

CHẨN ĐOÁN BỆNH TIM MẠCH SỬ DỤNG HỌC MÁY CÓ GIÁM SÁT

Vũ Xuân Hạnh^{}, Nguyễn Đình Dũng[†], Vũ Tuấn Hiệp[‡]*
Email: hanhvx@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 03/11/2023

Ngày phản biện đánh giá: 18/06/2024

Ngày bài báo được duyệt đăng: 26/06/2024

DOI: 10.59266/houjs.2024.410

Tóm tắt: Bệnh tim mạch (CVD) đã cướp đi khoảng 19,1 triệu người, chiếm 33% số ca tử vong toàn cầu vào năm 2022 (theo thống kê của WHO). ECG (Electrocardiogram) được sử dụng để chẩn đoán và theo dõi CVD thông qua điện tâm đồ. Gần đây, một phương pháp dựa trên máy học được đề xuất để phát hiện CVD sớm, hỗ trợ chẩn đoán nhanh và chính xác hơn. Trong nghiên cứu này, nhóm tác giả sử dụng kỹ thuật máy học có giám sát xây dựng mô hình hỗ trợ chẩn đoán bệnh tim mạch dựa trên một số triệu chứng cận lâm sàng trong hai tập dữ liệu [1] [2]. Thử nghiệm đạt độ chính xác toàn cục (Accuracy - ACC) là 98.21%, tỷ lệ âm tính giả (False Negative Rate – FNR), tỷ lệ dương tính giả (False Positive Rate – FPR) lần lượt là 0.63% và 2.82%. Tỷ lệ phát hiện của mô hình (Detect Rate - DR) đạt 99.02% cùng 10 đặc trưng sử dụng thuật toán Random Forest.

Từ khóa: Bệnh tim mạch, Phát hiện bệnh tim mạch, Máy học.

I. Đặt vấn đề

Công tác chăm sóc sức khỏe cộng đồng là một trong những vấn đề được quan tâm trên toàn thế giới. WHO tuyên bố rằng việc chăm sóc, tăng cường sức khỏe là quyền cơ bản của con người. Do đó, các bệnh mãn tính (CDs) có ảnh hưởng khá lớn và chiếm một tỷ lệ đáng kể trong tổng số các ca tử vong, một trong số đó là bệnh tim mạch (CVD) [3]. Nguyên nhân chính

gây ra là thói quen ăn uống không lành mạnh, hút thuốc lá, uống quá nhiều rượu và lối sống không chuẩn mực của một số nhóm người trong cộng đồng.

Tại Hoa Kỳ (US), bệnh tim mạch (CVD) là một trong những nguyên nhân chính gây tử vong [4]. Các quốc gia khác trên thế giới cũng đang giải quyết các vấn đề liên quan đến CVD. Theo một công bố gần đây, các bệnh mãn tính là nguyên

^{*} Trường Đại học Mở Hà Nội

[†] Sinh viên, Trường Đại học Mở Hà Nội

[‡] Sinh viên, Trường Đại học Thủy Lợi

nhân gây ra 86.5% số ca tử vong ở Trung Quốc [5].

Tại Việt Nam số ca tử vong bởi các bệnh tim mạch là hơn 200,000 ca mỗi năm chiếm khoảng 33% tổng số ca tử vong (theo thống kê của Bộ Y tế). Trong 1, 2 thập niên gần đây, tỷ lệ tăng huyết áp tăng khoảng 1.2 % mỗi năm và chiếm 26.7 % ở người trưởng thành. Hệ lụy là các bệnh lý tim mạch làm tăng nguy cơ tử vong do đột quỵ lên gấp 4 lần và nguy cơ tử vong gấp 3 lần [6]. Để khắc phục tình trạng trên, những người có khả năng mắc bệnh tim mạch cần được phát hiện và theo dõi sớm, từ đó đưa ra liệu trình phù hợp, tránh nguy cơ tử vong.

Với sự phát triển mạnh mẽ của khoa học và công nghệ, nhất là trong lĩnh vực trí tuệ nhân tạo (AI), rất nhiều ứng dụng đã và đang liên tục phát triển trong mọi lĩnh vực, trong đó phải kể đến các giải pháp tích hợp AI để hỗ trợ chẩn đoán, điều trị nhiều nhóm bệnh trong lĩnh vực y tế, một ví dụ cụ thể hơn đó là mô hình chẩn đoán khả năng mắc bệnh tim mạch dựa trên các số liệu được máy móc ghi nhận tự động liên tục với tần suất cao và các triệu chứng cận lâm sàng khác.

Trong nghiên cứu này, nhóm tác giả đưa ra mô hình đề xuất để xây dựng ứng dụng hỗ trợ chẩn đoán bệnh nhân có nguy cơ mắc các bệnh tim mạch dựa trên các triệu chứng lâm sàng và các thông tin bệnh án khác một cách nhanh chóng và hiệu quả. Bài viết này có cấu trúc như sau: mục II – Cơ sở lý thuyết, thảo luận về một số nghiên cứu liên quan đến chẩn đoán bệnh tim mạch sử dụng mô hình máy học; mục III – Phương pháp nghiên cứu, trình bày về mô hình đề xuất, chi tiết về các đặc

trung trong bệnh tim mạch và số đo đánh giá; Kết quả thử nghiệm được đánh giá và phân tích trong mục IV – Kết quả và đánh giá; Kết luận được trình bày trong mục V.

II. Cơ sở lý thuyết

Gần đây có nhiều công trình nghiên cứu công bố đề xuất mô hình hỗ trợ chẩn đoán bệnh tim mạch dựa trên máy học. Chithambaram T và cộng sự [7] đã sử dụng 4 thuật toán là SVM, cây quyết định, Random Forest và KNN vào phân loại bệnh tim đạt hiệu suất 86.2% khi sử dụng Gaussian SVM. Tương tự, Satabdi Swain và cộng sự [8] đã thí nghiệm trên dữ liệu y tế của người Framingham, trong đó có thông tin về thói quen, lối sống và tiền sử bệnh của họ với 16 đặc trưng. Thấy rằng, thuật toán Random Forest vượt trội đạt độ chính xác 84.89%, trong khi thuật toán hồi quy logistic, cây quyết định, ANN đạt độ chính xác lần lượt là 84.81%, 84.66% và 84.54%. Prof. Madhavi Tota và cộng sự [9] đã đề xuất mô hình máy học để chẩn đoán sử dụng một số thuật toán cùng tập dữ liệu với 14 đặc trưng của người bệnh, đạt hiệu suất tốt nhất với Random Forest là 86.89% và XGBoost là 78.69%. KNN hoạt động với độ chính xác kém nhất là 57.83%. Sujata Joshi và cộng sự [10] đã tiến hành phân loại bệnh tim với tập dữ liệu gồm 14 đặc trưng sử dụng Random Forest đạt 92.2% và Naïve Bayes đạt 84.2 %. Takci H [11] để chẩn đoán đã kết hợp hai phương pháp học máy và chọn lọc đặc trưng. ACC đạt 82.59% khi không chọn lọc đặc trưng và được cải thiện lên 84.81% khi chọn lọc, sử dụng thuật toán SVM-linear và Naïve Bayes.

Với kết quả đạt được từ các nghiên cứu trước đây, hiệu suất chẩn đoán, phát hiện chưa cao, điều này mở ra nhiều thách thức cho các nhà nghiên cứu tìm hiểu, cải tiến và phát triển mô hình, từ đó xây dựng

các hệ thống hỗ trợ chẩn đoán, tư vấn điều trị cho bệnh nhân nhanh chóng kịp thời. Với mục tiêu trên, nhóm tác giả xem xét các đặc trưng từ 2 bộ dữ liệu bệnh tim mạch [1] [2], lựa chọn đặc trưng để giảm thời gian huấn luyện, tăng hiệu năng sử dụng thiết bị và hiệu suất chẩn đoán.

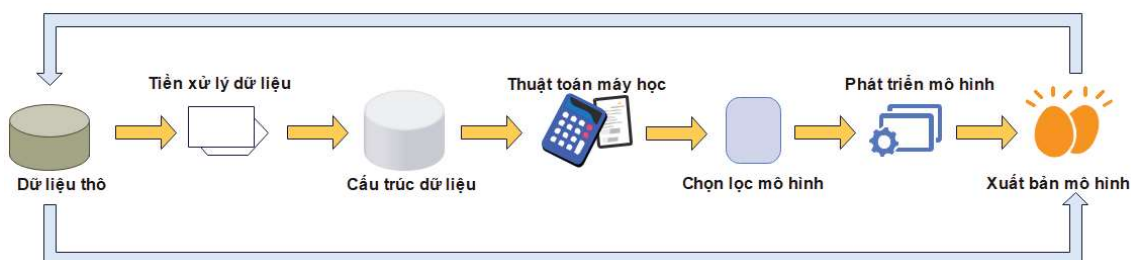
III. Phương pháp nghiên cứu

Nhóm tác giả triển khai dựa trên quy trình tổng thể, được mô tả trong Hình 1. Quy trình bao gồm 6 bước chính: thu thập dữ liệu, tiền xử lý dữ liệu, cấu trúc dữ liệu, thuật toán máy học, chọn lọc mô hình, phát triển mô hình, xuất bản mô hình.

3.1. Học máy có giám sát

Máy học có giám sát, máy học không giám sát, học tăng cường là ba nhóm kỹ

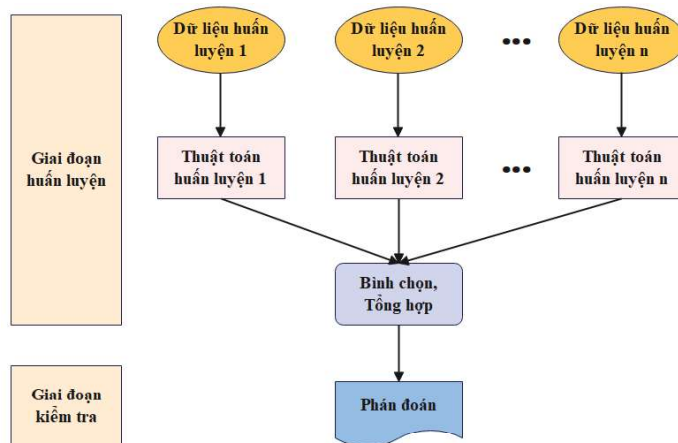
thuật trong lĩnh vực học máy. Tuy nhiên, với đặc thù bài toán là phân loại nhị phân, do đó nhóm tác giả nghiên cứu lựa chọn máy học có giám sát để huấn luyện mô hình. Máy học có giám sát cung cấp hàng loạt các thuật toán như Hồi quy tuyến tính, Cây quyết định, Random Forest, SVM, Naïve Bayes, v.v. Máy học có giám sát sử dụng tập huấn luyện để huấn luyện mô hình mang lại kết quả mong muốn. Tập dữ liệu huấn luyện này bao gồm đầu vào và đầu ra. Thuật toán đo độ chính xác của nó thông qua hàm mất mát, điều chỉnh cho đến khi lỗi được giảm thiểu [12]. Ngoài ra, nhóm tác giả sử dụng phương pháp kiểm tra chéo, cân bằng dữ liệu, giám nhiễu để tránh tình trạng quá khớp và sử dụng ma trận nhầm lẫn để đánh giá, hiệu chỉnh mô hình.



Hình 1: Tổng quan về quy trình máy học

Kỹ thuật máy học được nhóm tác giả sử dụng trong nghiên cứu này là Random Forest (100 cây), được chọn sau khi thử

nghiệm các thuật toán: Random Forest, cây J48, Naïve Bayes, máy hỗ trợ véc tơ (được trình bày tại mục 4.2).



Hình 2: Thuật toán rừng ngẫu nhiên [13]

Thuật toán Random Forest hoạt động bằng cách tạo ra một số cây quyết định trong giai đoạn huấn luyện. Mỗi cây được xây dựng bằng cách sử dụng một tập con ngẫu nhiên dữ liệu để đo lường ngẫu nhiên trong mỗi phân vùng. Tính ngẫu nhiên này tạo ra sự khác biệt giữa các cây riêng lẻ, giảm nguy cơ quá mức và cải thiện hiệu suất tổng thể. Trong dự đoán, thuật toán tổng hợp kết quả của tất cả các cây, bằng cách bỏ phiếu (đối với bài toán phân loại) hoặc lấy trung bình (đối với bài toán hồi quy). Quá trình ra quyết định hợp tác này, được hỗ trợ bởi nhiều cây với thông tin chi tiết của chúng, cung cấp một ví dụ về kết quả ổn định và chính xác. Random Forest được sử dụng rộng rãi cho các bài toán phân loại và hồi quy, được biết đến với khả năng xử lý dữ liệu phức tạp, giảm tình trạng quá mức và cung cấp dự báo đáng tin cậy trong các môi trường khác nhau.

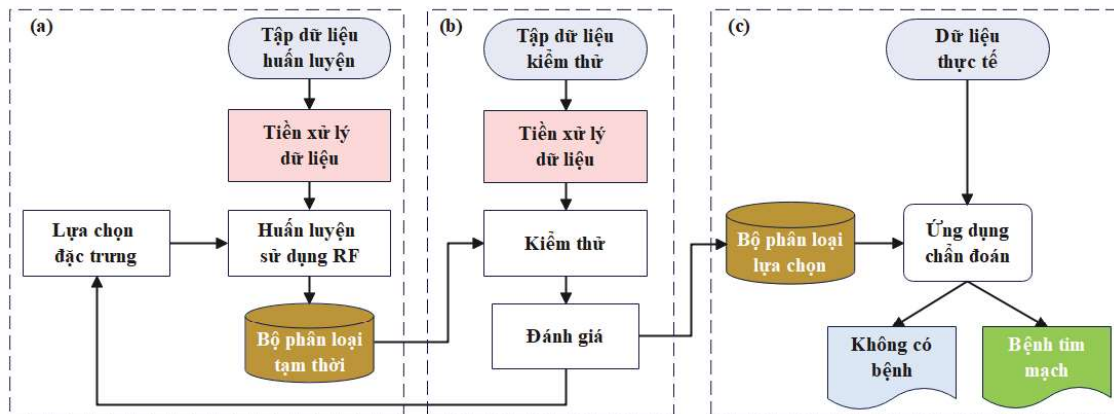
3.2. Lựa chọn đặc trưng

Trong nghiên cứu này, nhóm tác giả phân tích mức độ ảnh hưởng của các đặc trưng trong dữ liệu gốc. Từ đó, giảm chiều dữ liệu bằng cách lược bớt một số

đặc trưng ít ảnh hưởng mà tại đó hiệu suất mô hình giảm không đáng kể. Mục đích để giảm thời gian huấn luyện cũng như tiết kiệm hiệu năng của máy tính. Kết quả được trình bày tại mục 4.3.

3.3. Mô hình phát hiện

Mô hình nhóm tác giả đề xuất sử dụng thuật toán Random Forest được chia thành 3 giai đoạn được minh họa tại hình 3: (a) Giai đoạn huấn luyện: Tập dữ liệu huấn luyện bao gồm người bị bệnh và người không bị bệnh. Ở giai đoạn 1, 12 đặc trưng ban đầu sử dụng 4 thuật toán tại mục 3.1 để lựa chọn ra thuật toán có hiệu suất cao nhất. Sau đó, sử dụng thuật toán đã lựa chọn để đưa ra bộ phân loại tạm thời; (b) Giai đoạn kiểm thử: sử dụng bộ phân loại tạm thời để kiểm thử, sau đó đánh giá mô hình. Sử dụng thuật toán để kiểm tra mức ảnh hưởng của các đặc trưng từ đó lược bớt đặc trưng ít ảnh hưởng quay lại giai đoạn huấn luyện cho đến khi chọn lựa được mô hình tối ưu. (c) Giai đoạn phát triển: sử dụng mô hình để xây dựng ứng dụng hỗ trợ chẩn đoán người bệnh có nguy cơ mắc bệnh tim mạch.



Hình 3: Mô hình chẩn đoán bệnh tim mạch đề xuất

3.4. Phương pháp đánh giá

Nhóm tác giả sử dụng các độ đo FPR, FNR, ACC, F1 để đánh giá mô hình:

- Tỷ lệ dương tính giả:

$$FPR = \frac{FP}{FP + TN} \quad (1)$$

- Tỷ lệ âm tính giả:

$$FNR = \frac{FN}{FN + TP} \quad (2)$$

- Độ đo:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (3)$$

- Độ chính xác toàn cục:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

trong đó, TP là số lượng người bị bệnh được chẩn đoán đúng, TN là số lượng người không bị bệnh được chẩn đoán đúng, FP là số lượng người không bị bệnh bị chẩn đoán thành người bị bệnh (nhầm lẫn) và FN là số lượng người bị bệnh bị chẩn đoán thành người không bị bệnh (bỏ sót).

- Tỷ lệ kiểm thử:

$$DR = \frac{N_{Correct}}{N_{Test}} \quad (5)$$

trong đó, NCorrect là số mẫu chẩn đoán chính xác, NTest là tổng số mẫu kiểm thử.

IV. Kết quả và thảo luận

4.1. Tập dữ liệu huấn luyện và kiểm thử

Hai bộ dữ liệu nhóm tác giả sử dụng được lấy từ một trong những bệnh viện đa khoa ở Ấn Độ [1] và 4 cơ sở dữ liệu lấy từ Cleveland, Hungary, Switzerland, and Long Beach [2]. Từ hai bộ dữ liệu này trích chọn 12 đặc trưng cùng nhau được mô tả tại bảng 4 với 1,929 mẫu. Hai tập dữ liệu này hữu ích cho việc phát hiện bệnh tim ở giai đoạn đầu và sử dụng để tạo ra mô hình chẩn đoán. Tập dữ liệu huấn luyện chọn ra 840 mẫu bình thường và 840 mẫu bị bệnh. Tập dữ liệu kiểm thử với 160 người [1] và 89 người [2] bị bệnh được mô tả trong bảng 1.

Bảng 1: Dữ liệu huấn luyện và kiểm thử

Tập dữ liệu	Bình thường		Bị bệnh	
Nguồn dữ liệu	[1]	[2]	[1]	[2]
Huấn luyện	420	420	420	420
Kiểm thử			160	89

4.2. Lựa chọn thuật toán

Pha huấn luyện, sử dụng 4 thuật toán máy học, kiểm tra chéo 10 lần để xây dựng mô hình. Hiệu suất mô hình tốt nhất được mô tả tại bảng 2, với ACC và F1 lần lượt bằng 98,21% và 98,28% với thuật toán Random Forest (100 cây).

Bảng 2: Hiệu suất của một số kỹ thuật học máy

Thuật toán	ACC	F1
Random Forest (100 cây)	98,21%	98,28%
J48	95,29%	95,40%
Naïve Bayes	88,39%	88,69%
Support Vector Machine	69,64%	70,17%

Bảng 3: So sánh với các công bố trước đây

Đề xuất	Sử dụng	ACC
Chithambaram T và cộng sự [7]	KNN, Random Forest, SVM,...	86,20 %
Satabdi Swain và cộng sự [8]	Random Forest, ANN, ...	84.89 %
Madhavi Tota và cộng sự [9]	Random Forest, XGBoots, KNN,...	86.89 %
Sujata Joshi và cộng sự [10]	Random Forest, Naïve Bayes	92.20 %
Của chúng tôi	Random Forest (100)	98,21 %

Mặt khác, để tối ưu hóa thuật toán, lần lượt thay đổi số cây sử dụng là 50, 100, 150 đạt được ACC lần lượt là: 98.20%,

98.21%, 97.92%. Do đó, lựa chọn thuật toán Random Forest với số cây là 100 để huấn luyện mô hình và kiểm thử.

Bảng 4: 12 đặc trưng để chẩn đoán bệnh tim mạch trong dataset

Đặc trưng	Giá trị đặc trưng
Tuổi	Tuổi
Giới tính	1, 0 (0 = Nữ, 1 = Nam)
Kiểu đau ngực	0, 1, 2, 3 (0: đau thắt ngực điển hình, 1: đau thắt ngực không điển hình, 2: đau không đau thắt ngực, 3: bất tỉnh)
Huyết áp lúc nghỉ ngơi	94-200 (in mm HG)
Cholesterol huyết thanh	126-564 (in mg/dl)
Lượng đường trong máu	0,1 > 120 mg/dl (0 = Sai, 1 = Đúng)
Kết quả điện tâm đồ lúc nghỉ	0, 1, 2 (0: bình thường, 1: có bất thường sóng ST-T (sóng T đảo ngược và/hoặc ST chênh lên hoặc chênh xuống > 0,05 mV), 2: biểu thị phì đại thất trái có thể xảy ra hoặc xác định theo tiêu chuẩn Estes)
Nhịp tim tối đa đạt được	71-202
Đau thắt ngực do tập thể dục	0, 1 (0 = Không, 1 = Có)
Đỉnh cũ = ST	0-6.2
Độ dốc của đoạn ST gắng sức đỉnh điểm	1, 2, 3 (1 – dốc lên, 2 – bằng phẳng, 3 dốc xuống)
Số lượng mạch chính	0, 1, 2, 3

Sử dụng thuật toán Random Forest xây dựng mô hình cùng với tập dữ liệu huấn luyện tại bảng 1. Mô hình đề xuất với 10 đặc trưng người bị bệnh đạt độ chính xác ACC là 98.21%, FNR là 0.63%, FPR là 2.82% tốt hơn so với một số công bố trước đây, mô tả tại bảng 3.

4.3 Kết quả và đánh giá

Sử dụng mô hình đề xuất với thuật toán Random Forest sử dụng 100 cây kiểm thử trên tập dữ liệu 160 mẫu [1] và 89 mẫu [2] người bị bệnh cho tỷ lệ DR được thể hiện tại Bảng 5. Mặt khác, khi lược bớt 2 đặc trưng: lượng đường trong

máu và đau thắt ngực do tập thể dục có mức ảnh hưởng thấp nhất để huấn luyện lại và kiểm thử mô hình cho kết quả: ACC của mô hình là 98.21%, F1 là 98.28%, ở đây có sự thay đổi các số đo FNR và FPR cụ thể là 0.63% và 2.82% so với 1.84% và 1.73%. Như vậy, ACC và F1 là tương đương, tuy nhiên FNR giảm và FPR tăng theo chiều hướng tích cực. Với đặc thù các bài toán chẩn đoán, tỷ lệ âm tính giả càng thấp càng tốt. Mặt khác, với các tập dữ liệu huấn luyện lớn, thời gian huấn luyện, hiệu năng sử dụng thiết bị sẽ tốt hơn khi sử dụng 10 đặc trưng.

Bảng 5: Hiệu suất mô hình

Mô hình	ACC	F1	FNR	FPR	DR	
Bộ dữ liệu	840 mẫu [1] và 840 mẫu [2]				160 mẫu [1]	89 mẫu [2]
12 đặc trưng	98.21%	98.26%	1.84%	1.73%	96.87%	95.62%
10 đặc trưng	98.21%	98.28%	0.63%	2.82%	99.02%	99.01%

4.4 Xây dựng mô hình ứng dụng

Nhóm tác giả xây dựng một ứng dụng dựa trên mô hình đề xuất để hiện thực hóa hỗ trợ chẩn đoán, người dùng nhập các thông tin bệnh lý vào ứng dụng, ứng dụng sẽ chẩn đoán và đưa ra kết quả một người có bị bệnh tim mạch hay không, minh họa tại hình 4.

Hình 4: Màn hình ứng dụng

V. Kết luận

Chẩn đoán, phát hiện sớm người bị bệnh tim là nhu cầu rất cấp thiết hiện nay, với mục đích dự báo nguy cơ mắc bệnh tim mạch, đồng thời cải thiện độ chính xác trong chẩn đoán bệnh tim mạch. Nhóm tác giả đã nghiên cứu chi tiết 12 đặc trưng của người bị bệnh, sau đó lựa chọn 10 đặc trưng ảnh hưởng lớn để xây dựng mô hình. Việc giảm chiều dữ liệu trong đề xuất của nhóm tác giả làm tăng hiệu quả sử dụng tài nguyên, giảm thiểu thời gian huấn luyện và kiểm thử. Việc sử dụng hai bộ dữ liệu từ 2 nguồn khác nhau cho thấy độ tin cậy của mô hình.

Kết quả của nghiên cứu là khá khả quan khi đạt được ACC 98.21%, FNR 0.63% và DR 99.02%, tốt hơn so với các nghiên cứu trước đây. Tuy nhiên cũng tồn tại một số hạn chế: (i) hai tập dữ liệu dùng

để huấn luyện và kiểm thử không nhiều (1929 mẫu); (ii) không lấy được dữ liệu thực tế bệnh nhân bệnh tim mạch tại Việt Nam để thử nghiệm. Để tăng hiệu suất, độ tin cậy của mô hình, nhóm tác giả tiếp tục nghiên cứu, bổ sung các đặc trưng, sử dụng các tập dữ liệu lớn và đa dạng để xây dựng ứng dụng hỗ trợ chẩn đoán bệnh tim mạch chính xác, hiệu quả hơn.

Tài liệu tham khảo:

- [1]. J. Dumlaio, "Cardiovascular Disease Dataset," 16 4 2021. [Online]. Available: <https://www.kaggle.com/datasets/jocelyndumlaio/cardiovascular-disease-dataset>. [Accessed 24 3 2024].
- [2]. D. Lapp, "Heart Disease Dataset," Kaggle, 2019. [Accessed 25 3 2024].
- [3]. Available: www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download.
- [4]. W. H. Organization, "Cardiovascular Disease," WHO, [Online]. Available: <https://www.who.int/vietnam/vi/health-topics/cardiovascular-disease>. [Accessed 16 3 2024].
- [5]. V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," Healthc. Anal, vol. 2, no. doi:10.1016/j.health.2022.100016, 2022.
- [6]. M. Ganesan and N. Sivakumar, "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models," IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019, no. doi: 10.1109/ICSCAN.2019.8878850, pp. 1-5, 2019.
- [7]. B. Y. tế, "Bệnh tim mạch - Nguyên nhân gây tử vong hàng đầu," Bộ Y tế, 16 6 2023. [Online]. Available: https://moh.gov.vn/hoat-dong-cua-dia-phuong/-/asset_publisher/gHb1a8vOQDuS/content/khoang-200-000-nguoi-viet-tu-vong-vi-tim-mach-moi-nam-8-loi-khuyen-e-khong-mac-benh-nay. [Accessed 15 3 2024].

- [8]. Chithambaram T, Logesh Kannan N, Gowsalya M, “Heart disease detection using machine learning”, Research Square (2020) 1-5. <https://doi.org/10.21203/rs.3.rs97004/v1>,” Research Square, pp. DOI: <https://doi.org/10.21203/rs.3.rs-97004/v1>, 2020.
- [9]. Satabdi Swain, Sujata Chakravarty, Bijay Paikaray, Harshvardhan Bhojar, “Heart Disease Detection Using Machine Learning Techniques,” in Proceedings of the International Health Informatics Conference, DOI: 10.1007/978-981-19-9090-8_24, 2023.
- [10]. Prof. Madhavi Tota, Manthan Moom, Pranit Nagrale, Akshay Pandav, Gunjan Das, “Heart Diseases Prediction System using ML,” IJAR SCT ISSN (Online) 2581-9429, vol. 3, no. 10, 2023.
- [11]. Sujata Joshi, Mydhili K. Nair, “Prediction of Heart Disease Using Classification Based Data Mining Techniques,” Computational Intelligence in Data Mining, Springer, vol. 2, no. DOI: 10.1007/978-81-322-2208-8_46, 2015.
- [12]. H. Takci, “Improvement of heart attack prediction by the feature selection methods,” Turkish Journal of Electrical Engineering & Computer Sciences, no. DOI: 10.3906/elk-1611-235, 2018.
- [13]. V. H. Tiệp, “Phân nhóm các thuật toán Machine Learning,” in Machine Learning, <https://machinelearningcoban.com/2016/12/27/categories/#supervised-learning-hoc-co-giam-sat>, 2016.
- [14]. “Random Forest Algorithm,” Javatpoint, [Online]. Available: <https://www.javatpoint.com/machine-learning-random-forest-algorithm>. [Accessed 14 2 2024].

CARDIOVASCULAR DISEASE DIAGNOSIS USING SUPERVISED MACHINE LEARNING

*Vu Xuan Hanh[§], Nguyen Dinh Dung[¶], Vu Tuan Hiep^{**}*

Abstract: Cardiovascular diseases (CVD) claimed approximately 19.1 million lives, accounting for 33% of global deaths in 2022 (according to WHO statistics). ECG (Electrocardiogram) is used to diagnose and monitor CVD through electrocardiographic recordings. Recently, a machine learning-based method has been proposed to detect CVD early, aiding in faster and more accurate diagnoses. In this study, the authors utilized supervised machine learning techniques to build a model for diagnosing cardiovascular diseases based on certain clinical symptoms in two datasets [1] [2]. The experiment achieved an overall accuracy (Accuracy - ACC) of 98.21%, a false negative rate (FNR) of 0.63%, and a false positive rate (FPR) of 2.82%. The model's detection rate (DR) reached 99.02%, using 10 features with the Random Forest algorithm.

Keywords: Cardiovascular disease, Detection of Cardiovascular disease, Machine Learning.

[§] Hanoi Open University

[¶] Student, Hanoi Open University

^{**} Student, Thuyloi University