



(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN
GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ**

(11)



2-0003243

(51) **C12Q 1/68; G01N 15/14; C07H 21/02**
2020.01

(13) **Y**

(21) 2-2022-00129

(22) 31/03/2022

(45) 25/07/2023 424

(43) 25/05/2023 422

(73) Trường Đại học Y Hà Nội (VN)

Số 1, Tôn Thất Tùng, Đống Đa, Hà Nội

(72) Nguyễn Thị Trang (VN); Trần Danh Cường (VN); Nguyễn Chí Thành (VN); Đặng Anh Linh (VN); Ngô Toàn Anh (VN); Trần Trung Kiên (VN); Lê Thị Minh Phương (VN); Hoàng Thị Ngọc Lan (VN); Trương Thị Thu Hằng (VN); Phạm Quang Anh (VN); Lê Minh Giang (VN).

(54) **PHƯƠNG PHÁP ĐÁNH GIÁ CHỈ SỐ PHÂN TÍCH MÁU ĐỂ HỖ TRỢ CẢNH BÁO SỚM
NGUY CƠ MẮC BỆNH TAN MÁU BẨM SINH (THALASSEMIA)**

(57) Sáng chế đề cập đến phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh. Trong đó mẫu máu được phân tích bốn chỉ số HGB (lượng huyết sắc tố trong một thể tích máu), MCV (thể tích trung bình của hồng cầu), MCH (lượng huyết sắc tố trung bình có trong một hồng cầu) và RDW (độ phân bố hồng cầu) và đánh giá theo cơ sở dữ liệu được thiết lập và cân bằng trong mô hình mạng nơron truyền thẳng nhiều lớp. Phương pháp đánh giá cho phép cơ sở dữ liệu tự hiệu chỉnh theo mẫu thời gian thực và thích hợp để phân tích để đưa ra cảnh báo sớm nguy cơ mắc tan máu bẩm sinh của thai nhi trên cơ sở phân tích máu của thai phụ.

Lĩnh vực kỹ thuật được đề cập

Sáng chế thuộc lĩnh vực y tế và chăm sóc sức khỏe, cụ thể là sáng chế đề xuất phương pháp đánh giá chỉ số phân tích máu hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh (thalassemia) cho thai nhi dựa trên các chỉ số phân tích máu của thai phụ.

Tình trạng kỹ thuật của sáng chế

Bệnh tan máu bẩm sinh (thalassemia) là một dạng rối loạn tổng hợp huyết sắc tố và là rối loạn di truyền phổ biến nhất trên thế giới với ước tính khoảng 7% dân số thế giới mang gen bệnh. Nguyên nhân dẫn đến bệnh tan máu bẩm sinh là đột biến gen quy định tổng hợp chuỗi globin dẫn đến mất cân bằng các loại chuỗi globin, tạo nên bất thường về huyết sắc tố, làm giảm chất lượng hồng cầu, gây ra tình trạng thiếu máu ở bệnh nhân. Vì thiếu máu, cơ thể tăng hấp thu sắt và tăng tích lũy tại tổ chức trong cơ thể do phải truyền máu, từ đó gây ra nhiều biến chứng tại gan, tim, tuyến nội tiết làm giảm chất lượng cuộc sống và tuổi thọ của người bệnh.

Một trong những giải pháp giúp hạn chế số ca mắc bệnh tan máu bẩm sinh là sàng lọc để xác định sớm đối tượng mang gen gây tan máu bẩm sinh và điều trị sớm. Điện di hemoglobin là một trong những phương pháp chẩn đoán tan máu bẩm sinh phổ biến. Người trưởng thành bình thường trong máu có HbA chiếm từ 96,5 đến 98,5%, HbA2 chiếm từ 0,5 đến 3,5% và HbF chiếm khoảng 0,5%. Điện di hemoglobin cũng như các phương pháp như sắc ký lỏng hiệu năng cao (HPLC) giúp xác định các Hb bất thường như HbH, Hb Bart's, HbF trong hồng cầu để phát hiện tan máu bẩm sinh. Bên cạnh đó, xét nghiệm di truyền học phân tử cũng là một trong những phương pháp giúp chỉ ra chính xác vị trí đột biến trên 2 gen HbA và HbB để chẩn đoán chính xác thể bệnh và được coi là xét nghiệm cuối cùng nhằm chẩn đoán bệnh thalassemia. Ngoài ra, nhiều kỹ thuật sinh học phân tử khác nhau nhằm phát hiện đột biến gen globin, đa số là dựa trên phản ứng PCR như multiplex gap-PCR, ARSM-PCR, MLPA, giải trình tự gen, phương pháp lai ngược (reverse hybridization) đang được sử dụng trong chẩn đoán tan máu bẩm sinh.

Việc xét nghiệm sàng lọc tan máu bẩm sinh (thalassemia) được khuyến cáo cho thai phụ trước sinh trong quý đầu của thai kỳ dựa trên việc xét nghiệm tổng phân tích tế bào máu, kết hợp tiền sử bệnh và tiền sử gia đình, kết quả xét nghiệm máu của chồng để xác định sớm đối tượng mang gen gây tan máu bẩm sinh và có chính sách dự phòng và can thiệp kịp thời. Tuy nhiên, việc sàng lọc trước sinh bệnh thalassemia cho tất cả thai phụ tại Việt Nam còn gặp nhiều khó khăn do hiểu biết về hiệu quả của việc sàng lọc này của thai phụ, gia đình thai phụ và cũng như nhiều nhân viên y tế còn hạn chế. Ngoài ra, chi phí xét nghiệm đột biến gen chẩn đoán cả 2 dạng alpha và beta (thiếu hoặc không có chuỗi α -globin hoặc chuỗi β -globin) có thể tốn đến hàng chục triệu đồng và yêu cầu các cơ sở y tế đầu tư máy móc thiết bị chuyên dụng.

Các xét nghiệm sàng lọc sử dụng tổng phân tích tế bào máu ngoại vi hiện nay chưa cho phép xác định người lành mang gen gây tan máu bẩm sinh (gen bệnh thalassemia) mà chỉ phát hiện sớm những người có nguy cơ cao mắc bệnh. Để xác định chính xác có phải là mang gen bệnh thalassemia hay không vẫn cần thực hiện các xét nghiệm như điện di và di truyền phân tử để khẳng định. Tuy nhiên, tổng phân tích tế bào máu ngoại vi vẫn được yêu cầu thực hiện trước để sàng lọc tan máu bẩm sinh do có chi phí hợp lý. Dấu hiệu gợi ý người lành mang gen bệnh là tình trạng thiếu máu, hồng cầu nhỏ, nhược sắc mức độ nhẹ tương ứng với $Hb < 120g/l$, $MCV < 85fl$ hoặc $MCH < 28 pg$ là các chỉ số được Tổ chức Y tế Thế giới (WHO) và Liên đoàn Thalassemia thế giới (Thalassemia International Federation -TIF) khuyến cáo sử dụng. Hầu hết các quốc gia dùng chỉ số MCV và MCH để sàng lọc. Nếu $MCV < 85fl$ hoặc $MCH < 28pg$, bước tiếp theo là điện di hemoglobin xác định loại hemoglobin trong hồng cầu. Tất cả bệnh nhân β -thalassemia từ thể nhẹ đến nặng đều có MCV, MCH giảm nên đây là 2 chỉ số rất hữu hiệu để sàng lọc. Tuy nhiên đối với trường hợp α -thalassemia thể ẩn (chỉ mất 1 gen α) có MCV và MCH bình thường thì có thể bị bỏ sót. Với người mắc tan máu bẩm sinh, lượng sắt trong máu, độ bão hòa sắt và ferritin huyết thanh đều tăng do tình trạng ứ sắt trong cơ thể. Các chỉ số này được sử dụng để phân biệt với tình trạng thiếu máu thiếu sắt cũng có MCV và MCH thấp nhưng lại có Fe và Ferritin đều giảm. Bên cạnh đó, tan máu bẩm sinh phụ thuộc rất nhiều vào sự khác biệt về địa lý, sắc tộc nên một số chỉ số ngưỡng để kết luận nguy cơ mắc tan máu bẩm sinh thường được sử dụng không hoàn toàn phù hợp với người Việt Nam. Như vậy để đánh giá các kết quả xét nghiệm máu và kết hợp tiền sử gia đình nhằm kết luận

về nguy cơ mang gen gây tan máu bẩm sinh, cần phải có chuyên gia di truyền, huyết học hoặc chuyên gia sản khoa có kinh nghiệm.

Chính vì vậy, tồn tại nhu cầu đối với phương pháp hỗ trợ sàng lọc và xác định thai phụ mang gen gây tan máu bẩm sinh nhanh, có độ chính xác cao thông qua khai thác tiền sử gia đình, tiền sử bệnh và xét nghiệm tổng phân tích tế bào máu trước khi các nhân viên y tế chỉ định thai phụ thực hiện các xét nghiệm sinh học phân tử khác.

Bản chất kỹ thuật của sáng chế

Sáng chế nhằm khắc phục vấn đề nêu trên, cụ thể là sáng chế đề xuất phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh (thalassemia), trợ giúp cho phương pháp sàng lọc thai nhi đối với thai phụ mang gen tan máu bẩm sinh, từ đó giúp sàng lọc thai nhi có nguy cơ mắc bệnh.

Phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh theo sáng chế bao gồm các bước:

a) thu thập thông tin phân tích máu toàn phần bằng cách tiến hành phân tích bốn chỉ số của mẫu máu bao gồm HGB (lượng huyết sắc tố trong một thể tích máu), MCV (thể tích trung bình của hồng cầu), MCH (lượng huyết sắc tố trung bình có trong một hồng cầu) và RDW (độ phân bố hồng cầu) để thu nhận kết quả phân tích;

b) xử lý dữ liệu bằng cách nạp bốn chỉ số phân tích máu nêu trên vào hệ thống phân tích với cơ sở dữ liệu về nguy cơ mắc bệnh tan máu bẩm sinh được xây dựng dựa trên mô hình mạng nơron truyền thẳng nhiều lớp (Multi-layer perceptron- MLP) với đầu vào là các chỉ số HGB, MCV, MCH, RDW và ngưỡng (cut- off) so sánh các giá trị HGB, MCV, MCH, RDW này được thiết lập trong cơ sở dữ liệu lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%, trong đó hệ thống phân tích với cơ sở dữ liệu về bệnh tan máu bẩm sinh được xây dựng trên cơ sở bao gồm:

- thiết lập cơ sở dữ liệu bao gồm xây dựng thông tin dữ liệu phân tích máu của 10.000 thai phụ, trong đó thông tin dữ liệu phân tích máu của mỗi thai phụ được liệt kê tương ứng như một bộ giá trị thuộc tính bao gồm ít nhất bốn chỉ số phân tích máu HGB, MCV, MCH và RDW;

- xử lý dữ liệu bằng cách trích xuất dữ liệu gồm bốn chỉ số công thức máu trong cơ sở dữ liệu được xây dựng, cân bằng và chuẩn hóa dữ liệu bằng thuật toán chống mất cân bằng dữ liệu SMOTE (Synthetic Minority Over-sampling), sau đó phân chia

ngẫu nhiên cơ sở dữ liệu thành dữ liệu huấn luyện (train data) và dữ liệu thử nghiệm (test data) theo tỷ lệ 9:1;

- huấn luyện dữ liệu bằng cách so sánh dữ liệu từ dữ liệu huấn luyện và dữ liệu thử nghiệm theo mô hình mạng đa lớp truyền thẳng và kiểm tra theo dữ liệu giám sát được thiết lập lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%; và

- tự hiệu chỉnh giá trị vượt ngưỡng theo giá trị giám sát đối với dữ liệu huấn luyện và dữ liệu nạp để tăng hiệu quả và độ chính xác hệ thống phân tích;

c) đưa ra kết quả đánh giá chỉ số phân tích dựa trên tổ hợp bốn chỉ số phân tích máu được nạp và cảnh báo nguy cơ mẫu mắc tan máu bẩm sinh nếu các giá trị này vượt ngưỡng hiệu chỉnh thiết lập.

Mô tả vắn tắt các hình vẽ

Hình 1 là các sơ đồ cột thể hiện độ quan trọng của 9 chỉ số công thức máu khi a) sử dụng thuật toán MID, và b) sử dụng thuật toán MDA.

Hình 2 là hình vẽ thể hiện mô hình mạng MLP tổng quát.

Hình 3 là sơ đồ thể hiện các bước xây dựng mô hình cảnh cáo tan máu bẩm sinh theo sáng chế.

Mô tả chi tiết sáng chế

Cơ sở phân tử và phân loại bệnh tan máu bẩm sinh dựa trên hemoglobin và gen globin

Hemoglobin là một loại protein hồng cầu có chức năng vận chuyển oxy từ phổi đến các mô trong cơ thể. Về mặt cấu trúc, hemoglobin là một tetrame được cấu tạo từ 4 monome. Mỗi monome gồm 2 thành phần là nhân hemoglobin và globin. Hemoglobin bao gồm 1 vòng porphyrin gắn với Fe^{2+} ở chính giữa, có sắc tố đỏ, giống nhau ở các loài động vật. Globin có nhiều loại và mỗi phân tử hemoglobin sẽ gồm 4 chuỗi globin thuộc 2 loại nhất định, từ đó hình thành nên các loại hemoglobin khác nhau. Ở người trưởng thành có 4 loại globin là α , β , δ và γ , tương ứng tạo ra HbA (2 chuỗi α , 2 chuỗi β), HbA2 (2 chuỗi α , 2 chuỗi δ) và HbF (2 chuỗi α , 2 chuỗi γ).

Gen quy định tổng hợp globin alpha nằm trên nhánh ngắn nhiễm sắc thể số 16, gồm 4 gen, mỗi nhiễm sắc thể chứa 2 gen alpha 1 và gen alpha 2 (HBA1 và HBA2). Gen quy định tổng hợp globin β nằm trên nhánh ngắn nhiễm sắc thể số 11, cùng với các gen quy định chuỗi ϵ , γ và δ , và mỗi nhiễm sắc thể chỉ chứa 1 gen β (HBB). Ở người bình thường tất cả các gen globin đều cùng hoạt động nhằm tổng hợp lượng Hb

đầy đủ để thực hiện chức năng vận chuyển oxy trong cơ thể. Gen globin có thể bị đột biến làm giảm hoặc mất khả năng tổng hợp chuỗi globin, từ đó làm giảm số lượng chuỗi globin có chức năng bình thường gây ra bệnh thalassemia. Thiếu hoặc không có chuỗi α -globin gọi là α -thalassemia, thiếu hoặc không có chuỗi β -globin gọi là β -thalassemia. Các đột biến gây ra bệnh alpha-thalassemia chủ yếu là đột biến mất gen alpha-globin, bên cạnh đó còn có các đột biến điểm làm thay đổi cấu trúc hoặc chức năng của chuỗi globin alpha, dẫn đến hình thành các chuỗi polypeptit bất thường. Số lượng gen bị đột biến càng nhiều thì số lượng chuỗi alpha bình thường càng giảm và biểu hiện lâm sàng của bệnh càng nặng.

Theo một phương án cụ thể, sáng chế đề xuất phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh, trong đó phương pháp này bao gồm các bước:

a) thu thập thông tin phân tích máu toàn phần bằng cách tiến hành phân tích bốn chỉ số của mẫu máu bao gồm HGB (lượng huyết sắc tố trong một thể tích máu), MCV (thể tích trung bình của hồng cầu), MCH (lượng huyết sắc tố trung bình có trong một hồng cầu) và RDW (độ phân bố hồng cầu) để thu nhận kết quả phân tích;

b) xử lý dữ liệu bằng cách nạp bốn chỉ số phân tích máu nêu trên vào hệ thống phân tích với cơ sở dữ liệu về nguy cơ mắc bệnh tan máu bẩm sinh được xây dựng dựa trên mô hình mạng nơron truyền thẳng nhiều lớp (Multi-layer perceptron- MLP) với đầu vào là các chỉ số HGB, MCV, MCH, RDW và ngưỡng (cut- off) so sánh các giá trị HGB, MCV, MCH, RDW này được thiết lập trong cơ sở dữ liệu lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%, trong đó hệ thống phân tích với cơ sở dữ liệu về bệnh tan máu bẩm sinh được xây dựng trên cơ sở bao gồm:

- thiết lập cơ sở dữ liệu bao gồm xây dựng thông tin dữ liệu phân tích máu của 10.000 thai phụ, trong đó thông tin dữ liệu phân tích máu của mỗi thai phụ được liệt kê tương ứng như một bộ giá trị thuộc tính bao gồm ít nhất bốn chỉ số phân tích máu HGB, MCV, MCH và RDW;

- xử lý dữ liệu bằng cách trích xuất dữ liệu gồm bốn chỉ số công thức máu trong cơ sở dữ liệu được xây dựng, cân bằng và chuẩn hóa dữ liệu bằng thuật toán chống mất cân bằng dữ liệu SMOTE (Synthetic Minority Over-sampling), sau đó phân chia ngẫu nhiên cơ sở dữ liệu thành dữ liệu huấn luyện (train data) và dữ liệu thử nghiệm (test data) theo tỷ lệ 9:1;

- huấn luyện dữ liệu bằng cách so sánh dữ liệu từ dữ liệu huấn luyện và dữ liệu thử nghiệm theo mô hình mạng đa lớp truyền thẳng và kiểm tra theo dữ liệu giám sát được thiết lập lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%; và

- tự hiệu chỉnh giá trị vượt ngưỡng theo giá trị giám sát đối với dữ liệu huấn luyện và dữ liệu nạp để tăng hiệu quả và độ chính xác của hệ thống phân tích;

c) đưa ra kết quả đánh giá chỉ số phân tích dựa trên tổ hợp bốn chỉ số phân tích máu được nạp và cảnh báo nguy cơ mắc tan máu bẩm sinh nếu các giá trị này vượt ngưỡng hiệu chỉnh thiết lập.

Phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh theo sáng chế cho phép cảnh báo với độ chính xác trên 90%.

Xác định các chỉ số công thức máu đặc trưng cần sử dụng cho sàng lọc tan máu bẩm sinh và ngưỡng (cut-off) của các chỉ số này

Trong kết quả xét nghiệm công thức máu, 9 chỉ số thường được xem xét để các nhân viên y tế chỉ định thai phụ tiếp tục làm các xét nghiệm sinh học phân tử bao gồm MCH, MCV, HGB, RDW, RBC, HCT, ferritin huyết thanh, MCHC và sắt huyết thanh. Tuy nhiên, để tối ưu hơn phương pháp cảnh báo theo sáng chế, các tác giả sáng chế sử dụng thuật toán MID (Mutual information difference) và MDA (Mixture Discriminant Analysis) và nhận thấy rằng 4 chỉ số HGB, MCV, MCH và RDW có vai trò quan trọng nhất cho sàng lọc bệnh tan máu bẩm sinh từ 10.000 trường hợp thai phụ có kết quả công thức máu như được thể hiện trên Hình 1. Ngoài ra, các tác giả sáng chế cũng nhận thấy rằng ngưỡng để phát hiện tan máu bẩm sinh của các chỉ số này ở người Việt Nam như được thể hiện trong Bảng 1 khác so với các giá trị được tổ chức Y tế thế giới khuyến cáo về nguy cơ mắc tan máu bẩm sinh. Phát hiện cũng trùng với báo cáo của tác giả Nguyễn Khắc Hân Hoan tại bệnh viện Từ Dũ năm 2018, dù ngưỡng (cut-off) của MCV theo lý thuyết là 80fl nhưng thực tế, một số thai phụ dù có $MCV > 80fl$ nhưng trẻ sinh ra bởi các thai phụ này vẫn mắc tan máu bẩm sinh.

Bảng 1. Ngưỡng (cut-off) của các chỉ số HGB, MCV, MCH và RDW dựa trên bộ dữ liệu 10.000 sản phụ người Việt Nam và dùng các thuật toán MID và MDA

	HGB	MCV	MCH	RDW
Cut-off lý thuyết	110 g/l	80fl	28 pg	14,0%
Cut-off thực tế	113,5 g/l	83,15fl	27,65 pg	14,25%

Độ nhạy	0,907	0,938	0,948	0,914
Độ đặc hiệu	0,918	0,973	0,977	0,909
Tỷ lệ thay đổi so với ngưỡng lý thuyết (ROC)	0,902	0,977	0,978	0,906

Ngưỡng của các chỉ số HGB, MCV, MCH và RDW như được sử dụng theo sáng chế khác với ngưỡng lý thuyết có thể được lý giải là rối loạn tan máu bẩm sinh chịu ảnh hưởng bởi các yếu tố như chủng tộc, địa lý nên mỗi quốc gia có ngưỡng cut-off khác nhau. Theo đó, các giá trị ngưỡng này được sử dụng thay cho các trị số ngưỡng lý thuyết trong phương pháp cảnh báo theo sáng chế.

Trong số các mô hình học máy bao gồm giải thuật KNN (K-Nearest Neighbour: KNN), SVM (Support Vector Machine: SVM), RF (RandomForest: RF) và mạng đa lớp truyền thẳng (Multilayer Perceptron: MLP) được sử dụng phổ biến với độ tin cậy cao và phù hợp cho trường hợp số lượng đặc trưng ít và số bản ghi trong cơ sở dữ liệu vừa phải, các tác giả sáng chế nhận thấy rằng mô hình MLP là thích hợp nhất để ứng dụng trong phương pháp cảnh báo theo sáng chế.

Mạng nơ-ron truyền thẳng nhiều lớp (Multi-layer Perceptron: MLP)

Thuật ngữ mạng nơ-ron truyền thẳng nhiều lớp (Multi-layer Perceptron: MLP) là một trong những mạng nơ-ron nhân tạo phổ biến nhất được lấy cảm hứng từ các nơ-ron thần kinh sinh học gồm các nơ-ron (phần tử cấu tạo nên mạng) kết nối với nhau thông qua các liên kết (gọi là trọng số liên kết) để giải quyết các vấn đề như nhận dạng mẫu, phân loại dữ liệu, v.v... thông qua quá trình học từ tập các mẫu huấn luyện. Một mạng MLP tổng quát là mạng có nhiều hơn hai lớp trong đó bao gồm một lớp vào, một lớp ra và một hoặc nhiều lớp ẩn (hình 2). Mỗi đơn vị trong các lớp ngoại trừ lớp đầu vào là một nơ-ron sử dụng chức năng kích hoạt phi tuyến. Số lượng nơ-ron của lớp đầu vào và lớp đầu ra dựa trên bộ dữ liệu. Cụ thể là, số lượng nơ-ron ở lớp đầu vào tương ứng với số các đặc trưng của dữ liệu, còn số nơ-ron ở lớp đầu ra thể hiện nhiệm vụ của mô hình. Việc tìm ra số lượng nơ-ron tối ưu cho các lớp ẩn dựa trên phương pháp thử và sai. Tốc độ học tối ưu của MLP được tính toán dựa trên thử nghiệm, và điều kiện dừng được đưa ra dựa trên giá trị sai số bình phương trung bình. Ví dụ, số nơ-ron là 1 tương ứng với bài toán hồi quy tuyến tính, số nơ-ron là 2 thể hiện bài toán phân lớp nhị phân, v.v. Ở lớp đầu ra thường sử dụng hàm Softmax để tính xác suất của một điểm dữ liệu rơi vào các lớp thể nào (với trường hợp có 2 nơ-ron trở lên ở lớp đầu ra). Đối với các nhiệm vụ

cụ thể liên quan đến phân loại, nhận dạng, dự đoán và xấp xỉ mẫu, MLP được sử dụng cho cả các bài toán tuyến tính và phi tuyến tính.

Hoạt động của mạng MLP như sau: Tại lớp đầu vào, các nơron nhận tín hiệu vào xử lý (tính tổng trọng số, gửi tới hàm truyền) rồi cho ra kết quả (là kết quả của hàm truyền); kết quả này sẽ được truyền tới các nơron thuộc lớp ẩn thứ nhất; các nơron tại đây tiếp nhận như là tín hiệu đầu vào, xử lý và gửi kết quả đến lớp ẩn thứ 2, và lớp ẩn thứ n; quá trình tiếp tục cho đến khi các nơron thuộc lớp đầu ra cho kết quả.

Nơron đầu vào là các tham số chỉ báo nguy cơ mắc tan máu bẩm sinh được xác định bởi các ẩn $X_i = \{X_1, X_2, X_3 \dots X_i\}$ trong đó i là số tham số (nơron đầu vào). Hiệu quả của trọng số liên kết, W_i trên mỗi nơron đầu vào ở lớp j được biểu diễn theo công thức:

$$z_j = w_{1j}x_1 + w_{2j}x_2 + \dots \dots w_{3j}x_3 + b \quad (1)$$

Phương trình (1) được gửi đến hàm kích hoạt (hàm sigmoid/logistic) được áp dụng để giới hạn đầu ra với ngưỡng $[-1, +1]$, do đó:

$$V_j = \varphi(z) = \frac{1}{1 + e^{-z_j}} \quad (2)$$

Việc tính toán sự chênh lệch giữa đầu ra dự kiến (p) và đầu ra thực tế (y) được thực hiện bằng cách sử dụng phép toán bình phương sai số (E)

$$E = (p - y)^2 \quad (3)$$

Tuy nhiên độ phủ (recall) mà đầu ra (p) của nơron phụ thuộc vào tổng trọng số của tất cả đầu vào như được đề cập trong phương trình (1); cho thấy rằng sai số (E) cũng phụ thuộc vào trọng số đầu vào của nơron mà cần được thay đổi trọng mạng để cho phép học. Thuật toán lan truyền ngược được sử dụng để tìm tập hợp trọng số mà tối thiểu hóa lỗi. Thuật toán gradient descent (giảm gradient) được áp dụng để tối thiểu hóa lỗi và nhờ đó tìm ra các trọng số tối ưu mà thỏa mãn yêu cầu đề ra. Do lan truyền ngược sử dụng phương pháp gradient descent, cần tính toán đạo hàm của hàm sai số bình phương cho các trọng số của mạng. Do đó, hàm sai số bình phương được định nghĩa lại là ($\frac{1}{2}$ được yêu cầu để triệt tiêu mũ 2 khi vi phân):

$$E = \frac{1}{2}(p - y)^2 \quad (4)$$

Với mỗi nơron j , đầu ra O_j của nó được xác định là:

$$O_j = \varphi(\text{net}_j) = \varphi\left(\sum_{k=1}^n w_{kj}x_k\right) \quad (5)$$

Đầu vào của nơron là tổng trọng số của đầu ra của nơron trước. Số nơron đầu vào là n và biến để chỉ trọng số giữa các nơron i và j . Vì hàm kích hoạt nhìn chung là phi tuyến tính và có thể vi phân được nên đạo hàm của phương trình (2) là:

$$\frac{\partial \varphi}{\partial z} = \varphi(1 - \varphi) \quad (6)$$

Đạo hàm riêng của sai số (E) tương ứng với trọng số w_{ij} được tính bằng cách sử dụng quy tắc chuỗi hai lần như sau:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial O_j} \frac{\partial O_j}{\partial \text{net}_j} \frac{\partial \text{net}_j}{\partial w_{ij}} \quad (7)$$

Đạo hàm của đầu ra của nơron j tương ứng với đầu vào của nó là đạo hàm riêng của hàm kích hoạt như đã nêu trong phương trình (6). Do đó, đạo hàm của O_j có thể được tính toán nếu tất cả các đạo hàm của đầu ra O_j của lớp kế tiếp- lớp gần hơn với nơron đầu ra được biết đến nhờ phương trình

$$\delta_j = \frac{\partial E}{\partial O_j} \frac{\partial O_j}{\partial \text{net}_j} = \begin{cases} (O_j - p_j)\varphi(\text{net}_j)(1 - \varphi(\text{net}_j)) & \text{nếu } j \text{ là nơron đầu ra,} \\ (\sum_{l \in L} \delta_l w_{jl})\varphi(\text{net}_j)(1 - \varphi(\text{net}_j)) & \text{nếu } j \text{ là nơron bên trong} \end{cases} \quad (8)$$

Do đó, để cập nhật trọng số w_{ij} sử dụng phương pháp giảm gradient, phải chọn tốc độ học α . Thay đổi trọng số, mà cộng thêm vào trọng số cũ, bằng với kết quả của tốc độ học và gradient, nhân với -1:

$$\Delta w_{ij} = -\alpha \frac{\partial E}{\partial w_{ij}} \quad (9)$$

Phương trình này được sử dụng bởi thuật toán lan truyền ngược để điều chỉnh giá trị của các trọng số liên kết gắn với đầu vào ở mỗi nơron trong phương trình tương ứng với lớp bên trong của bộ phân loại đa lớp truyền thẳng.

Theo sáng chế, các giá trị của các tham số bao gồm kết quả công thức máu toàn phần gồm ít nhất 4 chỉ số HGB (lượng huyết sắc tố trong một thể tích máu), MCV (thể tích trung bình của hồng cầu), MCH (lượng huyết sắc tố trung bình có trong một hồng cầu) và RDW (độ phân bố hồng cầu) được nhập vào lớp đầu vào. Theo một phương án, các tham số khác bao gồm tuổi, giới tính, tiền sử bệnh, tiền sử gia đình, nồng độ sắt trong huyết thanh, tuổi thai, kết quả siêu âm thai của thai phụ và các thông tin và kết

quả công thức máu toàn phần của chồng thai phụ cũng được coi là các đặc trưng được nhập vào lớp đầu vào. Số lượng noron của lớp đầu vào dựa trên số lượng tham số dữ liệu đầu vào, theo sáng chế là ít nhất 4 noron đầu vào tương ứng với 4 tham số nêu trên. Theo một phương án cụ thể, số lượng noron của lớp ẩn thứ nhất là 50 và của lớp ẩn thứ hai là 65, và hàm sigmoid được sử dụng cho lớp ẩn là lớp đầu ra. Theo một phương án được ưu tiên, mạng MLP theo sáng chế bao gồm lớp ẩn có 128 noron hoặc 256 noron. Theo sáng chế, mạng MLP bao gồm 2 noron trong lớp đầu ra tương ứng với nhóm có nguy cơ tan máu bẩm sinh và nhóm bình thường.

Chuẩn hóa dữ liệu:

Chuẩn hóa dữ liệu là bước giúp tất cả các đặc trưng có khoảng trị số bằng nhau và tránh sự tác động của các đặc trưng có khoảng trị số lớn hơn trong quá trình huấn luyện dữ liệu và quy đổi tất cả các thuộc tính mang tính mô tả và định danh của bộ dữ liệu được thành các giá trị số,. Trong phương pháp theo sáng chế, tất cả các đặc trưng dạng số được chuẩn hóa bằng cách sử dụng chuẩn min-max.

Mất cân bằng dữ liệu:

Một bộ dữ liệu bị coi là mất cân bằng nếu số lượng trường hợp của một nhóm cụ thể cao hơn nhiều so với các nhóm còn lại. Điển hình là trong lĩnh vực y tế-sức khỏe, số ca gặp vấn đề về sức khỏe luôn nhỏ hơn số ca bình thường. Trong một số trường hợp, điều này dẫn đến khi áp dụng mô hình phân loại, mô hình phân loại này dự đoán mọi trường hợp đều thuộc nhóm đa số và bỏ qua nhóm thiểu số.

Để giải quyết vấn đề mất cân bằng dữ liệu này, thuật toán lấy mẫu quá mức cho nhóm thiểu số bằng cách tổng hợp (Synthetic Minority Over-sampling Technique - SMOTE) được sử dụng trong phương pháp theo sáng chế. SMOTE tạo ra dữ liệu tổng hợp mới cho mẫu thiểu số thay vì nhân đôi dữ liệu thật và được sử dụng rộng rãi trong lĩnh vực kỹ thuật.

Huấn luyện dữ liệu

Mạng noron MLP được huấn luyện sao cho với một tập các vector đầu vào X, mạng có khả năng tạo ra tập các vector đầu ra mong muốn Y của nó. Tập X được sử dụng cho huấn luyện mạng được gọi là tập huấn luyện (training set). Các phần tử x thuộc X được gọi là các mẫu huấn luyện (training example). Quá trình huấn luyện bản chất là sự thay đổi các trọng số liên kết của mạng. Trong quá trình này, các trọng số của mạng sẽ hội tụ dần tới các giá trị sao cho với mỗi vector đầu vào x từ tập huấn

luyện, mạng sẽ cho ra vector đầu ra y như mong muốn. Có ba phương pháp học phổ biến là học có giám sát (supervised learning), học không giám sát (unsupervised learning) và học tăng cường (Reinforcement learning).

Học có giám sát là quá trình học có sự tham gia giám sát của “giáo viên”, trong đó số lớp cần phân loại đã được biết trước. Nhiệm vụ của thuật toán là phải xác định được cách thức phân lớp sao cho với mỗi vector đầu vào sẽ được phân loại chính xác vào lớp của nó.

Học không giám sát là việc học không cần có bất kỳ một sự giám sát nào. Trong bài toán học không giám sát, tập dữ liệu huấn luyện được cho dưới dạng: $D = \{(x_1, x_2, \dots, x_N)\}$, với $(x_1, x_2, \dots, x_N)$ là vector đặc trưng của mẫu huấn luyện. Nhiệm vụ của thuật toán là phải phân chia tập dữ liệu D thành các nhóm con, mỗi nhóm chứa các vector đầu vào có đặc trưng giống nhau. Như vậy với học không giám sát, số lớp phân loại chưa được biết trước, và tùy theo tiêu chuẩn đánh giá độ tương tự giữa các mẫu để có các lớp phân loại khác nhau.

Học tăng cường còn được gọi là học thưởng-phạt (reward- penalty learning) là sự tổ hợp của cả hai mô hình trên. Phương pháp này cụ thể như sau: với vector đầu vào, quan sát vector đầu ra do mạng tính được. Nếu kết quả được xem là “tốt” thì mạng sẽ được thưởng theo nghĩa tăng các trọng số kết nối lên; ngược lại mạng sẽ bị phạt, các trọng số kết nối không thích hợp sẽ được giảm xuống. Do đó học tăng cường là học theo góp ý phê bình (critic), khác với học có giám sát là học theo “giáo viên” chỉ định.

Theo một phương án, mô hình mạng được sử dụng trong phương pháp theo sáng chế được huấn luyện và thử nghiệm bằng cách theo phương pháp có giám sát đánh giá chéo 10 lần. Theo đó, bộ dữ liệu được chia ngẫu nhiên thành 10 phần bằng nhau mà sử dụng 9 phần và thử nghiệm phần còn lại. Quá trình được lặp lại mười lần, trong đó ở mỗi lần thực hiện, bộ phân loại được thử nghiệm ở phần dữ liệu khác nhau. Ở cuối của quá trình, độ lệch chuẩn và độ lệch trung bình của tất cả kết quả thử nghiệm được đánh giá.

Đánh giá mô hình mạng MLP sau khi huấn luyện dữ liệu theo sáng chế

Ba chỉ số thường được sử dụng để xác định hiệu suất của phương pháp cảnh báo nhờ mô hình phân loại theo sáng chế so với các mô hình phân loại khác bao gồm độ phủ (recall), khả năng dự báo (precision) và độ chính xác (accuracy).

Bảng 2: Các chỉ số đánh giá mô hình theo sáng chế

Chỉ số	Công thức
Độ chính xác (Accuracy)	$\frac{(TP + TN). 100\%}{TP + TN + FP + FN}$
Khả năng dự báo (Precision)	$\frac{TP. 100\%}{TP + FP}$
Độ phủ (Recall)	$\frac{TP. 100\%}{TP + FP}$
F1-score	$2 \times \frac{Precision.Recall}{Precision + Recall}$

Ví dụ thực hiện sáng chế

Phương pháp cảnh báo nguy cơ mắc tan máu bẩm sinh theo sáng chế đã được xây dựng để cảnh báo nguy cơ mắc tan máu bẩm sinh đối với 10000 đối tượng là các sản phụ của Bệnh viện Phụ sản trung ương (PSTW) và được kiểm chứng bằng kết quả chẩn đoán sinh học phân tử xác định đột biến gen gây tan máu bẩm sinh.

Đối tượng nghiên cứu:

10.000 bệnh án có kết quả xét nghiệm công thức máu và kết quả xét nghiệm đột biến gen gây tan máu bẩm sinh của 10.000 thai phụ và chồng thai phụ được lưu tại Bệnh viện Phụ sản trung ương (PSTW) từ năm 2012-2021 bao gồm cả các trường hợp có đột biến gen và không có đột biến gen gây tan máu bẩm sinh, và được chia làm 2 tập dữ liệu: dữ liệu huấn luyện mô hình (9.000) và dữ liệu thử nghiệm mô hình (1.000).

Thiết lập cơ sở dữ liệu:

Nhập thông tin về tuổi, tuổi thai, mã số, tiền sử bệnh và tiền sử gia đình của thai phụ, các kết quả xét nghiệm của thai phụ, chồng thai phụ và thai nhi bao gồm kết quả xét nghiệm công thức máu của thai phụ, chồng thai phụ và kết quả xét nghiệm đột biến gen gây tan máu bẩm sinh của thai phụ, chồng thai phụ và thai nhi bao gồm: các chỉ số Hb, MCV, MCH của thai phụ, chồng thai phụ; chỉ số Fe, ferritin huyết thanh của thai phụ, chồng thai phụ; kết quả điện di Hb của thai phụ, chồng thai phụ; kết quả siêu âm thai.

Xử lý dữ liệu

Tất cả các đặc trưng dạng số được chuẩn hóa bằng cách sử dụng chuẩn min-max và sử dụng thuật toán SMOTE để cân bằng dữ liệu.

Huấn luyện dữ liệu theo phương pháp học có giám sát

Đầu vào của mô hình là 4 chỉ số công thức máu của thai phụ gồm HGB (lượng huyết sắc tố trong một thể tích máu), MCV (thể tích trung bình của hồng cầu), MCH (lượng huyết sắc tố trung bình có trong một hồng cầu) và RDW (độ phân bố hồng cầu) và so sánh dữ liệu từ dữ liệu huấn luyện và dữ liệu thử nghiệm theo mô hình mạng đa lớp truyền thẳng và kiểm tra theo dữ liệu giám sát được thiết lập lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%.

Hiệu chỉnh giá trị theo mô hình AI sau khi huấn luyện dữ liệu theo sáng chế

Đánh giá độ chính xác của phần mềm bằng cách so sánh kết quả máy đưa ra so với chẩn đoán phân tử xác định đột biến gen gây tan máu bẩm sinh từ bệnh án. Bên cạnh đó, độ chính xác và hiệu quả của mô hình theo sáng chế được so sánh với các mô hình KNN, SVM, Random Forest cũng được huấn luyện bộ dữ liệu theo sáng chế. Hiệu quả của thuật toán SMOTE cân bằng dữ liệu cũng được đánh giá như được thể hiện trong bảng 3.

Bảng 1: So sánh kết quả các mô hình theo sáng chế sử dụng 4 đặc trưng được lựa chọn theo sáng chế với dữ liệu của thai phụ

Mô hình	Accuracy (%)		Precision* (%)		Recall* (%)		F1-Score	
	DL gốc	SMOTE	DL gốc	SMOTE	DL gốc	SMOTE	DL gốc	SMOTE
KNN	97,53	97,53	93,33	93,33	98,21	100	96,55	96,55
SVM	96,91	90,74	91,8	79,71	100	98,21	95,73	88
RF	96,91	91,36	91,8	80	98,21	100	95,65	88,9
MLP	97,81	97,53	93,22	93,33	100	100	95,72	96,55

* các giá trị Precision và Recall được tính cho nhóm thai phụ mang gen gây tan máu bẩm sinh

Từ những trị số nêu trong bảng 3 có thể thấy rằng độ chính xác của mô hình được sử dụng trong phương pháp cảnh báo theo sáng chế là trên 90%. Ngoài ra, sau khi thực hiện cân bằng dữ liệu bằng thuật toán SMOTE, độ chính xác được cải thiện đáng kể.

Hiệu quả đạt được của sáng chế

Phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh (thalassemia) cho phép hỗ trợ hiệu quả trong việc cảnh báo

sớm nguy cơ mắc bệnh tan máu bẩm sinh của thai nhi dựa trên bốn chỉ số phân tích mẫu máu của thai phụ. Phương pháp theo sáng chế dựa trên cơ sở học máy cho phép cơ sở dữ liệu tự điều chỉnh theo cơ sở dữ liệu xây dựng cho các nhóm nguy cơ khác nhau.

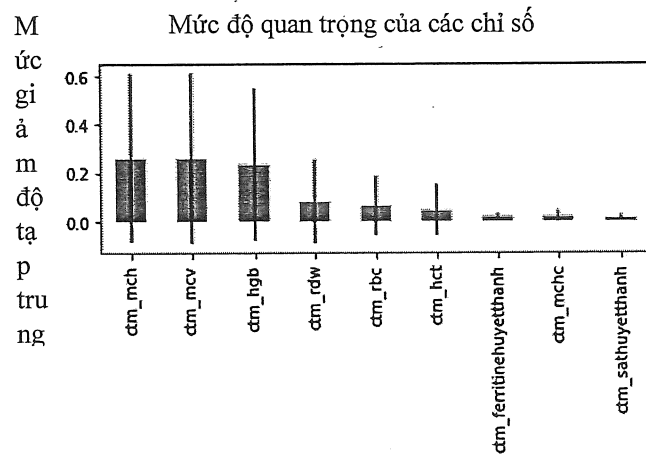
Phương pháp đánh giá chỉ số phân tích máu theo sáng chế tối ưu số lượng đặc trưng sử dụng giúp làm giảm thời gian và số phép toán cần thực hiện, trợ giúp hiệu quả cho công tác sàng lọc trước khi sinh với chi phí thấp, tốc độ nhanh và độ chính xác trên 90%. Phương pháp theo sáng chế là công cụ hỗ trợ cho nhân viên y tế và thai phụ trước khi thực hiện các xét nghiệm sinh học phân tử sàng lọc tiếp theo như điện di Hg và/hoặc PCR để đưa ra kết luận chính xác đối với thai nhi mắc bệnh tan máu bẩm sinh.

YÊU CẦU BẢO HỘ

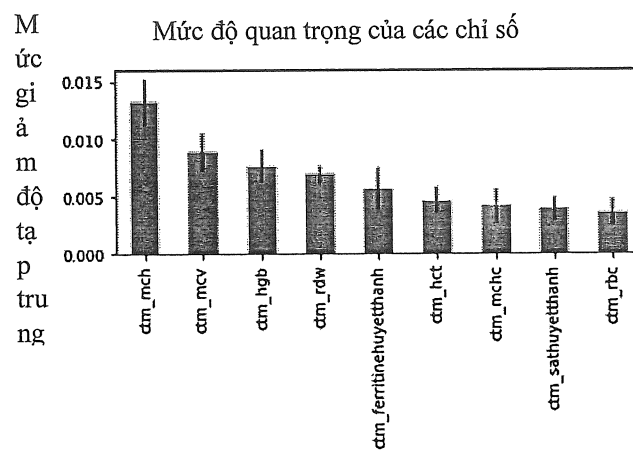
1. Phương pháp đánh giá chỉ số phân tích máu để hỗ trợ cảnh báo sớm nguy cơ mắc bệnh tan máu bẩm sinh, trong đó phương pháp này bao gồm các bước:

- a) thu thập thông tin phân tích máu toàn phần bằng cách tiến hành phân tích bốn chỉ số của mẫu máu bao gồm HGB (lượng huyết sắc tố trong một thể tích máu), MCV (thể tích trung bình của hồng cầu), MCH (lượng huyết sắc tố trung bình có trong một hồng cầu) và RDW (độ phân bố hồng cầu) để thu nhận kết quả phân tích;
- b) xử lý dữ liệu bằng cách nạp bốn chỉ số phân tích máu nêu trên vào hệ thống phân tích với cơ sở dữ liệu về nguy cơ mắc bệnh tan máu bẩm sinh được xây dựng dựa trên mô hình mạng nơron truyền thẳng nhiều lớp (Multi-layer perceptron- MLP) với đầu vào là các chỉ số HGB, MCV, MCH, RDW và ngưỡng (cut- off) so sánh các giá trị HGB, MCV, MCH, RDW này được thiết lập trong cơ sở dữ liệu lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%, trong đó hệ thống phân tích với cơ sở dữ liệu về bệnh tan máu bẩm sinh được xây dựng trên cơ sở bao gồm:
 - thiết lập cơ sở dữ liệu bao gồm xây dựng thông tin dữ liệu phân tích máu của 10.000 thai phụ, trong đó thông tin dữ liệu phân tích máu của mỗi thai phụ được liệt kê tương ứng như một bộ giá trị thuộc tính bao gồm ít nhất bốn chỉ số phân tích máu HGB, MCV, MCH và RDW;
 - xử lý dữ liệu bằng cách trích xuất dữ liệu gồm bốn chỉ số công thức máu trong cơ sở dữ liệu được xây dựng, cân bằng và chuẩn hóa dữ liệu bằng thuật toán chống mất cân bằng dữ liệu SMOTE (Synthetic Minority Over-sampling), sau đó phân chia ngẫu nhiên cơ sở dữ liệu thành dữ liệu huấn luyện (train data) và dữ liệu thử nghiệm (test data) theo tỷ lệ 9:1;
 - huấn luyện dữ liệu bằng cách so sánh dữ liệu từ dữ liệu huấn luyện và dữ liệu thử nghiệm theo mô hình mạng đa lớp truyền thẳng và kiểm tra theo dữ liệu giám sát được thiết lập lần lượt là 113,5g/l, 83,15fL, 27,65pg và 14,25%; và
 - tự hiệu chỉnh giá trị vượt ngưỡng theo giá trị giám sát đối với dữ liệu huấn luyện và dữ liệu nạp để tăng hiệu quả và độ chính xác của hệ thống phân tích;
- c) đưa ra kết quả đánh giá chỉ số phân tích dựa trên tổ hợp bốn chỉ số phân tích máu được nạp và cảnh báo nguy cơ mắc bệnh tan máu bẩm sinh nếu các giá trị này vượt ngưỡng hiệu chỉnh thiết lập.

HÌNH 1

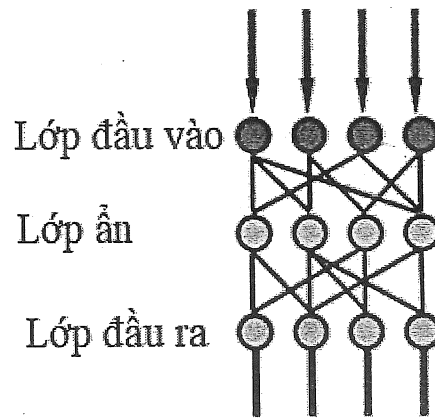


(A)



(B)

HÌNH 2



HÌNH 3

