Mặc dù tiêu chuẩn chọn bệnh chỉ thu thập loại B, C nhưng nghiên cứu của chúng tôi không gặp trường hợp ở dạng C3. Trong 4 BN bị gãy xương loại C là do TNGT, với lực chấn thương lớn, mức chấn thương năng lượng cao, có 2 BN kèm theo tổn thương kết hợp tuy nhiên có thể do C3 là dạng dãy phức tạp, không gặp trong 63 BN của mẫu nghiên cứu này.

V. KẾT LUẬN

Tuổi của các BN dao động từ 19 tuổi đến 65 tuổi, độ tuổi trung bình của nhóm nghiên cứu là 34,19 ± 12,63 tuổi. Tỉ lệ nam/nữ là 3,2/1 với 15 BN nữ (23,81%) và 48 BN nam (76,19%). Nguyên nhân thường do tai nạn giao thông (TNGT), tai nạn lao động (TNLĐ), tai nạn sinh hoạt (TNSH) và tai nạn thể thao (TNTT), trong đó tỉ lệ bị TNGT nhiều hơn. Tần suất bị gãy xương cẳng tay bên trái cao hơn bên phải, với gãy cùng mức là 51 trường hợp, chiếm 80,95% các trường hợp bị gãy xương. Gãy loại B theo phân loại AO chiếm 93,65%.

TÀI LIỆU THAM KHẢO

1. **Lê Văn Hiệu** (2019), Đánh giá kết quả điều trị gãy kín thân hai xương cẳng tay bằng kết xương nẹp vít tại Bệnh viện Quân y 175, Luận văn thạc sĩ y học, Học viện Quân y.
2. **Huỳnh Văn Lem** (2016), Đánh giá kết quả điều trị phẫu thuật gãy kín hai xương cẳng tay ở người lớn bằng nẹp vít nén ép động tại bệnh viện đa khoa khu vực Hóc Môn, Luận văn bác sĩ chuyên khoa cấp II, Trường Đại học y khoa Phạm Ngọc Thạch.
3. **Lê Ngọc Thường** (2010), Đánh giá kết quả điều trị gãy kín thân hai xương cẳng tay bằng phương pháp kết xương nẹp vít tại Bệnh viện Bưu điện, Luận án tiến sĩ y học, Học viện Quân y.
4. **Nguyễn Công Trình** (1995), Nhận xét 149 trường hợp gãy kín thân hai xương cẳng tay ở người lớn được điều trị tại bệnh viện Việt Đức trong 2 năm 1993 -1994, Luận văn thạc sĩ y học, Trường Đại học Y Hà Nội.
5. **Bot A.G.** (2011), "Long-term outcomes of fractures of both bones of the forearm", The Journal of bone and joint surgery. American volume, vol. 93 (6), pp. 527-532.
6. **Tran T.D.** (2017), "The surgical outcomes of diaphyseal fractures of radius and ulna treated by plate and screws fixation in Vietnam", Open Journal of Trauma, vol. 1, pp. 066-068.
7. **Truntzer J.** (2014), "Forearm diaphyseal fractures in the adolescent population: treatment and management", European Journal of Orthopaedic Surgery & Traumatology, vol. 25, pp. 201-209.

DỰ ĐOÁN BỆNH LÝ TIM MẠCH BẰNG PHƯƠNG PHÁP KẾT HỢP BẰNG CHỨNG SỬ DỤNG LÝ THUYẾT DEMPSTER SHAFER

Nguyễn Thái Hà Dương[1], Lê Đình Khiết[1], Lê Trần Đạt[1], Phạm Thị Thu Phương[1], Ngô Thị Huế[1], Phan Thị Ngọc Lan[1], Phạm Thanh Xuân[1]

TÓM TẮT

Nhóm bệnh lý tim mạch là nguyên nhân gây tử vong hàng đầu trên thế giới, chiếm 31% tổng số ca tử vong. Việc chẩn đoán sớm bệnh và giai đoạn bệnh hỗ trợ rất nhiều cho quá trình điều trị, hạn chế sự tiến triển cũng như biến chứng và tỷ lệ tử vong. Quá trình này được thực hiện thông qua sự phân tích những thông tin, bằng chứng, triệu chứng thăm khám lâm sàng, cận lâm sàng bởi các chuyên gia, y bác sĩ. Gần đây, để góp phần hỗ trợ cho quá trình chẩn đoán, phương pháp tiếp cận trí tuệ nhân tạo đã được áp dụng để tăng tốc quá trình phân tích và xử lý. Các phương pháp này hầu hết sử dụng lý thuyết xác suất với vai trò trung tâm là định lý Bayes. Trong nghiên cứu này, chúng tôi cũng dự đoán bệnh lý tim mạch theo hướng tiếp cận khoa học dữ liệu, nhưng đi theo một nhánh khác – kết hợp bằng chứng sử dụng lý thuyết Dempster Shafer. Cụ thể, mỗi triệu chứng được xem là một bằng chứng để kết luận về bệnh với một mức độ không chắc chắn nào đó. Phép kết hợp Dempster được dùng để tổng hợp các bằng chứng. Mức độ không chắc chắn của mỗi bằng chứng sẽ được tìm bởi thuật toán tối ưu sườn dốc (gradient descent). Kết quả bước đầu cho thấy phương pháp mới này không chỉ có sự cải thiện đáng kể về khả năng dự đoán khi so sánh với các phương pháp Bayes mà còn chỉ ra được mức độ chắc chắn của từng triệu chứng trong quá trình chẩn đoán. Những kết quả này cho phép sự kỳ vọng vào khả năng hỗ trợ lâm sàng của phương pháp cũng như tiềm năng ứng dụng của khoa học dữ liệu vào lĩnh vực y học.

Từ khóa: Dempster Shafer Theory, Machine learning, Bệnh lý tim mạch

[1]*Trường đại học Y Dược, Đại học quốc gia Hà Nội*
Chịu trách nhiệm chính: Nguyễn Thái Hà Dương
Email: duongnth.ump@vnu.edu.vn
Ngày nhận bài: 3.7.2023
Ngày phản biện khoa học: 18.8.2023
Ngày duyệt bài: 7.9.2023

SUMMARY

PREDICTING CARDIOVASCULAR DISEASES BY COMBINING EVIDENCES USING DEMPSTER SHAFER THEORY

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for 31% of all deaths. The early diagnosis and stage of the diseases greatly support the treatment process, limiting the evolutions, complications and deadth rate. This process through the analysis of information, evidence, clinical examination symptoms, subclinical by experts, medical doctors. Recently, to contribute to the diagnostic process, artificial intelligence has been applied to speed up the analysis and processing process. These methods mostly use probability theory with the central role being Bayes' theorem. In this study, we also predicted cardiovascular diseases with data science approach, but followed another way – evidence-based integration using Dempster Shafer theory. In particular, each symptom is considered a evidence about the disease with some degree of uncertainty. Dempster combine is used to synthesize the evidence. The degree of uncertainty of each piece of evidence will be optimized by the gradient descent optimization algorithm. Preliminary results show that this new method not only has a significant improvement in predictability when compared with Bayesian but also shows the certainty of each symptom in the diagnostic process. These results allow expectations for the clinical support of the method as well as the potential application of data science to the field of medicine.

Keywords: Dempster Shafer Theory, Machine learning, Cardiovascular diseases

I. ĐẶT VẤN ĐỀ

Nhóm bệnh lý tim mạch là nguyên nhân gây tử vong hàng đầu trên thế giới. Theo thống kê của WHO năm 2018 số lượng ca tử vong thuộc về nhóm tim mạch là 2380 trường hợp mỗi ngày [1], chiếm 25% tổng số ca tử vong tại Mỹ (không kể những ca tử vong do tai nạn giao thông, dịch bệnh) [2]. Tại Việt Nam, theo thống kê của WHO, số lượng ca tử vong do tim mạch chiếm 31%, đứng đầu nhuyên nhân gây tử vong. Với tình trạng đó, nhu cầu phát hiện và chẩn đoán sớm bệnh cũng như giai đoạn bệnh của các bệnh lý tim mạch trở nên cấp thiết.

Gần đây, khoa học dữ liệu phát triển mạnh, cùng với sự hỗ trợ tích cực từ khả năng tính toán của máy tính, lĩnh vực trí tuệ nhân tạo ra đời thúc đẩy quá trình khai phá tri thức trong hầu hết các lĩnh vực của cuộc sống. Tiêu biểu như trong lĩnh vực xử lý ảnh, dịch máy, hay các trò chơi trí tuệ, trí thông minh nhân tạo đã ban đầu cho thấy những năng lực vượt trội hơn cả khả năng của con người. Tại một thử nghiệm, hệ thống AI của Google-DeepMind AlphaGo đã đánh bại nhà vô địch cờ vây thế giới Lee Sedol với tỉ số 4-1 [3].

Cũng như thế ở lĩnh vực y học, sự kết hợp tri thức liên ngành tạo ra hướng mới như "health-informatics" hay "medicine informatics" kỳ vọng xử lý được các bài toán khó trong y học và gây ra sự bùng nổ tri thức khi phân tích các nguồn dữ liệu khổng lồ. Các nghiên cứu ứng dụng cụ thể cũng đã được triển khai, như máy mornitoring kết hợp với theo dõi bất thường tự động [4], các mô hình chẩn đoán bệnh [5]. Hầu hết, các phương pháp này dựa trên lý thuyết xác suất thống kê kinh điển Bayes. Mặc dù, lý thuyết Bayes vẫn đóng vai trò trung tâm trong phương pháp luận của khoa học dữ liệu, nhưng một hướng khác – lý thuyết Dempster Shafer theory về kết hợp bằng chứng dường như tương thích hơn với dữ liệu y học khi mô phỏng hành vi phân tích của con người và có xem xét đến mức độ không chắc chắn (uncertainty) của dữ liệu.

Trong nghiên cứu ngày, chúng tôi xây dựng mô hình chẩn đoán bệnh lý tim mạch bằng lý thuyết Dempster Shafer. Mỗi triệu chứng được xem là một bằng chứng để kết luận về các bệnh khả dĩ với các trọng số đóng vai trò như xác suất. Phép kết hợp Dempster được sử dụng để kết hợp các bằng chứng này lại, từ đó đưa ra quyết định dự đoán cuối cùng. Phương pháp chi tiết được trình bày cụ thể ở phần 2, kết quả thử nghiệm được trình bày ở phần 3, các phân tích về ưu-nhược điểm được nêu ở phần bàn luận và kết luận.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu. Bộ dữ liệu được thu thập từ 4 nguồn: Cleveland, Hungary, Switzerland và the VA Long Beach, được đặt tên là Heart Disease Data Set, và công bố tại UC Irvine Machine Learning Repository vào ngày 01/07/1988 [6]. Tập dữ liệu thứ nhất cung cấp bởi Cleveland với 164 người nguy cơ thấp, 139 người nguy cơ cao. Hungary cung cấp tập dữ liệu thứ hai có 188 người nguy cơ thấp và 106 người nguy cơ cao. Tập dữ liệu thứ ba chứa 8 đối tượng nguy cơ thấp và 115 đối tượng nguy cơ cao được cung cấp bởi Switzerland. Cuối cùng, Long Beach VA cung cấp tập dữ liệu thứ 4 với 51 người nguy cơ thấp và 149 người nguy cơ cao. Như vậy, tập dữ liệu tổng mà chúng tôi sử dụng có 920 người tham gia, trong đó có 411 đối tượng nguy cơ thấp và 509 người nguy cơ cao.

Mỗi cơ sở dữ liệu bao gồm 76 thuộc tính nhưng trong nghiên cứu này, chúng tôi cũng chỉ sử dụng 9 thuộc tính có thể phân loại được. Ngoại trừ thuộc tính "num" đại diện cho mức độ nguy cơ tim mạch, chúng tôi xử lý 8 thuộc tính còn lại thành 23 triệu chứng cụ thể như sau:

Bảng 1: Mô tả 23 triệu chứng phân loại

STT	Viết tắt	Thuộc tính	Thông tin chi tiết
1 2	sex = 1 sex = 0	Giới tính	1 = nam 0 = nữ
3 4 5 6	cp = 1 cp = 2 cp = 3 cp = 4	Loại đau thắt ngực	1 = đau thắt ngực điển hình 2 = đau thắt ngực không điển hình 3 = không đau thắt ngực 4 = không có triệu chứng
7 8	fbs = 1 fbs = 0	Đường huyết lúc đói > 120 mg/dl	1 = đúng 0 = sai
9 10 11	restecg = 0 restecg = 1 restecg = 2	Kết quả điện tâm đồ lúc nghỉ	0 = bình thường 1 = có bất thường sóng ST-T 2 = phì đại thất trái xác định theo tiêu chuẩn của Estes
12 13	exang = 0 exang = 1	đau thắt ngực do tập thể dục	0 = không 1 = có
14 15 16	slope = 1 slope = 2 slope = 3	Độ dốc của đoạn ST khi tập thể dục	1 = chênh lên 2 = bằng 3 = chênh xuống
17 18 19 20	ca = 0 ca = 1 ca = 2 ca = 3	Số mạch chính được tô màu bằng phương pháp soi huỳnh quang	0/1/2/3: Số lượng mạch chính được tô màu
21 22 23	thal = 3 thal = 6 thal = 7	Bệnh thalassemia	3 = Bình thường 6 = Thalassemia không hồi phục 7 = Thalassemia có hồi phục
	num	Mức độ hẹp lòng mạch, tương ứng với nguy cơ tim mạch	0 = hẹp < 50% → Nguy cơ tim mạch thấp 1 = hẹp > 50% → Nguy cơ tim mạch cao

2.2. Phương pháp nghiên cứu:

2.2.1. Dempster Shafer Theory. Lý thuyết Dempster Shafer (DST) là lý thuyết về độ tin cậy, là sự tổng quát hóa lý thuyết Bayes khi có tính đến mức độ không chắc chắn của dữ liệu [7]. Ví dụ, từ một triệu chứng X, có thể hướng tới một số khả năng về bệnh lý thuộc tập giả thuyết Y = {Y1, Y2,…, Yn}. Lý thuyết DST cho phép gán trọng số về mức độ liên quan của X tới tập bệnh Y. Cụ thể, một khả năng bất kỳ của Y (thể hiện là một tập con của Y) được gán một giá trị trọng số p_i thuộc khoảng [0, 1] thể hiện khả năng mắc bệnh đó khi có triệu chứng X. Và tổng trọng số của tất cả các khả năng khả dĩ bằng 1:

$$\sum_i^{2^n} p_i = 1$$

Việc gán trọng số là vấn đề quan trọng bậc nhất trong lý thuyết DST. Thông thường, nó được xử lý bằng lý thuyết xác suất hoặc bằng ý kiến của chuyên gia. Một nghiên cứu gần đây phát triển phương pháp gán bằng các thuật toán tối ưu (Gradient Descent) [7].

Sau khi gán trọng số cho từng khả năng, DST cho phép kết hợp các khả năng riêng lẻ tạo thành khả năng chung. Đồng thời, DST cũng thực hiện tính toán trọng số của khả năng chung đó, bỏ qua sự xung đột giữa các khả năng. Trọng số của khả năng chung A được kết hợp dựa trên khả năng B và C theo công thức:

$$p_A = \frac{1}{1-k} \sum_{B \cap C = A \neq \emptyset} p_B \times p_C$$

Trong đó, k là đại lượng đại diện cho sự xung đột giữa khả năng B và C, k được tính bằng:

$$k = \sum_{B \cap C = \emptyset} p_B \times p_C$$

2.2.2. Gradient descent. Trong Machine Learning nói chung hay các thuật toán tối ưu nói riêng, Gradient descent là phương pháp được sử dụng nhiều nhất để tối ưu hóa mô hình [8]. Gradient descent xử lý tìm giá trị cực đại hay cực tiểu của một hàm số nhằm tối ưu hóa thuật toán bằng cách khởi tạo một giá trị ban đầu cho các biến của hàm số. Sau đó, dùng một phép toán lặp để tiến dần đến điểm cần tìm, tức đến khi đạo hàm gần với 0. Sau khi xử lý, Gradient Descent trả về bộ trọng số tối ưu của các khả năng.

2.2.3. Phương pháp Bayes classification (logistics regression). Bayes là một công thức toán học đơn giản được sử dụng để tính toán các xác suất có điều kiện, là phương pháp được sử dụng phổ biến và có tính ứng dụng cao. Trên thực tế, mỗi triệu chứng X cũng có thể gặp trong các khả năng Y = {Y1, Y2, …, Yn}. Lý thuyết Bayes tính toán xác suất mà triệu chứng X gặp

trong tất cả các khả năr$\frac{\square(X/Y).P(Y)}{\square(X)}$

P(Y/X) =

Mỗi triệu chứng X gặp trong 2^n khả năng bệnh Y. Các xác suất này có tổng bằng 1:

$$\sum_1^{2^n} P(Y/X) = 1$$

Sau đó, Logistic Regression là mô hình hồi quy được sử dụng để xây dựng mô hình phân biệt 2 nhóm nguy cơ tim mạch. Đây là một trong những thuật toán phân loại thuộc học máy có giám sát [10] được áp dụng phổ biến trong Machine Learning.

2.2.4. K-fold cross-validation. Cross validation là một phương pháp thống kê được sử dụng để ước lượng hiệu quả của các mô hình học máy. Trong nghiên cứu này, chúng tôi sử dụng kỹ thuật kiểm tra chéo với 10 phân nhóm (ten-folds cross-validation). Cụ thể, dữ liệu được chia ngẫu nhiên thành 10 nhóm. Mỗi lần chạy, dùng 9 nhóm để dựng mô hình và nhóm còn lại để kiểm tra. Độ chính xác của phép dự đoán là trung bình cho cả 10 lần chạy.

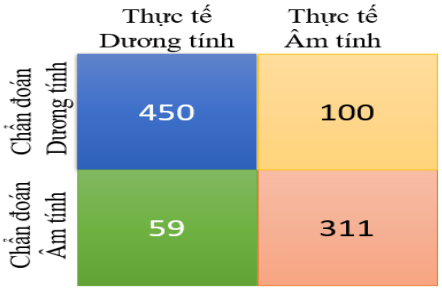
III. KẾT QUẢ NGHIÊN CỨU

3.1. Lý thuyết Dempster Shafer kết hợp với đánh trọng số bằng Bayes. Ở đây, mỗi triệu chứng sẽ đưa ra các đánh giá của nó về khả năng không bị bệnh (nhóm 0), khả năng bị bệnh (nhóm 1), hoặc chưa rõ về cả 2 (nhóm {0, 1}). Trọng số của nhóm {0,1} được khởi tạo ngẫu nhiên là p thuộc [0, 1]. Trọng số của 2 nhóm còn lại tương ứng là xác suất không bị bệnh và bị bệnh khi đã có triệu chứng và nhân với hệ số (1-p). Như vậy, với 1 triệu chứng, mô hình sẽ kết luận vào 3 khả năng: [{0}, {1}, {0,1}] với các hệ số tương ứng: [(1–p)P(0), (1–p)P(1), p]. Với 23 triệu chứng riêng rẽ, mô hình có 23 tham số p tương ứng. Bộ tham số p tối ưu được tìm bằng thuật toán tối ưu sườn dốc. Kết quả chạy cross validation với 10 nhóm cho độ chính xác 83%, độ nhạy 88%, và độ đặc hiệu 76%. Kết quả này có sự cải thiện đáng kể khi so sánh với nghiên cứu tương đương của Ram Kumar (2020) sử dụng RFC (Random Forest Classifier) dự đoán 2 mức độ nguy cơ với độ chính xác 80%. Kết quả chi tiết được trình bày ở Hình 1.

Bộ tham số p thể hiện mức độ không chắc chắn của từng triệu chứng được chỉ ra ở Bảng 2. Ở đây, dễ nhận thấy các thuộc tính bình thường bao gồm: không có triệu chứng đau thắt ngực (p = 0.23), đường huyết lúc đói < 120mg/dl (p = 0.58), kết quả điện tâm đồ lúc nghỉ bình thường (p = 0.37) có mức độ không chắc chắn cao, tham số p lớn. Điều này phù hợp với thực tế khi các triệu chứng bình thường không có ý nghĩa cao trong chẩn đoán bệnh.

Bảng 2: Trọng số trong phương pháp DST + GD tối ưu [p]

STT	Trọng số	STT	Trọng số	STT	Trọng số
1	[0.31, 0.53, 0.16]	9	[0.31, 0.32, 0.37]	17	[0.66, 0.24, 0.1]
2	[0.57, 0.2, 0.23]	10	[0.16, 0.32, 0.52]	18	[0.2, 0.44, 0.36]
3	[0.25, 0.19, 0.56]	11	[0.21, 0.28, 0.51]	19	[0.11, 0.46, 0.43]
4	[0.68, 0.11, 0.21]	12	[0.51, 0.29, 0.19]	20	[0.07, 0.4, 0.53]
5	[0.49, 0.28, 0.23]	13	[0.09, 0.44, 0.48]	21	[0.63, 0.27, 0.1]
6	[0.16, 0.61, 0.23]	14	[0.43, 0.27, 0.3]	22	[0.12, 0.5, 0.38]
7	[0.22, 0.46, 0.32]	15	[0.21, 0.69, 0.1]	23	[0.11, 0.36, 0.53]
8	[0.21, 0.2, 0.58]	16	[0.12, 0.41, 0.48]		



Hình 1: Confusion matrix mô tả kết quả của DST + GD [p]

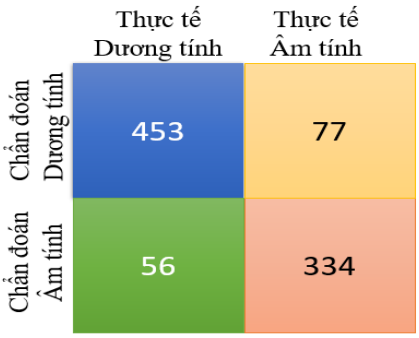
3.2. Lý thuyết Dempster Shafer kết hợp với Gradient Descent. Trong mô hình này, mỗi triệu chứng cũng đưa ra 3 khả năng bệnh [Nhóm 0, nhóm 1, Nhóm {0, 1}]. Trọng số của nhóm 0 và nhóm 1 lần lượt là xác suất không mắc bệnh và có mắc bệnh. Gradient Descent thực hiện tối ưu hóa 2 trọng số này và đưa ra hệ số lần lượt là p_0 và p_1. p_0 và p_1 thuộc [0, 1]. Hệ số của nhóm {0, 1} đại diện cho sự không chắc chắn được đặt là $p = (1 - p_0 - p_1)$. Như vậy, 3 khả năng [{0}, {1}, {0, 1}] sẽ được đại diện bởi 3 hệ số [p_0, p_1, $1 - p_0 - p_1$]. 23 triệu chứng sẽ đưa ra 23 bộ trọng số tương ứng. Cross Validation cho 10 nhóm đưa ra kết quả độ chính xác 86%, độ nhạy 89% và độ đặc hiệu 81%. Kết quả này chỉ ra sự cải tiến đáng kể khi so sánh với các nghiên cứu trong cùng tập dữ liệu, tiêu biểu như nghiên cứu của Bemando sử dụng Naive Bayes and Random

Forest Algorithms đưa ra độ chính xác 85% [13]. Kết quả chi tiết được thể hiện ở Hình 2. Bộ trọng số tương ứng được mô tả trong Bảng 2.

Theo kết quả mô tả, các thuộc tính bình thường đều có độ không chắc chắn cao. Cụ thể, các triệu chứng: Không có triệu chứng đau thắt ngực (p = 0.69), đường huyết lúc đói < 120mg/dl (p = 0.96), kết quả điện tâm đồ lúc nghỉ bình thường (p = 0.98), không đau thắt ngực do tập thể dục (p = 0.98), không có mạch chính được tô màu (p = 0.57), không có Thalassemia (p = 0.56) đều đưa ra khả năng {0, 1} với tham số p > 0.55. Do độ không chắc chắn cao nên các thuộc tính này ít có ảnh hưởng đến chẩn đoán. Kết quả này cũng phù hợp với thực tế sinh bệnh của cơ thể.

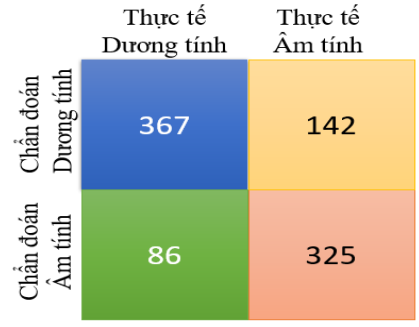
Bảng 3: Trọng số trong phương pháp DST + GD tối ưu [p_0, p_1]

STT	Trọng số	STT	Trọng số	STT	Trọng số
1	[0.01, 0.01, 0.98]	9	[0.01, 0.01, 0.98]	17	[0.33, 0.1, 0.57]
2	[0.14, 0.01, 0.85]	10	[0.13, 0.14, 0.73]	18	[0.23, 0.26, 0.51]
3	[0.28, 0.25, 0.47]	11	[0.23, 0.22, 0.55]	19	[0.35, 0.38, 0.27]
4	[0.2, 0.01, 0.79]	12	[0.01, 0.01, 0.98]	20	[0.34, 0.34, 0.32]
5	[0.12, 0.08, 0.8]	13	[0.22, 0.33, 0.45]	21	[0.25, 0.19, 0.56]
6	[0.11, 0.2, 0.69]	14	[0.21, 0.2, 0.59]	22	[0.34, 0.37, 0.29]
7	[0.14, 0.18, 0.68]	15	[0.12, 0.24, 0.64]	23	[0.1, 0.2, 0.7]
8	[0.03, 0.01, 0.96]	16	[0.26, 0.32, 0.42]		



Hình 2: Confusion matrix mô tả kết quả của DST + GD [p_0, p_1]

3.3. Lý thuyết xác suất Bayes kết hợp Logistic Regression.



Hình 3: Confusion matrix mô tả kết quả của Bayes

Từ một triệu chứng, xác suất thống kê Bayes tính toán khả năng mắc bệnh, tương ứng với nhóm {1} là p. Khả năng không mắc bệnh khi có triệu chứng đó là (1 – p). Bayes bỏ qua phần không chắc chắn của dữ liệu, tức là khả năng nhóm {0, 1}. Như vậy, 23 triệu chứng riêng lẻ sẽ đưa ra 23 bộ xác suất cho đại diện cho khả năng mắc bệnh của 2 nhóm [{1}, {0}] là [p, (1-p)]. Logistic Regression tính toán và đưa ra kết quả phân loại với độ chính xác 75%, độ nhạy 81% và độ đặc hiệu 70%. Kết quả chi tiết được trình bày ở Hình 3.

Như vậy, sau khi sử dụng cả 3 phương pháp nghiên cứu, kết quả cuối cùng được tổng hợp như sau:

Bảng 4: Tổng quan kết quả ở 3 phương pháp

Phương pháp	DST+GD [p]	DST+GD [p_0,p_1]	Bayes
Độ chính xác	83%	86%	75%
Độ nhạy	88%	89%	81%
Độ đặc hiệu	76%	81%	70%
Precision	82%	85%	72%
Độ lệch chuẩn	p < 0.05	p < 0.05	p=0.05

IV. BÀN LUẬN

Trong nghiên cứu này, chúng tôi sử dụng 3 phương pháp để xây dựng mô hình dự đoán nguy cơ tim mạch ở 920 bệnh nhân. Kết quả của cả 3 phương pháp được mô tả ở Bảng 4. Kết quả này hoàn toàn phù hợp với thực tế khi tập dữ liệu các triệu chứng lâm sàng có độ không đảm bảo cao. Bayes bỏ qua tất cả sự không chắc chắn của dữ liệu, kết quả thu được độ chính xác thấp nhất (75%), với độ lệch chuẩn cao (p = 0.05). DST có tính toán đến phần dữ liệu liệu không chắc chắn và gán trọng số cho nó, xây dựng mô hình với độ chính xác cao (81%), độ lệch chuẩn p < 0.05. Khi kết hợp thuật toán tối ưu Gradient Descent với DST, độ chính xác được tối ưu hóa rõ ràng (86%) với p < 0.05. Mô hình này hoàn toàn có thể ứng dụng trên lâm sàng do độ chính xác cao, tính hợp lý và tính khoa học của nó.

So với các nghiên cứu trên cùng tập dữ liệu, mô hình DST kết hợp với Gradient Descent đạt độ chính xác cao hơn. Ram Kumar và các cộng sự (2020) đã chỉ ra được rằng RFC (Random Forest Classifier) dự đoán 2 mức độ nguy cơ với độ chính xác 80.327%. Một số mô hình khác như Naive Bayes and Random Forest Algorithms của Bemando hoặc K neighbors cũng đạt độ chính xác tương đương (85%). Nghiên cứu của chúng tôi, sử dụng DST kết hợp Gradient Descent đưa ra mô hình có độ chính xác cao, đồng thời phù hợp với tính chất không đảm bảo của bộ dữ liệu, đưa ra được lời giải thích hợp lý trên lâm sàng. Trong tương lai, khi khai thác các hướng đi này sâu hơn, DST và Gradient Descent sẽ đem lại kết quả ứng dụng cao hơn nữa.

Trong mô hình của DST, mỗi phần không chắc chắn của chẩn đoán đều được đánh giá mức độ tin cậy, thể hiện qua tham số p của nhóm {0, 1}. Nó đại diện cho khả năng chẩn đoán không chắc chắn, bệnh nhân có thể thuộc mức nguy cơ cao, cũng có thể là nguy cơ thấp. p càng thấp thì sự ảnh hưởng của triệu chứng đến việc chẩn đoán càng cao và ngược lại. Sau khi sử dụng Gradient Descent để tối ưu hóa trọng số, đưa ra giá trị p cho từng triệu chứng, ta có thể đánh giá được sự đóng góp của triệu chứng trong việc chẩn đoán là cao hay thấp. Đây là một ưu điểm lớn của DST vì đã xem xét và gắn trọng số cho từng triệu chứng, làm tăng độ tin cậy của chẩn đoán và phù hợp hơn trong thực tế.

Tuy nhiên, DST khi kết hợp với Gradient Descent cũng có một số hạn chế nhất định. Thuật toán hồi quy của quá trình xử lý khá cồng kềnh phức tạp. Nghiệm của quá trình cũng không đồng nhất giữa các lần chạy mô hình. Hơn nữa, như đã biết, Gradient Descent chọn một điểm ở gần local minimum (cực tiểu địa phương), sử dụng các phép toán lặp để tiến tới điểm tối ưu hóa mô hình. Vì vậy, hạn chế tiếp theo của mô hình chính là việc dễ chọn điểm rơi vào đúng local minimum, khiến phép toán lặp trở nên vô hạn. Việc xử lý thuật toán hồi quy ở Gradient Descent còn tồn tại một số khó khăn như trên cần khắc phục.

Bên cạnh những khó khăn trên, mô hình DST kết hợp với Gradient Descent cũng có nhiều mặt tích cực. Đây là một phương pháp mới, có chứa nhiều tiềm năng lớn. Cách thức hoạt động và giải thích của nó linh động và phù hợp với cách nghĩ của các chuyên gia hơn phương pháp xác suất thống kê Bayes. Ngoại trừ trả về độ chính xác khi dự đoán nguy cơ tiến triển bệnh, phương pháp này còn đánh giá độ tin cậy của từng thuộc tính. Đặc điểm này rất phù hợp với những bộ dữ liệu không đồng nhất như dữ liệu triệu chứng lâm sàng. Như vậy, tính khoa học và tính hợp lý của DST rất phù hợp để áp dụng và nghiên cứu sâu thêm nữa cho các vấn đề lâm sàng nói chung và dự đoán nguy cơ tim mạch nói riêng.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã xây dựng mô hình dự đoán bệnh lý tim mạch sử dụng lý thuyết Dempster Shafer. Kết quả chạy thử nghiệm trên bộ dữ liệu UCI với 920 bệnh nhân phân thành 2 nhóm hẹp mạch vành dưới 50% và trên 50%, với 8 thuộc tính thăm khám lâm sàng, cận lâm sàng kết quả dự đoán đạt 83% - 86%. Kết quả này có sự cải thiện đáng kể khi so sánh với các mô hình hồi quy Bayes cơ bản (logistics regression, decision tree,…). Bên cạnh đó, kết quả còn chỉ ra mức độ không chắc chắn của từng thuộc tính và cả mức độ không chắc chắn trong kết quả chẩn đoán của từng người. Những thông tin này cho phép phân tích đầy đủ hơn về khả năng mắc bệnh của người bệnh, từ đó xây dựng các phác đồ điều trị phù hợp. Kết quả nghiên cứu cũng cho phép sự kỳ vọng khả năng hỗ trợ lâm sàng của phương pháp này nói riêng và của lĩnh vực liên ngành medicine informatics.

TÀI LIỆU THAM KHẢO

1. **Virani, Salim S., et al.** "Heart disease and stroke statistics—2021 update: a report from the American Heart Association." Circulation 143.8 (2021): e254-e743.
2. **Centers for Disease Control and Prevention.** "Heart Disease Facts" (2022).
3. **Chouard, T.** (2016). The Go Files: AI computer wraps up 4-1 victory against human champion. Nature News.
4. **Sorkin, R. D., & Woods, D. D.** (1985). Systems with human monitors: A signal detection analysis. Human-computer interaction, 1(1), 49-75.
5. **Fatima, M., & Pasha, M.** (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.
6. **Jackins, V., Vimal, S., Kaliappan, M. et al.** AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J Supercomput 77, 5198–5219 (2021)
7. **Peñafiel, Sergio, et al.** "Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier." Expert Systems with Applications 148 (2020): 113262.
8. **Ruder, Sebastian.** "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).