗi lần xác thực. Cách tiếp cận này có nhiều hạn chế khi muốn chuyển đổi, mở rộng hoặc tái sử dụng hệ thống.
- Không phụ thuộc vào nội dung lời nói (Text-Independent): Nội dung và số lượng các câu nói trong hệ thống là không cố định và không có giới hạn. Nội dung lời nói khi xác thực và dữ liệu đối sánh trong cơ sở dữ liệu không cần phải giống nhau. Có thể dễ dàng chuyển đổi và tái sử dụng. Đây là xu hướng nghiên cứu của lĩnh vực nhận dạng người nói (Speaker Recognition) và ngày càng phát triển.

Nội dung bài báo chủ yếu tập trung vào việc xác thực người nói (Speaker Verification) dựa trên hướng tiếp cận không phụ thuộc vào nội dung lời nói (Text-Independent).

Bài báo xây dựng mô hình AM-RawNet dựa trên mô hình RawNet2 và hàm loss Additive Margin Soft-max (AM-Softmax). Trong giải pháp đề xuất, mô hình xác thực giọng nói có baseline là mô hình RawNet2 (Jung et al., 2020), minh họa ở Hình 3. Bài báo sử dụng phương pháp tranfer learning bằng cách tinh chỉnh (fine-tuning) lớp fully connected cuối cùng của mô hình sau đó huấn luyện trên tập dữ liệu VoxCeleb1. Thêm vào đó, bài báo cải tiến bằng hàm Additive Margin Soft-max (AM-Softmax) để tăng hiệu quả phân lớp cho mô hình (gọi là AM- RawNet). Đầu ra của mô hình là vector embedding có số chiều là 512, vector này dùng để tính độ tương đồng cosine với dữ liệu sinh trắc học trong giai đoạn đăng kí (The enrollment phase).



Hình 3. *Kiến trúc mô hình AM-RawNet*

Hàm AM-softmax được định nghĩa như sau:

$$Loss = -\frac{1}{n}\sum_{i=1}^{n} log\frac{\emptyset_i}{\emptyset_i + \sum_{j=1,\ j=y_i}^{c} exp(s(W_j^T f_i))} \quad (1)$$

$$\emptyset_i = exp(s(W_{y_j}^T f_i - m))$$

trong đó:

- f_i là đầu vào từ mẫu thứ *i* của lớp fully connected cuối cùng
- W là ma trận hệ số
- s là biến hệ số nhân (scale)
- m là tham số additive margin tương ứng, bài báo sử dụng m = 3.

2.3.2. Mô hình nhận dạng người dùng dựa trên yếu tố khuôn mặt

Dựa trên mức độ khó, nhận dạng khuôn mặt chia thành:

- Dễ (easy with high quality): Khuôn mặt với góc trực diện, độ sáng tốt, thấy rõ khuôn mặt, ít đeo phụ kiện. Không có sự thay đổi theo độ tuổi.
- Trung bình (medium with regular quality): Ảnh khuôn mặt nghiêng, có đeo phụ kiện, bị ảnh hưởng bởi ánh sáng chiếu vào. Thay đổi theo độ tuổi không quá lớn.
- Khó (hard with low quality): Ảnh khuôn mặt có góc nghiêng lớn, điều kiện ánh sáng kém, độ nhiễu lớn, đeo nhiều phụ kiện. Có sự thay đổi lớn theo độ tuổi.

Bài báo kế thừa lại mô hình MagFace (Meng et al., 2021) và fine-tuning cho phù hợp với tập dữ liệu. Nội dung bài báo chủ yếu tập trung vào giải quyết bài toán xác thực khuôn mặt ở mức trung bình, có thể giải quyết được các trường hợp người học ngồi trong điều kiện thiếu sáng, đeo kính, camera nhiễu ở mức trung bình.

Trong giải pháp đề xuất, mô hình xác khuôn mặt có baseline là mô hình MagFace (Meng et al., 2021). Sau đó, sử dụng phương pháp tranfer learning bằng cách tinh chỉnh (fine-tuning) lớp fully connected cuối cùng của mô hình sau đó huấn luyện trên tập dữ liệu VoxCeleb1. Mô hình cho phép nhận dạng khuôn mặt ở mức trung bình và trung bình-khó, hiệu quả trong các khi chịu ảnh hưởng từ môi trường, thay đổi góc nhìn. Mô hình có kiến trúc là IResNet50 sử dụng hàm loss MagFace. Hàm loss của mô hình được triển khai dựa trên làm loss của mô hình ArcFace kết hợp với hàm ý thức về độ khó $m(a_i)$ (magnitude aware angular margin), hàm loss được định nghĩa như sau:

$$L_{Mag} = \frac{1}{N}\sum_{i=0}^{N} L_i \tag{2}$$

$$L_i = -log\frac{e^{s\,cos(\theta_{y_i} + m(a_i))}}{e^{s\,cos(\theta_{y_i} + m(a_i))} + \sum_{j \neq y_i} e^{s\,cos\theta_j}} + \lambda_g g(a_i) \tag{3}$$

trong đó:

- a_i được tính dựa trên công thức $a_i = \|f_i\|$, *với* f_i là vector kết quả của lớp fully-connected cuối cùng của mô hình
- y_i là nhãn của mẫu x_i
- θ_j là góc giữa đặc trưng f_i và tâm của lớp đó
- $g(a_i)$ là hàm lồi giảm dần, trong mô hình MagFace $g(a_i)$:

$$g(a_i) = \frac{1}{a_i} + \frac{1}{u_a^2}a_i \tag{4}$$

- Với u_a = 120
- $m(a_i)$ là một hàm tăng dần, trong mô hình MagFace $m(a_i)$ là một hàm tuyến tính:

$$m(a_i) = \frac{u_m - l_m}{u_a - l_a}(a_i - l_a) + l_m$$

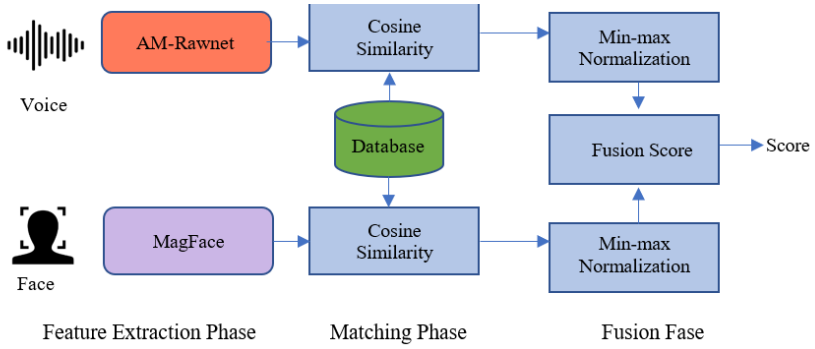
trong đó:

- Hằng số của bài báo là l_a = 15, u_a = 120 (so với nghiên cứu gốc l_a = 10, u_a = 110).
- $l_m = m(l_a),\ u_m = m(u_a)$

- λ_g dùng để điều chỉnh trọng số của hàm $g(a_i)$.

2.3.3. Đề xuất phương pháp kết hợp điểm dựa trên chuẩn hóa lớn nhất và nhỏ nhất (min-max score-fusion)

Nghiên cứu đề xuất phương pháp kết hợp các yếu tố bằng phương pháp kết hợp điểm (score fusion) chuẩn hóa lớn nhất và nhỏ nhất. Trong nghiên cứu sử dụng giá trị lớn nhất (đặt là *MAX*) của các độ tương đồng và sử dụng một ngưỡng (đặt là *THRESHOLD*) để chuẩn hóa. Với phương pháp kết hợp điểm (Score Level Fusion), minh họa ở Hình 4.



Hình 4. *Mô hình xác thực người học bằng phương pháp đa yếu tố sinh trắc học khuôn mặt và giọng nói*

Để trích xuất đặc trưng (Feature Extraction phase), ta sử dụng 2 mô hình để tính vector embedding cho 2 yếu tố:

$$f_{Face} = MagFace(x_{Face})$$

$$f_{Voice} = AM - RawNet(x_{Voice})$$

trong đó: x_{Face} là ảnh khuôn mặt đầu vào, x_{Voice} là bản ghi âm đầu vào

- MagFace và AM − RawNet lần lượt là mô hình MagFace mà mô hình AM-RawNet
- f_{Face} là vector embedding của ảnh đầu vào có kích thước là 512
- f_{Voice} là vector embedding của bản ghi âm đầu vào có kích thước là 256.

Giai đoạn trước khi xác thực (chỉ tiến hành một lần sau khi huấn luyện): cần xác định các tham số $THRESHOLD_k$, MAX_k, MIN_k của yếu tố thứ k:

- Tìm ngưỡng xác thực $THRESHOLD_k$ của từng mô hình dựa vào phương pháp tỉ lệ lỗi cân bằng (Equal Error Rate). Một ngưỡng đạt được tỉ lệ lỗi cân bằng khi: tỉ lệ chấp nhận sai (False Accept Rate, FAR) bằng hoặc xấp xỉ với tỉ lệ từ chối sai (False Reject Rate, FRR).

$$THRESHOLD_k = T_k$$

trong đó: T_k là ngưỡng của yếu tố thứ k sao cho $FAR \approx FRR$

- Tìm độ tương đồng lớn nhất của yếu tố thứ k:

$$MAX_k = max(sim_k)$$

trong đó: sim_k là tập hợp các độ tương đồng cosine của các cặp trong tập thử nghiệm của yếu tố thứ k gồm n phần tử. Ví dụ: sim_{Face} là tập hợp độ tương đồng của n cặp dữ liệu thử nghiệm của yếu tố khuôn mặt gồm n phần tử.

- Giá trị MIN_k của yếu tố thứ k được xác định như sau:

$$MIN_k = \text{THRESHOLD}_k - (MAX_k - \text{THRESHOLD}_k)$$

Giai đoạn tiến hành xác thực: Tính độ tương đồng và chuẩn hóa min-max

- Dữ liệu xác thực của mỗi cá nhân gồm ảnh khuôn mặt và bản ghi âm.
- Sau khi thu được kết quả là các vector embedding, ta đến giai đoạn đối sánh (Matching Phase). Ở giai đoạn này ta đối sánh các vector embedding ở thời điểm hiện tại với các vector embedding trong cơ sở dữ liệu (ở giai đoạn đăng kí). Bài báo sử dụng độ tương đồng cosine để tính khoản cách giữa các vector embedding.

$$SIM_{Face} = S_C(f_{Face}, f_{FaceBase}) = \frac{f_{Face} + f_{FaceBase}}{\|f_{Face}\| . \|f_{FaceBase}\|} \quad (5)$$

$$SIM_{Voice} = S_C(f_{Voice}, f_{VoiceBase}) = \frac{f_{Voice} + f_{VoiceBase}}{\|f_{Voice}\| . \|f_{VoiceBase}\|} \quad (6)$$

trong đó: S_C là hàm tính độ tương đồng cosine, $f_{FaceBase}$ và $f_{VoiceBase}$ là các vector embedding trong cơ sở dữ liệu.

- Độ tượng đồng của yếu tố k sẽ được chuẩn hóa về thang điểm 0 đến 100 dựa trên khoảng $[MIN_k; MAX_k]$ của chính yếu tố k đó, hàm chuẩn hóa của mỗi yếu tố k được xác định trên hàm không liên tục $NORM_k$ như sau:

$$SCORE_k = NORM_k(SIM_k, MIN_k, MAX_k) = \begin{cases} 0, & SIM_k \leq MIN_k \\ 100, & SIM_k \geq MAX_k \\ \dfrac{(SIM_k - MIN_k) * 100}{MAX_k - MIN_k} \end{cases} \quad (7)$$

trong đó: SIM_k là độ tương đồng cosine giữa dữ liệu hiện tại và dữ liệu trong cơ sở dữ liệu của yếu tố thứ k. Ví dụ điểm của yếu tố khuôn mặt được kí hiệu là $SCORE_{Face}$.

- Ở giai đoạn kết hợp điểm (Fusion Fase), điểm của các yếu tố được kết hợp theo một tỉ lệ λ:

$$SCORE_{Fusion} = \lambda . SCORE_{Face} + (1 - \lambda) . SCORE_{Voice} \quad (8)$$

trong đó: $SCORE_{Voice}$ là điểm chuẩn hóa của giọng nói

$SCORE_{Face}$ là điểm chuẩn hóa của yếu tố khuôn mặt

λ là tham số điều chỉnh tỉ lệ của điểm chuẩn hóa của các yếu.

Sau khi chuẩn hóa và tính điểm tổng $SCORE_{Fusion}$, ta sử dụng kết quả này để đưa ra kết quả xác thực: xác thực thành công nếu điểm $SCORE_{Fusion} > 50$, ngược lại xác thực thất bại. Tính hiệu quả của phương pháp kết hợp điểm dựa trên chuẩn hóa lớn nhất và nhỏ nhất được trình bày ở Bảng 4.

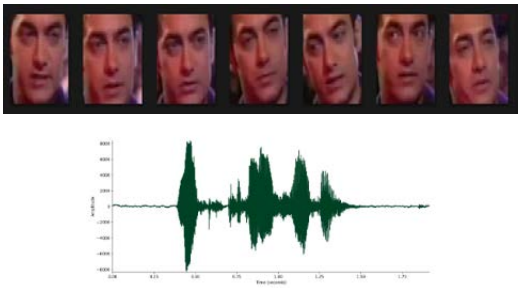
2.4. Tập dữ liệu thử nghiệm

Nghiên cứu sẽ đánh giá kết quả xác thực ở 4 trường hợp sau: đơn yếu tố giọng nói, đơn yếu tố khuôn mặt, phương pháp kết hợp Decision Fusion, phương pháp kết hợp Score Fusion. Tập dữ liệu được sử dụng là tập dữ liệu VoxCeleb1, trong tập dữ liệu này gồm có 2 phần riêng biệt là giọng nói và khuôn mặt, được minh họa ở Hình 5. Để kiểm tra tính hiệu quả trong thực tế, bài báo sử dụng dữ liệu thực tế từ sinh viên trong kì thi Tin học cơ bản, được trình bày ở Bảng 3. Mỗi cá nhân đều có dữ liệu ảnh khuôn mặt và bản ghi âm giọng nói.

Bảng 3. *Phân bố tập dữ liệu trong tập Voxceleb1 cho giọng nói và khuôn mặt*

	VoxCeleb1		**Dữ liệu thực tế sinh viên**	
	Train	Test	Train	Test
Số người	1,000	251	320	80
Bản ghi âmm	122,813	30,703	4165	1043
Khuôn mặt	989,925	263,196	1925	482

a) Tập dữ liệu Voxceleb1 gồm nhiều ảnh khuôn mặt từ nhiều chủng tộc trên thế giới



b) Mỗi cá nhân trong tập dữ liệu bao gồm nhiều ảnh và nhiều bản ghi âm giọng nói của chính họ

Hình 5. *Minh họa tập dữ liệu VoxCeleb1*

2.5. Kết quả và đánh giá

Qua kết quả thực nghiệm cho thấy phương pháp tổng hợp điểm (Score Fusion) cho độ chính xác tăng đáng kể và thể hiện rõ sự hỗ trợ giữa hai yếu tố sinh trắc học giảm tỉ lệ ở cả FA và FR, được thể hiện ở Bảng 4. Trong khi đó, phương pháp kết hợp quyết định Decision OR giảm tối đa tỉ lệ từ chối sai nhưng lại tăng đáng kể tỉ lệ chấp nhận sai. Ngược lại, phương pháp kết hợp quyết định Decision AND giảm tối đa tỉ lệ chấp nhận sai nhưng lại tăng đáng kể tỉ lệ từ chối sai. Qua đối sánh kết quả với các phương pháp khác, phương pháp tổng hợp điểm Score Fusion cho bài toán xác thực người dùng dựa trên đa yếu tố sinh trắc học giọng nói và khuôn mặt mang lại hiệu quả cân bằng và độ chính xác cao hơn.

Bảng 4. *Kết quả mô hình đa yếu tố khuôn mặt và giọng nói trên tập dữ liệu Voxceleb1 bằng phương pháp Score Fusicon Normalization*

	VoxCeleb1					**Dữ liệu thực tế sinh viên**
Phương pháp	TA(%)	TR(%)	FA(%)	FR(%)	Accur (%)	Accur (%)
Giọng nói	49.26	49.40	0.59	0.75	98.66	97.85
Khuôn mặt	47.82	47.80	2.19	2.19	95.62	97.66
Decision OR	49.98	47.43	2.55	0.04	97.41	97.53
Decision AND	47.48	49.97	0.02	2.54	97.45	97.67
Score Fusion	49.56	49.54	0.32	0.58	99.10	99.31

Không ít các mô hình kết hợp các yếu tố sinh trắc học khuôn mặt và giọng nói và cũng đạt được kết quả đáng ghi nhận, được trình bày trong Bảng 5. Trong nghiên cứu của mình Y. Qian và cộng sự (Qian et al., 2019) sử dụng mô hình AVN-F có kiến trúc dựa trên

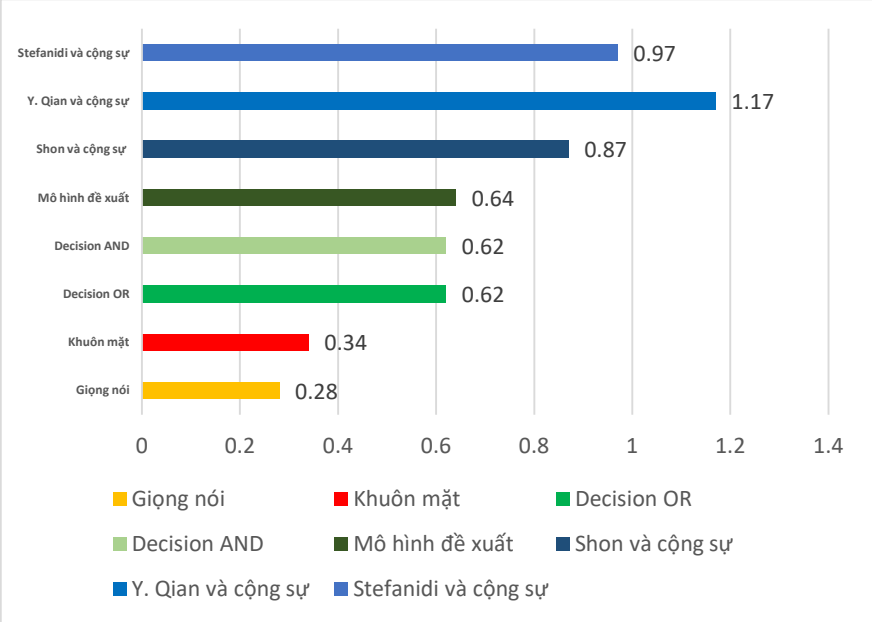
ResNet34 để kết hợp đặc trưng sau đó sử dụng độ tương đồng cosine để đưa ra kết quả xác thực. Trong một nghiên cứu khác, Stefanidi và cộng sự sử dụng mô hình tích chập gồm nhiều lớp convolutional 2D để kết hợp đặt trưng với nhau (Stefanidi et al., 2020). Shon và cộng sự sử dụng mô hình mạng học sâu dựa trên Attention, sau đó đã kết hợp các yếu tố ở mức đặc trưng và ở mức điểm (Shon et al., 2018) sau đó đưa ra kết quả xác thực. Các nghiên cứu này sử dụng mạng học sâu để kết hợp và đưa ra kết quả nên mỗi khi có sự thay đổi hoặc mở rộng thì cần huấn luyện lại mô hình. Mặt khác, các mô hình này đòi hỏi các hệ thống có khả năng tính toán cao hơn nên khó tối ưu khi triển khai trên nền tảng trực tuyến hoặc các thiết bị cầm tay.

Bảng 5. *Kết quả so sánh của giải pháp đề xuất với các mô hình trong thời gian gần đây trên tập dữ liệu Voxceleb1*

Mô hình	ERR(%)	Thời gian xác thực (giây)
Giọng nói	2.32	0.28
Khuôn mặt	3.23	0.34
Shon và cộng sự	0.97	0.87
Y. Qian và cộng sự	1.15	1.17
Stefanidi và cộng sự	1.31	0.97
Mô hình đề xuất	1.45	0.64

Mặc dù có độ lỗi cao hơn các nghiên cứu trước, nhưng bài báo đã đóng góp một mô hình xác thực đa yếu tố khuôn mặt và giọng nói với các ưu điểm:

- Độ chính xác ở mức cao và cao hơn đáng kể so với đơn yếu tố.
- Thời gian xác thực ngắn hơn, thích hợp áp dụng trên các nền tảng trực tuyến hoặc di động.
- Các phương pháp được so sánh mất nhiều thời gian triển khai do sử dụng thêm các mô hình học sâu để kết hợp kết quả, kết hợp đặc trưng. Giải pháp đề xuất có thể mở rộng số lượng yếu tố mà không cần huấn luyện lại mô hình.



Hình 6. *So sánh về mặt thời gian trên trên mỗi xác thực (giây) của các mô hình xác thực đơn yếu tố và đa yếu tố*

Mặc dù, đem lại hiệu quả xác thực đáng kể nhưng thời gian xác thực dài hơn vẫn luôn là một thử thách cho các mô hình sử dụng đa yếu tố. Với thời gian khoảng 0.647 giây đáp ứng được nhu cầu của một kì thi trong thực tế. Tuy nhiên, với số lượng lớn hơn thì hệ thống xác thực cần phải đáp ứng được yêu cầu về mặt phần cứng nhất định.

3. Kết luận

Hiện nay, các hệ thống quản lí đào tạo, nền tảng kiểm tra và các hệ thống học tập trực tuyến của các trường đại học quan tâm việc cung cấp kiến thức cho người học thông những khóa học có nội dung đặc sắc, tăng tính linh hoạt trong kiểm tra đánh giá bằng các kì thi trực tuyến. Nhưng chưa có hệ thống nào tập trung vào việc ngăn chặn các hiện tượng tiêu cực trong quá trình tham gia các khóa học trực tuyến và kì thi trực tuyến. Việc ngăn chặn các hiện tượng tiêu cực trong học tập và kiểm tra trực tuyến là một vấn đề cấp thiết, cần được triển khai kịp thời trên các hệ thống học và kiểm tra trực tuyến nhằm đảm bảo chất lượng và tính công bằng của quá trình dạy và học. Bài báo đã trình bày một giải pháp xác thực đa yếu tố sinh trắc học có hiệu quả tốt hơn so với phương pháp xác thực khác, đồng thời cũng ngăn chặn nguy cơ giả mạo dữ liệu sinh trắc học. Giải pháp cũng đạt độ chính xác là 99.1 %, vượt trội so với 98.66% khi chỉ sử dụng giọng nói và 95.62% khi chỉ sử khuôn mặt. Giải pháp đã được thực nghiệm và tích hợp vào hệ thống ETEST (Hệ thống kiểm tra trực tuyến của Trung tâm Tin học Trường Đại học Sư phạm Thành phố Hồ Chí Minh). Sau khi tích hợp, triển khai cho sinh viên thi đạt được hiệu quả với độ chính xác đạt 99.31%, không có sinh viên nào không thể hoàn thành quá trình xác thực. Thời gian tới, các tác giả sẽ tiếp tục giải quyết bài toán xác thực người dùng trên các nền tảng học tập trực tuyến với mục tiêu nâng cao hiệu quả xác thực và tập trung vào xây dựng các giao thức xác thực chống gian lận.

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

TÀI LIỆU THAM KHẢO

Asha, S., & Chellappan, C. (2008). Authentication of e-learners using multimodal biometric technology. In *2008 International Symposium on Biometrics and Security Technologies* (pp. 1-6). Islamabad, Pakistan. https://doi.org/10.1109/ISBAST.2008.4547640

Dinca, L. M., & Hancke, G. P. (2017). The Fall of One, the Rise of Many: A Survey on Multi-Biometric Fusion Methods. *IEEE Access, 5*, 6247-6289. https://doi.org/10.1109/ACCESS.2017.2694050

Drozdowski, P., Rathgeb, C., Mokroß, B. -A., & Busch, C. (2020). Multi-Biometric Identification With Cascading Database Filtering. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 2*(3), 210-222. https://doi.org/10.1109/TBIOM.2020.2977215

Jung, J., Kim, S., Shim, H., Kim, J., & Yu, H. (2020). Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2004.00526

Gavrilova, M. L., & Monwar, M. M. (2011). Current Trends in Multimodal Biometric System—Rank Level Fusion. In P. S. P. Wang (Ed.), *Pattern Recognition, Machine Intelligence and Biometrics*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22407-2_25

Gad, R., El-Fishawy, N., El-Sayed, A., & Zorkany, M. (2015). Multi-Biometric Systems: A State of the Art Survey and Research Directions. *International Journal of Advanced Computer Science and Applications (IJACSA), 6*(6), 128-138.

Meng, Q., Zhao, S., Huang, Z., & Zhou, F. (2021). MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14225-14234). https://doi.org/10.48550/arXiv.2103.06627

Sabhanayagam, T., Venkatesan, V. P., & Senthamaraikannan, K. (2018). A Comprehensive Survey on Various Biometric Systems. *International Journal of Applied Engineering Research, 13*(5), 2276-2297.

Sharma, S. (2014). An Improved Iris Recognition System Based on 2-D DCT and Hamming Distance Technique. *ICRTEDC-2014, GV/ICRTEDC/08, 1*(2), 32-34.

Shon, S., Oh, T. -H., & Glass, J. (2019). Noise-tolerant Audio-visual Online Person Verification Using an Attention-based Neural Network Fusion. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3995-3999). Brighton, UK. https://doi.org/10.1109/ICASSP.2019.8683477

Siddiqui, A. M. N., Telgad, R., & Deshmukh, P. D. (2014). Multimodal Biometric Systems: Study to Improve Accuracy and Performance. *International Journal of Current Engineering and Technology, 4*(1), 165-171.

Stefanidi, A., Topnikov, A., Tupitsin, G., & Priorov, A. (2020). Application of Convolutional Neural Networks for Multimodal Identification Task. In *2020 26th Conference of Open Innovations Association (FRUCT)* (pp. 423-428). Yaroslavl, Russia. https://doi.org/10.23919/FRUCT48808.2020.9087458

Qian, Y., Chen, Z., & Wang, S. (2021). Audio-Visual Deep Neural Network for Robust Person Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 1079-1092. https://doi.org/10.1109/TASLP.2021.3057230

LEARNER AUTHENTICATION VIA BIOMETRIC-BASED MULTI-FACTOR

Nguyen Quoc Trung[1*], Nguyen Vo Phi Long[1], Le Duc Long[1], Nguyen Dinh Thuc[2]

[1]Ho Chi Minh City University of Education, Vietnam

[2]University of Science, Vietnam National University Ho Chi Minh City, Vietnam

[*]Corresponding author: Nguyen Quoc Trung – Email: trungnq@hcmue.edu.vn

Received: February 16, 2023; Revised: March 18, 2023; Accepted: March 20, 2023

ABSTRACT

Online learning systems, particularly systems for remote student assessment, face a challenge in student authentication. Currently, the most common authentication method for these systems is the username-password approach, which is easy to use but requires users to create strong and complex passwords that are often difficult to remember. Some online learning systems have adopted biometric authentication methods, which enable learners to log in without a password. However, relying on a single biometric factor presents several issues, particularly the stability of biometric information. This article proposes a multi-factor authentication solution that combines facial and voice biometrics. The proposed approach was tested on the VoxCeleb1 dataset and achieved exceptional results, with an accuracy of 99.6%. By comparison, facial recognition alone yielded 95.62% accuracy, and voice recognition alone achieved 98.66%.

Keywords: authentication; biometrics; biometric multi-factor; biometric single-factor; learners authentication; online learning system