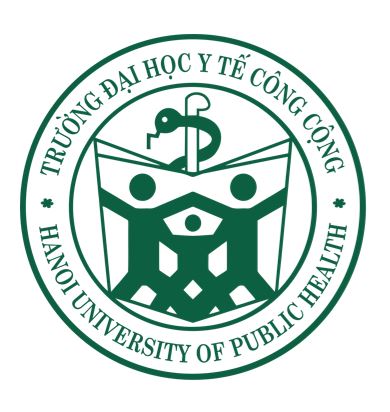
**TRƯỜNG ĐẠI HỌC Y TẾ CÔNG CỘNG**

**KHOA CÁC KHOA HỌC CƠ BẢN**



**BÀI TẬP HẾT MÔN**

**HỌC SÂU**

**Lớp CNCQKHDL1-1A**

**NHÓM SỐ: 1**

**Chủ đề:**

**XÂY DỰNG MÔ HÌNH HỖ TRỢ CHẨN ĐOÁN BỆNH TIM**

**Thành viên:**

**Trần Đức Hải - 2211090009**

**Nguyễn Thảo Trang - 2211090035**

**Nguyễn Minh Huyền - 2211090019**

**Hà Nội, 2025**

**MỤC LỤC**

[1. Phân Tích và Xác Định Vấn Đề 3](#_Toc211948643)

[1.1. Giới thiệu bối cảnh 3](#_Toc211948644)

[1.2. Mục tiêu nghiên cứu 3](#_Toc211948645)

[1.3. Nguồn dữ liệu và đặc trưng 3](#_Toc211948646)

[1.4. Đặc điểm bài toán 3](#_Toc211948647)

[1.5. Thách thức và định hướng 4](#_Toc211948648)

[2. Xử Lý và Khám Phá Dữ Liệu 4](#_Toc211948649)

[2.1. Tổng quan dữ liệu 4](#_Toc211948650)

[2.2. Thống kê mô tả 4](#_Toc211948651)

[2.3. Phân tích tương quan 4](#_Toc211948652)

[2.4. Phân tích trực quan và nhóm đặc trưng 5](#_Toc211948653)

[3. Phân Tích Mô Hình 5](#_Toc211948654)

[3.1. Tiền xử lý dữ liệu 5](#_Toc211948655)

[3.2. Phân tích và lựa chọn các mô hình Machine Learning 5](#_Toc211948656)

[3.3. Phân tích và lựa chọn các mô hình Deep Learning 7](#_Toc211948657)

[4. Cài Đặt Thử Nghiệm 8](#_Toc211948658)

[4.1. Cài đặt môi trường 8](#_Toc211948659)

[4.2. Quy trình tiền xử lý dữ liệu 8](#_Toc211948660)

[4.3. Huấn luyện mô hình 8](#_Toc211948661)

[4.3.1. Thực nghiệm mô hình Machine Learning 8](#_Toc211948662)

[4.3.2. Thực nghiệm mô hình Deep Learning (MLP) 9](#_Toc211948663)

[5. Truyền Đạt và Đưa Ra Kiến Nghị 9](#_Toc211948664)

[5.1. Trình bày và diễn giải kết quả 9](#_Toc211948665)

[5.2. Kiến nghị ứng dụng 10](#_Toc211948666)

[6. Kết Luận 11](#_Toc211948667)

# 1. Phân Tích và Xác Định Vấn Đề

## 1.1. Giới thiệu bối cảnh

Bệnh tim mạch là một trong những nguyên nhân gây tử vong hàng đầu trên toàn cầu, chiếm tỷ lệ đáng kể trong gánh nặng y tế của nhân loại. Việc phát hiện sớm các dấu hiệu và yếu tố nguy cơ đóng vai trò quan trọng trong việc giảm tỷ lệ tử vong cũng như cải thiện chất lượng cuộc sống cho bệnh nhân. Trong bối cảnh đó, ứng dụng khoa học dữ liệu và học máy (Machine Learning) trở thành công cụ hữu ích giúp bác sĩ dự đoán nguy cơ mắc bệnh tim dựa trên dữ liệu lâm sàng.

## 1.2. Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là xây dựng mô hình học máy có khả năng dự đoán khả năng mắc bệnh tim dựa trên các đặc trưng y khoa thu thập từ bệnh nhân. Bài toán được mô hình hóa dưới dạng phân loại nhị phân (binary classification):

* 0: Không mắc bệnh tim.
* 1: Có mắc bệnh tim.

Kết quả của mô hình có thể hỗ trợ các chuyên gia y tế trong việc:

* Sàng lọc và phân loại bệnh nhân theo mức độ rủi ro.
* Tối ưu hóa quy trình khám sàng lọc và chỉ định xét nghiệm.
* Hỗ trợ ra quyết định lâm sàng dựa trên dữ liệu thực chứng.

## 1.3. Nguồn dữ liệu và đặc trưng

Bộ dữ liệu được sử dụng là Cleveland Heart Disease Dataset từ UCI Machine Learning Repository, được tải về thông qua Kaggle. Bộ dữ liệu gồm 303 bản ghi và 14 thuộc tính, trong đó biến target là nhãn kết quả.

Các đặc trưng bao gồm: tuổi (age), giới tính (sex), loại đau ngực (cp), huyết áp đo được (trestbps), cholesterol (chol), đường huyết lúc đói (fbs), kết quả ECG lúc nghỉ (restecg), nhịp tim tối đa (thalach), đau thắt ngực do vận động (exang), độ trũng ST (oldpeak), độ dốc ST (slope), số mạch máu chính được nhuộm bằng fluoroscopy (ca), và chỉ số thalassemia (thal).

## 1.4. Đặc điểm bài toán

* Loại bài toán: Phân loại có giám sát (Supervised Learning).
* Đầu vào: Các đặc trưng định lượng và định tính của bệnh nhân.
* Đầu ra: Biến nhị phân target.
* Mục tiêu mô hình: Dự đoán chính xác tình trạng bệnh, đồng thời đảm bảo cân bằng giữa độ nhạy (Sensitivity) và độ đặc hiệu (Specificity).

## 1.5. Thách thức và định hướng

Bài toán gặp phải một số thách thức đặc trưng của dữ liệu y học:

* Phân bố không cân bằng nhẹ giữa hai lớp bệnh/không bệnh.
* Một số biến có mối tương quan phức tạp hoặc phi tuyến, ảnh hưởng đến hiệu quả mô hình tuyến tính.
* Cần cân bằng giữa độ chính xác và khả năng diễn giải, vì các quyết định trong y học đòi hỏi tính minh bạch.

Từ đó, nghiên cứu hướng đến việc thử nghiệm nhiều mô hình học máy khác nhau để xác định mô hình phù hợp nhất cho mục tiêu dự đoán lâm sàng.

# 2. Xử Lý và Khám Phá Dữ Liệu

## 2.1. Tổng quan dữ liệu

Dữ liệu được tải trực tiếp từ Kaggle (heart.csv) và kiểm tra bằng pandas để đảm bảo tính toàn vẹn. Bộ dữ liệu bao gồm 303 hàng và 14 cột, trong đó:

* Biến mục tiêu: target (0 hoặc 1).
* Các biến đầu vào: kết hợp giữa định lượng và định tính, đã được mã hóa sẵn bằng số, do đó không cần xử lý chuỗi hoặc mã hóa thêm.

Kết quả kiểm tra bằng df.info() cho thấy không có giá trị khuyết (missing values), giúp đơn giản hóa giai đoạn tiền xử lý.

## 2.2. Thống kê mô tả

Kết quả từ df.describe() cho thấy:

* Tuổi trung bình là 54.36, dao động từ 29 đến 77.
* 68% là nam giới, phản ánh thực tế bệnh tim thường phổ biến ở nam.
* Huyết áp trung bình khoảng 131 mmHg, cholesterol trung bình 246 mg/dl.
* Khoảng 15% bệnh nhân có đường huyết cao (>120 mg/dl).
* 54% bệnh nhân trong tập dữ liệu được chẩn đoán có bệnh tim (target=1).

Nhìn chung, dữ liệu có phân bố tương đối cân bằng, phù hợp cho các mô hình phân loại không yêu cầu cân bằng lại lớp.

## 2.3. Phân tích tương quan

Ma trận tương quan (df.corr()) và heatmap trực quan hóa mối quan hệ giữa các biến cho thấy:

* Các biến có tương quan dương mạnh với target gồm: cp (0.43), thalach (0.42), slope (0.35) – biểu thị rằng người có loại đau ngực điển hình, nhịp tim cao và độ dốc ST thuận lợi có khả năng ít mắc bệnh hơn.
* Các biến có tương quan âm mạnh với target gồm: exang (-0.44), oldpeak (-0.43), ca (-0.39), thal (-0.34) – thể hiện rằng các đặc trưng này gia tăng nguy cơ bệnh tim.
* Các biến như fbs và chol có tương quan yếu, cho thấy ảnh hưởng hạn chế đến kết quả mô hình.

Không có cặp biến nào có tương quan >0.8, do đó không xuất hiện đa cộng tuyến nghiêm trọng (multicollinearity).

## 2.4. Phân tích trực quan và nhóm đặc trưng

Một số bước khám phá dữ liệu được thực hiện qua biểu đồ:

* Biểu đồ phân bố độ tuổi: cho thấy đa số bệnh nhân thuộc nhóm trung niên (40–55 tuổi).
* Biểu đồ giới tính: tỷ lệ nam giới cao hơn rõ rệt, phù hợp với dịch tễ học bệnh tim.
* Biểu đồ phân bố theo nhóm tuổi: ba nhóm (trẻ, trung niên, cao tuổi) có tỷ lệ mắc bệnh khác nhau, với nhóm trung niên và cao tuổi chiếm ưu thế.

Những trực quan hóa này giúp hiểu sâu hơn về cấu trúc dữ liệu và yếu tố nguy cơ, đồng thời hỗ trợ chọn đặc trưng quan trọng cho giai đoạn mô hình hóa.

# 3. Phân Tích Mô Hình

## 3.1. Tiền xử lý dữ liệu

Dữ liệu được phân chia thành 2 bộ dữ liệu riêng biệt cho việc đánh giá hiệu quả của mô hình:

* Tập huấn luyện (Training set): Dạy cho mô hình về các mối quan hệ giữa các dữ liệu với tỷ lệ 80% bộ dữ liệu gốc.
* Tập kiểm tra (Testing set): Đánh giá độ chính xác dự đoán của mô hình trên những dữ liệu mà nó chưa từng thấy trước đây, sử dụng 20% còn lại của bộ dữ liệu gốc.

## 3.2. Phân tích và lựa chọn các mô hình Machine Learning

Để giải quyết bài toán phân loại nhị phân, một loạt các thuật toán học máy cổ điển đại diện cho các phương pháp tiếp cận khác nhau đã được lựa chọn. Việc lựa chọn đa dạng mô hình mang đến một góc nhìn toàn diện và so sánh được ưu, nhược điểm của từng phương pháp trên cùng một bộ dữ liệu.

Các mô hình được lựa chọn bao gồm:

* Logistic Regression (Hồi quy Logistic):
  + Phân tích: Đây là một mô hình tuyến tính, tuy đơn giản nhưng rất hiệu quả và thường được sử dụng làm mô hình cơ sở trong các bài toán phân loại. Nó hoạt động bằng cách ước tính xác suất một mẫu thuộc về một lớp cụ thông qua hàm sigmoid. Ưu điểm của mô hình là dễ diễn giải và tính toán nhanh.
  + Lý do lựa chọn: Phù hợp để thiết lập một ngưỡng hiệu suất ban đầu để so sánh với các mô hình phức tạp hơn.
* K-Nearest Neighbors (KNN - K Láng giềng gần nhất):
  + Phân tích: Thuật toán phân loại một điểm dữ liệu mới bằng cách dựa vào số lượng của K điểm dữ liệu gần nhất trong không gian đặc trưng. Đây là một thuật toán trực quan và dễ triển khai. Tuy nhiên, hiệu suất của nó phụ thuộc nhiều vào việc lựa chọn tham số K và độ đo khoảng cách.
  + Lý do lựa chọn: Để đánh giá một phương pháp tiếp cận phi tuyến, dựa trên sự tương đồng thay vì tìm một hàm phân tách rõ ràng.
* Support Vector Machine (SVM - Máy Véc-tơ Hỗ trợ):
  + Phân tích: SVM là một thuật toán mạnh mẽ, hoạt động bằng cách tìm ra một siêu phẳng trong không gian nhiều chiều để phân tách các lớp dữ liệu một cách tối ưu nhất. Mục tiêu là tối đa hóa khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất của mỗi lớp. SVM đặc biệt hiệu quả trong không gian có số chiều cao.
  + Lý do lựa chọn: Để thử nghiệm một mô hình dựa trên việc tối ưu hóa biên phân loại, có khả năng xử lý tốt các mối quan hệ phức tạp.
* Naive Bayes:
  + Phân tích: Là một thuật toán phân loại dựa trên xác suất, áp dụng định lý Bayes với một giả định "ngây thơ" rằng các đặc trưng là độc lập với nhau. Mặc dù giả định này hiếm khi đúng trong thực tế, Naive Bayes vẫn hoạt động tốt một cách đáng ngạc nhiên trong nhiều bài toán, đặc biệt với các bộ dữ liệu nhỏ. Nó có ưu điểm là tốc độ huấn luyện rất nhanh.
  + Lý do lựa chọn: Để kiểm tra hiệu quả của một phương pháp tiếp cận dựa trên xác suất thống kê.
* Random Forest (Rừng Ngẫu nhiên):
  + Phân tích: Đây là một thuật toán học tập hợp thuộc loại bagging. Random Forest xây dựng một "rừng" gồm nhiều cây quyết định trong quá trình huấn luyện. Để phân loại một mẫu mới, nó lấy kết quả dự đoán từ mỗi cây và chọn ra lớp được "bỏ phiếu" nhiều nhất. Kỹ thuật này giúp cải thiện đáng kể độ chính xác và giảm thiểu hiện tượng học vẹt so với việc chỉ dùng một cây quyết định đơn lẻ.
  + Lý do lựa chọn: Để đánh giá sức mạnh của các phương pháp học tập hợp, vốn thường cho kết quả hàng đầu trong nhiều cuộc thi học máy.

## 3.3. Phân tích và lựa chọn các mô hình Deep Learning

Để khám phá khả năng nâng cao hiệu suất và tìm ra các mối quan hệ phi tuyến phức tạp hơn mà các mô hình cổ điển có thể bỏ qua, nhóm chúng em đã quyết định xây dựng và bổ sung một mô hình Deep Learning, cụ thể là mạng Multi-layer Perceptron (MLP).

* Phân tích: MLP là một mạng nơ-ron truyền thẳng (Feedforward Network), rất phù hợp cho các bài toán với dữ liệu có cấu trúc dạng bảng như bộ dữ liệu bệnh tim. Với nhiều lớp ẩn, nó có khả năng học các đặc trưng ở nhiều mức độ trừu tượng khác nhau.
* Lý do lựa chọn:
  + Khả năng học các mẫu phức tạp: MLP có thể mô hình hóa các mối quan hệ phi tuyến tính phức tạp giữa các chỉ số y tế và nguy cơ mắc bệnh tim.
  + Kiến trúc linh hoạt:
    - Sử dụng hàm kích hoạt ReLU ở các lớp ẩn để giúp mô hình học nhanh hơn và tránh được vấn đề triệt tiêu đạo hàm (Vanishing Gradients).
    - Sử dụng hàm kích hoạt Sigmoid ở lớp đầu ra để đưa ra kết quả dưới dạng xác suất (từ 0 đến 1), rất phù hợp cho bài toán phân loại nhị phân.
  + Tích hợp các kỹ thuật hiện đại:
    - Sử dụng thuật toán tối ưu Adam, một lựa chọn phổ biến và mạnh mẽ, giúp mô hình hội tụ nhanh và ổn định.
    - Áp dụng kỹ thuật Dropout như một phương pháp điều chuẩn (Regularization) hiệu quả để chống lại hiện tượng học vẹt (Overfitting), đặc biệt hữu ích khi làm việc với các bộ dữ liệu có kích thước không quá lớn.

# 4. Cài Đặt Thử Nghiệm

Mã nguồn code được đẩy lên github tại link sau:

## 4.1. Cài đặt môi trường

* Ngôn ngữ: Python 3.
* Thư viện: TensorFlow, Keras, Scikit-learn, Pandas, NumPy, Matplotlib.

## 4.2. Quy trình tiền xử lý dữ liệu

* Phân chia dữ liệu: Chia bộ dữ liệu thành 2 tập:
  + Tập huấn luyện (Training set): 80%
  + Tập kiểm thử (Test set): 20%
* Chuẩn hóa dữ liệu (Data Scaling): Áp dụng StandardScaler từ Scikit-learn cho các thuộc tính số. Việc này giúp đưa tất cả các thuộc tính về cùng một thang đo, giúp quá trình huấn luyện mô hình hội tụ nhanh và ổn định hơn.

## 4.3. Huấn luyện mô hình

### 4.3.1. Thực nghiệm mô hình Machine Learning

Kết quả đánh giá độ chính xác trên tập kiểm tra của các mô hình Machine Learning cơ sở như sau:

|  |  |  |
| --- | --- | --- |
| STT | Mô hình | Độ chính xác trên tập Test |
| 1 | Logistic Regression | 85.25% |
| 2 | K-Nearest Neighbors (KNN) | 63.93% |
| 3 | Support Vector Machine (SVM) | 68.85% |
| 4 | Naive Bayes | 85.25% |
| 5 | Random Forest | 85.25% |

Phân tích: Các mô hình Logistic Regression, Naive Bayes, và Random Forest cho kết quả tốt nhất và đồng đều. Điều này cho thấy các phương pháp khác nhau có thể đạt được hiệu suất tương đương trên bộ dữ liệu này. Hiệu suất thấp của KNN và SVM có thể do sự nhạy cảm của chúng với việc co giãn các dữ liệu đặc trưng.

### 4.3.2. Thực nghiệm mô hình Deep Learning (MLP)

Mô hình MLP được xây dựng và huấn luyện bằng thư viện PyTorch.

* Cài đặt thực nghiệm:
  + Kiến trúc: Mạng nơ-ron truyền thẳng gồm 1 lớp đầu vào (13 nơ-ron), 2 lớp ẩn (64 và 32 nơ-ron với hàm kích hoạt ReLU), và 1 lớp đầu ra (1 nơ-ron).
  + Hàm mất mát: Binary Cross-Entropy (BCEWithLogitsLoss).
  + Thuật toán tối ưu: Adam với learning rate là 0.001.
  + Kỹ thuật chống quá khớp: Áp dụng Dropout với tỷ lệ 0.3.
  + Huấn luyện: Mô hình được huấn luyện trong 50 epochs với kích thước batch là 32.
* Kết quả từ mô hình Deep Learning: Sau khi huấn luyện, mô hình MLP đã cho thấy sự vượt trội so với các mô hình Machine Learning, đạt độ chính xác trên tập kiểm tra là 88.52%.

# 5. Truyền Đạt và Đưa Ra Kiến Nghị

## 5.1. Trình bày và diễn giải kết quả

Kết quả thực nghiệm cho thấy các mô hình học máy truyền thống như Logistic Regression, Naive Bayes và Random Forest đều đạt độ chính xác cao (~85.25%), trong khi mô hình Deep Learning (MLP) thể hiện hiệu suất vượt trội hơn, đạt 88.52% trên tập kiểm tra.

Sự cải thiện này phản ánh khả năng của mạng nơ-ron trong việc học các mối quan hệ phi tuyến tính phức tạp giữa các chỉ số y sinh học và tình trạng bệnh tim — điều mà các mô hình tuyến tính hoặc dựa trên giả định độc lập thường khó nắm bắt.

Khi so sánh về khả năng tổng quát hóa:

* Logistic Regression: Đơn giản, dễ diễn giải, phù hợp cho mục đích chẩn đoán lâm sàng đòi hỏi minh bạch.
* Random Forest: Cân bằng giữa độ chính xác và khả năng chống quá khớp (overfitting), thích hợp cho các hệ thống hỗ trợ quyết định thực tế.
* Deep Learning (MLP): Có hiệu suất tốt nhất, đặc biệt trong việc phát hiện các mẫu dữ liệu phi tuyến, tuy nhiên khó diễn giải và yêu cầu tài nguyên tính toán cao hơn.

Nhìn chung, các mô hình đều cho thấy khả năng dự đoán đáng tin cậy, mở ra tiềm năng ứng dụng trong hệ thống hỗ trợ chẩn đoán sớm bệnh tim.

## 5.2. Kiến nghị ứng dụng

Dựa trên kết quả mô hình, nhóm đề xuất các hướng ứng dụng và cải thiện như sau:

* **Ứng dụng thực tiễn trong y tế**

Mô hình có thể được tích hợp vào các hệ thống hỗ trợ quyết định lâm sàng (Clinical Decision Support System – CDSS), nhằm:

* + Hỗ trợ bác sĩ đánh giá nhanh nguy cơ mắc bệnh tim dựa trên dữ liệu đầu vào của bệnh nhân.
  + Gợi ý mức độ ưu tiên xét nghiệm hoặc khám chuyên sâu cho những người có nguy cơ cao.
  + Nâng cao khả năng sàng lọc chủ động và dự phòng sớm trong cộng đồng.
* **Mở rộng dữ liệu và chuẩn hóa mô hình**
  + Thu thập thêm dữ liệu từ nhiều nguồn bệnh viện khác nhau để tăng tính khái quát.
  + Chuẩn hóa quy trình nhập liệu và xử lý dữ liệu y học nhằm giảm sai lệch.
  + Ứng dụng kỹ thuật cân bằng lớp (SMOTE, ADASYN) để cải thiện độ nhạy với các trường hợp mắc bệnh hiếm.
* **Cải tiến kỹ thuật mô hình**
  + Kết hợp các kỹ thuật Feature Selection hoặc SHAP/LIME để tăng khả năng diễn giải và nhận diện biến quan trọng.
  + Tối ưu siêu tham số (Hyperparameter Tuning) bằng Grid Search hoặc Bayesian Optimization.
  + Xây dựng mô hình lai (Hybrid Model) giữa Random Forest và MLP để tận dụng cả tính ổn định và độ chính xác.
* **Triển khai và đánh giá thực tế**
  + Cần thử nghiệm mô hình trên dữ liệu mới (external validation) để kiểm chứng tính ổn định.
  + Đánh giá thêm bằng các chỉ số y học như Precision, Recall, F1-score, ROC–AUC, thay vì chỉ dựa trên Accuracy.
  + Kết hợp ý kiến chuyên gia y tế trong việc điều chỉnh ngưỡng phân loại phù hợp với mục tiêu lâm sàng (ưu tiên phát hiện đúng ca bệnh thật sự).

# 6. Kết Luận

Nghiên cứu này đã thực hiện phân tích dữ liệu và xây dựng mô hình dự đoán bệnh tim dựa trên bộ dữ liệu Cleveland Heart Disease, theo quy trình khoa học dữ liệu chuẩn gồm bốn giai đoạn:(1) xác định vấn đề, (2) xử lý và khám phá dữ liệu, (3) huấn luyện và đánh giá mô hình, (4) truyền đạt kết quả và kiến nghị ứng dụng.

Kết quả cho thấy mô hình Multi-layer Perceptron (MLP) đạt độ chính xác cao nhất (88.52%), vượt trội so với các mô hình học máy cổ điển. Điều này chứng minh tiềm năng của Deep Learning trong việc khai thác các mối quan hệ phi tuyến tính giữa các biến y học.

Tuy nhiên, nghiên cứu cũng chỉ ra rằng những mô hình dễ diễn giải như Logistic Regression và Random Forest vẫn giữ vai trò quan trọng trong thực hành lâm sàng, nơi tính minh bạch và khả năng giải thích kết quả là yêu cầu bắt buộc.

Từ kết quả đạt được, có thể rút ra các kết luận chính:

* Các yếu tố cp (loại đau ngực), thalach (nhịp tim tối đa), oldpeak (độ trũng ST) và exang (đau thắt ngực do vận động) là những chỉ số có ảnh hưởng mạnh đến nguy cơ mắc bệnh tim.
* Các mô hình học máy có thể trở thành công cụ hỗ trợ đắc lực cho bác sĩ, đặc biệt trong giai đoạn sàng lọc và theo dõi bệnh nhân.
* Việc mở rộng quy mô dữ liệu, tăng tính đa dạng mẫu và kết hợp các kỹ thuật học sâu sẽ giúp cải thiện hơn nữa hiệu suất và độ tin cậy của mô hình.

Nghiên cứu không chỉ minh chứng cho giá trị ứng dụng của khoa học dữ liệu trong y học, mà còn mở ra hướng phát triển cho các hệ thống dự đoán bệnh dựa trên dữ liệu thực — góp phần thúc đẩy chẩn đoán sớm, điều trị cá thể hóa và chăm sóc sức khỏe cộng đồng thông minh.