DATA SECURITY IN GENE MUTATION ANALYSIS: VISUALIZATION AND GAN FOR CANCER CLASSIFICATION

Nguyen Thi Thuy Quynh*, Phan Thi Hai Hong
Institute of Information and Communications Technology - Military Technical Academy

ARTICLE INFO

Received: 23/10/2024
Revised: 18/12/2024
Published: 18/12/2024

KEYWORDS

Gene mutations
Visualization
GAN
CNN
Information security

ABSTRACT

In this paper, we investigate and propose a novel deep learning-based approach for detecting gene mutations associated with several common cancer types, while ensuring information security during the analysis process. The method begins by visualizing gene mutation data as grayscale images, a crucial step in safeguarding patients sensitive information. Following this, deep learning models are employed to more effectively extract latent features from gene data, while ensuring that personal data remains protected. After visualization, Generative Adversarial Networks are applied to enhance data diversity, generating new image samples from the original gene data without compromising information security. This process not only highlights key features of gene mutations but also improves the generalization apabilities of model while maintaining patient privacy. The critical features learned by the Discriminator are used as input for a Convolutional Neural Network to classify 12 common cancer types. Experimental results demonstrate that the proposed method achieves superior performance in detecting and classifying cancer gene mutations, while ensuring that personal genetic data is safeguarded. This research not only introduces a novel deep learning approach for gene mutation analysis but also ensures information security, effectively supporting cancer diagnosis and treatment.

ĐẢM BẢO AN TOÀN THÔNG TIN TRONG PHÂN TÍCH ĐỘT BIẾN GEN: TRỰC QUAN HOÁ DỮ LIỆU VÀ ỨNG DỤNG GAN ĐỂ PHÂN LOẠI UNG THƯ

Nguyễn Thị Thuý Quỳnh*, Phan Thị Hải Hồng
Viện Công nghệ thông tin và truyền thông – Học viện Kỹ thuật quân sự

THÔNG TIN BÀI BÁO

Ngày nhận bài: 23/10/2024
Ngày hoàn thiện: 18/12/2024
Ngày đăng: 18/12/2024

TỪ KHÓA

Đột biến gen
Trực quan hoá
GAN
CNN
An toàn thông tin

TÓM TẮT

Trong bài báo này, chúng tôi nghiên cứu và đề xuất một phương pháp ứng dụng học sâu mới trong việc phát hiện đột biến gen liên quan đến một số loại ung thư phổ biến, đồng thời đảm bảo an toàn thông tin trong quá trình phân tích. Phương pháp này bắt đầu bằng việc trực quan hóa dữ liệu gen đột biến thành hình ảnh xám, một bước quan trọng giúp bảo vệ thông tin nhạy cảm của bệnh nhân. Tiếp theo, các mô hình học sâu được sử dụng để khai thác đặc trưng tiềm ẩn trong dữ liệu gen một cách hiệu quả hơn, trong khi vẫn đảm bảo rằng các thông tin cá nhân không bị lộ. Sau quá trình trực quan hóa, mạng đối nghịch tạo sinh được áp dụng để tăng cường sự đa dạng của dữ liệu, sinh ra các mẫu ảnh mới từ các hình ảnh dữ liệu gen ban đầu mà không làm mất đi tính bảo mật của thông tin gốc. Quá trình này không chỉ làm nổi bật các đặc điểm quan trọng của đột biến gen mà còn cải thiện khả năng tổng quát hóa của mô hình, trong khi bảo vệ quyền riêng tư của bệnh nhân. Các đặc trưng quan trọng được bộ phân biệt (Discriminator) học và được sử dụng làm đầu vào cho mạng nơ-ron tích chập để phân loại 12 loại ung thư phổ biến. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt hiệu suất vượt trội trong việc phát hiện và phân loại các loại đột biến gen ung thư, đồng thời đảm bảo rằng dữ liệu gen cá nhân được bảo vệ. Nghiên cứu này không chỉ mở ra một hướng tiếp cận mới trong việc ứng dụng học sâu cho phân tích gen đột biến, mà còn đảm bảo an toàn thông tin, hỗ trợ hiệu quả trong chẩn đoán và điều trị ung thư.

DOI: https://doi.org/10.34238/tnu-jst.11388

* Corresponding author. *Email: nguyenthuyquynhcvp@gmail.com*

1. Giới thiệu

Việc phát hiện và phân loại ung thư dựa trên đột biến gen là một bước quan trọng trong quá trình chẩn đoán và lựa chọn phương pháp điều trị hiệu quả. Những đột biến trong gen có thể làm thay đổi chức năng của tế bào, dẫn đến sự hình thành và tiến triển của các loại ung thư [1]. Do đó, việc xác định chính xác các đột biến gen không chỉ cung cấp thông tin cần thiết cho việc phát triển các phương pháp điều trị cá nhân hóa mà còn yêu cầu sự đảm bảo an toàn thông tin khi xử lý và phân tích dữ liệu gen nhạy cảm. Dữ liệu đột biến gen có thể chứa thông tin cá nhân quan trọng, và việc bảo vệ dữ liệu này trong suốt quá trình phân tích là điều bắt buộc. Phân tích các đột biến gen là một thách thức lớn do sự phức tạp của dữ liệu và sự đa dạng của các loại đột biến [2].

Học sâu (Deep Learning) đã trở thành một công cụ mạnh mẽ trong lĩnh vực phân tích dữ liệu sinh học, đặc biệt trong việc phát hiện các mẫu phức tạp trong dữ liệu gen. Tuy nhiên, việc áp dụng học sâu trực tiếp trên dữ liệu gen thô gặp phải nhiều hạn chế do bản chất không đồng nhất và kích thước lớn của dữ liệu [3]. Các phương pháp truyền thống dựa vào học sâu thường yêu cầu tiền xử lý dữ liệu phức tạp và phụ thuộc nhiều vào việc thiết kế đặc trưng thủ công, điều này không chỉ ảnh hưởng đến hiệu suất của mô hình mà còn có thể tạo ra các lỗ hổng bảo mật, dễ dẫn đến rò rỉ thông tin cá nhân nếu không có các biện pháp bảo mật thích hợp [4].

Trong những năm gần đây, nhiều nghiên cứu đã tập trung vào việc áp dụng các phương pháp học sâu để phân tích và phân loại các đột biến gen liên quan đến ung thư. Các phương pháp này giúp nâng cao khả năng phát hiện và phân loại các đột biến gen, từ đó hỗ trợ hiệu quả cho việc chẩn đoán và điều trị ung thư.

Nghiên cứu [5] đã đề xuất một mô hình học sâu để phân loại các đột biến gen. Mô hình này sử dụng các đặc trưng từ dữ liệu gen và áp dụng mạng nơ-ron sâu (DNN) để phân loại chính xác các đột biến liên quan đến ung thư. Phương pháp này đạt được những kết quả ấn tượng, đặc biệt trong việc nhận diện các loại ung thư khác nhau từ các đột biến gen, góp phần quan trọng trong lĩnh vực y sinh học.

Tiếp tục phát triển trong hướng nghiên cứu này, Sanjaya và cộng sự [6] đã giới thiệu cơ chế chú ý (attention mechanism) trong mạng học sâu để học đại diện của các đột biến gen. Mô hình MuAt không chỉ dựa vào việc phân loại mà còn cho phép phân loại phụ ung thư (tumor subtyping) dựa trên các mẫu đột biến. Kết quả từ nghiên cứu này đã chứng minh rằng việc kết hợp giữa học sâu và cơ chế chú ý giúp cải thiện đáng kể hiệu suất phân loại và dự đoán.

Nhóm tác giả của nghiên cứu [7] đã đề xuất việc áp dụng học sâu giám sát yếu để phát hiện chính xác các đột biến gen. Trong nghiên cứu này, tác giả đã sử dụng các mô hình học sâu không hoàn toàn dựa trên dữ liệu được gán nhãn rõ ràng mà còn khai thác các dữ liệu không gán nhãn để cải thiện hiệu suất phát hiện đột biến. Phương pháp giám sát yếu đã tỏ ra hiệu quả trong việc giảm sự phụ thuộc vào dữ liệu có gán nhãn và vẫn đạt được kết quả phân loại cao.

Kết quả trong nghiên cứu [8] đã mở rộng việc ứng dụng học sâu trong phân loại nhiều loại ung thư khác nhau thông qua việc phân tích dữ liệu gen. Nghiên cứu này đã xây dựng một mô hình học sâu có khả năng phân loại 12 loại ung thư dựa trên các mẫu gen đặc trưng của mỗi loại. Kết quả nghiên cứu đã khẳng định tầm quan trọng của học sâu trong việc khai thác các đặc trưng phức tạp của dữ liệu gen để phục vụ cho chẩn đoán ung thư. Mặc dù những phương pháp này đã đạt được nhiều kết quả đáng khích lệ, nhưng chúng chủ yếu tập trung vào việc phân tích trực tiếp dữ liệu gen mà chưa xem xét đến các phương pháp trực quan hóa dữ liệu kết hợp với học sâu xử lý ảnh. Để cải thiện hiệu suất phân loại và khai thác toàn diện các thông tin từ dữ liệu gen, nghiên cứu của chúng tôi đề xuất một phương pháp mới kết hợp giữa trực quan hóa dữ liệu đột biến gen và sử dụng các mô hình học sâu xử lý ảnh, nhằm tối ưu hóa khả năng phát hiện và phân loại đột biến gen các bệnh ung thư phổ biến.

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp mới kết hợp trực quan hóa dữ liệu gen với học sâu, đồng thời đảm bảo an toàn thông tin của người bệnh. Phương pháp này bắt đầu bằng việc chuyển đổi dữ liệu gen đột biến thành các hình ảnh xám, giúp bảo vệ thông tin nhạy

cảm bằng cách ẩn đi các chi tiết cá nhân trong dữ liệu gốc. Sau đó, chúng tôi sử dụng mạng đối nghịch tạo sinh (GAN) để tăng cường sự đa dạng của dữ liệu và tạo ra các mẫu ảnh mới từ các hình ảnh dữ liệu gen ban đầu mà không làm mất đi tính bảo mật của thông tin gốc [9]. GAN không chỉ giúp khai thác các đặc trưng tiềm ẩn trong dữ liệu gen một cách hiệu quả mà còn đảm bảo rằng các thông tin cá nhân của bệnh nhân được bảo vệ trong suốt quá trình tạo dữ liệu mới. Các đặc trưng quan trọng được bộ phân biệt (Discriminator) của GAN học và sau đó được sử dụng làm đầu vào cho mạng nơ-ron tích chập (CNN) để phân loại 12 loại ung thư phổ biến.

Phương pháp mới đề xuất này gồm ba bước chính: đầu tiên, các đột biến gen được chuyển đổi thành ảnh xám để bảo mật thông tin; thứ hai, GAN được sử dụng để tăng cường dữ liệu và trích xuất đặc trưng mà không làm lộ thông tin cá nhân; cuối cùng, các đặc trưng này được đưa vào CNN để phân loại các loại ung thư. Bằng cách này, phương pháp của chúng tôi không chỉ cải thiện độ chính xác trong phân loại ung thư mà còn đảm bảo an toàn thông tin, bảo vệ dữ liệu nhạy cảm của bệnh nhân trong suốt quá trình phân tích. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt hiệu suất vượt trội trong việc phát hiện và phân loại các loại đột biến gen ung thư so với các phương pháp truyền thống, đồng thời đảm bảo rằng dữ liệu gen cá nhân được bảo vệ nghiêm ngặt. Nghiên cứu này không chỉ mở ra một hướng tiếp cận mới trong việc ứng dụng học sâu cho phân tích gen đột biến mà còn đặt ra tiêu chuẩn cao về an toàn thông tin, góp phần hỗ trợ hiệu quả trong chẩn đoán và điều trị ung thư một cách bảo mật và tin cậy.

Các đóng góp chính của bài báo này bao gồm:

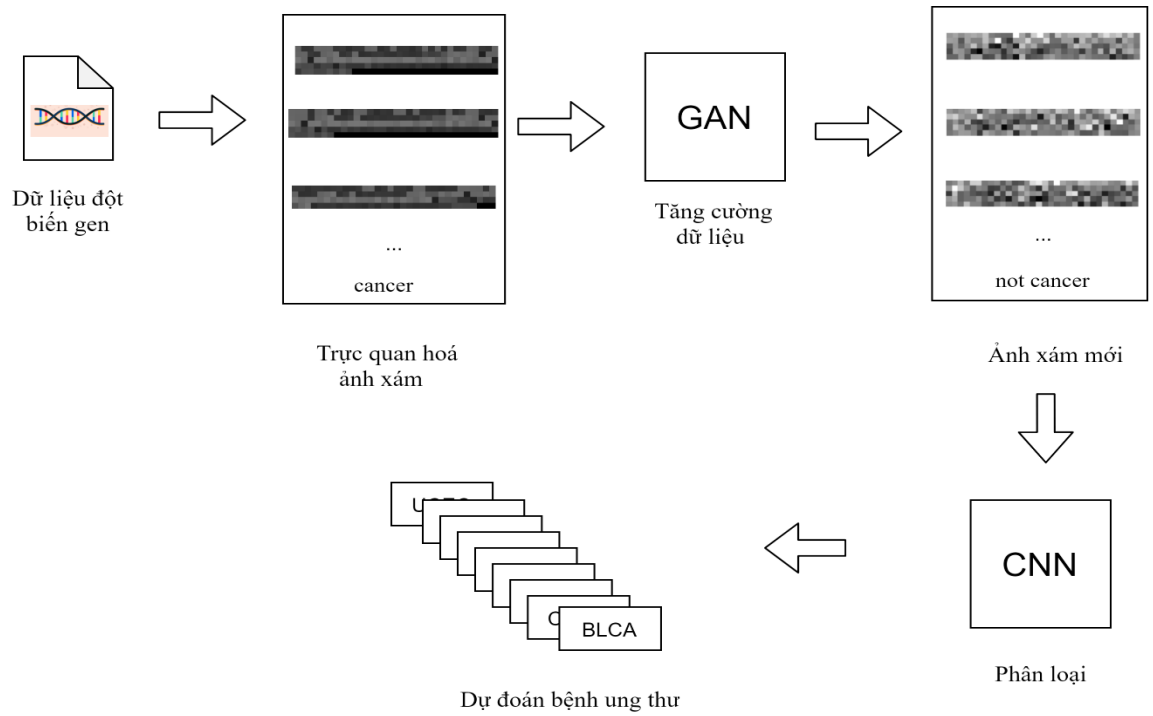
- Đề xuất một phương pháp mới để trực quan hóa dữ liệu gen đột biến thành hình ảnh, không chỉ tạo điều kiện cho việc áp dụng các mô hình học sâu trong phân tích đột biến gen, mà còn đảm bảo an toàn thông tin của bệnh nhân bằng cách ẩn đi các thông tin nhạy cảm trong dữ liệu gốc.
- Giới thiệu một kiến trúc kết hợp giữa GAN và CNN, trong đó GAN đóng vai trò trích xuất đặc trưng và tạo ra các mẫu dữ liệu mới nhằm tăng cường sự đa dạng của dữ liệu, đồng thời bảo vệ quyền riêng tư và tính bảo mật của dữ liệu gen trong suốt quá trình phân tích.
- Chúng tôi thực hiện các thí nghiệm trên bộ dữ liệu gen đột biến từ các bệnh ung thư và chứng minh rằng phương pháp đề xuất không chỉ cải thiện đáng kể độ chính xác và khả năng tổng quát hóa của mô hình so với các phương pháp truyền thống, mà còn đảm bảo an toàn thông tin, giúp bảo vệ dữ liệu nhạy cảm của bệnh nhân trong quá trình phân tích.

2. Phương pháp nghiên cứu

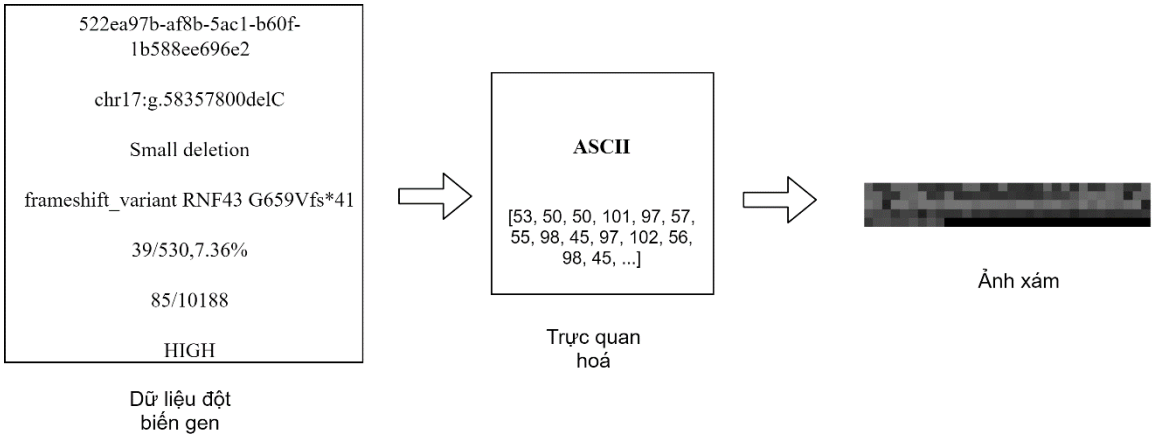
Bài báo này dựa trên khung tổng thể về dự đoán các bệnh ung thư dựa trên dữ liệu đột biến gen, trong đó kết hợp xử lý hình ảnh và CNN như minh họa trong Hình 1. Quá trình chính bao gồm hai giai đoạn: (1) Tăng cường dữ liệu thông qua chuyển đổi dữ liệu đột biến gen thành ảnh xám và áp dụng GAN; và (2) Sử dụng CNN để dự đoán đột biến gen. Ở giai đoạn đầu tiên, dữ liệu đột biến gen đã được chuyển đổi thành các hình ảnh xám để phục vụ cho quá trình xử lý hình ảnh. Sau đó, để khắc phục vấn đề thiếu sự đa dạng trong dữ liệu, mạng GAN được sử dụng để tạo ra thêm các hình ảnh mới từ ảnh xám ban đầu, tăng cường tập dữ liệu. Ở giai đoạn thứ hai, các hình ảnh đã được tăng cường bằng GAN sẽ được đưa vào mạng CNN để thực hiện nhiệm vụ dự đoán đột biến gen. Các bước cụ thể sẽ được trình bày trong các phần tiếp theo.

2.1. Tăng cường dữ liệu trực quan hoá gen

Trong nghiên cứu này, dữ liệu đột biến gen chứa các đặc trưng cụ thể của các loại ung thư được chuyển đổi thành hình ảnh xám nhằm trực quan hóa và khai thác thông tin thể hiện trong Hình 2. Quá trình này bắt đầu bằng việc mã hóa các chuỗi đột biến gen thành các giá trị số thông qua mã ASCII, trong đó mỗi ký tự trong chuỗi đột biến được chuyển đổi thành một giá trị ASCII tương ứng. Điều này giúp biến các đặc trưng của đột biến gen từ chuỗi ký tự trở thành các giá trị số học có thể trực quan hóa.



Hình 1. *Mô hình GAN - CNN*



Hình 2. *Mô hình trực quan hoá dữ liệu*

Cụ thể, mỗi chuỗi đột biến được chuyển đổi thành một danh sách các giá trị ASCII, và những giá trị này sau đó được sắp xếp thành một ma trận 2D, trong đó mỗi giá trị đại diện cho một pixel trong ảnh xám. Độ rộng của ảnh được cố định là 32 pixel, và chiều cao của ảnh được tính toán dựa trên số lượng giá trị ASCII cần chuyển đổi. Nếu số lượng giá trị không chia hết cho độ rộng, ảnh sẽ được đệm thêm bằng các giá trị pixel màu đen (giá trị 0) để đảm bảo rằng kích thước của ảnh là một ma trận hình chữ nhật.

Mỗi pixel trong ảnh sẽ có giá trị từ 0 (màu đen) đến 255 (màu trắng), tạo thành một ảnh xám 8-bit. Điều này giúp trực quan hóa các đặc trưng phức tạp của đột biến gen và tạo điều kiện thuận lợi cho các mô hình học sâu trong việc khai thác các đặc trưng này. Phương pháp chuyển đổi này đảm bảo rằng không có thông tin quan trọng nào bị mất mát trong quá trình chuyển đổi từ chuỗi dữ liệu đột biến thành ảnh.

Sau khi chuyển đổi, các hình ảnh xám này được đưa vào mạng GAN để tăng cường dữ liệu. Cấu trúc GAN bao gồm hai thành phần: bộ tạo và bộ phân biệt. Bộ tạo có nhiệm vụ tạo ra các

hình ảnh mới dựa trên các hình ảnh xám ban đầu, trong khi bộ phân biệt được huấn luyện để phân biệt xem hình ảnh là thật hay được tạo ra bởi bộ tạo. Mạng GAN trong nghiên cứu được xây dựng với bộ tạo có đầu vào là một vectơ nhiễu ngẫu nhiên có phân phối chuẩn với 100 chiều, và sử dụng các lớp chập chuyển vị (Conv2DTranspose) để tạo ra ảnh xám có kích thước 5x32. Quá trình này bao gồm một lớp kết nối đầy đủ và các lớp LeakyReLU để cải thiện khả năng sinh ảnh.

Bộ phân biệt nhận đầu vào là các ảnh kích thước 5x32x1 và sử dụng một kiến trúc phân loại nhị phân. Ảnh được chuyển thành vectơ 1 chiều thông qua lớp Flatten, sau đó được xử lý bởi hai lớp kết nối đầy đủ với 512 và 256 node, cả hai lớp đều sử dụng hàm kích hoạt LeakyReLU. Lớp đầu ra cuối cùng sử dụng hàm kích hoạt sigmoid để đưa ra quyết định liệu ảnh đó là thật hay giả.

Quá trình đối kháng giữa bộ tạo và bộ phân biệt trong quá trình huấn luyện giúp tạo ra các hình ảnh giả có độ chân thực cao hơn. Các hình ảnh mới do mạng tạo tạo ra được đưa vào quá trình huấn luyện cùng với dữ liệu thật, từ đó tăng cường sự đa dạng và phong phú của tập dữ liệu.

Sự kết hợp mạng GAN để tăng cường dữ liệu không chỉ cải thiện chất lượng tập dữ liệu đầu vào mà còn mở rộng quy mô dữ liệu, giúp tối ưu hóa khả năng của các mô hình học trong việc dự đoán và phân tích đột biến gen. Các hình ảnh sinh ra từ quá trình này sau đó được sử dụng trong các bước huấn luyện tiếp theo, đảm bảo rằng mô hình có thể khai thác đầy đủ các đặc trưng của đột biến gen trong quá trình dự đoán.

2.2. Sử dụng mạng CNN để dự đoán đột biến gen

Sau khi các hình ảnh đột biến gen được tăng cường thông qua mạng GAN, quá trình phân loại đột biến gen được thực hiện thông qua CNN.

Cấu trúc của mô hình CNN bao gồm ba thành phần chính: lớp đầu vào, các lớp tích chập và gộp, và lớp đầu ra. Lớp đầu vào tiếp nhận hình ảnh xám dưới dạng ma trận các giá trị số, ví dụ, hình ảnh có kích thước 5x32 pixel được biểu diễn dưới dạng ma trận 5 x 32.

Lớp tích chập có nhiệm vụ trích xuất các đặc trưng quan trọng từ hình ảnh, mỗi lớp tích chập sử dụng các bộ lọc với kích thước 3 x 3. Kết quả của phép tính này tạo ra một bản đồ đặc trưng (feature map), biểu diễn các đặc trưng không gian quan trọng từ ảnh gốc. Sau mỗi lớp tích chập, hàm kích hoạt ReLU được sử dụng để thêm tính phi tuyến cho mô hình:

Các lớp gộp (Pooling layers), thường là MaxPooling với kích thước 1 x 2, giúp giảm kích thước không gian của các bản đồ đặc trưng mà vẫn giữ lại các đặc trưng quan trọng.

Sau các lớp tích chập và gộp, các bản đồ đặc trưng được chuyển thành vectơ 1 chiều thông qua lớp Flatten, chuẩn bị cho các lớp. Lớp kết nối đầy đủ đầu tiên bao gồm 128 node, sử dụng hàm kích hoạt ReLU. Lớp kết nối đầy đủ cuối cùng sử dụng hàm softmax để phân phối xác suất cho từng loại ung thư.

3. Kết quả và bàn luận

Chúng tôi tiến hành thực nghiệm trên bộ dữ liệu bao gồm thông tin về đột biến gen liên quan đến 12 loại bệnh ung thư phổ biến thể hiện trên Bảng 1. Tổng cộng có hơn 88000 bản ghi, mỗi bản ghi chứa các thông tin về đột biến gen của bệnh nhân mắc các loại ung thư khác nhau. Dữ liệu gen được thu thập từ các tệp biến thể gen của dự án gen 1000 (1000 Genome Project) và cơ sở dữ liệu trực tuyến từ chương trình bản đồ bộ gen ung thư (TCGA). Cả hai nguồn dữ liệu này đều được công nhận rộng rãi là các tiêu chuẩn vàng trong nghiên cứu di truyền và ung thư, nhờ vào quy trình kiểm định chặt chẽ và sự đóng góp của các tổ chức khoa học hàng đầu thế giới.

Các thông tin trong một bản ghi ví dụ bao gồm các trường dữ liệu sau:

- Mã bệnh nhân: Một mã định danh duy nhất cho mỗi bệnh nhân (ví dụ: 9e42ae44).
- Vị trí gen: Thông tin về vị trí đột biến trên bộ gen, VD: chrX:g.positionC>T, trong đó "C>T" biểu thị sự thay đổi từ cytosine (C) sang thymine (T) tại vị trí cụ thể.
- Loại đột biến: Xác định loại đột biến, ví dụ như "Single base substitution" (thay thế một bazơ), mô tả sự thay đổi một nucleotide trong chuỗi DNA.

- Hiệu ứng: Hậu quả của đột biến đối với protein, ví dụ như "stop_gained KDM6A Q555*).", biểu thị việc đột biến tạo ra một mã kết thúc sớm trong quá trình dịch mã, dẫn đến protein ngắn hơn bình thường (mất chức năng hoặc chức năng bị thay đổi).

- Số mẫu/Phần trăm: Tổng số mẫu phát hiện đột biến này cùng với tỷ lệ phần trăm so với tổng số mẫu (ví dụ: 8/412, 1,94%).

- Tần suất toàn cầu: Tổng số trường hợp được ghi nhận trên toàn cầu với đột biến này (ví dụ: 8/10188).

- Mức độ tác động: Đánh giá mức độ ảnh hưởng của đột biến, như "HIGH" (cao), chỉ ra rằng đột biến này có khả năng gây ra những thay đổi quan trọng đối với chức năng của tế bào hoặc protein.

Bảng 1. *Số lượng bản ghi cho từng loại ung thư*

Loại ung thư	Tên đầy đủ	Số lượng
BLCA	Ung thư bàng quang	5000
BRCA	Ung thư vú	5000
COAD	Ung thư đại tràng	7000
GBM	U nguyên bào thần kinh đệm	10000
KIRC	Ung thư thận (carcinoma tế bào thận)	6000
LGG	Ung thư não cấp thấp	8000
LUSC	Ung thư phổi tế bào vảy	10000
OV	Ung thư buồng trứng	9000
PRAD	Ung thư tuyến tiền liệt	8000
SKCM	Ung thư hắc tố da (melanoma)	6000
THCA	Ung thư tuyến giáp	10000
UCEC	Ung thư nội mạc tử cung	8000

Dữ liệu từ ví dụ cho thấy đột biến của bệnh nhân 9e42ae44 xảy ra tại vị trí chrX:g.45063557C>T, là một thay đổi bazơ đơn lẻ từ cytosine (C) sang thymine (T), dẫn đến việc tạo ra mã kết thúc sớm ở protein KDM6A tại vị trí Q555. Tỷ lệ xuất hiện đột biến này là 1,94% trong số 412 mẫu đã được phân tích, và trên toàn cầu đã ghi nhận 8 trường hợp trong số 10188. Mức độ tác động của đột biến này được đánh giá là "HIGH", nghĩa là đột biến có khả năng ảnh hưởng lớn đến chức năng của protein hoặc cơ thể.

Kết quả phân loại tổng thể của mô hình, thể hiện ở Bảng 2, đạt độ chính xác (Accuracy) 97,87%. Độ chính xác này được tính trên bộ dữ liệu kiểm tra (test) theo tỷ lệ 7:1:2 (huấn luyện: xác thực: kiểm tra), với 70% dữ liệu dành cho huấn luyện, 10% dành cho xác thực và 20% dành cho kiểm tra. Các chỉ số khác như Precision (0,9705), Recall (0,9692) và F1-score (0,9751) cũng cho thấy mô hình hoạt động rất tốt, đặc biệt trong các bài toán phân loại bệnh ung thư, độ nhạy (Recall) và độ chụm (Precision) đóng vai trò quan trọng trong việc phát hiện và tránh bỏ sót các ca bệnh. Kết quả này chứng tỏ mô hình có khả năng tổng quát cao và có thể áp dụng hiệu quả trong thực tế, hỗ trợ việc phân loại và chẩn đoán các loại ung thư.

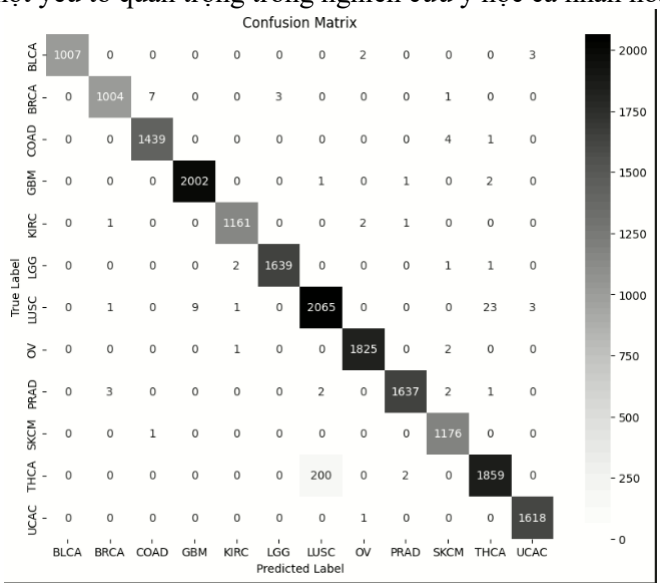
Bảng 2. *Các thông số kết quả phân loại trên tập dữ liệu test*

Accuracy	Precision	Recall	F1 - score
0,9787	0,9705	0,9692	0,9751

Hình 3 là ma trận nhầm lẫn thể hiện kết quả phân loại của mô hình trên từng loại ung thư. Ma trận nhầm lẫn cho thấy: GBM và THCA là hai loại bệnh được phân loại chính xác cao nhất, với số lượng mẫu đúng lần lượt là 1994 và 1842. LUSC và OV cũng đạt kết quả tốt, với số lượng mẫu đúng là 1963 và 1814. Một số nhầm lẫn xảy ra đối với BRCA và KIRC, với BRCA có 38 mẫu bị nhầm lẫn với các loại khác như PRAD và COAD.

Phương pháp của chúng tôi thể hiện khả năng phân loại vượt trội đối với nhiều loại ung thư khác nhau. GAN đã giúp tạo ra các mẫu đa dạng hơn, cải thiện đáng kể hiệu quả của mô hình.

Đồng thời, việc trực quan hóa dữ liệu gen dưới dạng hình ảnh cũng giúp bảo mật thông tin cá nhân của bệnh nhân, một yếu tố quan trọng trong nghiên cứu y học cá nhân hóa.



Hình 3. *Ma trận nhầm lẫn trên 12 loại ung thư*

Ngoài ra, điểm mạnh quan trọng của phương pháp này là khả năng bảo mật thông tin cá nhân của bệnh nhân. Việc chuyển đổi dữ liệu gen thành hình ảnh xám không chỉ giúp khai thác sức mạnh của học sâu mà còn bảo vệ được các thông tin nhạy cảm trong suốt quá trình phân tích và xử lý, đảm bảo cho thông tin không bị rò rỉ, sửa xoá trong quá trình nghiên cứu phân tích. Kết quả nghiên cứu này khẳng định phương pháp đề xuất có tiềm năng ứng dụng cao trong các nghiên cứu y học cá nhân hóa và các nghiên cứu liên quan đến dữ liệu sinh học, giúp hỗ trợ hiệu quả trong việc phát hiện và phân loại đột biến gen nhằm chẩn đoán và điều trị bệnh.

4. Kết luận

Trong nghiên cứu này, chúng tôi đã đề xuất một phương pháp mới kết hợp giữa việc chuyển đổi dữ liệu đột biến gen thành hình ảnh xám và sử dụng GAN để tăng cường dữ liệu, sau đó áp dụng CNN để phân loại 12 loại bệnh ung thư. Phương pháp này không chỉ mang lại hiệu quả vượt trội trong việc cải thiện độ chính xác của mô hình phân loại mà còn đảm bảo an toàn thông tin trong quá trình xử lý dữ liệu.

Việc chuyển đổi dữ liệu đột biến gen thành hình ảnh xám giúp ẩn đi các thông tin nhạy cảm, từ đó bảo vệ quyền riêng tư của bệnh nhân trong quá trình phân tích và xử lý dữ liệu. Kết hợp với GAN, phương pháp này tạo ra các mẫu dữ liệu mới, phong phú hơn, làm tăng tính đa dạng của tập dữ liệu và cải thiện khả năng tổng quát hóa của mô hình khi làm việc với các loại ung thư khác nhau. Điều này giúp mô hình học sâu không chỉ giảm thiểu hiện tượng quá khớp mà còn cải thiện đáng kể độ chính xác ngay cả trên các tập dữ liệu không đồng nhất.

Bên cạnh việc cải thiện hiệu suất phân loại, phương pháp này còn đáp ứng các yêu cầu nghiêm ngặt về bảo mật dữ liệu y tế, một yếu tố đặc biệt quan trọng trong các nghiên cứu liên quan đến dữ liệu gen. Những kết quả đạt được khẳng định tiềm năng của phương pháp trong việc hỗ trợ phát hiện và phân loại bệnh ung thư dựa trên dữ liệu đột biến gen, đồng thời có thể được áp dụng rộng rãi trong các lĩnh vực liên quan đến y học cá nhân hóa và bảo mật thông tin.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] I. Tomlinson, P. Sasieni, and W. Bodmer, "How many mutations in a cancer?" *The American Journal of Pathology*, vol. 160, no. 3, pp. 755-758, 2020.

[2] S. Wang, J. Shi, Z. Ye, D. Dong, D. Yu, M. Zhou, Y. Liu, O. Gevaert, K. Wang, Y. Zhu *et al.*, "Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning," *European Respiratory Journal*, vol. 53, no. 3, 2019, doi: 10.1183/13993003.00986-2018.

[3] M. K. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine learning in genomic medicine: a review of computational problems and data sets," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 176-197, 2015.

[4] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature Communications*, vol. 7, no. 1, 2016, Art. no. 12474.

[5] H. Wang, C. Wang, and H. Qu, "Deep neural network for somatic mutation classification," *Scientific Programming*, vol. 2021, no. 1, 2021, Art. no. 5529202.

[6] P. Sanjaya, K. Maljanen, R. Katainen, S. M. Waszak, L. A. Aaltonen, O. Stegle, J. O. Korbel, and E. Pitk¨anen, "Mutation-attention (muat): deep representation learning of somatic mutations for tumour typing and subtyping," *Genome Medicine*, vol. 15, no. 1, 2023, Art. no. 47.

[7] K. Krishnamachari, D. Lu, A. Swift-Scott, A. Yeraliyev, K. Lee, W. Huang, S. N. Leng, and A. J. Skanderup, "Accurate somatic variant detection using weakly supervised deep learning," *Nature Communications*, vol. 13, no. 1, 2022, Art. no. 4248.

[8] Y. Sun, S. Zhu, K. Ma, W. Liu, Y. Yue, G. Hu, H. Lu, and W. Chen, "Identification of 12 cancer types through genome deep learning," *Scientific Reports*, vol. 9, no. 1, 2019, Art. no. 17256.

[9] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4-21, 2016.

[10] M. Durgadevi *et al.*, "Generative Adversarial Network (GAN): A general review on different variants of GAN and applications," in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, 2021, pp. 1-8.

[11] P. Chaudhari, H. Agrawal, and K. Kotecha, "Data augmentation using mg-gan for improved cancer classification on gene expression data," *Soft Computing*, vol. 24, pp. 11 381-11 391, 2020.