

MÔ HÌNH HỌC MÁY PHÂN LOẠI RỐI LOẠN NHỊP TIM SỬ DỤNG DỮ LIỆU ECG

Nguyễn Thị Thu Hiền

Học viện Công nghệ Bưu chính Viễn thông, Hà nội, Việt nam

Tóm tắt: Bệnh tim mạch hiện là một trong những nguyên nhân gây tử vong hàng đầu trên toàn thế giới, do đó, việc phát hiện sớm và can thiệp kịp thời các bệnh lý tim mạch trở thành vấn đề cấp thiết. Trong bối cảnh này, theo dõi điện tâm đồ (ECG) được xem là công cụ quan trọng để đánh giá nhịp tim và phòng ngừa bệnh tim mạch (CVD). Với sự hỗ trợ của các phương pháp học máy, độ chính xác trong chẩn đoán rối loạn nhịp tim đã được cải thiện đáng kể. Bài báo giới thiệu một mô hình học máy hiệu quả để phân loại rối loạn nhịp tim dựa trên dữ liệu điện tâm đồ. Nghiên cứu tiến hành khảo sát và so sánh các thuật toán học máy điển hình, điều chỉnh tham số tối ưu trên tập dữ liệu ECG trong tập dữ liệu Chapman ECG 12 chuyển đạo. Phương pháp tối ưu nhất, mô hình RF, được lựa chọn dựa trên hiệu năng của mô hình thông qua các chỉ số đánh giá như độ chính xác, điểm F1, độ nhạy, khả năng thu hồi và độ đặc hiệu. Kết quả thực nghiệm cho thấy, mô hình đề xuất đạt độ chính xác tổng thể 96,20% trong việc phân loại các loại rối loạn nhịp tim khác nhau. Các kết quả đạt được cho thấy tính khả thi của việc ứng dụng các thuật toán học máy tiên tiến trong chẩn đoán tim mạch, đồng thời mở ra tiềm năng ứng dụng trong các hệ thống chăm sóc sức khỏe từ xa và giám sát sức khỏe cá nhân.

Từ khóa: Bệnh tim mạch, tín hiệu điện tâm đồ, học máy, phân loại rối loạn nhịp tim, thuật toán rừng ngẫu nhiên.

I. GIỚI THIỆU CHUNG

Hiện nay, bệnh tim mạch được coi là một trong những vấn đề sức khỏe nghiêm trọng nhất, gây ra tỷ lệ tử vong cao trên toàn cầu [1]. Theo báo cáo của Tổ chức Y tế Thế giới, vào năm 2019, bệnh tim mạch đã khiến 17,9 triệu người tử vong, chiếm 32% tổng số ca tử vong toàn cầu, với 85% trong số đó liên quan đến các cơn đau tim và đột quỵ [1]. Những số liệu này nhấn mạnh tầm quan trọng của việc phát triển các công cụ chẩn đoán hiện đại và phương pháp phát hiện sớm các bất thường về tim [2], [3], [4], [5]. Trong số các phương pháp chẩn đoán, điện tâm đồ (ECG) được xem như tiêu chuẩn vàng trong việc phát hiện loạn nhịp tim, mang lại thông tin chính xác về hoạt động điện của tim và

các bất thường tiềm ẩn [6], [7]. Là một phương pháp không xâm lấn, ECG cung cấp dữ liệu hữu ích cho cả việc chẩn đoán tức thì và giám sát lâu dài [8], [9]. Vì vậy, việc xây dựng các giải pháp hiệu quả nhằm phát hiện sớm nhịp tim bất thường là rất quan trọng trong việc giảm tác động của các bệnh lý tim mạch, với ECG được công nhận rộng rãi là công cụ đáng tin cậy nhất cho nhiệm vụ này [10], [11], [12].

Trong khi đó, nhờ các thành tựu công nghệ, đặc biệt là trong lĩnh vực trí tuệ nhân tạo, sự tiến bộ trong chẩn đoán bệnh tim đã được nâng cao đáng kể [13], [14]. Trong những năm gần đây, các phương pháp học máy (ML) và học sâu đã được áp dụng ngày càng nhiều để phân loại loạn nhịp tim. Các nhà khoa học đã thử nghiệm nhiều cách tiếp cận thuật toán, bao gồm các kỹ thuật học máy truyền thống như K-Nearest Neighbors (KNN), Random Forest (RF) và Support Vector Machine (SVM) [9], [15]. Những phương pháp này đã được sử dụng riêng lẻ hoặc kết hợp theo những cách sáng tạo, mang lại kết quả khả quan trong việc phát hiện các bất thường nhịp tim [9], [16]. Việc tích hợp các công nghệ tính toán hiện đại với phân tích ECG truyền thống đã đánh dấu một bước tiến quan trọng, giúp cải thiện cả độ chính xác lẫn hiệu suất trong chẩn đoán loạn nhịp tim, mở ra một giai đoạn mới cho lĩnh vực chẩn đoán tim mạch.

Hơn nữa, với tư cách là tiêu chuẩn lâm sàng, nhờ những lợi thế cung cấp cái nhìn toàn diện hơn về hoạt động điện của tim bằng cách thu tín hiệu từ 12 góc độ khác nhau và tăng cường khả năng chẩn đoán loạn nhịp phức tạp, nhồi máu cơ tim và các tình trạng tim khác với độ chính xác và độ đặc hiệu cao hơn, ECG 12 chuyển đạo rất quan trọng để đánh giá tim kỹ lưỡng, cho phép chẩn đoán chính xác và đáng tin cậy hơn so với các cấu hình ECG đơn giản hóa khác [17], [18], [19]. Nhiều nghiên cứu gần đây đã tập trung vào phân loại loạn nhịp dựa trên ECG 12 chuyển đạo [20], [21]. Phương pháp CNN chuyển đổi tín hiệu ECG thành hình ảnh RGB để trích xuất đặc điểm và phân loại đa lớp. Ngoài ra, MLP đã được chứng minh vượt trội hơn SVM, đạt độ chính xác 84,2%, trong khi một nghiên cứu khác đạt tới 97,0% khi phân loại bốn lớp loạn nhịp tim [10], [22]. Tuy nhiên, các phương pháp này thường yêu cầu thêm đặc điểm lâm sàng, gây khó khăn cho việc tích hợp vào hệ thống tự động hóa.

Trên cơ sở đó, bài báo này trình bày một mô hình học máy hiệu quả để phân loại rối loạn nhịp tim dựa trên dữ liệu

Tác giả liên hệ: Nguyễn Thị Thu Hiền,

Email: hiennt@ptit.edu.vn

Đến tòa soạn: 12/2024, chỉnh sửa: 01/2025, chấp nhận đăng: 02/2025.

điện tâm đồ (ECG). Nghiên cứu đã thực hiện khảo sát, so sánh các thuật toán học máy tiêu biểu và tối ưu hóa tham số trên tập dữ liệu Chapman ECG 12 chuyển đạo. Trong đó, mô hình Random Forest (RF) được lựa chọn nhờ hiệu năng vượt trội dựa trên các chỉ số đánh giá như độ chính xác, F1-score, độ chuẩn xác, độ nhạy, và độ đặc hiệu. Các kết quả mô phỏng số đạt được không chỉ khẳng định tính khả thi của việc áp dụng các thuật toán học máy tiên tiến trong chẩn đoán tim mạch mà còn mở ra triển vọng ứng dụng trong các hệ thống chăm sóc sức khỏe từ xa và giám sát sức khỏe cá nhân.

Các nội dung trình bày của bài báo được bố cục như sau. Phương pháp áp dụng trong nghiên cứu này, bao gồm giải pháp học máy phân loại rối loạn nhịp tim sử dụng ECG, tập dữ liệu và tiền xử lý dữ liệu và mô hình học máy, được trình bày trong Mục II. Mục III trình bày các kết quả mô phỏng số và các thảo luận về các kết quả đạt được. Cuối cùng, các kết luận đúc kết từ nghiên cứu được tổng hợp trong Mục IV.

II. PHƯƠNG PHÁP

A. Giải pháp học máy phân loại rối loạn nhịp tim sử dụng ECG

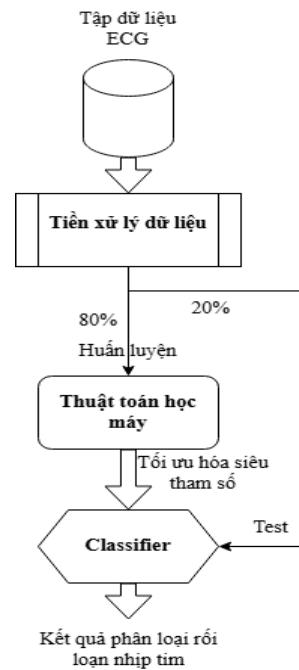
Quy trình thực hiện phân loại rối loạn nhịp tim sử dụng bộ dữ liệu ECG được mô tả trên Hình 1. Việc xây dựng mô hình học máy nhằm phân loại rối loạn nhịp tim dựa trên tập dữ liệu ECG được thực hiện theo các bước sau:

❖ **Tiền xử lý dữ liệu:** Tập dữ liệu ECG Chapman chứa các tín hiệu ECG từ 12 chuyển đạo của nhiều bệnh nhân khác nhau. Dữ liệu được làm sạch để loại bỏ các tín hiệu nhiễu và dữ liệu không hợp lệ. Các tín hiệu ECG được chuẩn hóa về độ dài và biên độ. Tiếp theo, áp dụng các kỹ thuật làm mịn tín hiệu và lọc nhiễu (như bộ lọc Butterworth) để cải thiện chất lượng tín hiệu. Từ tín hiệu ECG, các đặc trưng quan trọng được trích xuất để phục vụ quá trình học máy. Các đặc trưng có thể bao gồm các đặc điểm trong miền thời gian (khoảng RR, độ rộng của sóng P, QRS, T), miền tần số (biến đổi Fourier) và các đặc trưng thống kê (trung bình, phương sai, độ lệch chuẩn). Đặc biệt, sử dụng các thuật toán phát hiện đỉnh (peak detection) để xác định vị trí các sóng P, QRS và T, từ đó trích xuất các khoảng PR, QT và QRS. Việc phân chia tập dữ liệu thành tập huấn luyện, tập kiểm định và tập kiểm thử được thực hiện để đảm bảo tính khách quan trong đánh giá mô hình.

❖ **Xây dựng và huấn luyện mô hình học máy:** Các mô hình học máy phổ biến như Random Forest, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), ... được sử dụng để phân loại rối loạn nhịp tim. Mô hình được huấn luyện trên tập huấn luyện bằng cách tối ưu hóa và điều chỉnh các siêu tham số.

❖ **Đánh giá mô hình:** Sau khi mô hình được huấn luyện, tập kiểm thử sẽ được sử dụng để đánh giá hiệu năng của mô hình. Các chỉ số đánh giá bao gồm độ chính xác (accuracy), độ chuẩn xác (precision), độ nhạy (recall/sensitivity), độ đặc hiệu (specificity) và F1-score. Đánh giá mô hình không chỉ tập trung vào độ chính xác tổng thể mà còn kiểm tra khả năng nhận diện từng loại rối loạn nhịp tim riêng biệt.

Trong nghiên cứu này, từng thuật toán học máy phổ biến được thử nghiệm và thuật toán cho hiệu năng cao nhất sẽ được lựa chọn để khuyến nghị xây dựng mô hình phân loại rối loạn nhịp tim.



Hình 1. Xây dựng mô hình học máy phân loại rối loạn nhịp tim dựa trên tín hiệu điện tâm đồ.

B. Tập dữ liệu và tiền xử lý dữ liệu

Bộ dữ liệu ECG Chapman, một nguồn tài nguyên toàn diện được phát triển thông qua sự hợp tác giữa Đại học Chapman, Bệnh viện Nhân dân Thiệu Hưng và Bệnh viện số 1 Ninh Ba, được coi là bộ dữ liệu chính để phát triển phương pháp và xác nhận phân loại loạn nhịp trong nghiên cứu này [17]. Bộ dữ liệu, một cơ sở dữ liệu được công nhận rộng rãi, cập nhật và mới nhất, cung cấp các chú thích chất lượng cao cho nhiều loại loạn nhịp, bao gồm các bản ghi ECG 12 chuyển đạo được lấy mẫu ở tần số 500 Hz với thời lượng 10 giây từ 10.646 bệnh nhân. Các bác sĩ lâm sàng chuyên gia đã cung cấp các chú thích cho 11 nhịp tim phổ biến và 67 tình trạng tim mạch bổ sung. Các tính năng này làm cho bộ dữ liệu trở thành một nguồn tài nguyên có giá trị để phát triển, so sánh và tối ưu hóa các phương pháp thống kê và học máy trong nghiên cứu về loạn nhịp và các rối loạn tim mạch [16,20]. Sử dụng công cụ Neurokit, 20 đặc trưng của tín hiệu ECG liên quan đến các đặc điểm trong miền thời gian (khoảng RR, độ rộng của sóng P, QRS, T), miền tần số và các đặc trưng thống kê (trung bình,

phương sai, độ lệch chuẩn) được trích xuất cho mỗi đầu (chuyển đạo) của dữ liệu thu được. Các dữ liệu thu được sẽ được loại bỏ các thông tin không rõ ràng hay không đầy đủ, tổng số đặc trưng thu được là 212 đặc trưng.

Dựa trên ý nghĩa lâm sàng của chúng, 11 nhịp tim được các bác sĩ lâm sàng chuyên gia xác định đã được hợp nhất thành 4 loại chính: Rung tâm nhĩ/Rung nhĩ (AFIB), Nhịp chậm xoang (SB), Nhịp xoang (SR) và Nhịp nhanh trên thất chung (GSVT) [12,16,17]. Loại AFIB bao gồm cả tình trạng Rung tâm nhĩ (Atrial Fibrillation) và Rung nhĩ (Atrial Flutter). SB biểu thị các trường hợp Nhịp chậm xoang, trong khi SR bao gồm cả Nhịp xoang đều và không đều. Phân loại GSVT kết hợp năm tình trạng riêng biệt: nhịp nhanh trên thất, nhịp nhanh nhĩ, nhịp nhanh tái nhập nút nhĩ thất, nhịp nhanh tái nhập nhĩ thất và máy tạo nhịp nhĩ lang thang. Bên cạnh việc phản ánh các giao thức điều trị chung và cơ chế sinh lý trong mỗi loại, sự hợp nhất này còn giải quyết một số thách thức kỹ thuật trong việc triển khai máy học: nó làm giảm các vấn đề mất cân bằng lớp, tăng cường sự khác biệt về tính năng giữa các loại, cải thiện hiệu quả đào tạo mô hình và thúc đẩy khả năng khái quát hóa tốt hơn vào các tình huống thực tế. Hơn nữa, hệ thống phân loại đơn giản hóa cung cấp những lợi thế thực tế như giảm độ phức tạp của chú thích, độ tin cậy giữa các bác sĩ lâm sàng cao hơn, ứng dụng thời gian thực hiệu quả hơn và giảm nguy cơ phân loại sai giữa các phân nhóm tương tự. Cách tiếp cận cân bằng này đảm bảo cả tính liên quan lâm sàng và tính khả thi về mặt kỹ thuật trong khi vẫn duy trì tiện ích chẩn đoán thiết yếu của hệ thống phân loại.

C. Mô hình học máy

Trong nghiên cứu này, mười thuật toán học máy điển hình được áp dụng và đánh giá trên tập dữ liệu ECG nói trên. Các mô hình học máy này được huấn luyện trên bộ dữ liệu huấn luyện và kiểm thử hiệu năng trên bộ dữ liệu test. Việc điều chỉnh và tối ưu hóa siêu tham số được thực hiện bằng kỹ thuật Grid Search. Thuật toán đạt hiệu năng cao nhất sẽ được lựa chọn. Các thuật toán học máy được khảo sát trong nghiên cứu này bao gồm:

- ❖ *AdaBoost (Ada)* là thuật toán Boosting kết hợp nhiều mô hình phân loại yếu thành một mô hình mạnh bằng cách điều chỉnh trọng số của các mẫu bị phân loại sai qua từng vòng lặp. Thuật toán này hiệu quả trong việc giảm lỗi phân loại nhưng dễ bị ảnh hưởng bởi nhiễu và dữ liệu ngoại lai.
- ❖ *Bagging* sử dụng phương pháp lấy mẫu ngẫu nhiên có lặp để tạo ra nhiều mô hình độc lập (thường là cây quyết định) và tổng hợp kết quả qua bỏ phiếu hoặc trung bình. Kỹ thuật này giảm phương sai và hiện tượng overfitting, nhưng đòi hỏi nhiều tài nguyên tính toán.
- ❖ *Decision Tree (DT)* phân loại dữ liệu thông qua các điều kiện dạng "có/không" tại các nút. Mô hình này

dễ hiểu và triển khai nhưng dễ bị overfitting và nhạy cảm với dữ liệu nhiễu.

- ❖ *k-Nearest Neighbors (KNN)* là mô hình phân loại phi tham số, dựa trên khoảng cách của một điểm mới so với k điểm lân cận gần nhất. Mô hình này đơn giản và không cần huấn luyện, nhưng chậm với dữ liệu lớn và nhạy cảm với outliers.
- ❖ *Gradient Boosting (GB)* là thuật toán Boosting tối ưu hóa gradient của hàm lỗi qua nhiều mô hình con (thường là cây quyết định). So với AdaBoost, Gradient Boosting tối ưu hiệu năng dựa trên gradient, mang lại kết quả chính xác cao nhưng tốn tài nguyên và dễ bị overfitting.
- ❖ *Logistic Regression (LR)* là mô hình phân loại tuyến tính, áp dụng hàm sigmoid để đưa xác suất về khoảng $[0,1]$. Mô hình này đơn giản, dễ áp dụng, nhưng không hoạt động tốt với dữ liệu phi tuyến và bị ảnh hưởng bởi đa cộng tuyến.
- ❖ *Naive Bayes (NB)* là thuật toán phân loại dựa trên định lý Bayes, giả định rằng các đặc trưng là độc lập có điều kiện. Mô hình này nhanh, đơn giản và hiệu quả với dữ liệu lớn, nhưng giả định về tính độc lập thường không thực tế trong nhiều bài toán.
- ❖ *Random Forest (RF)* là tập hợp của nhiều cây quyết định, mỗi cây huấn luyện trên các tập con dữ liệu ngẫu nhiên, sau đó tổng hợp kết quả bằng bỏ phiếu. Mô hình này cải thiện độ chính xác và kháng nhiễu tốt, nhưng tốn tài nguyên và khó diễn giải.
- ❖ *Support Vector Machine (SVM)* tìm siêu phẳng tối ưu để phân chia dữ liệu, tối đa hóa khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất. SVM hoạt động tốt trong các bài toán phi tuyến thông qua "kernel trick", nhưng tốn thời gian huấn luyện và dễ bị ảnh hưởng bởi nhiễu.
- ❖ *XGBoost (Extreme Gradient Boosting)* là phiên bản cải tiến của Gradient Boosting, được tối ưu về tốc độ và bộ nhớ, đồng thời sử dụng *regularization* (L1, L2) để tránh *overfitting*. XGBoost nổi bật với hiệu năng cao nhưng đòi hỏi tài nguyên lớn và quá trình điều chỉnh siêu tham số phức tạp.

Các thuật toán trên khác nhau về nguyên lý hoạt động, ưu nhược điểm và phạm vi ứng dụng. Các thuật toán như Logistic Regression, Naive Bayes và KNN thích hợp cho các bài toán đơn giản, dễ triển khai và thời gian tính toán nhanh. Ngược lại, các thuật toán như Gradient Boosting, XGBoost và Random Forest có khả năng xử lý tốt các bài toán phức tạp nhưng yêu cầu nhiều tài nguyên tính toán và thời gian huấn luyện lớn hơn.

III. ĐÁNH GIÁ HIỆU NĂNG VÀ THẢO LUẬN

Để đánh giá hiệu năng của mô hình học máy phân loại rối loạn nhịp tim, các tham số hiệu năng cơ bản được sử dụng để đánh giá các thuật toán và lựa chọn giải pháp tốt nhất bao gồm: độ chính xác (accuracy), độ chuẩn xác

(precision), độ nhạy (sensitivity), điểm F1 (F1-score), và độ đặc hiệu (specificity). Độ chính xác đề cập đến tỷ lệ chính xác trong cả kết quả dương tính thật và kết quả âm tính thật so với tổng số dự đoán. Điểm F1 thể hiện sự hài hòa giữa độ chính xác và thu hồi, điều này hữu ích khi đối mặt với dữ liệu đầu vào mất cân bằng. Độ chính xác là số trường hợp dương tính thực tế để xem dữ liệu có liên quan như thế nào. Mô hình thu hồi tất cả các trường hợp dương tính có liên quan để trả lời câu hỏi về số lượng trường hợp được dự đoán chính xác. Đo lường độ đặc hiệu được tính toán để đo lường mức độ hiệu quả của mô hình trong việc tránh các kết quả dương tính giả, trong đó việc nhầm lẫn kết quả ngược lại sẽ dẫn đến hậu quả nghiêm trọng.

Giá sử ký hiệu TP, TN, FP và FN lần lượt là số mẫu dương tính thật, số mẫu âm tính thật, số mẫu dương tính giả và số mẫu âm tính giả. Trong bối cảnh phân loại loạn nhịp tim dựa trên ECG, TP là những trường hợp mà mô hình xác định đúng loạn nhịp tim khi nó thực sự có mặt, TN là những trường hợp mà mô hình xác định đúng nhịp tim bình thường khi không có loạn nhịp tim, trong khi FP là những trường hợp mà mô hình xác định sai loạn nhịp tim khi nhịp tim thực sự bình thường và FN là những trường hợp mà mô hình không phát hiện ra loạn nhịp tim khi nó thực sự có mặt. Các tham số hiệu năng này được tính theo các công thức như sau:

$$\text{Độ chính xác} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Độ chuẩn xác} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Độ nhạy} = \frac{TP}{TP+FN} \quad (3)$$

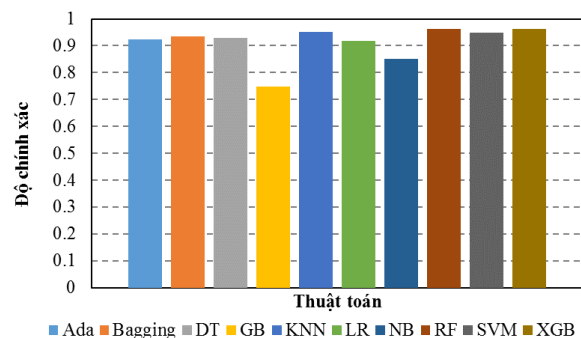
$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Độ đặc hiệu} = \frac{TN}{TN+FP} \quad (5)$$

Dữ liệu huấn luyện (80% dữ liệu tổng) thu được sau tiền xử lý được đưa vào 10 thuật toán theo thứ tự lần lượt là Ada, Bagging, DT, GB, KNN, LR, NB, RF, SVM và XGB. Các mô hình học máy được huấn luyện, điều chỉnh và tối ưu các siêu tham số. Bảng 1 thể hiện các siêu tham số tương ứng của mỗi mô hình học máy được khảo sát. Sau đó, các mô hình với siêu tham số tối ưu được áp dụng trên bộ dữ liệu test (20% dữ liệu còn lại) để đánh giá hiệu năng và so sánh với nhau. Trong khuôn khổ bài báo này, tham số hiệu năng chính được lựa chọn làm mục tiêu so sánh và lựa chọn mô hình cho bài toán phân loại rối loạn nhịp tim là độ chính xác. Như thể hiện trên Hình 2, độ chính xác của các mô hình đạt được đều tương đối cao, độ chính xác thấp nhất là 74,69% (đạt được với thuật toán GB) và độ chính xác cao nhất là 96,20% (đạt được với thuật toán RF). Trên thực tế, các thuật toán XGB, KNN và SVM cũng đạt được độ chính xác rất cao, lần lượt là 96,15%, 95,26% và 94,98%.

Bảng 1. Siêu tham số của các mô hình học máy.

Số thứ tự	Thuật toán	Siêu tham số tối ưu
1	Ada	{'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 61}
2	Bagging	{'estimator': XGBClassifier(...), 'max_features': 10, 'max_samples': 100, 'n_estimators': 100, 'n_jobs': -1}
3	DT	{'criterion': 'entropy', 'max_depth': 6, 'max_features': 'sqrt', 'splitter': 'best'}
4	GB	{'criterion': 'friedman_mse', 'learning_rate': 0.1, 'loss': 'log_loss', 'n_estimators': 300}
5	KNN	{'algorithm': 'auto', 'n_neighbors': 8, 'p': 1, 'weights': 'uniform'}
6	LR	{'C': 0.14, 'max_iter': 30, 'multi_class': 'auto', 'n_jobs': -1, 'penalty': 'l2', 'solver': 'sag', 'tol': 0.015}
7	NB	{'var_smoothing': 1e-08}
8	RF	{'criterion': 'entropy', 'max_depth': 15, 'max_features': 'sqrt', 'n_estimators': 40}
9	SVM	{'C': 15, 'gamma': 'scale', 'kernel': 'rbf', 'probability': True}
10	XGB	{'gamma': 0, 'learning_rate': 0.05, 'max_depth': 5, 'min_child_weight': 1, 'n_estimators': 950}



Hình 2. So sánh độ chính xác đạt được của các mô hình học máy được khảo sát.

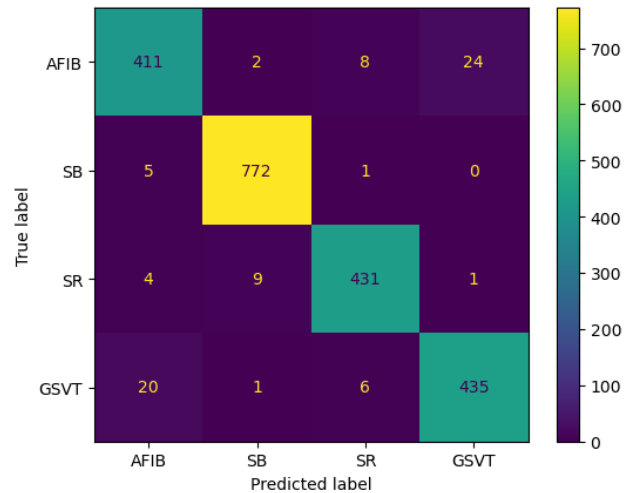
Dựa trên các kết quả đạt được về độ chính xác, thuật toán RF được lựa chọn là thuật toán tốt nhất trong số các thuật toán được khảo sát trong nghiên cứu này để giải quyết bài toán phân loại rối loạn nhịp tim. Bảng 2 tổng hợp chi tiết các tham số hiệu năng cho từng loại rối loạn nhịp tim thành phần bao gồm: AFIB, SB, SR và GSVT. Độ chính xác tổng thể đạt được của mô hình này là 96,20%. Trong khi đó, trung bình macro, trung bình micro và trung bình theo trọng lượng (theo số lượng các mẫu tương ứng với từng loại rối loạn nhịp tim) của F1-score lần lượt là 95,71%, 96,20% và 96,19%. Các kết quả đạt được của mô hình phân loại rối loạn nhịp tim cho thấy hiệu năng rất cao đối với cả 4 loại rối loạn. Cụ thể:

Bảng 2. Tham số hiệu năng theo từng thành phần của mô hình học máy RF.

Loại rối loạn nhịp tim	Tham số hiệu năng				
	Độ chính xác	F1-score	Độ chuẩn xác	Độ nhạy	Độ đặc hiệu
AFIB	0.9704	0.9236	0.9341	0.9288	0.9828
SB	0.9915	0.9923	0.9847	0.9885	0.9911
SR	0.9864	0.9685	0.9664	0.9675	0.9911
GSVT	0.9756	0.9416	0.9457	0.9436	0.9850

- ❖ Loại AFIB: Độ chính xác đạt 0.9704 cùng với F1-score 0.9236, thể hiện sự cân bằng giữa độ nhạy (0.9288) và độ đặc hiệu (0.9828). Điều này cho thấy mô hình có khả năng phân biệt tốt giữa AFIB và các loại khác, mặc dù F1-score còn thấp hơn so với các loại khác do sự ảnh hưởng của dữ liệu mất cân bằng.
- ❖ Loại SB: Hiệu năng mô hình cao nhất ở tất cả các tham số, với độ chính xác lên đến 0.9915 và F1-score 0.9923. Độ nhạy (0.9885) và độ đặc hiệu (0.9911) đều rất cao, chứng tỏ mô hình nhận diện SB rất ổn định và đáng tin cậy.
- ❖ Loại SR: Kết quả cho loại SR cũng rất tốt với độ chính xác 0.9864 và F1-score 0.9685. Độ nhạy và độ đặc hiệu lần lượt là 0.9675 và 0.9911, cho thấy mô hình có khả năng nhận diện SR hiệu quả, với tỷ lệ dương tính giả thấp.
- ❖ Loại GSVT: Hiệu năng của mô hình đối với GSVT cũng đáng khen ngợi với độ chính xác 0.9756 và F1-score 0.9416. Độ nhạy (0.9436) và độ đặc hiệu (0.9850) đảm bảo rằng mô hình có thể phát hiện chính xác các trường hợp GSVT, dù kết quả F1-score thấp hơn một chút so với SB và SR.

Hình 3 cho thấy ma trận nhầm lẫn đạt được của mô hình học máy phân loại 4 loại rối loạn nhịp tim với thuật toán RF. Các giá trị trên đường chéo chính (từ trên trái xuống dưới phải) của hình biểu thị số lượng mẫu được phân loại đúng cho mỗi lớp. Giá trị cao trên đường chéo chính (411, 772, 431, 435) cho thấy mô hình phân loại phần lớn đúng các nhãn của dữ liệu. Sự nhầm lẫn giữa AFIB và GSVT (20 và 24 mẫu) có thể là do các đặc trưng tín hiệu sinh lý tương tự trong dữ liệu. Kết quả này cho thấy còn có thể cải thiện khả năng phân biệt hai loại này thông qua việc sử dụng thêm các đặc trưng hoặc tinh chỉnh mô hình. Trong khi đó, vẫn có sự nhầm lẫn nhỏ giữa SB và SR (9 mẫu), mặc dù ít, nhưng việc cải thiện vẫn có thể tăng độ chính xác tổng thể. Mô hình cho thấy hiệu năng vượt trội, đặc biệt là với loại SB và SR. Tuy nhiên, các chỉ số F1-score của AFIB và GSVT có thể được cải thiện hơn nữa bằng cách cân bằng dữ liệu hoặc tinh chỉnh mô hình để xử lý tốt hơn các loại có số lượng dữ liệu ít hơn.



Hình 3. Ma trận nhầm lẫn.

IV. KẾT LUẬN

Bài báo đã đề xuất và xây dựng mô hình học máy hiệu quả nhằm phân loại rối loạn nhịp tim dựa trên dữ liệu ECG, sử dụng tập dữ liệu Chapman ECG 12 chuyển đạo. Mười thuật toán học máy điển hình bao gồm Ada, Bagging, DT, GB, KNN, LR, NB, RF, SVM và XGB được khảo sát và đánh giá hiệu năng, trên cơ sở đó, thuật toán RF được lựa chọn nhờ vào hiệu năng vượt trội, đạt độ chính xác tổng thể 96.20%. Mô hình phân loại rối loạn nhịp tim cũng cho thấy hiệu quả phân loại rất cao đối với từng loại rối loạn thành phần. Các kết quả thực nghiệm không chỉ chứng minh khả năng ứng dụng hiệu quả của các phương pháp học máy trong chẩn đoán rối loạn nhịp tim mà còn mở ra triển vọng lớn trong việc phát triển các hệ thống chăm sóc sức khỏe từ xa.

TÀI LIỆU THAM KHẢO

- [1] World Health Organization, "Cardiovascular Diseases," 2024.
- [2] Future market insights, "AI in IoT Market Outlook (2023 to 2033)."
- [3] M. S. Alauddin, A. S. Baharuddin, and M. I. M. Ghazali, "The modern and digital transformation of oral health care: A mini review," 2021. doi: 10.3390/healthcare9020118.
- [4] E. al. Abhijeet Shelke, "Role of Artificial Intelligence in Cardiovascular Risk Prediction and Prevention," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 8, 2023, doi: 10.17762/ijritcc.v11i8.9437.
- [5] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci Rep*, vol. 10, no. 1, p. 16057, 2020, doi: 10.1038/s41598-020-72685-1.
- [6] J. Liu *et al.*, "A review of arrhythmia detection based on electrocardiogram with artificial intelligence," 2022. doi: 10.1080/17434440.2022.2115887.
- [7] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and Challenges of Deep Learning Methods for Electrocardiogram Data: A Systematic Review," Dec. 2019, [Online]. Available: <http://arxiv.org/abs/2001.01550>

- [8] M. Merone, P. Soda, M. Sansone, and C. Sansone, "ECG databases for biometric systems: A systematic review," 2017. doi: 10.1016/j.eswa.2016.09.030.
- [9] E. J. da S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Comput Methods Programs Biomed*, vol. 127, pp. 144–164, Apr. 2016, doi: 10.1016/j.cmpb.2015.12.008.
- [10] S. K. Pandey, V. R. Sodum, R. R. Janghel, and A. Raj, "ECG Arrhythmia Detection with Machine Learning Algorithms," in *Advances in Intelligent Systems and Computing*, 2020. doi: 10.1007/978-981-15-1097-7_34.
- [11] S. Sahoo, M. Dash, S. Behera, and S. Sabut, "Machine Learning Approach to Detect Cardiac Arrhythmias in ECG Signals: A Survey," 2020. doi: 10.1016/j.irbm.2019.12.001.
- [12] G. H. Tison, J. Zhang, F. N. Delling, and R. C. Deo, "Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery," *Circ Cardiovasc Qual Outcomes*, vol. 12, no. 9, 2019, doi: 10.1161/CIRCOUTCOMES.118.005289.
- [13] G. S. Patel, S. Nayak, and S. K. Chaudhary, *Machine Learning, Deep Learning, Big Data, and Internet of Things for Healthcare*. 2022. doi: 10.1201/9781003227595.
- [14] P. Manickam *et al.*, "Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare," 2022. doi: 10.3390/bios12080562.
- [15] V. Gupta, M. Mittal, V. Mittal, and N. K. Saxena, "A Critical Review of Feature Extraction Techniques for ECG Signal Analysis," Oct. 01, 2021, *Springer*. doi: 10.1007/s40031-021-00606-5.
- [16] J. Li, S. peng Pang, F. Xu, P. Ji, S. Zhou, and M. Shu, "Two-dimensional ECG-based cardiac arrhythmia classification using DSE-ResNet," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-18664-0.
- [17] J. S. Taggar, T. Coleman, S. Lewis, C. Heneghan, and M. Jones, "Accuracy of methods for diagnosing atrial fibrillation using 12-lead ECG: A systematic review and meta-analysis," *Int J Cardiol*, vol. 184, no. 1, 2015, doi: 10.1016/j.ijcard.2015.02.014.
- [18] J. M. Kwon *et al.*, "Artificial Intelligence-Enhanced Smartwatch ECG for Heart Failure-Reduced Ejection Fraction Detection by Generating 12-Lead ECG," *Diagnostics*, vol. 12, no. 3, 2022, doi: 10.3390/diagnostics12030654.
- [19] J. Lai *et al.*, "Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset," *Nat Commun*, vol. 14, no. 1, Dec. 2023, doi: 10.1038/s41467-023-39472-8.
- [20] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network," *Information Fusion*, vol. 53, pp. 174–182, Jan. 2020, doi: 10.1016/j.inffus.2019.06.024.
- [21] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-15432-4.
- [22] J. Zheng *et al.*, "Optimal Multi-Stage Arrhythmia Classification Approach," *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-59821-7.

MACHINE LEARNING-BASED ARRHYTHMIA CLASSIFICATION USING ECG DATASET

Abstract: Cardiovascular disease is currently one of the leading causes of death worldwide, so early detection and timely intervention of cardiovascular diseases have become urgent issues. In this context, electrocardiogram (ECG) monitoring is considered an important tool for assessing heart rate and preventing cardiovascular disease (CVD). With the support of machine learning methods, the accuracy in diagnosing arrhythmias has been significantly improved. This paper introduces an effective machine learning model for classifying arrhythmias based on electrocardiogram data. The study investigates and compares typical machine learning algorithms, optimizing parameter tuning on HRV datasets extracted from the 12-lead Chapman ECG dataset. The most optimal method is selected based on the model performance through evaluation indexes such as accuracy, F1 score, sensitivity, recall, and specificity. Experimental results show that the proposed model achieves an accuracy of over 86% in classifying different types of arrhythmias. The results demonstrate the feasibility of applying advanced machine learning algorithms in cardiovascular diagnosis, while opening up potential applications in remote healthcare systems and personal health monitoring.

Keywords: Cardiovascular disease, Electrocardiogram, machine learning, arrhythmia classification, Random Forest.



Nguyễn Thị Thu Hiền, nhận bằng tốt nghiệp đại học Khoa học Tự nhiên năm 1996. Năm 2004 nhận bằng thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông, ngành Tin học- Điện tử- Viễn thông. Hiện đang giảng dạy tại Trung tâm Thí nghiệm thực hành, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Học máy, Xử lý tín hiệu, Kỹ thuật Điện tử.