BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

Lê Mạnh Hùng

NÂNG CAO HIỆU QUẢ DỰ ĐOÁN LIÊN KẾT THUỐC-BỆNH DỰA TRÊN SIÊU ĐƯỜNG DẪN VÀ SUY LUẬN BAYES TRÊN MẠNG KHÔNG ĐỒNG NHẤT

Ngành: Công nghệ thông tin
Mã số: 9480201

TÓM TẮT LUẬN ÁN TIẾN SĨ
CÔNG NGHỆ THÔNG TIN

Hà Nội – 2026

MỞ ĐẦU

Tính cấp thiết của đề tài Phát triển thuốc mới truyền thống đòi hỏi chi phí khổng lồ, thời gian kéo dài (12-17 năm) và tỷ lệ rủi ro cao. Tái định vị thuốc - khai thác các công dụng mới từ các thuốc hiện có - nổi lên như một chiến lược đầy hứa hẹn nhằm rút ngắn thời gian và giảm thiểu chi phí. Tuy nhiên, việc xác định các liên kết thuốc-bệnh tiềm năng một cách hiệu quả vấp phải những thách thức lớn về mặt tính toán, bao gồm: (i) tính phức tạp và không đồng nhất của dữ liệu sinh học đa nguồn; (ii) sự mất cân bằng nghiêm trọng và đặc tính thưa của dữ liệu; (iii) sự tồn tại của các mẫu âm tính giả làm sai lệch mô hình; và (iv) hạn chế về khả năng giải thích của các mô hình dự đoán. Xuất phát từ thực tiễn đó, luận án "Dự đoán liên kết thuốc – bệnh trong mạng không đồng nhất" được thực hiện nhằm đề xuất các giải pháp tính toán toàn diện, khắc phục các hạn chế nêu trên và góp phần đẩy nhanh quá trình tái định vị thuốc.

Mục tiêu nghiên cứu Để hoàn thành luận án, nghiên cứu sinh (NCS) đề ra các mục tiêu cụ thể sau đây:

- **Mục tiêu 1:** Phát triển một mô hình dựa trên việc khai thác siêu đường dẫn và áp dụng suy luận Bayes trên mạng thông tin không đồng nhất. Mục tiêu này nhằm (i) tích hợp hiệu quả các nguồn dữ liệu sinh học đa dạng (thuốc, bệnh, protein...) và (ii) ước lượng xác suất tồn tại liên kết thuốc–bệnh, qua đó nâng cao khả năng diễn giải của mô hình.
- **Mục tiêu 2:** Xây dựng và triển khai các kỹ thuật xử lý dữ liệu chuyên sâu, bao gồm: (i) thuật toán lựa chọn mẫu âm tính đáng tin cậy (thông qua tiền xử lý và gán nhãn thông minh) để hạn chế âm tính giả, và (ii) kỹ thuật cân bằng dữ liệu để giảm thiểu tác động của mất cân bằng lớp.

Nội dung và phạm vi nghiên cứu

- *Nội dung chính:*

  1. Tổng quan các phương pháp dự đoán liên kết thuốc-bệnh.
  2. Khai thác siêu đường dẫn trên HIN.
  3. Đề xuất sử dụng suy luận Bayes xây dựng quan hệ tiềm ẩn thuốc-bệnh và mất cân bằng dữ liệu.

- *Phạm vi:* Tập trung vào khía cạnh tính toán, sử dụng dữ liệu chuẩn (DrugBank, OMIM), không bao gồm thử nghiệm lâm sàng.

Đóng góp chính của luận án Luận án trình bày hai đóng góp cốt lõi:

- Đóng góp 1: Khai thác các siêu đường dẫn trong HIN để tính toán mối quan hệ tiềm ẩn giữa thuốc và bệnh từ nhiều nguồn dữ liệu phức tạp, từ đó đề xuất mô hình dự đoán liên kết thuốc–bệnh tin cậy hơn.

- Đóng góp 2:Đề xuất dùng lý thuyết Bayes phân tích mối liên kết thuốc–bệnh dựa trên HIN thuốc–bệnh và xuất phương pháp trích xuất mẫu âm tính chất lượng cao kết hợp kỹ thuật cân bằng dữ liệu cũng như xử lý vấn đề sự mất cân bằng dữ liệu. Phương pháp này góp phần giảm thiểu tỷ lệ sai lệch trong phân loại và nâng cao khả năng tổng quát hóa của mô hình.

Cấu trúc luận án Luận án được tổ chức thành 03 chương với nội dung chính như sau:

- **Chương 1: Tổng quan** - Giới thiệu bài toán, cơ sở lý thuyết về HIN và tổng quan các phương pháp nghiên cứu liên quan.

- **Chương 2: Khai thác siêu đường dẫn trong dự đoán thuốc–bệnh** - Trình bày các mô hình dự đoán dựa trên siêu đường dẫn trong HIN.

- **Chương 3: Suy luận Bayesian và xử lý mất cân bằng dữ liệu** - Đề xuất khung suy luận Bayes và các kỹ thuật xử lý dữ liệu để nâng cao chất lượng mô hình.

- **Kết luận và hướng phát triển** - Tóm tắt kết quả, đánh giá ưu điểm/hạn chế và đề xuất hướng nghiên cứu tiếp theo.

CHƯƠNG 1. TỔNG QUAN

1.1. Giới thiệu bài toán và bối cảnh

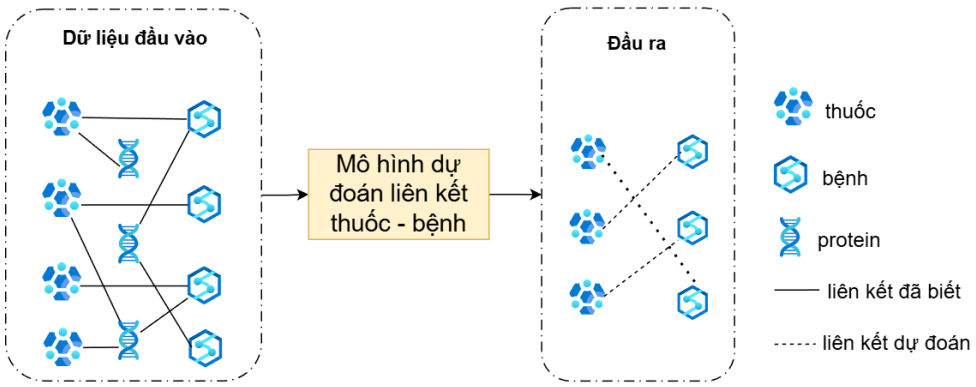
Dự đoán liên kết thuốc–bệnh là thành phần cốt lõi của tái định vị thuốc, nhằm phát hiện chỉ định mới cho thuốc đã có. Bài toán dự đoán liên kết thuốc – bệnh trong HIN đóng một vai trò quan trọng trong việc phát hiện các chỉ định mới tiềm năng cho các dược phẩm hiện có. Mục tiêu cốt lõi của bài toán là đánh giá khả năng tồn tại của một liên kết giữa một cặp thuốc–bệnh dựa trên các thông tin quan hệ gián tiếp có sẵn trong mạng (ví dụ: thông qua protein, gene), ngay cả khi dữ liệu trực tiếp về tương tác đó còn thiếu hoặc chưa được xác minh.

Quy trình tổng quan của bài toán được minh họa trong Hình 1.1 và có thể mô tả như sau:

- **Đầu vào:** Một mạng dữ liệu không đồng nhất $G = (V, E)$, trong đó V là tập các đối tượng (ví dụ: thuốc, bệnh, gen, protein, ...) và E là tập các loại quan hệ sinh học đã biết giữa chúng (ví dụ: thuốc - bệnh, disease - protein, thuốc - protein Associations, ...).
- **Xử lý:** Áp dụng các phương pháp khai thác đồ thị, kỹ thuật phân rã hoặc hoàn thiện ma trận, hoặc các mô hình học sâu để phân lớp nhằm dự đoán khả năng tồn tại của liên kết mục tiêu.
- **Đầu ra:** Với mỗi cặp thuốc–bệnh tiềm năng (v_d, v_b), mô hình đưa ra dự đoán về sự tồn tại của cạnh:

$$p(E_k) = \begin{cases} 1, & \text{nếu liên kết } E_k \text{ có khả năng tồn tại;} \\ 0, & \text{ngược lại.} \end{cases}$$

Mạng thông tin không đồng nhất (HIN) cho phép tích hợp và mô hình hóa nhiều loại quan hệ khác nhau, như giữa thuốc, bệnh, protein, và gen. Tuy nhiên, việc sử dụng HIN cũng gặp phải những thách thức như: dữ liệu không đồng nhất, thiếu dữ liệu, mất cân bằng trong các lớp dữ liệu, âm tính giả, và khó khăn trong việc giải thích kết quả..



Hình 1.1: *Mô tả bài toán dự đoán liên kết thuốc–bệnh.*

1.2. Khái niệm nền tảng

Một mạng thông tin đại diện cho một trừu tượng hóa của thế giới thực, tập trung vào các đối tượng và các tương tác giữa các đối tượng này. Một cách hình thức, mạng thông tin được hiểu như sau:

1.2.1. Mạng thông tin

Một mạng thông tin được biểu diễn là đồ thị $G = (V, E)$ với một hàm ánh xạ loại đối tượng $\phi : V \to A$ và một hàm ánh xạ loại liên kết $\psi : E \to R$. Mỗi đối tượng $v \in V$ thuộc về một loại đối tượng cụ thể trong tập đối tượng $A : \phi(v) \in A$, mỗi liên kết $e \in E$ thuộc về một loại quan hệ cụ thể trong tập các loại quan hệ $R : \psi(e) \in R$. Nếu hai liên kết thuộc cùng một loại quan hệ, hai liên kết này sẽ chia sẻ cùng một loại đối tượng bắt đầu và loại đối tượng kết thúc.

1.2.2. Mạng thông tin không đồng nhất và không đồng nhất

Một mạng thông tin được gọi là mạng thông tin không đồng nhất khi tồn tại nhiều hơn một loại đối tượng ($|A| > 1$) hoặc nhiều hơn một loại quan hệ ($|R| > 1$). Ngược lại, mạng thông tin được gọi là mạng thông tin đồng nhất.

1.2.3. Siêu đường dẫn (meta-path)

Một siêu đường dẫn P [24] là một con đường được định nghĩa trên sơ đồ $TG = (A, R)$, và được ký hiệu dưới dạng $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$, trong đó định nghĩa một quan hệ tổng hợp $R = R_1 \circ R_2 \circ \cdots \circ R_l$ giữa các đối tượng $A_1, A_2, \ldots, A_{l+1}$, trong đó $\circ$ biểu thị toán tử hợp nhất các quan hệ.

1.2.4. Ma trận kết hợp của siêu đường dẫn

*Cho một mạng $G = (V, E)$ và lược đồ mạng [**Wu2019**] của nó là TG, một ma trận kết hợp cho một siêu đường dẫn $P = T_1 \to T_2 \to \cdots \to T_k$ được định nghĩa là:*

$$X = A_{T_1T_2} A_{T_2T_3} \cdots A_{T_{k-1}T_k} \tag{1.1}$$

trong đó $A_{T_iT_j}$ là ma trận kề (tương tác) giữa loại T_i và loại T_j. Giá trị $X(i, j)$ biểu thị số lượng các thể hiện con đường giữa đối tượng $u_i \in T_1$ và đối tượng $v_j \in T_k$ theo siêu đường dẫn P.

1.3. Tổng quan nghiên cứu

Theo các tổng quan gần đây Bagherian (2021),Kim (2022) và Cai (2023) và các công sự, các hướng dự đoán liên kết thuốc–bệnh có thể phân thành ba loại:

1. **Dựa trên độ tương đồng/khoảng cách**: khai thác từ tương đồng cấu trúc/kiểu hình. *Ưu điểm*: đơn giản, dễ triển khai. *Hạn*: phụ thuộc dữ liệu đã biết, yếu với thuốc/bệnh mới.

2. **Dựa trên ma trận (MC/MF/CMF/RMF)**: khai thác cấu trúc hạng thấp của ma trận tương tác để suy luận phần tử thiếu; hiệu quả với dữ liệu thưa nhưng dễ mắc cực tiểu cục bộ, chi phí tối ưu cao.

3. **Khai thác đồ thị (graph mining)**: làm việc trực tiếp trên HIN (phân cụm, meta-path, meta-graph, random walk, lan truyền, khuếch tán). *Ưu điểm*: tích hợp đa nguồn, tận dụng cấu trúc mạng. *Hạn chế* : cần chọn meta-path thủ công, khó mở rộng mạng rất lớn.

Xu hướng hiệu quả là kết hợp phương pháp ma trận và khai thác đồ thị để vừa tận dụng cấu trúc HIN vừa học quan hệ phi tuyến.

1.4. Khoảng trống và hướng tiếp cận của luận án

Khoảng trống. (a) Tích hợp đa nguồn trên HIN còn thủ công, thiếu chuẩn hoá đặc trưng meta-path;

(b) Xử lý mất cân bằng dự liệu, âm tính giả chưa triệt để;

(c) Cuối cùng, Khả năng diễn giải và độ tin cậy của kết quả dự đoán vẫn là một thách thức lớn.

Hướng tiếp cận.

(1) Khai thác sâu meta-path và ma trận kết hợp + SVD để nâng chất lượng biểu diễn;

(2) Áp dụng suy luận Bayes để ước lượng xác suất liên kết và giảm âm tính giả;

(3) Phát triển cân bằng dữ liệu và lựa chọn mẫu âm chất lượng cao (hard negatives) để tăng độ tin cậy dự đoán.

1.5. Dữ liệu và phương pháp đánh giá

1.5.1. Tập dữ liệu thử nghiệm

Bộ dữ liệu A_dataset:

Bộ dữ liệu kế thừa từ nghiên cứu của Wu và cộng sự (2019), tích hợp từ các nguồn uy tín như OMIM (bệnh-protein), Gottlieb (thuốc-bệnh) và DrugBank (thuốc-protein). Quy mô của bộ dữ liệu này bao gồm 1.186 thuốc, 449 bệnh và 1.147 protein. Các tương tác đã biết trong bộ dữ liệu này là 1.827 thuốc-bệnh, 4.642 thuốc-protein và 1.365 bệnh-protein. Đặc điểm nổi bật của bộ dữ liệu là tính mất cân bằng cao, với tỷ lệ tương tác dương tính thấp, dao động từ 0.207% đến 0.344%.

Bộ dữ liệu B_dataset:

Bộ dữ liệu được xây dựng bởi Zhang và Zhao, với quy mô gồm 269 thuốc, 598 bệnh và 1.021 protein. Các tương tác đã biết trong bộ dữ liệu này là 18.416 thuốc-bệnh, 3.110 thuốc-protein và 5.898 bệnh-protein. Bộ dữ liệu này cung cấp một lượng lớn thông tin tương tác, phục vụ cho việc nghiên cứu và đánh giá các mô hình dự đoán liên kết thuốc-bệnh.

1.5.2. Phương pháp đánh giá

Để đánh giá hiệu suất các mô hình, luận án sử dụng các chỉ số dưới đây:

Để đánh giá hiệu suất các mô hình, luận án sử dụng các chỉ số đánh giá đa dạng. Cụ thể, các chỉ số cơ bản bao gồm Độ chính xác (Accuracy - ACC), Độ thu hồi (Recall - REC), và Độ chính xác dương (Precision - PRE). Đối với dữ liệu mất cân bằng, luận án sử dụng Điểm F1 (F1-Score), Hệ số tương quan Matthews (MCC), và Giá trị trung bình hình học (G-Mean). Bên cạnh đó, các chỉ số dựa trên đường cong như Diện tích dưới đường cong ROC (AUC) và Diện tích dưới đường cong Precision-Recall (AUPR) cũng được áp dụng để đánh giá hiệu suất mô hình.

1.5.3. Thiết lập thực nghiệm

- **Ngôn ngữ lập trình**: Python 3.x; Thư viện chính: Scikit-learn; LightGBM; NetworkX;

  Phương pháp đánh giá: Đánh giá chéo 5-fold

- Áp dụng kỹ thuật loại bỏ thông tin (information removal) để tránh rò rỉ dữ liệu; Đảm bảo tính công bằng trong so sánh bằng cách sử dụng cùng tập dữ liệu và phương pháp đánh giá; Thực hiện kiểm định thống kê để xác nhận ý nghĩa của kết quả

Các phương pháp đánh giá kết hợp với hai bộ dữ liệu độc lập giúp đánh giá khách quan và toàn diện hiệu suất của các phương pháp đề xuất.

1.6. Kết luận

Chương này thiết lập nền tảng lý thuyết và hiện trạng, làm rõ hạn chế của các phương pháp hiện có. Luận án vì vậy tập trung vào: (i) Phương pháp khai thác thông tin trên đa nguồn dựa trên meta-path; (ii) xử lý mất cân bằng và âm tính giả bằng cân bằng dữ liệu. Áp dụng suy luận Bayes trong dự đoán thuốc-bệnh; (iii) đánh giá nghiêm ngặt qua phương pháp cắt bỏ và thông kê.

CHƯƠNG 2. KHAI THÁC SIÊU ĐƯỜNG DẪN TRONG DỰ ĐOÁN LIÊN KẾT THUỐC–BỆNH

Từ những thách thức như đã phân tích ở Chương 1, chương này tập trung vào việc khai thác mạng thông tin không đồng nhất (HIN) để giải quyết bài toán dự đoán liên kết thuốc-bệnh. Cụ thể, chương này nhằm khắc phục những hạn chế quan trọng của mô hình EMP-SVD, bao gồm: (i) chưa khai thác triệt để vai trò trung gian của protein, (ii) thiên lệch trong việc sử dụng các quan hệ, (iii) bỏ qua các quan hệ đồng nhất, và (iv) xử lý dữ liệu chưa tối ưu. Để giải quyết các điểm yếu này, luận án đề xuất một phương pháp tiếp cận dựa trên ba hướng cải tiến chính, nhằm khai thác toàn diện ngữ nghĩa phức tạp trong HIN để nâng cao hiệu quả dự đoán.

2.1. Phân tích mô hình EMP-SVD và định hướng cải tiến

2.1.1. Giới thiệu mô hình EMP-SVD

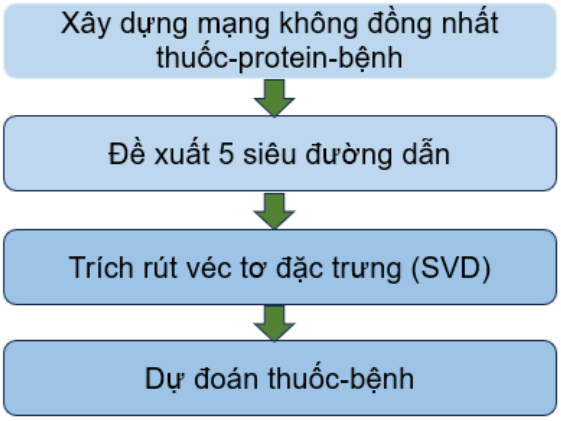
Mô hình EMP-SVD kết hợp các kỹ thuật meta-path, SVD và Random Forest, tích hợp dữ liệu tương tác giữa các thực thể sinh học như thuốc, protein và bệnh để dự đoán mối liên kết giữa thua thuốc và bệnh. Mô hình được thực hiện qua các bước chính như Hình 2.1

- **Bước 1**: Xây dựng mạng dữ liệu không đồng nhất với ba loại nút: thuốc, protein và bệnh
- **Bước 2**: Đề xuất 5 siêu đường dẫn với độ dài nhỏ hơn 4
- **Bước 3**: Trích xuất đặc trưng bằng SVD để giảm chiều dữ liệu
- **Bước 4**: Xây dựng và tổng hợp các mô hình phân lớp sử dụng Random Forest

2.1.2. Phân tích hạn chế của mô hình EMP-SVD

Qua phân tích và so sánh hiệu suất của EMP-SVD, có thể rút ra một số hạn chế chính:

Thứ nhất, các meta-path không đi qua protein (Meta-Path-1 và Meta-Path-4) cho kết quả dự đoán thấp hơn đáng kể. Nguyên nhân chính là do protein đóng vai trò



Hình 2.1: *Mô tả phương pháp EMP-SVD*

trung gian quan trọng trong nhiều quá trình sinh học, việc bỏ qua thông tin này làm suy giảm khả năng khái quát hóa của mô hình.

Thứ hai, trong năm meta-path được đề xuất, số lần khai thác quan hệ thuốc–bệnh (6 lần) vượt trội so với quan hệ thuốc–protein và bệnh–protein (mỗi loại chỉ 3 lần). Sự thiên lệch này dẫn đến việc chưa tận dụng đầy đủ các nguồn thông tin tiềm ẩn trong mạng dữ liệu.

Thứ ba, mô hình EMP-SVD chủ yếu tập trung vào quan hệ không đồng nhất mà chưa khai thác quan hệ đồng nhất như thuốc–thuốc, bệnh–bệnh hay protein–protein. Điều này hạn chế khả năng mô hình hóa đầy đủ cấu trúc và ngữ nghĩa phức tạp trong HIN.

2.1.3. Định hướng phát triển

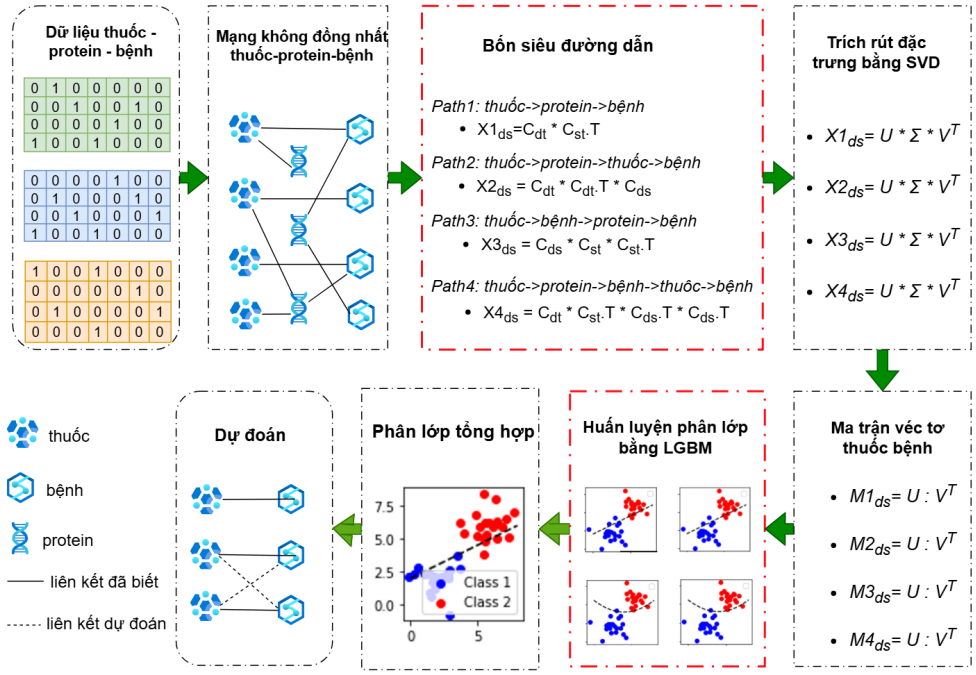
Trong Chương 2, luận án kế thừa khung mô hình EMP-SVD và đề xuất ba hướng cải tiến nhằm mở rộng khả năng mô hình hóa ngữ nghĩa trong HIN, từ đó nâng cao độ chính xác trong dự đoán liên kết thuốc–bệnh cụ thể:

- **Đề xuất mô hình DR-LGBM-MH**: Xây dựng 01 siêu đường dẫn mới và 3 siêu đường dẫn đề xuất bởi EMP-SVD dựa trên HIN thuốc–protein–bệnh, kết hợp với LightGBM
- **Đề xuất mô hình HS-TMP**: Xây dựng sáu ma trận đồng nhất thuốc-thuốc, bệnh-bệnh, protein-protein, từ đó xây dựng ba nhóm siêu đường dẫn mới, phản ánh tương quan nội tại giữa thuốc, bệnh và protein

2.2. Đề xuất mô hình DR-LGBM-MH với 4 siêu đường dẫn

2.2.1. Cơ sở phương pháp

Mô hình DR-LGBM-MH kế thừa kiến trúc EMP-SVD nhưng tối ưu hóa việc lựa chọn meta-path dựa trên phân tích vai trò sinh học của protein. Phương pháp này tập trung vào việc xây dựng các siêu đường dẫn có ý nghĩa sinh học rõ ràng và khả năng tính toán hiệu quả, khung phương pháp được thể hiện như Hình 2.2



Hình 2.2: *Sơ đồ luồng công việc của mô hình*

2.2.2. Thiết kế 4 siêu đường dẫn

- **Meta-path-1 (d–t–s)**: Thuốc → Protein → Bệnh

$$X_1 = C_{dt} \times C_{st}^T$$

Đường dẫn trực tiếp nhất, phản ánh cơ chế tác động cơ bản của thuốc thông qua protein mục tiêu.

- **Meta-path-2 (d–t–d–s)**: Thuốc → Protein → Thuốc → Bệnh

$$X_2 = C_{dt} \times C_{dt}^T \times C_{ds}$$

Khai thác các thuốc có chung mục tiêu protein, hỗ trợ tái định vị thuốc.

- **Meta-path-3 (d–s–t–s)**: Thuốc → Bệnh → Protein → Bệnh

$$X_3 = C_{ds} \times C_{st} \times C_{st}^T$$

Phát hiện các bệnh có chung cơ chế phân tử thông qua protein trung gian.

- **Meta-path-4 (d–t–s–d–s)**: Thuốc → Protein → Bệnh → Thuốc → Bệnh

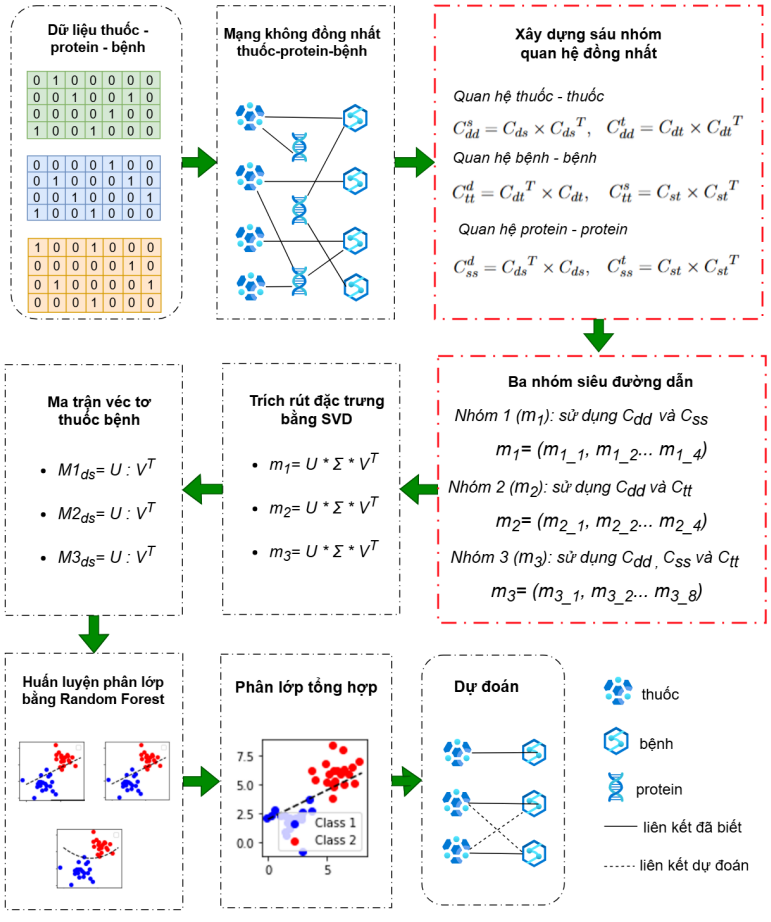
$$X_4 = C_{dt} \times C_{st} \times C_{ds}^T \times C_{ds}$$

Mô hình hóa mối quan hệ đa bước phức tạp trong mạng sinh học.

2.3. Mô hình HS-TMP: Quan hệ đồng nhất và siêu đường dẫn

2.3.1. Cơ sở phương pháp

Đề xuất này mở rộng HIN bằng cách tích hợp các quan hệ đồng nhất dựa trên lý thuyết về độ tương đồng cấu trúc. Hệ thống gồm 6 ma trận đồng nhất phản ánh mối quan hệ nội bộ giữa các thực thể cùng loại, mô hình phương pháp được thể hiện như Hình 2.3.



Hình 2.3: *Sơ đồ luồng công việc 3 nhóm meta-path*

2.3.2. Đề xuất 6 ma trận đồng nhất

Trong việc thiết kế HIN, luận án đề xuất ba loại cạnh mới ngoài các cạnh hiện có (thuốc-protein, bệnh-protein, thuốc-bệnh), bao gồm thuốc-thuốc, bệnh-bệnh và protein-protein. Các ma trận liên kết này được tính toán như sau:

- **Thuốc-thuốc:** Mối quan hệ thuốc-thuốc được thiết lập qua bệnh (ma trận C^s_{dd}) hoặc protein (ma trận C^t_{dd}) làm trung gian:

$$C^s_{dd} = C_{ds} \times C^T_{ds}, \quad C^t_{dd} = C_{dt} \times C^T_{dt}$$

- **Protein-protein:** Mối quan hệ này có thể được xác định qua liên kết thuốc-protein hoặc bệnh-protein, tạo thành các ma trận C^d_{tt} và C^s_{tt}:

$$C^d_{tt} = C^T_{dt} \times C_{dt}, \quad C^s_{tt} = C^T_{ds} \times C_{ds}$$

- **Bệnh-bệnh:** Tương quan bệnh-bệnh được xây dựng qua bệnh-thuốc và bệnh-protein, tạo ra các ma trận C^d_{ss} và C^t_{ss}:

$$C^d_{ss} = C^T_{ds} \times C_{ds}, \quad C^t_{ss} = C_{st} \times C^T_{st}$$

Các ma trận này cho phép mở rộng khả năng mô hình hóa mối quan hệ giữa các thực thể trong mạng và cải thiện độ chính xác của dự đoán.

2.3.3. Ba nhóm siêu đường dẫn

Nhóm 1: Khai thác quan hệ thuốc-bệnh cốt lõi Nhóm này khai thác các quan hệ thuốc-thuốc, thuốc-bệnh và bệnh-bệnh, với ma trận kết hợp tổng quát:

$$m_1 = C_{dd} \times C_{ds} \times C_{ss}.$$

Các lựa chọn liên kết thuốc-thuốc và bệnh-bệnh cho phép tạo ra bốn meta-path con sau:

$$m1\_1 : C^s_{dd} \times C_{ds} \times C^d_{ss} \quad m1\_2 : \quad C^s_{dd} \times C_{ds} \times C^t_{ss}$$
$$m1\_3 : C^t_{dd} \times C_{ds} \times C^d_{ss} \quad m1\_4 : \quad C^t_{dd} \times C_{ds} \times C^t_{ss}$$

Nhóm 2: Tích hợp thông tin protein trung gian Nhóm này tích hợp quan hệ thuốc-protein và bệnh-protein, tạo ra ma trận kết hợp:

$$m_2 = C_{dd} \times C_{dt} \times C_{tt} \times C^T_{st}.$$

Bốn meta-path con từ các lựa chọn thuốc-thuốc và protein-protein được mô tả như sau:

$$m2\_1 : C_{dd}^{s} \times C_{dt} \times C_{tt}^{s} \times C_{st}^{T} \quad m2\_2 : \quad C_{dd}^{s} \times C_{dt} \times C_{tt}^{d} \times C_{st}^{T}$$
$$m2\_3 : C_{dd}^{t} \times C_{dt} \times C_{tt}^{s} \times C_{st}^{T} \quad m2\_4 : \quad C_{dd}^{t} \times C_{dt} \times C_{tt}^{d} \times C_{st}^{T}$$

Nhóm 3: Khai thác toàn diện đa quan hệ Nhóm này kết hợp ba mối quan hệ đồng nhất (thuốc-thuốc, protein-protein, bệnh-bệnh) và không đồng nhất (thuốc-protein, protein-bệnh) với ma trận kết hợp:

$$m_3 = C_{dd} \times C_{dt} \times C_{tt} \times C_{st}^{T} \times C_{ss}.$$

Có tám meta-path con trong nhóm này:

$$m3\_1 : C_{dd}^{s} \times C_{dt} \times C_{tt}^{s} \times C_{st}^{T} \times C_{ss}^{d} \quad m3\_2 : \quad C_{dd}^{s} \times C_{dt} \times C_{tt}^{s} \times C_{st}^{T} \times C_{ss}^{t}$$
$$m3\_3 : C_{dd}^{s} \times C_{dt} \times C_{tt}^{d} \times C_{st}^{T} \times C_{ss}^{d} \quad m3\_4 : \quad C_{dd}^{s} \times C_{dt} \times C_{tt}^{d} \times C_{st}^{T} \times C_{ss}^{t}$$
$$m3\_5 : C_{dd}^{t} \times C_{dt} \times C_{tt}^{s} \times C_{st}^{T} \times C_{ss}^{d} \quad m3\_5 : \quad C_{dd}^{t} \times C_{dt} \times C_{tt}^{s} \times C_{st}^{T} \times C_{ss}^{t}$$
$$m3\_7 : C_{dd}^{t} \times C_{dt} \times C_{tt}^{d} \times C_{st}^{T} \times C_{ss}^{d} \quad m3\_8 : \quad C_{dd}^{t} \times C_{dt} \times C_{tt}^{d} \times C_{st}^{T} \times C_{ss}^{t}$$

2.4. Kết quả thực nghiệm và đánh giá

2.4.1. Xây dựng kịch bản

Kịch bản 1: Mô hình DR-LGBM-MH

- **Bước 1:** Trích xuất đặc trưng từ 4 meta-path biểu diễn quan hệ thuốc–bệnh, sau đó giảm chiều bằng SVD (β tối ưu = 0,04).
- **Bước 2:** Huấn luyện 4 LightGBM riêng biệt và kết hợp bằng phương pháp bỏ phiếu.
- **Bước 3:** Đánh giá toàn diện:
  - LightGBM vượt trội so với các bộ phân loại khác (RF, XGB, SVM, KNN, AC)
  - Ensemble giảm đáng kể FN và FP so với meta-path đơn lẻ
  - Nghiên cứu cắt bỏ: SVD và LightGBM cải thiện hiệu suất rõ rệt
  - Kiểm định t-test ($p < 0.05$) khẳng định ý nghĩa thống kê
  - Vượt trội các chỉ số về Recall, MCC, F1-score

Kịch bản 2: Mô hình HS-TMP:

- Xây dựng 6 loại tương quan (thuốc–thuốc, bệnh–bệnh, protein–protein) để làm giàu mạng tri thức.
- Tạo 3 nhóm mô hình cơ sở sử dụng Random Forest trên các ma trận.
- Kết hợp 3 mô hình cơ sở (tổng cộng 128 cách kết hợp) để đánh giá hiệu quả.
- So sánh kết quả với các nghiên cứu trước và chọn 20 cặp thuốc–bệnh có điểm dự đoán cao nhất để phân tích điển hình.

2.4.2. Kết quả

Kết quả thực nghiệm mô hình DR-LGBM-MH

So sánh hiệu suất với các mô hình phân loại truyền thống và hiện đại

Bảng 2.1: *Kết quả nghiên cứu cắt bỏ về ảnh hưởng của SVD*

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
RF không SVD	0,955	0,949	0,905	0,861	0,877	0,756	0,882
RF có SVD	0,959	0,953	0,888	0,891	0,887	0,776	0,889
LightGBM không SVD	0,960	0,955	0,898	0,893	0,894	0,787	0,895
LightGBM có SVD	**0,969**	**0,966**	**0,915**	**0,921**	**0,917**	**0,834**	**0,918**

Các giá trị tốt nhất được in đậm.

Kết quả cho thấy, LightGBM với SVD đạt hiệu suất cao nhất trên toàn bộ các thước đo, vượt trội so với cả RF và LightGBM không dùng SVD.

Kiểm định thống kê sự khác biệt hiệu suất

Để xác nhận rằng sự cải thiện hiệu suất không phải ngẫu nhiên, luận án tiến hành kiểm định t-test hai mẫu cho các thước đo chính. Kết quả được trình bày trong Bảng 2.2.

Bảng 2.2: *So sánh hiệu suất bằng kiểm định t-test hai mẫu*

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
RF không SVD	0,0128	0,00997	0,6240	0,0089	0,00236	0,0020	0,0017
RF có SVD	0,0055	0,0002	0,1180	0,0960	0,0044	0,0046	0,0024
LightGBM không SVD	0,0330	0,0150	0,1240	0,0793	0,0039	0,0038	0,0027

Kết quả kiểm định cho thấy, phương pháp đề xuất đạt cải thiện có ý nghĩa thống kê trên các chỉ số AUPR, AUC, ACC, MCC và F1-score (p-value < 0.05). Điều này đảm bảo rằng các cải thiện về hiệu suất không chỉ là kết quả ngẫu nhiên, mà phản ánh tác động thực sự của phương pháp.

So sánh với các phương pháp trong nghiên cứu trước

Cuối cùng, luận án so sánh mô hình DR-LGBM với các phương pháp được công bố trước đây. Nhóm phương pháp cổ điển bao gồm PREDICT, TL-HGBI, LRSSL, SCMFDD, MBiRW; trong khi nhóm phương pháp gần đây có EMP-SVD

Bảng 2.3: *Kết quả so sánh hiệu suất giữa các phương pháp*

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
PREDICT	0,908	0,895	0,809	0,850	0,830	0,662	0,828
TL-HGBI	0,852	0,846	0,829	0,750	0,774	0,552	0,787
LRSSL	0,881	0,861	0,864	0,732	0,770	0,553	0,790
SCMFDD	0,836	0,854	**0,926**	0,713	0,774	0,575	0,805
MBiRW	0,952	0,942	0,867	0,901	0,884	0,769	0,884
EMP-SVD	0,956	0,951	0,913	0,854	0,876	0,755	0,882
AICI2023	0,968	0,966	0,930	0,882	0,903	0,806	0,906
DR-LGBM-MH	**0,969**	**0,966**	0,915	**0,921**	**0,917**	**0,834**	**0,918**

Các giá trị tốt nhất được in đậm.

Nhìn chung, DR-LGBM-MH không chỉ đạt được giá trị cao ở từng chỉ số riêng lẻ mà còn duy trì được sự cân bằng giữa các tiêu chí. Sự vượt trội của DR-LGBM-MH so với EMP-SVD có thể được lý giải bởi việc đề xuất các meta-path tập trung vào vai trò trung gian của protein (MP-1, MP-2, MP-3, MP-4), khắc phục được hạn chế của EMP-SVD khi bỏ qua các quan hệ đồng nhất. So với các phương pháp dựa trên ma trận (SCMFDD) hoặc lan truyền (MBiRW), phương pháp của chúng tôi linh hoạt hơn trong việc nắm bắt các mối quan hệ phi tuyến phức tạp thông qua LightGBM.

Kết quả thực nghiệm Mô hình HS-TMP:

Luận án so sánh với một số mô hình tiên tiến trước đó, bao gồm EMP-SVD, LRSSL, MBiRW, MPG-DDA, PREDICT, SCMFDD và TL-HGBI. Kết quả so sánh được trình bày trong Bảng 2.4 theo các chỉ số AUPR, AUC, PRE, REC, ACC, MCC và F1-score.

Bảng 2.4: *Hiệu suất của các phương pháp liên quan*

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
EMP-SVD	0,956	0,951	**0,913**	0,854	0,876	0,755	0,882
LRSSL	0,881	0,861	0,864	0,732	0,770	0,553	0,790
MBiRW	0,952	0,942	0,867	0,901	0,884	0,769	0,884
MPG-DDA	0,944	0,930	0,886	0,842	0,867	—	0,863
PREDICT	0,908	0,895	0,809	0,850	0,830	0,662	0,828
SCMFDD	0,836	0,854	0,926	0,713	0,774	0,575	0,805
TL-HGBI	0,852	0,846	0,829	0,750	0,774	0,552	0,787
Phương pháp của luận án	**0,968**	**0,963**	0,895	**0,922**	**0,908**	**0,816**	**0,908**

2.4.3. Các nghiên cứu điển hình

Phần này kiểm chứng các cặp thuốc – bệnh có xác suất dự đoán cao (0,99), cho thấy mô hình dự đoán chính xác nhiều mối liên hệ phù hợp với bằng chứng y học, như nhóm β-blocker cho suy thận tiến triển và TCA cho bệnh thần kinh di truyền. Một số liên kết mới như Fludrocortisone – hạ huyết áp tư thế đứng và Oseltamivir – bệnh não gợi mở tiềm năng tái định vị thuốc.

2.5. Kết luận và đóng góp chính

- Đề xuất 1 siêu đường dẫn mới giúp khai thác sâu vai trò trung gian của protein và khắc phục hạn chế của EMP-SVD.
- Phát triển khung 6 ma trận đồng nhất và 3 nhóm siêu đường dẫn cho phép khai thác toàn diện cấu trúc HIN, tạo ra 128 tổ hợp meta-path con.
- Phương pháp này mở ra hướng tiếp cận mới trong khai thác HIN cho bài toán dự đoán thuốc-bệnh, với ứng dụng thực tiễn rộng rãi.
- Kết quả chương 2 được công bố trong bài báo ISI Q2 [CT01], 2 bài hội nghị SCOPUS [CT02], [CT03], và 1 bài hội nghị quốc gia [CT04], chứng minh tính khả thi và ứng dụng của phương pháp.

CHƯƠNG 3. SUY LUẬN BAYESIAN VÀ XỬ LÝ DỮ LIỆU TRONG DỰ ĐOÁN LIÊN KẾT THUỐC–BỆNH

3.1. Bối cảnh và mục tiêu

Kế thừa các mô hình từ Chương 2, chương này giải quyết ba thách thức then chốt nhằm củng cố độ tin cậy cho các dự đoán: (i) âm tính giả trong dữ liệu huấn luyện, (ii) mất cân bằng dữ liệu nghiêm trọng giữa các lớp, và (iii) nhu cầu về một khung suy luận có thể lượng hoá mức độ tin cậy.

Để đồng thời giải quyết các thách thức này, luận án đề xuất một quy trình tích hợp ba kỹ thuật chính: suy luận Bayesian, thuật toán chọn mẫu âm tính chất lượng cao (HNS) và kỹ thuật cân bằng dữ liệu Gaussian–SMOTE. Chương này sẽ trình bày chi tiết các bước thực hiện, bắt đầu từ việc thiết lập mô hình, lựa chọn mẫu âm tính, cân bằng tập dữ liệu, cho đến bước suy luận cuối cùng. Khung phương pháp được minh hoạ trong Hình 3.1.

3.2. Phương pháp đề xuất

3.2.1. Cơ sở Lý thuyết Bayesian

Xác suất liên kết thuốc–bệnh thông qua các protein trung gian được xác định như sau:

$$p(d \mid s) = \sum_{t} p(d \mid t)\, p(t \mid s). \tag{3.1}$$

Công thức mô tả xác suất dự đoán liên kết giữa thuốc d và bệnh s thông qua các protein trung gian t, trong đó $p(d \mid t)$ và $p(t \mid s)$ được ước lượng từ dữ liệu huấn luyện.

3.2.2. Thuật toán chọn mẫu âm tính chất lượng cao (HNS)

Thay vì sử dụng tất cả các cặp thuốc-bệnh chưa được chứng minh, luận án đề xuất tính toán xác suất tương tác $p^*(d|s)$ dựa trên xác suất tiên nghiệm kết hợp của protein $p^*(t)$. Chỉ những cặp có xác suất $p^*(d|s) = 0$ mới được chọn vào tập mẫu âm mới Z_*^-. Tập dữ liệu huấn luyện cuối cùng $Z = Z^+ \cup Z_*^-$ nhờ đó được cân bằng và có chất lượng cao hơn, giúp mô hình học tập hiệu quả và cho kết quả dự đoán chính xác hơn.

Những mẫu này được xem là chất lượng cao (HNS) vì có xác suất liên kết rất thấp, làm giảm thiểu khả năng chúng thực chất là các liên kết tích cực chưa được phát hiện (false negative)

Thuật toán 3.1 Chọn mẫu âm tính tiềm năng cao

Input : $A_{dt}[n \times k]$ (ma trận thuốc-protein)
$A_{st}[m \times k]$ (ma trận bệnh-protein-)
$A_{tt}[k \times k]$ (ma trận protein-protein);

Output: Z^- (tập mẫu âm tính)

Begin Algorithm

$A_{ds} \leftarrow A_{dt} \cdot A_{tt} \cdot A_{st}^T$;

$Z^- \leftarrow \emptyset$

$i \leftarrow 1$

while $i \leq m$ **do**
  $j \leftarrow 1$
  while $j \leq n$ **do**
    if $A_{ds}(i,j) = 0$ **then**
      $Z^- \leftarrow Z^- \cup \{(i,j)\}$
    end
    $j \leftarrow j + 1$
  end
  $i \leftarrow i + 1$
end

return Z^-

Kết thúc thuật toán

3.2.3. Suy luận Bayesian trong dự đoán liên kết thuốc–bệnh

Luận án xây dựng năm mô hình suy luận Bayesian khác nhau, mỗi mô hình thể hiện một góc nhìn riêng về mối quan hệ thuốc–protein–bệnh. Dự đoán cuối cùng được tính bằng cách lấy giá trị lớn nhất trong năm mô hình, giúp khai thác tối đa thông tin và cung cấp xác suất hậu nghiệm minh bạch cho từng cặp thuốc–bệnh.

$$p_1(d \mid s) = \frac{p(s \mid d)\, p(d)}{p(s)} \qquad \text{(dựa trên dữ liệu huấn luyện trực tiếp)}, \tag{3.2}$$

$$p_2(d \mid s) = \sum_t p(d \mid t)\, p(t \mid s) \qquad \text{(thuốc–protein–bệnh)}, \tag{3.3}$$

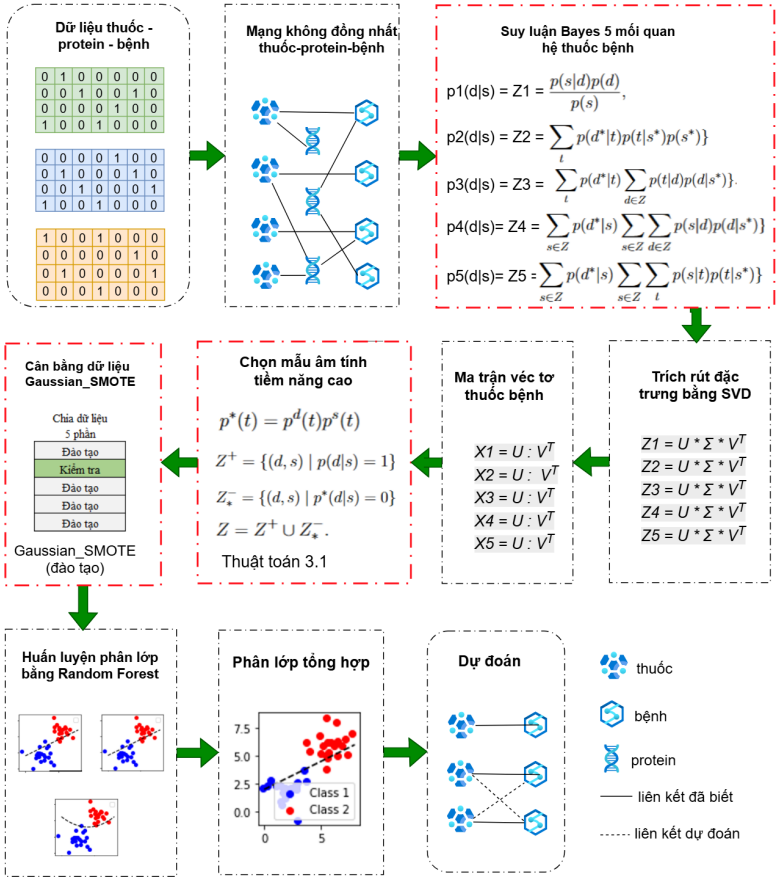
$$p_3(d \mid s) = \sum_{t} p(d \mid t) \sum_{d' \in Z} p(t \mid d')\, p(d' \mid s) \quad \text{(mở rộng qua thuốc trung gian)}, \tag{3.4}$$

$$p_4(d \mid s) = \sum_{s' \in Z} p(d \mid s') \sum_{d' \in Z} p(s' \mid d')\, p(d' \mid s) \quad \text{(quan hệ thuốc–bệnh–thuốc)}, \tag{3.5}$$

$$p_5(d \mid s) = \sum_{s' \in Z} p(d \mid s') \sum_{t} p(s' \mid t)\, p(t \mid s) \quad \text{(kết hợp protein–bệnh)}. \tag{3.6}$$

Dự đoán tổng hợp cuối cùng:

$$p(d \mid s) = \max_{i=1,\dots,5} p_i(d \mid s). \tag{3.7}$$



Hình 3.1: *Khung phương pháp đề xuất*

3.3. Kết quả thực nghiệm và đánh giá

3.3.1. Dữ liệu

Hai bộ dữ liệu được sử dụng: **A_dataset** (OMIM, Gottlieb, DrugBank) gồm 551 thuốc, 302 bệnh, 1.147 protein; và **B_dataset** (Zhang, 2018; Zhao, 2022) gồm 269 thuốc, 598 bệnh, 1.021 protein. Cả hai đều mất cân bằng nghiêm trọng, là nền tảng để đánh giá các kỹ thuật xử lý và mô hình dự đoán.

Mục tiêu thử nghiệm gồm: (i) đánh giá hiệu quả phương pháp đề xuất trong phát hiện liên kết thuốc–bệnh; (ii) so sánh với các phương pháp hiện tại bằng bộ chỉ số F1, G-mean, PR_AUC; và (iii) kiểm định thống kê để chứng minh ý nghĩa cải thiện.

Dữ liệu được biểu diễn trên mạng không đồng nhất thuốc–protein–bệnh; mẫu âm tính chọn theo Wu (2019) và thuật toán đề xuất, tạo các biến thể A_FNdataset, A_HNdataset, B_FNdataset, B_HNdataset; sau đó chia train/test theo tỷ lệ 80:20.

Các **thực nghiệm chính gồm**: (1) So sánh nhiều kỹ thuật cân bằng dữ liệu (SPY, NCR, SMOTE, KMeans_SMOTE, Gaussian_SMOTE, ...); (2) Phân tích ablation để đánh giá vai trò của Gaussian_SMOTE; (3) Kiểm chứng phương pháp trên B_dataset.

Kết quả cho thấy **Gaussian–SMOTE** kết hợp với mẫu âm tính chất lượng cao (HNS) mang lại hiệu quả tốt nhất và được dùng cho các thử nghiệm tiếp theo.

3.3.2. Hiệu quả của Gaussian–SMOTE

Các thí nghiệm lấy mẫu quá mức (SMOTE và biến thể) trên A_FNdataset và A_HNdataset được tổng hợp trong Bảng 3.1. Kết quả cho thấy HNdataset luôn vượt trội FNdataset trên mọi thước đo. Đáng chú ý, trên A_HNdataset, Gaussian_SMOTE đạt F1 = 0.8509, G-Mean = 0.8980 (cao nhất), và PR-AUC = 0.8839; CURE_SMOTE cũng cạnh tranh với F1 = 0.8453, PR-AUC = 0.8836.

Bảng 3.1: *Hiệu suất mô hình theo các kỹ thuật lấy mẫu quá mức.*

Phương pháp	A_FNdataset (mẫu quá mức)			A_HNdataset (mẫu quá mức)*		
	F1	**G-Mean**	**PR-AUC**	**F1**	**G-Mean**	**PR-AUC**
SMOTE	0.7308	0.7909	0.7531	0.7985	0.7586	0.8395
Borderline-SMOTE	0.7206	0.7854	0.7345	0.8217	0.8761	0.8351
CURE-SMOTE	0.7920	0.8556	0.8320	0.8453	0.8858	0.8836
SMOTE-TomekLinks	0.7292	0.7907	0.7516	0.7988	0.8528	0.8393
AND-SMOTE	0.7057	0.7708	0.7181	0.7794	0.8565	0.8063
SMOTE-D	0.7911	0.8589	0.8288	0.8424	0.8848	0.8831
Random-SMOTE	0.7384	0.8061	0.7728	0.8168	0.8842	0.8541
KMeans-SMOTE	**0.7940**	**0.8686**	0.8045	0.8426	0.8851	0.8809
SMOTEWB	0.7757	0.8527	0.8152	0.8371	0.8970	0.8767
Gaussian_SMOTE	0.7936	0.8667	**0.8327**	**0.8509**	**0.8980**	**0.8839**

*Điểm số cao nhất được in đậm.

3.3.3. Đề xuất 3: Đánh giá hiệu suất của phương pháp đề xuất

Luận án tiến hành so sánh bốn cấu hình gồm: dữ liệu gốc, chỉ áp dụng Gaussian_SMOTE, chỉ áp dụng High Negative (HN), và mô hình kết hợp HN + Gaus-

sian_SMOTE. Mỗi cấu hình được kiểm định bằng 5-fold cross-validation nhằm bảo đảm tính khách quan và ổn định. Kết quả ở Bảng 3.2 cho thấy mô hình kết hợp đạt hiệu suất cao nhất trên cả ba thước đo (F1 = 0.8509, G-Mean = 0.8980, AUPR = 0.8839), vượt trội so với các cấu hình còn lại. Đặc biệt, việc kết hợp HN và Gaussian_SMOTE giúp cải thiện rõ rệt khả năng nhận diện mẫu dương hiếm, cân bằng ranh giới phân loại và tăng độ ổn định trên dữ liệu mất cân bằng.

Bảng 3.2: *So sánh F1, G-mean và PR-AUC giữa các phương pháp*

Phương pháp	**F1**	**G-mean**	**PR-AUC**
Original	0.5264	0.6976	0.5067
High Negative	0.8339	0.8779	0.8746
Gaussian_SMOTE	0.5413	0.6933	0.5248
Our method	**0.8509**	**0.8980**	**0.8839**

Sự kết hợp HN + Gaussian-SMOTE hoạt động hiệu quả vì: HN loại bỏ nhiễu từ âm tính giả, trong khi Gaussian-SMOTE cân bằng không gian đặc trưng, giúp mô hình học được biên quyết định tối ưu cho cả lớp thiểu số và đa số.

3.3.4. Kết quả thống kê t-test

Kết quả t-test hai mẫu độc lập cho thấy phương pháp Gaussian–SMOTE + HNS có sự khác biệt có ý nghĩa thống kê so với các phương pháp khác (tất cả các chỉ số cho $p < 0{,}05$), chứng minh tính ưu việt của phương pháp, kết quả thể hiện Bảng 3.3.

Bảng 3.3: *Kết quả thống kê t-test hai mẫu*

So sánh	**F1**	**G-mean**	**PR-AUC**
Original vs Gaussian–SMOTE	0,000010	0,00012	0,000006
Original vs High Negative	0,370000	0,780000	0,066000
Original vs Our Method	0,000005	0,000051	0,000006
Gaussian–SMOTE vs High Negative	0,000005	0,000370	0,000005
Gaussian–SMOTE vs Our Method	0,000002	0,000190	0,000004
High Negative vs Our Method	0,000310	0,000440	0,000180

3.3.5. So sánh hiệu suất với các nghiên cứu gần đây

Phương pháp đề xuất được so sánh với các mô hình hiện có như deepDR, HINGRL, AMDDT, v.v. Kết quả trên tập Bdataset cho thấy phương pháp của luận án đạt AUPR = 0,915, AUC = 0,966 và F1 = 0,856, cao nhất trong các mô hình so sánh xem Bảng 3.4.

Bảng 3.4: *So sánh hiệu suất với các nghiên cứu gần đây trên Bdataset*

Phương pháp	**AUPR**	**AUC**	**PRE**	**REC**	**ACC**	**MCC**	**F1**
deepDR	0,804	0,820	0,883	0,233	0,601	0,299	0,369
DDAGDL	0,831	0,842	0,761	0,770	0,764	0,529	0,765
HINGRL	0,877	0,884	0,800	0,808	0,803	0,607	0,804
HNet-DNN	0,891	0,892	0,782	0,828	0,810	0,621	0,804
DRWBNCF	0,901	0,900	0,981	0,202	0,599	0,326	0,335
DRHGCN	0,910	0,909	0,867	0,771	0,826	0,658	0,816
AMDDT	0,930	0,933	0,861	0,865	0,862	0,725	0,863
Model with FNS	0,892	0,959	0,835	0,843	0,932	0,793	0,835
Phương pháp luận án	**0,915**	**0,966**	**0,862**	**0,851**	**0,938**	**0,817**	**0,856**

** Các giá trị cao nhất được in đậm.*

Phương pháp đề xuất trong luận án này kết hợp mô hình dựa trên xử lý mất cân bằng dữ liệu với kỹ thuật Gaussian_SMOTE, nhằm cải thiện độ chính xác trong dự đoán liên kết thuốc–bệnh. Không chỉ chứng minh hiệu quả qua các chỉ số định lượng, mô hình còn dự đoán chính xác các thuốc có tiềm năng thực tiễn, mở ra hướng tiếp cận mới cho bài toán tái định vị thuốc (drug repositioning) trong lĩnh vực dược phẩm, giúp rút ngắn thời gian và chi phí phát triển thuốc.

3.4. Kết luận và đóng góp chính

Chương này đề xuất phương pháp kết hợp suy luận Bayesian với xử lý dữ liệu thông minh cho bài toán dự đoán liên kết thuốc-bệnh, với các đóng góp chính:

- Thuật toán HNS dựa trên Bayesian để lọc mẫu âm tính chất lượng cao
- Kết hợp Gaussian-SMOTE với HNS tạo bộ dữ liệu cân bằng tối ưu
- Hệ thống năm công thức suy luận Bayesian cho dự đoán đa góc nhìn

Kết quả đã được công bố trong bài báo ESCI Q2 [CT05], 01 tạp chí trong nước[CT06], và 01 hội nghị SCOPUS [CT07], khẳng định tính ứng dụng thực tiễn của phương pháp.

KẾT LUẬN VÀ HƯỚNG NGHIÊN CỨU

Luận án đóng góp quan trọng trong dự đoán mối quan hệ thuốc–bệnh, đặc biệt trong tái định vị thuốc, qua việc ứng dụng học máy và suy luận Bayesian để giải quyết các vấn đề như dữ liệu không đồng nhất, mất cân bằng và mẫu âm tính giả.

Đóng góp chính của luận án

1. Đề xuất giải pháp sử dụng các siêu đường dẫn (meta-path) để nâng cao hiệu quả dự đoán liên kết thuốc-bệnh trong mạng không đồng nhất thuốc-protein-bệnh. - Đề xuất 1 siêu đường dẫn mới kết hợp với 3 meta-path của Wu và cộng sự nhằm khai thác tối đa thông tin từ các nút trung gian như protein - Đề xuất 6 quan hệ đồng nhất thuốc-thuốc, protein-protein, bệnh-bệnh từ đó xây đề xuất xây dựng 3 nhóm siêu đường dẫn dựa trên 6 mối quan hệ đồng nhất và lựa chọn sự kết hợp tốt nhất từ mỗi nhóm 1 siêu đường dẫn con

2. Đề xuất phương pháp suy luận Bayes để phân tích mối quan hệ giữa thuốc và bệnh từ mạng thuốc – protein – bệnh không đồng nhất. Đồng thời, phương pháp đề xuất cũng sử dụng kỹ thuật cân bằng dữ liệu, chọn các mẫu âm tính có độ tin cậy cao để dự đoán mối quan hệ giữa thuốc và bệnh.

Hướng nghiên cứu tiếp theo

Hướng nghiên cứu tiếp theo sẽ tập trung vào:

- Tự động hóa việc lựa chọn và kết hợp các meta-path, nâng cao tính chính xác và linh hoạt của mô hình.
- Phát triển cơ chế gán trọng số để tối ưu hóa khả năng kết hợp thông tin từ các mạng dữ liệu phức tạp.

DANH MỤC CÔNG TRÌNH CÔNG BỐ

[**CT01**] Anh Dao, N, Le MH, Tho Dang, X. (2024). "Label Transfer for Drug Disease Association in Three Meta-Paths", Evol Bioinform Online. 2024 Sep 13;20:11769343241272414. doi: 10.1177/11769343241272414. PMID: 39279816; PMCID: PMC11401013.

[**CT02**] Tho Dang, X., Hung Le, M., Anh Dao, N. (2023). "Drug Repositioning for Drug Disease Association in Meta-paths", In: Phuong, N.H., Kreinovich, V. (eds) Deep Learning and Other Soft Computing Techniques. Studies in Computational Intelligence, vol 1097. Springer, Cham. https://doi.org/10.1007/978-3-031-29447-1_4

[**CT03**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2025). "Drug Repositioning by XGBoost for Meta-Paths in Heterogeneous Networks", In: Hoang Phuong, N., Huyen Chau, N.T., Vladik Kreinovich, (editors), Explainable AI and Other Soft Computing Techniques: Biomedical and Related Applications, Springer, (To appear in 2026) (indexed in Scopus)

[**CT04**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2025). "Drug Repositioning by LightGBM for Meta-Paths in Heterogeneous Networks", The National Conference on Fundamental and Applied IT Research, 2025.

[**CT05**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2025). "Bayes Inference for Drug Discovery by High Negative Samples and Oversampling", Bioinformatics and Biology Insights. 2025;19. doi:10.1177/11779322251328269

[**CT06**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2024). "Enhancing Drug Discovery Through A Meta-Path Based Oversampling Approach For Imbalanced Data", Journal of Science and Technique-Section on Information and Communication Technology 13.01 (2024).

[**CT07**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2024). "High Potential Negative Sampling for Drug Disease Association Prediction", In: Hoang Phuong, N., Huyen Chau, N.T., Kreinovich, V. (eds) Machine Learning and Other Soft Computing Techniques: Biomedical and Related Applications. Studies in Systems, Decision and Control, vol 543. Springer, Cham. https://doi.org/10.1007/978-3-031-63929-6_7

[**CT08**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2026). Drug repositioning by belief networks and ensemble method. Biomed Phys Eng Express. 2026 Feb 19;12(2). doi: 10.1088/2057-1976/ae43f0. PMID: 41666480.

[**CT09**] Hung Le, M., Anh Dao, N., Tho Dang, X. (2026). Drug Repositioning by Multilayer Perceptron with KernelPCA in Heterogeneous Networks, In: Hoang Phuong, N., Huyen Chau, N.T., Vladik Kreinovich, (editors), Explainable AI and Other Soft Computing Techniques: Biomedical and Related Applications, Springer, (To appear in 2026), (indexed in Scopus).