TRƯỜNG ĐẠI HỌC CẦN THƠ

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**NIÊN LUẬN CƠ SỞ NGÀNH**

**NGÀNH HỆ THỐNG THÔNG TIN**

**Đề tài**

**PHÂN TÍCH DỮ LIỆU Y HỌC CÁ THỂ HÓA VỚI GAUSSIAN MIXTURE MODELS**

**Sinh viên: Đặng Ngọc Xuân Đài**

**Mã số: B2003731**

**Khóa: K46**

**Cần Thơ, 11/2023**

TRƯỜNG ĐẠI HỌC CẦN THƠ

**TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**NIÊN LUẬN CƠ SỞ NGÀNH**

**NGÀNH HỆ THỐNG THÔNG TIN**

**Đề tài**

**PHÂN TÍCH DỮ LIỆU Y HỌC CÁ THỂ HÓA VỚI GAUSSIAN MIXTURE MODELS**

**Người hướng dẫn**

**TS. Nguyễn Thanh Hải**

**Sinh viên: Đặng Ngọc Xuân Đài**

**Mã số: B2003731**

**Khóa: K46**

**Cần Thơ, 11/2023**

**LỜI CẢM ƠN**

Để hoàn thành niên luận với đề tài là “Phân tích dữ liệu y học cá thể hóa với Gaussian Mixture Models”, bên cạnh sự cố gắng nỗ lực không ngừng của bản thân, trước tiên em xin được bày tỏ lòng biết ơn chân thành sâu sắc đến cha mẹ đã cổ vũ, động viên, hỗ trợ về tinh thần cũng như tạo mọi điều kiện thuận lợi cho em hoàn thành tốt niên luận này.

Em xin chân thành cảm ơn thầy Nguyễn Thanh Hải đã luôn quan tâm chỉ dạy, theo dõi, giúp đỡ tận tình trong suốt khoảng thời gian em thực hiện niên luận của mình tốt nhất.

Và hơn hết, em xin bày tỏ lòng biết ơn trân trọng đến quý thầy cô Trường Công Nghệ Thông Tin và Truyền Thông nói chung và quý thầy cô bộ môn hệ thống thông tin nói riêng đã tận tình chỉ dạy, truyền đạt những kiến thức quý báu cho em trong thời gian vừa qua để em có đủ kiến thức, điều kiện để thực hiện niên luận này. Đồng thời em cũng rất biết ơn các cán bộ trực ở thư viện, trung tâm học liệu, phòng máy... đã hỗ trợ giúp đỡ em trong thời gian qua.

Đồng cảm ơn đến các tác giả trong các quyển sách báo, Internet, anh chị đi trước đã tìm tòi, nghiên cứu đúc kết kinh nghiệm làm tài liệu để em có thể tham khảo trong quá trình thực hiện niên luận.

Sau cùng xin cảm ơn các bạn cùng lớp Hệ Thống Thông Tin, Trường Công Nghệ Thông Tin và Truyền Thông, Trường Đại học Cần Thơ đã tận tình giúp đỡ, hỗ trợ cho em thực hiện niên luận này.

**MỤC LỤC**

[CHƯƠNG 1. GIỚI THIỆU VÀ MÔ TẢ BÀI TOÁN 1](#_Toc152152150)

[1.1. Mục tiêu đề tài 1](#_Toc152152151)

[1.2. Mô tả chi tiết đề tài 3](#_Toc152152152)

[1.3. Hướng tiếp cận giải quyết của đề tài 12](#_Toc152152153)

[CHƯƠNG 2. THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP 15](#_Toc152152154)

[2.1. Kiến trúc tổng quát hệ thống 15](#_Toc152152155)

[2.2. Xây dựng các mô hình 15](#_Toc152152156)

[2.3. Giải pháp cài đặt 18](#_Toc152152157)

[CHƯƠNG 3. KIỂM THỬ VÀ ĐÁNH GIÁ 19](#_Toc152152158)

[3.1. Kịch bản kiểm thử 19](#_Toc152152159)

[3.2. Kết quả kiểm thử 19](#_Toc152152160)

[CHƯƠNG 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 27](#_Toc152152161)

[4.1. Kết luận 27](#_Toc152152162)

[4.2 Hướng phát triển 28](#_Toc152152163)

CHƯƠNG [5. TÀI LIỆU THAM KHẢO 29](#_Toc152152164)

**DANH MỤC BẢNG**

[Bảng 1 Dataset 16](#_Toc152444346)

[Bảng 2. So sánh độ đo AUC và ACC bài báo của [39], [38], và GMM 19](#_Toc152444347)

[Bảng 3. So sánh độ đo AUC và ACC của GMM, binning GMM với RF và SVM 21](#_Toc152444348)

[Bảng 4. So sánh ACC và AUC dữ liệu gốc và GMM 21](#_Toc152444349)

[Bảng 5. So sánh độ đo AUC của các bài nghiên cứu về dữ liệu metagenomics đường ruột 22](#_Toc152444350)

[Bảng 6. So sánh độ đo ACC của các bài nghiên cứu về dữ liệu metagenomics đường ruột 24](#_Toc152444351)

**DANH MỤC HÌNH**

[Hình 1. Sơ đồ các bước tạo nhóm bằng GMM 17](#_Toc152335927)

[Hình 2. Biểu đồ box plots thể hiện từng loại bệnh của GMM 20](#_Toc152335928)

[Hình 3. Biểu đồ box plots thể hiện từng loại bệnh của GMM, MetAML, EQW 20](#_Toc152335929)

[Hình 4. Biểu đồ box plots thể hiện từng loại bệnh của GMM, PopPhy-CNN, DeepMicro, RegMIL 23](#_Toc152335930)

[Hình 5. Biểu đồ box plots thể hiện từng loại bệnh theo RF của GMM, MetAML, RegMIL, Met2Img 25](#_Toc152335931)

[Hình 6. Biểu đồ box plots thể hiện từng loại bệnh theo SVM của GMM, MetAML, RegMIL, Met2Img 25](#_Toc152335932)

**DANH MỤC TỪ CHUYÊN NGÀNH**

|  |  |
| --- | --- |
| **Viết tắt** | **Giải thích** |

|  |  |
| --- | --- |
| 3D-SWIG | Xung cân bằng không đổi trạng thái 3D với góc vàng trên mặt phẳng (3D Steady-State Free Precession with In-plane Golden-angle) |
| ACC  ACE  ATLAS  AUC  BALF  BIPLS  B-NMI  CAR  CARS  CC  CIR  CRB  CRC  CT  DBSCAN  DMSP  DNA  DP-Net  FASTQ  GMHI  GMM  GNN  GSMD  GITM  HDBSCAN  Hi-C  HiFine  hiPCA  HIV  HMP2  IBD  IBS  ID  IMF  LDPC  LSH  LSTM  Ma  MD-graph  MetaBAT-LR  MetaCoAG  MetAML  MLAT  MLT  mNGS  NID  NNC  NSR  NTMPD  OBE  Opal  PCA  PCIB  PCR  PJP  PMI  PHAMB  PHATE  RF  SCG  SSFP  SVM  T2D  t-SNE  UC  UD  UMAP  UVE  VAE  VAMB  VIP  WGS  WWTP | Độ chính xác (Accuracy)  Ước lượng phủ sóng dựa trên sự hiện diện (Abundance-based Coverage Estimation)  Công cụ tự động cho Phân tích Quy mô Lớn của Các Trình tự (Automated Tool for Large-scale Analysis of Sequences)  Diện tích dưới đường cong (Area Under the Curve)  Dịch rửa phế quản phế nang (Bronchoalveolar Lavage Fluid)  Hồi quy bậc phần tử khoảng cách ngược (Backward Interval Partial Least Squares)  Thông tin chung được chuẩn hóa bằng phân loại (Binning-Normalized Mutual Information)  Nghiên cứu Hành động trong lớp học (Classroom Action Research)  Lấy mẫu có trọng số cạnh tranh thích ứng (Competitive Adaptive Reweighted Sampling)  Hệ số Cramér (Cramér’s Coefficient)  Xơ gan (Cirrhosis)  Ranh giới liên quan đến mỏng (Cusp-Related Boundary)  Ung thư đại trực tràng (Colorectal Cancer)  Tầm soát hình ảnh cắt lớp (Computed Tomography)  Phân cụm không gian dựa trên mật độ với nhiễu (Density-Based Spatial Clustering of Applications with Noise)  Chương trình Vệ tinh Khí tượng Quốc phòng (Defense Meteorological Satellite Program)  Axít Deoxyribonucleic (Deoxyribonucleic Acid)  Mạng dự đoán bệnh lý (Disease prediction Network)  Tệp Trình tự Định dạng nhanh (Fastq Format Sequence Files)  Chỉ số sức khỏe vi sinh vật đường ruột (gut microbiota health index)  Mô hình hỗn hợp Gaussian (Gaussian Mixture Model)  Mạng Nơ-ron Đồ Thị (Graph Neural Network)  Dữ liệu mạng xã hội có tham chiếu địa lý (Georeferenced Social Media Data)  Mô hình tầng điện ly-nhiệt quyển toàn cầu (Global Ionosphere-Thermosphere Model)  Phân cụm không gian dựa trên mật độ phân cấp với nhiễu (Hierarchical Density-Based Spatial Clustering of Applications with Noise)  Chỉ bắt sự cấu trúc của nhiễm sắc thể (Chromatin Conformation Capture)  Dây chuyền tạo nhóm có cấu trúc phân cấp và được điều chỉnh tỉ mỉ (rarchical and Finely-tuned Binning Pipeline)  Chỉ số sức khỏe với PCA (Health index with PCA)  Vi rút suy giảm miễn dịch người (Human Immunodeficiency Virus)  Dự án Vi sinh vật học của con người 2 (Human Microbiome Project 2)  Bệnh viêm ruột (Inflammatory Bowel Disease)  Hội chứng ruột kích thích (Irritable Bowel Syndrome)  Bệnh truyền nhiễm (Infectious Disease)  Từ trường liên hành tinh (Interplanetary Magnetic Field)  Kiểm tra tính đồng dư mật độ thấp (Low-Density Parity Check)  Băm nhạy cảm với địa phương (Locality-Sensitive Hashing)  Bộ nhớ Ngắn hạn Dài (Long Short-Term Memory)  Triệu năm (Mega-annum)  Đồ thị bệnh lý vi sinh vật đa nguyên (Metagenomic Disease graph)  Mở rộng đọc dài của MetaBAT (Long-read extension of MetaBAT)  Đồ thị hợp nhất metagenomic (Metagenomic Co-assembly Graph)  Phân tích Dự đoán Metagenomic dựa trên Học máy (Metagenomic Analysis using Machine Learning)  Vĩ độ từ tính (Magnetic Latitude)  Thời gian địa phương từ tính (Magnetic Local Time)  Công nghệ Sequencing thế hệ tiếp theo trong nghiên cứu Metagenomics (Metagenomic Next-Generation Sequencing)  Bệnh không truyền nhiễm (Non-Infectious Disease)  Bộ phân loại hàng xóm gần nhất (Nearest Neighbor Classifier)  Tỷ lệ Nhiễu trên Tín hiệu (Noise-to-Signal Ratio)  Bệnh phổi do Mycobacteria không lao (Non-Tuberculous Mycobacterial Pulmonary Disease)  Béo phì (Obesity)  Tối ưu hóa cho việc gán họ phylogenetic của contig dài (Optimization for Phylogenetic Assignment of Long contigs)  Phân tích thành phần chính (Principal Component Analysis)  Phân loại bằng thành phần chính isometric (Principal Components Isometric Binning)  Phản ứng chuỗi polymerase (Polymerase Chain Reaction)  viêm phổi do Pneumocystis jirovecii (Pneumocystis jirovecii Pneumonia)  Chỉ số quản lý mua hang (Purchasing Managers' Index)  Tạo nhóm virus từ dữ liệu metagenomic (Phages from Metagenomics Binning)  Tiềm năng lan truyền nhiệt độ cho Nhúng Chuyển đổi dựa trên Sự Gần gũi (otential of Heat-diffusion for Affinity-based Transition Embedding)  Rừng Ngẫu Nhiên (Random Forest)  Đồ thị biểu diễn thành phần chuỗi (Sequence Composition Graph)  Xung cân bằng không đổi trạng thái (Steady-State Free Precession)  Máy Vector Hỗ trợ (Support Vector Machine)  Tiểu đường loại 2 (Type 2 Diabetes)  Nhúng Lân cận Ngẫu nhiên theo Phân phối t (t-distributed Stochastic Neighbor Embedding)  Viêm đại tràng loét (Ulcerative Colitis)  Bệnh Crohn (Crohn's Disease)  Xấp xỉ và Chiếu Manifold Đồng Nhất (Uniform Manifold Approximation and Projection)  Loại bỏ biến không thông tin (Uninformative Variable Elimination)  Mã hóa tự biến thiên (Variational Autoencoder)  Bộ mã hóa tự động biến thiên cho việc tạo nhóm metagenomic (Variational Autoencoder for Metagenomic Binning)  Tầm quan trọng của biến trong phép chiếu (Variable Importance in Projection)  giải trình tự toàn bộ bộ gen (Whole Genome Sequencing)  Nhà máy xử lý nước thải (Wastewater Treatment Plant) |
|  |  |

# GIỚI THIỆU VÀ MÔ TẢ BÀI TOÁN

## Mục tiêu đề tài

Cùng với sự tiến bộ của khoa học kỹ thuật hiện đại, metagenomics đang là một trong những lĩnh vực nghiên cứu phát triển nhanh nhất hiện nay. Mặc dù không phải là một lĩnh vực tương đối mới đối với khoa học, nhưng vài năm qua đã có ​​sự bùng nổ trong các phương pháp tính toán áp dụng cho nghiên cứu dựa trên metagenomics. Các nghiên cứu về bộ gen của vi sinh vật bắt đầu vào cuối những năm 1970 nhưng đến năm 1998 thuật ngữ "metagenomics" mới được giới thiệu đến cộng đồng bởi Jo Handelsman và các cộng sự [1]và được định nghĩa là việc đánh giá tất cả các vật liệu di truyền được phân lập trực tiếp từ các mẫu môi trường. Metagenomic (có thể được hiểu là di truyền học môi trường, di truyền học sinh thái hay di truyền học quần xã.) là phương pháp nghiên cứu (phân tích) đa hệ gen (metagenome) của tất cả vi sinh vật thu nhận trực tiếp mẫu từ môi trường tự nhiên mà không qua nuôi cấy. Thực chất, metagenomics là thuật ngữ được sử dụng mô tả lĩnh vực nghiên cứu khoa học cho phép phân tích toàn thể vi sinh vật sống trong bất kỳ một trường tự nhiên nào. Kevin Chen và cộng sự [2] đã định nghĩa metagenomics là " việc ứng dụng các kỹ thuật di truyền hiện đại trong nghiên cứu về quần xã vi sinh vật một cách trực tiếp trong môi trường tự nhiên của chúng mà không cần phải phân lập và nuôi cấy chúng trong phòng thí nghiệm". Chúng ta có khả năng thu được thông tin bộ gen trực tiếp từ các cộng đồng vi sinh vật trong môi trường sống tự nhiên của chúng, thay vì xem xét một vài loài riêng lẻ, ta có thể nghiên cứu hàng chục nghìn loài cùng nhau.

Metagenomics đã cho chúng ta khả năng nghiên cứu, ở cấp độ gen cơ bản nhất, mối quan hệ giữa vi khuẩn với quần thể và môi trường sống nơi chúng sinh sống. Ban đầu, metagenomics chủ yếu được thúc đẩy bởi việc tìm kiếm các phân tử sinh học mới trong quần thể vi sinh vật có nguồn gốc từ môi trường ôn đới. Sự phát triển của các phương pháp phân lập DNA cải tiến [3], chiến lược nhân bản và kỹ thuật sàng lọc cho phép đánh giá và khai thác các tập hợp vi sinh vật từ các môi trường khắc nghiệt.

Với sự ra mắt của các kỹ thuật giải trình tự thế hệ tiếp theo, các bộ dữ liệu trình tự môi trường ngày càng phức tạp đã được tạo ra, từ đó dẫn đến sự phát triển của các công cụ sinh học khác nhau để phân tích và so sánh các bộ dữ liệu này. Các metagenome khác nhau có thể được tìm thấy trong đất, biển, bùn, rừng, thực vật, động vật không gian và thậm chí trong cơ thể con người,… đôi khi được thu thập từ các mẫu vật thủy sinh, trên cạn, trong đường ruột, phân.Các vi sinh vật có thể được tìm thấy trong các môi trường này có thể bao gồm vi khuẩn, virus, nấm, giun,… Thành công của metagenomics phụ thuộc rất lớn vào các phần mềm tin sinh học và nguồn dữ liệu thu thập được.

Metagenomics có rất nhiều ứng dụng tiềm năng trong cả y học và vi sinh vật môi trường. Nghiên cứu Metagenomics được thực hiện nhiều nhất trong vi sinh môi trường là nghiên cứu của [Atif Khurshid Wan](https://link.springer.com/article/10.1007/s11356-023-25192-5#auth-Atif_Khurshid-Wani-Aff1) và các cộng sự [4] về đa dạng của các quần xã vi sinh vật trong môi trường cụ thể thông qua phân tích các gene rRNA và khám phá các cộng đồng vi sinh vật chưa được khai thác khi điều kiện lý hóa của môi trường biến đổi.

Như trong lĩnh vực khoa học trái đất: phát triển mô hình genome trên cơ sở sinh thái vi sinh vật nhằm mô tả và dự báo các quá trình môi trường toàn cầu, thay đổi khí hậu và sự bền vững của trái đất theo nghiên cứu của Ủy ban về Metagenomics [5]: Những thách thức và Ứng dụng chức năng, Quốc gia Hội Đồng Nghiên Cứu

Trong lĩnh vực khoa học sự sống: các thuyết mới trình độ tiên tiến dự báo trước năng lực sinh học quần thể vi sinh vật gốc, sinh thái tiến hóa bằng nghiên cứu mô hình trao đổi chất của các cộng đồng trong kỷ nguyên bộ gen được lắp ráp metagenome của Clémence Frioux và các cộng sự [6].

Trong lĩnh vực năng lượng: phát triển hệ vi sinh vật và các quá trình cho nguồn năng lượng sinh học mới có hiệu quả kinh tế, môi trường bền vững hơn và có khả năng phục hồi nhanh với các biến động xấu trên trái đất như nghiên cứu Làm sáng tỏ phương pháp tiếp cận Metagenomics để sản xuất nhiên liệu sinh học vi sinh vật của [Km Sartaj](https://pubmed.ncbi.nlm.nih.gov/?term=Sartaj%20K%5BAuthor%5D) và các cộng sự [7].

Trong lĩnh vực môi trường: Phát triển công cụ để quan trắc môi trường khi có sự cố ở tất cả các mức độ, từ thay đổi khí hậu đến rò rỉ khí đốt, hóa chất, dầu từ kho chứa và các phương pháp vi sinh vật cơ bản được gọi là phương pháp canh phục vụ việc duy trì các hệ sinh thái như Phương pháp tiếp cận metagenomic để nghiên cứu sự đa dạng vi khuẩn không phụ thuộc vào nuôi cấy trong môi trường bị ô nhiễm - một trường hợp nghiên cứu ở bờ biển phía đông bắc của Vịnh Bengal, Ấn Độ của Jaya Chakraborty và cộng sự [8].

Hầu hết mọi khía cạnh của cuộc sống con người, cũng như cuộc sống của mọi sinh vật khác. đang ở trên hành tinh, bị ảnh hưởng bởi vi khuẩn. Vì thế Metagenomics cũng được ứng dụng nghiên cứu trong chăm sóc sức khỏe khi cơ thể con người có rất nhiều hệ vi sinh vật sinh sống, trên da, trong ruột, khoang miệng, khoang mũi, mắt, …. Các vi sinh vật này có ảnh hưởng lớn đến sức khỏe cả hai mặt tốt lẫn xấu, hiệu quả sử dụng thuốc, khả năng hấp thu dinh dưỡng, … của con người. Các nghiên cứu ước tính rằng số lượng vi sinh vật có trên cơ thể lớn gấp 10 lần số tế bào mà cơ thể bạn có, và có hơn 1000 loài vi sinh vật khác nhau trên cơ thể. Nghiên cứu dự đoán các đặc điểm BMI trong dân số Trung Quốc dựa trên metagenome ruột của Yu Liang và các cộng sự [9].

Metagenomics hiện nay cũng được áp dụng cho điều tra y tế hoặc pháp y như tạo điều kiện cho bước đột phá trong bệnh lý pháp y, độc tính học và xét nghiệm lạm dụng chất gây nghiện. Do đó, những tiến bộ trong metagenomics sẽ cung cấp những hiểu biết mới về lĩnh vực pháp y ứng dụng., suy luận vị trí địa lý của các vi sinh vật, ước tính PMI,… như trong bài báo của Jun Zhang và các cộng sự vào năm 2022 [10].

Cùng với việc xác định rõ đóng góp của hệ vi sinh vật cho sức khỏe, phát hiện, điều trị bệnh cho các cá thể bao gồm cả người và động vật, thực vật; phát triển các phương pháp điều trị mới trên cơ sở các kiến thức metagenomics. Theo nghiên cứu của [Kirsty TT Kwok](https://pubmed.ncbi.nlm.nih.gov/?term=Kwok%20KT%5BAuthor%5D) và cộng sự [11] vào năm 2020 về Metagenomics của virus ở động vật trang trại, các loại virus động vật, virus lây truyền từ động vật sang người và các loại virus mới đã được biết đến đã được báo cáo trong các tài liệu hiện có, chứng minh khả năng của mNGS trong việc xác định cả loại virus đã biết và loại virus mới.

Hiện nay người ta đã áp dụng Metagenomics lâm sàng ở các nước thu nhập thấp và trung bình để phát hiện, tiếp cận rộng rãi hơn cả mầm bệnh đã biết và mầm bệnh mới như nghiên cứu của Gert Marais và các cộng sự [12].

Trong bài niên luận này này, chúng tôi sẽ nghiên cứu về binning trong khai thác tổng hợp các tập dữ liệu của metagenomics. Một giải pháp gom cụm trình tự metagenomic được đề xuất đó là Gaussian Mixture Models nhằm giải quyết một tập hợp các điểm dữ liệu cần được nhóm thành nhiều phần hoặc cụm dựa trên mức độ giống nhau của chúng. Nghiên cứu này cũng đánh giá tổng quan hiệu suất phân cụm, so sánh với các giải pháp gom cụm trình tự metagenomic hiện nay.

## Mô tả chi tiết đề tài

Đề tài “Phân cụm metagenomics dựa trên thuật toán Gaussian Mixture Models” với mục tiêu phân cụm các mẫu metagenomics. Để thực hiện mục tiêu trên, nghiên cứu đã xây dựng áp dụng thuật toán Gaussian Mixture Models và kết hợp huấn luyện các mô hình như CNN, RF.

Trong đó Binning là một trong những vấn đề quan trọng trong quá trình phân tích dữ liệu metagenomic. Binning, còn được gọi là phân nhóm dữ liệu rời rạc hoặc phân nhóm dữ liệu, là một kỹ thuật tiền xử lý dữ liệu được sử dụng trong khai thác dữ liệu. Nó liên quan đến việc chia một biến liên tục thành một tập hợp các khoảng hoặc “thùng” nhỏ hơn và thay thế các giá trị ban đầu bằng nhãn thùng tương ứng. Việc tạo nhóm có thể được áp dụng cho cả biến số và biến phân loại và mục đích chính của nó là đơn giản hóa dữ liệu và giúp phân tích dễ quản lý và quan sát hơn. Việc kết hợp các biến liên tục tạo ra tính phi tuyến tính và có xu hướng cải thiện hiệu suất của mô hình. Nó cũng có thể được sử dụng để xác định các giá trị bị thiếu hoặc các giá trị ngoại lệ. Việc gộp nhóm các giá trị liên quan lại với nhau trong các thùng cũng để giảm số lượng giá trị riêng biệt. Ngoài ra, việc tạo nhóm có thể cải thiện chất lượng mô hình bằng cách tăng cường mối quan hệ giữa các thuộc tính. Nó cũng có thể được sử dụng trong phân tích đa biến bằng cách gộp nhiều đặc điểm cùng một lúc để đơn giản hóa việc phân tích của mình.

Trong khoa học Binning thường được sử dụng như một từ đồng nghĩa với việc nhóm hoặc phân loại như [Data binning](https://en.wikipedia.org/wiki/Data_binning) (một kỹ thuật tiền xử lý dữ liệu), [Binning (metagenomics)](https://en.wikipedia.org/wiki/Binning_(metagenomics)) (quá trình phân loại các lần đọc thành các nhóm hoặc phân loại khác nhau) [Phân loại sản phẩm](https://en.wikipedia.org/wiki/Product_binning)  (trong chế tạo thiết bị bán dẫn, quá trình phân loại thành phẩm). Trong xử lý ảnh kỹ thuật số, "binning" có một ý nghĩa rất khác. Bining pixel là quá trình kết hợp các khối pixel liền kề trong toàn bộ hình ảnh, bằng cách tính tổng hoặc lấy trung bình các giá trị của chúng, trong hoặc sau khi đọc.

Đối với binning dữ liệu các mẫu metagenomic có thể chứa các kết quả đọc từ một số lượng lớn sinh vật. Trong hầu hết các trường hợp, bản chất không hoàn chỉnh của các trình tự thu được khiến việc tập hợp các gen riêng lẻ trở nên khó khăn, càng khó phục hồi toàn bộ bộ gen của mỗi sinh vật. Do đó, các kỹ thuật binning là cách thể hiện tốt nhất để xác định và quan sát các dữ liệu.

Metagenomic đóng một vai trò rất quan trọng đối với nền y học. Chúng ta có thể tìm thấy các nghiên cứu về phần mềm tự động để phân tích dữ liệu giải trình tự 16S rRNA như MetaDP của Xilin Xu et al [13]. Bài viết nghiên cứu phương pháp bao gồm kiểm soát chất lượng dữ liệu, phân cụm đơn vị phân loại vận hành, phân tích đa dạng và mô hình dự đoán rủi ro bệnh tật. Ngoài ra, một mô hình dự đoán dựa trên SVM được cấu hình đã được xây dựng cho hội chứng ruột kích thích (IBS). Mục đích thiết yếu của dịch vụ web MetaDP là cung cấp hệ thống phân tích tự động thân thiện với người dùng, trong đó người dùng chỉ cần tải lên dữ liệu thô được tạo từ nền tảng giải trình tự thông lượng cao. Kết quả bài viết cho thấy độ chính xác và điểm AUC lần lượt là 0,93 và 0,95. Thành công của mô hình dự đoán IBS cho thấy nền tảng này cũng có thể được áp dụng cho các bệnh khác liên quan đến vi khuẩn đường ruột, chẳng hạn như béo phì, hội chứng chuyển hóa hoặc ung thư đường ruột, cùng nhiều bệnh khác.

Hoặc nghiên cứu về DisBalance của Fenglong Yang et al [14], được thiết kế như một nền tảng trực tuyến để giải quyết các vấn đề chính liên quan đến phân loại nhị phân dựa trên hệ vi sinh vật, bao gồm xây dựng mô hình bệnh tật, dự đoán rủi ro bệnh tật và khám phá dấu ấn sinh học vi khuẩn. Dựa trên mô hình bệnh tật, người ta có thể dự đoán nguy cơ mắc bệnh và khám phá các dấu ấn sinh học của vi khuẩn. Các mối liên hệ giữa vi khuẩn-bệnh mới có thể được suy ra bằng cách kết hợp các số dư hàng đầu và các vi khuẩn liên quan đến bệnh đã được chứng minh. Tính đến tháng 9/2021, DisBalance được xem là giải pháp tốt nhất cho các vấn đề liên quan đến phân loại nhị phân dựa trên hệ vi sinh vật. Tuy nhiên, vì nó dựa trên DBA-distal, vẫn chưa giải quyết được hoàn toàn vấn đề đếm dữ liệu hệ vi sinh vật, nên DisBalance được khuyến nghị áp dụng cho các tập dữ liệu có cỡ mẫu lớn hơn.

Một chương trình khác cũng chẩn đoán IBD dựa trên dữ liệu hệ vi sinh vật đường ruột của con người đó là LightCUD [15]. Dữ liệu giải trình tự toàn bộ bộ gen (WGS) thông lượng cao để phân tích thành phần vi sinh vật của các mẫu hệ vi sinh vật đường ruột. Ngoài ra, bài báo đã so sánh hiệu suất của năm thuật toán học máy khác nhau, là hồi quy logistic, rừng ngẫu nhiên, phân loại tăng cường độ dốc, máy vectơ hỗ trợ và LightGBM để đào tạo từng mô hình của mô-đun tương ứng.Với dữ liệu WGS hoặc dữ liệu giải trình tự 16S rRNA của các mẫu hệ vi sinh vật đường ruột làm đầu vào, LightCUD có thể phân biệt IBD với các biện pháp kiểm soát lành mạnh với độ chính xác cao và xác định thêm loại IBD cụ thể. LightCUD dự đoán khả năng mắc IBD và mẫu được xác định là IBD sau đó sẽ được phân loại là UC hoặc CD. Gần đây, một nghiên cứu về tự động dự đoán bệnh từ dữ liệu metagenomic trong ruột người bằng cách tăng cường GraphSAGE. Khung đề xuất có hai thành phần chính, module xây dựng Metagenomic Disease graph (MD-graph) và Disease prediction Network (DP-Net). Biểu đồ thể hiện mối quan hệ giữa các mẫu bằng thước đo độ gần. DP-Net bao gồm một mô hình GraphSAGE tăng cường dự đoán trạng thái của mẫu là ốm hay khỏe mạnh. Nghiên cứu gồm 6 bộ dữ liệu metagenomic bao gồm 5 bệnh (tiểu đường loại 2 (T2D), béo phì, xơ gan, CRC và IBD). Hiệu quả của phương pháp đề xuất được xác minh bằng cách sử dụng bộ dữ liệu thực và tổng hợp tương ứng với các bệnh như bệnh viêm ruột và ung thư đại trực tràng. Mô hình đề xuất đạt AUC cao nhất là 93%, Độ chính xác là 95%, điểm F1 là 95%.

Cũng có một vài nghiên cứu tổng hợp về các mô hình thống kê hệ vi sinh vật đường ruột để theo dõi tình trạng sức khỏe cá nhân như bài viết của Jinlin Zhu et al [16]. Mô hình giám sát thống kê được thiết lập, chỉ số sức khỏe được xây dựng và ranh giới sức khỏe được xác định, sau đó các mô hình lành mạnh được xác định ở những người khỏe mạnh và được phân tích bằng cách sử dụng phương pháp học tương phản để thấy được đóng góp của từng loại vi khuẩn vào chỉ số sức khỏe của dân số mắc bệnh đã được phân tích. Bài viết đã nghiên cứu tổng hợp các kết quả dự đoán sức khỏe bằng các phương pháp khác nhau liên quan đến hiPCA (Health index with PCA) và GMHI (gut microbiota health index) nhờ đó tìm ra các nguyên nhân gây bệnh được quan tâm nhiều nhất. Mặc dù hiPCA có một số hạn chế tuy nhiên, hiPCA tạo điều kiện thuận lợi rất nhiều cho việc đánh giá cá nhân về tình trạng sức khỏe và xác định các dấu ấn sinh học tiềm năng. Thêm vào đó một nghiên cứu hoàn toàn mới để tăng cường học sâu metagenomic cho bệnh tật dự đoán và nhận dạng chữ ký nhất quán bằng cách tái cấu trúc biểu diễn microbiome 2D. Ở đây, một thuật toán nhúng, nhóm và ánh xạ vi sinh vật không được giám sát (MEGMA) đã được phát triển để chuyển đổi dữ liệu metagenomic thành biểu diễn microbiome 2D đa kênh được cá nhân hóa bằng cách học đa dạng và phân cụm các cấu hình vi sinh vật. Các biểu diễn 2D này cho phép dự đoán bệnh nâng cao bằng các mô hình AggMapNet dựa trên ConvNet đã được thiết lập, vượt trội so với các mô hình học máy và học sâu thường được sử dụng trong bộ dữ liệu điểm chuẩn metagenomic. Kết quả cho thấy rằng những ưu điểm nổi bật của đường dẫn MEGMAAggMapNet-GFI trong dữ liệu metagenomic có thể có khả năng tìm thấy ứng dụng rộng rãi trong các lĩnh vực y sinh khác nhau.

Những năm gần đây, xuất hiện phần lớn các nghiên cứu về phương pháp Metagenomic Next-Generation Sequencing (mNGS). Vào năm 2022, Mengyi Zhao et al [17] đã cho ra bài viết về Giá trị chẩn đoán của giải trình tự metagenomic thế hệ tiếp theo để xác định nhiễm trùng Pneumocystis jirovecii ở bệnh nhân suy giảm miễn dịch không nhiễm HIV. 34 bệnh nhân suy giảm miễn dịch không nhiễm HIV được chẩn đoán là PJP qua biểu hiện lâm sàng. Các mẫu huyết tương và các mẫu dịch cơ thể được xử lý mẫu và tách chiết DNA. Sau đó, thư viện DNA được xây dựng bằng cách phân mảnh DNA, sửa chữa các đầu cùn do phân mảnh gây ra, nối các bộ điều hợp để khuếch đại và giải trình tự, cũng như khuếch đại PCR để làm phong phú các đoạn mục tiêu. Ngoài ra, bài báo còn so sánh một số yếu tố có thể ảnh hưởng đến mức độ đọc P. jirovecii được phát hiện bởi mNGS. So với xét nghiệm vi sinh lâm sàng, tỷ lệ phát hiện mNGS đối với P. jirovecii ở bệnh nhân PJP không nhiễm HIV cao hơn đáng kể so với nhuộm bạc Methenamine và huyết thanh 1-3-β-D-glucan. Xuan Zhang et al [18] dã nghiên cứu Hiệu suất chẩn đoán của trình tự thế hệ tiếp theo Metagenomic ở bệnh nhi. Nghiên cứu đã tiến hành giải trình tự mNGS (metaDNA-seq) và meta-transcriptomic (metaRNA-seq) cho 640 mẫu lâm sàng. Phân tích tin sinh học và định nghĩa độ nhạy, độ đặc hiệu, giá trị tiên đoán dương và giá trị tiên đoán âm trên các mẫu. Kết quả cho thấy mNGS có thể cung cấp độ nhạy cao hơn, từ đó cho phép chẩn đoán sớm hơn các mầm bệnh khó khăn và tốn thời gian, với ưu điểm rõ ràng là có khả năng phát hiện nhiễm trùng đa vi khuẩn chỉ trong một xét nghiệm. Đối với bệnh nhân nhi trong tình trạng nguy kịch bị nghi ngờ nhiễm trùng, xét nghiệm mNGS có thể cung cấp thông tin chẩn đoán có giá trị để giải quyết các kết quả xét nghiệm thường quy âm tính hoặc không kết luận, phân biệt ID với các trường hợp NID và tạo điều kiện cho việc ra quyết định điều trị lâm sàng chính xác và hiệu quả.

Cũng trong năm này, Jun Li et al [19] cũng cho ra bài viết về Giá trị chẩn đoán của trình tự metagenomic thế hệ tiếp theo đối với bệnh viêm phổi ở bệnh nhân suy giảm miễn dịch. Mẫu vật được chia thành các phần nhỏ và được ghép nối để xét nghiệm CT và mNGS. Chẩn đoán lâm sàng cuối cùng dựa trên biểu hiện lâm sàng, xét nghiệm, X quang ngực, xét nghiệm vi sinh (bao gồm CT và mNGS). Với bộ dữ liệu gồm 137 bệnh nhân thực hiện phát hiện mNGS, trong đó 53 bệnh nhân được sàng lọc. Nghiên cứu này mô tả sự phân bố của mầm bệnh ở những bệnh nhân bị suy giảm miễn dịch và có khả năng miễn dịch bình thường, đồng thời so sánh giá trị chẩn đoán của mNGS với CT trong nguyên nhân gây viêm phổi. mNGS có tỷ lệ chính xác tương đương với CT trong chẩn đoán nhiễm trùng và đồng nhiễm do vi khuẩn, virus, nấm và nấm. Tuy nhiên, ở những bệnh nhân bị suy giảm miễn dịch, mNGS có độ nhạy cao hơn CT trong việc phát hiện nhiễm trùng do vi khuẩn và nhiễm trùng hỗn hợp. Nó có thể phát hiện nhiều loại vi khuẩn gây bệnh mà CT không thể phát hiện được. Mới đây nhất, tác giả et al chẩn đoán bệnh phổi do Mycobacteria không lao bằng phương pháp giải trình tự thế hệ tiếp theo Metagenomic trên dịch rửa phế quản phế nang. Bằng cách xét nghiệm mNGS dựa trên DNA đối với các mẫu BALF của từng bệnh. Hiệu suất chẩn đoán của mNGS, AFS và nuôi cấy NTMPD đã được tính toán, bao gồm độ nhạy, độ đặc hiệu, giá trị tiên đoán dương (PPV) và giá trị tiên đoán âm (NPV). Hơn nữa, các đường cong đặc tính vận hành máy thu (ROC) đã được vẽ và diện tích dưới đường cong (AUC) đã được tính toán. Trong chẩn đoán NTMPD, độ nhạy chẩn đoán của mNGS (90,2%) cao hơn đáng kể so với phương pháp nuôi cấy phương pháp chẩn đoán thông thường. mNGS có thời gian xử lý ngắn hơn, từ đó góp phần chẩn đoán sớm và điều trị chính xác. Hơn nữa, mNGS của mẫu BALF có thể phát hiện đồng nhiễm phổi.

Dựa vào phương pháp binning, người ta có thể nghiên cứu nhiều bài báo về kỹ thuật này và áp dụng vào đa dạng lĩnh vực. Như trong lĩnh vực du lịch, bài báo của Malika Acharya et al [20] đề xuất mô hình dựa trên việc tạo nhóm không gian tập trung vào vị trí bằng cách sử dụng mạng xã hội. Việc phân nhóm không gian tập trung vào vị trí là bước đầu tiên của phương pháp, trong đó họ tính đến vị trí địa lý của người dùng và sau đó nhóm các địa điểm gần với vị trí đó. Tiếp theo bước này là nhiệm vụ đề xuất dựa trên Long Short-Term Memory (LSTM). Đề xuất bao gồm POI đủ điều kiện và danh mục địa điểm dựa trên vị trí. Phương pháp được đề xuất là một nỗ lực nhằm giảm bớt nỗ lực nhận thức của con người và tính đa dạng của đề xuất mô hình nhằm nâng cao hệ thống đề xuất POI. Nó xem xét các POI khác nhau trong các vùng không gian dựa trên mức độ phổ biến của các lượt truy cập. Độ chính xác cao và tiêu tốn ít thời gian hơn của nó vượt trội hơn tất cả các phương pháp khác. Một nghiên cứu khác cùng lĩnh vực đề xuất sử dụng cách tạo nhóm hình lục giác như một phương pháp trực quan hóa tương tác và đánh giá ba chỉ số tạo nhóm nhanh chóng khác nhau để lập bản đồ georeferenced social media data (GSMD) như các chức năng phân loại và phân loại màu sắc khác nhau, mỗi chức năng phù hợp với các trường hợp sử dụng khác nhau. Bài viết đã đánh giá ưu điểm và nhược điểm của số liệu được đề xuất cũng như hình ảnh trực quan trong việc hình ảnh hóa GSMD, đặc biệt là mạng xã hội Instagram. Nghiên cứu điển hình này sử dụng mẫu gồm 946.955 bài đăng từ Instagram từ năm 2011 đến tháng 10/2022. Dấu hiệu trực quan cho thấy Misclassification NNC có khả năng làm sạch các thiên hà bị ô nhiễm tốt hơn được định lượng bằng phần ô nhiễm.

Với lĩnh vực nông nghiệp, để giảm nhu cầu về kích cỡ của mẫu, bài viết của Zhe Ma et al [21] đề xuất một phương pháp phân loại không giám sát dựa trên thành phần chính là principal components isometric binning (PCIB). Cụ thể, việc giảm kích thước của principal component analysis (PCA) được áp dụng cho các tính năng phân loại, sau đó là phân chia k thành phần chính hàng đầu thành các ngăn cách đều nhau. Các thùng cùng loại sau đó sẽ được hợp nhất thành nhãn lớp. PCIB có độ chính xác phân loại cao nhất, đạt 82% với 322 và 455 mẫu vào năm 2016 ở phía tây nam thành phố Hulin và 81% với 383 mẫu tại huyện Luobei năm 2016. Phương pháp PCIB được đề xuất trong bài báo có thể thu được thông tin phân bố không gian của cây trồng một cách chính xác và kịp thời, có ý nghĩa rất lớn trong việc định hướng sản xuất nông nghiệp.

Trong lĩnh vực thiên văn học, nổi bật là bài báo tối ưu hóa việc temporal binning sao so sánh theo thời gian cho phép đo trắc quang vi sai của Kathryn E Hartley et al [22]. Bài viết mô tả quy trình giảm dữ liệu để triển khai phương pháp này nhằm tối ưu hóa số lượng khung hình được tạo cho ngôi sao so sánh và trình bày các kết quả mẫu cho dữ liệu trắc quang được phân giải theo thời gian. Kỹ thuật giảm dữ liệu được mô tả ở đây làm giảm đáng kể Tỷ lệ nhiễu trên tín hiệu (NSR) cho phép đo quang vi sai bằng cách tận dụng thực tế là công suất trong các xu hướng hệ thống thường ở tần số thấp so với nhịp của đường cong ánh sáng. Một trong những ưu điểm chính của phương pháp tạo nhóm này là nó cho phép sử dụng các ngôi sao so sánh mờ hơn nhiều. Ngoài ra, bài viết đã chứng minh rằng kỹ thuật này hoạt động với phạm vi giá trị nhịp lên tới 30 giây. Đối với một mục tiêu sáng, có độ nhấp nháy hạn chế và ngôi sao so sánh, việc tạo khối tạm thời cho ngôi sao so sánh có thể làm giảm NSR của đường cong ánh sáng đã hiệu chỉnh theo hệ số √2. Theo các tác giả nghiên cứu về phương pháp Binning với High-Latitude Electrodynamic Forcing vào năm 2020 [23]. Trong nghiên cứu này, dữ liệu lượng mưa điện tử và điện thế từ các vệ tinh của Chương trình Vệ tinh Khí tượng Quốc phòng (DMSP) được phân tích trong các điều kiện khi từ trường liên hành tinh (IMF) mạnh vừa phải và chủ yếu hướng về phía nam. Đầu tiên, các vĩ độ từ tính thống kê (MLAT) của ranh giới cực quang và ranh giới đảo ngược đối lưu (CRB) như là một hàm của thời gian địa phương từ tính (MLT) đã được nghiên cứu. Thứ hai, dữ liệu kết tủa hạt và điện thế được xử lý thông qua các phương pháp tĩnh và định hướng biên và kết quả được so sánh về mặt định lượng. Cuối cùng, cả kết quả tạo thùng tĩnh và định hướng ranh giới đều được sử dụng để điều khiển mô hình tầng điện ly-nhiệt quyển (GITM) toàn cầu nhằm đánh giá tác động của các kiểu lực điện động thu được từ các phương pháp tạo thùng khác nhau đối với hệ thống sưởi Joule. Độ lớn của điện thế và điện trường trong kết quả tạo nhóm định hướng ranh giới tăng lên gần ranh giới đảo ngược đối lưu, dẫn đến tăng cường 11% điện thế nắp cực chéo. Hệ thống sưởi Joule tích hợp bán cầu trong mô phỏng được điều khiển bởi các mẫu tạo nhóm theo hướng ranh giới cao hơn 18% so với hệ số được điều khiển bởi các mẫu tạo nhóm tĩnh.

Trong lĩnh vực Paleontology và khoa học trái đất, bài báo của Christopher D. Dean et al [24] nghiên cứu về việc tạo nhóm hình thành, một phương pháp mới để tăng độ phân giải thời gian trong các nghiên cứu khu vực, áp dụng cho hồ sơ hóa thạch khủng long kỷ Phấn trắng muộn ở Bắc Mỹ. Bài viết đề xuất một phương pháp mới để phân chia thời gian địa chất cho các nghiên cứu trong khu vực về bản chất kết hợp tính không đồng nhất về thời gian địa tầng của các sự hình thành khác nhau để tạo ra các thùng địa tầng độc đáo. Việc tạo nhóm hình thành sử dụng ngày bắt đầu và ngày kết thúc của các hình thành để đánh giá vị trí phù hợp nhất để vẽ ranh giới ngăn thời gian. Các lần xuất hiện hóa thạch từ toàn bộ hệ tầng được chỉ định vào các thùng tùy thuộc vào tiêu chí đã chọn của người dùng. Kết quả cho thấy sự suy giảm đa dạng khủng long Bắc Mỹ ở Maastrichtian xảy ra trong khoảng thời gian từ 68 đến 66 Ma, đồng thời với việc tái tổ chức địa động lực rộng rãi của các lưu vực trầm tích và sự rút lui tổng thể của Western Interior Seaway.

Với lĩnh vực của khoa học dữ liệu, trong nghiên cứu của Liang Zhong et al [15], lần đầu tiên, một phương pháp lựa chọn biến dựa trên entropy thông tin của “binning-normalized mutual information” (B-NMI) đã được đề xuất để hiệu chuẩn quang phổ đa thành phần. Sự kết hợp của hai phương pháp cho phép tính toán tối đa mối quan hệ giữa các biến phổ và giá trị tham chiếu, bao gồm cả các quan hệ tuyến tính và phi tuyến tính. Hơn nữa, để minh họa tính ưu việt của nó, phương pháp B-NMI đã được so sánh với năm phương pháp chọn biến cổ điển (BIPLS, VIP, CC, UVE, CARS). Phương pháp B-NMI cho thấy khả năng dự đoán tốt hơn trong các bộ dữ liệu này nhờ trích xuất tính năng hiệu quả và phát triển mô hình có mức độ phù hợp cao, đặc biệt là trong việc xử lý các mẫu phức tạp trong thế giới thực. Phương pháp B-NMI không chỉ có thể xác định và loại bỏ các biến không liên quan một cách hiệu quả mà còn loại bỏ các biến dư thừa bằng cách đánh giá tất cả các kết quả xác suất được tính toán bằng tìm kiếm toàn diện.

Trong lĩnh vực hình ảnh y học, Alexander Fyrdahl et al [25], chuỗi xung steady-state free precession (SSFP) cân bằng đã được triển khai với thứ tự xuyên tâm góc vàng (3D-SWIG) mới theo ngành 3D. Trong đó không gian k được chia thành các phần và mỗi phần được thu thập trong một nhịp tim riêng biệt. Tính đồng nhất cao của vùng bao phủ không gian k sau khi tạo nhóm sinh lý có thể được sử dụng để thực hiện hình ảnh chức năng bằng cách sử dụng thời gian thu nhận rất ngắn. Mô phỏng số cho thấy rằng 3D-SWIG đều yêu cầu các bước nhỏ hơn giữa các lần đọc liên tiếp và đạt được độ đồng đều lấy mẫu không gian k tốt hơn sau khi tạo thùng so với quỹ đạo góc vàng kép hoặc quỹ đạo phyllotaxis xoắn ốc. Đánh giá in vivo cho thấy rằng các phép đo phân suất tống máu thất trái được tính toán từ dữ liệu thu được 3D-SWIG nhịp thở tự do trong 48 nhịp tim có khả năng tái sản xuất cao và phù hợp với phim 2D-Cartesian được giữ trong hơi thở tự do với thời gian thu nhận dưới 1 phút. Hay bài viết của Irene Moskowitz et al [26] về cải tiến việc gộp ảnh chụp cắt lớp của các mẫu thấu kính 3 × 2 pt bằng Neural Network Classifiers và Optimal Bin Assignments. Bài báo đề xuất một phương pháp tối ưu hóa lựa chọn tạo ảnh chụp cắt lớp cho mẫu thấu kính của các thiên hà. Chia danh mục thiên hà mô phỏng thành tập huấn luyện và tập ứng dụng, sau đó ước tính độ lệch đỏ trắc quang cho các bộ ứng dụng. Các thiên hà được sắp xếp vào các thùng dịch chuyển đỏ bao gồm các khoảng dịch chuyển đỏ hoặc khoảng cách di chuyển bằng nhau hoặc với số lượng thiên hà bằng nhau trong mỗi thùng và xem xét phần mở rộng tổng quát của các phương pháp này. Sau đó, đào tạo bộ phân loại mạng thần kinh để xác định các thiên hà có nhiều khả năng có ước tính dịch chuyển đỏ trắc quang chính xác hoặc có nhiều khả năng được sắp xếp vào thùng dịch chuyển đỏ chính xác. Nhìn chung, Misclassification NNC loại bỏ các thiên hà gây ô nhiễm tốt hơn Outlier NNC. Các neural network classifiers có thể cải thiện giá trị thành tích lên 13% và có thể phục hồi 25% tổn thất về giá trị xảy ra khi mẫu huấn luyện không mang tính đại diện được sử dụng.

Phương pháp binning cũng áp dụng vào lĩnh vực gần gũi nhất với con người đó là giáo dục. Bài viết của tác giả D. Weckmüller et al [27] nghiên cứu sự hiệu quả của việc học tập dựa trên trò chơi đối với việc hiểu về Binning ở học sinh lớp 1. Nghiên cứu này nhằm mô tả các giai đoạn lập kế hoạch, thực hiện, quan sát và phản ánh trong quá trình học tập phân nhóm ở học sinh lớp 1 tiểu học. Loại nghiên cứu này là nghiên cứu Hành động trong lớp học (CAR) được thực hiện để xác định kết quả học tập, khả năng sáng tạo của học sinh, thực hiện việc học tập dựa trên trò chơi trên các tài liệu đóng thùng. Dựa trên kết quả quan sát việc học tập dựa trên trò chơi có thể nâng cao khả năng sáng tạo, hứng thú và hiểu biết của học sinh. Bên cạnh đó, khả năng hiểu bài của học sinh khá tốt, có thể thấy điều này qua kết quả bài kiểm tra đánh giá các câu hỏi xếp hình, 80% học sinh trả lời đúng các câu hỏi được đưa ra.

Phổ biến nhất là các nghiên cứu về phương pháp binning metagenomic. Tác giả Jakob Nybo Nissen et al [28] đã đề xuất cải thiện quá trình binning và lắp ráp metagenome bằng cách sử dụng deep variational autoencoders (Deep VAE). Họ sử dụng những tiến bộ gần đây trong học sâu để phát triển một thuật toán sử dụng bộ mã hóa tự động biến thiên để mã hóa sự phong phú và thành phần thông tin trước khi phân cụm và nhóm các biểu diễn tiềm ẩn thu được bằng cách sử dụng phép lặp thuật toán phân cụm medoid. Phương pháp đã áp dụng cho danh mục gen của gần 10 triệu gen và 1.270 mẫu từ ruột người hệ vi sinh vật. Kết quả cho thấy rằng có thể được thực hiện với dữ liệu chuỗi và dữ liệu đồng phong phú mà không cần phát triển các mô hình thống kê cho từng kiểu dữ liệu. Ngoài ra, Variational Autoencoder for Metagenomic Binning (VAMB) còn có thể tái tạo lại các thùng ít phổ biến hơn so với các thùng chứa khác. Ở đây bài viết đã xác minh hiệu suất tăng lên của VAMB hiển thị nhiều ngăn hơn để phân công phân loại cụ thể ở tất cả các cấp. Tác giả Georges P Schmartz et al [29] đã đề xuất nghiên cứu thực hiện bằng cách sử dụng các kỹ thuật học sâu hiện đại, đặc biệt là GNN (Graph Neural Network). Ý tưởng chính đằng sau GraphMB là tạo ra các phần nhúng dựa trên các tính năng cụ thể tiếp giáp và biểu đồ tập hợp. Mô hình GNN tìm hiểu các phần nhúng mới bằng cách tổng hợp các đường viền lân cận (các nút trong biểu đồ tập hợp). Các phần nhúng cuối cùng được nhóm lại và thu được các thùng sau đó được nhóm vào các thùng và được đánh giá theo mức độ đầy đủ và mức độ ô nhiễm. Bài viết cải thiện các thùng hiện có bằng cách kết hợp biểu đồ lắp ráp vào quá trình đào tạo với 6 tập dữ liệu về Nhà máy xử lý nước thải (WWTP) và một mẫu đất. Hơn nữa, nó còn thu được những chiếc thùng độc đáo mà những chiếc thùng khác không thể phục hồi được. Đặc biệt, đã thu được trung bình nhiều hơn 17,5% thùng HQ khi so sánh với các thùng chứa hiện đại. Những kết quả này chỉ ra rằng một mô hình học sâu có thể tích hợp thông tin về cấu trúc biểu đồ và đặc trưng của contig để cải thiện việc tạo nhóm metagenomic. Tác giả đã tìm hiểu về BusyBee Web, một công cụ tạo binning metagenomic dựa trên thành phần toàn diện và khác biệt. Đường dẫn cơ bản đào tạo bộ phân loại trên một tập hợp con của dữ liệu đầu vào, sau đó được sử dụng để gán các chuỗi vào các thùng. Hơn nữa, một biểu đồ sunburst đã được thêm vào và một số cài đặt chuyên môn mới cho các phương pháp phân cụm và nhúng đã được triển khai. Cụ thể đã bao gồm t-SNE, Fit-SNE, UMAP, PHATE và TriMap làm nhúng và DBSCAN, HDBSCAN, k-means và phân cụm quang phổ là phương pháp phân cụm mới. Một mặt, với các phương pháp phân cụm, thuật toán nhúng và API mới được bổ sung, BusyBee Web đã tăng khả năng phân tích dữ liệu cho người dùng chuyên gia và tùy chọn nén thông tin trước khi nhúng tính toán, giảm bớt một số hạn chế về thời gian và bộ nhớ. Tuy nhiên, việc trực quan hóa việc nhúng vào trình duyệt cục bộ cho nhiều điểm dữ liệu có thể trở nên chậm hoặc không phản hồi trên phần cứng kém mạnh mẽ hơn. Vào năm 2020, tác giả Silas Kieser et al [30] đã nghiên cứu về ATLAS, một quy trình làm việc của Snakemake để lắp ráp, chú thích và binning gen của dữ liệu trình tự metagenome. Khung ATLAS tổ chức các công cụ xử lý dữ liệu metagenome theo trình tự thành bốn mô-đun phân tích riêng biệt: kiểm soát chất lượng, lắp ráp, binning gen và chú thích; mỗi mô-đun có thể được chạy độc lập hoặc kết hợp cả bốn mô-đun trong một quy trình phân tích hoàn chỉnh. ATLAS cũng cung cấp khung tin sinh học mạnh mẽ cho dữ liệu chuỗi có thông lượng cao, trong đó các tệp FASTQ thô có thể được xử lý hoàn toàn thành các tệp dạng bảng có chú thích để phân tích và hiển thị tiếp theo. Vào năm 2022, tác giả Yuxuan Du et al [31] đã giới thiệu về HiFine, một quy trình tạo nhóm mới để tinh chỉnh kết quả binning của các đường viền metagenomic. Đầu tiên, thiết kế chiến lược phân mảnh bằng cách chọn ra các điểm giao nhau dưới dạng các thùng phân mảnh giữa hai bộ thùng được xây dựng bằng phương pháp tạo thùng dựa trên Hi-C và công cụ tạo thùng dựa trên shotgun. Sau đó, HiFine hợp nhất các thùng phân mảnh có khả năng thuộc cùng một loài và tuyển dụng các nhánh không có trong các thùng phân mảnh bằng cách sử dụng lại bản đồ liên hệ Hi-C đã chuẩn hóa. Bộ dữ liệu được sử dụng nghiên cứu gồm bộ dữ liệu nấm men, bộ dữ liệu bia và bộ dữ liệu về ruột người được lấy từ mẫu phân. HiFine đã đạt được sự cải thiện đáng kể về kết quả tạo thùng trên mẫu thực có độ phức tạp thấp hiện có của cả hai loại phương pháp tạo thùng và hoạt động tốt hơn trong việc truy xuất bộ gen của loài. Tuy nhiên, phương pháp cũng có những hạn chế riêng rằng HiFine có thể không có sự cải thiện đáng kể nếu một trong các phương pháp tạo nhóm hiện có chỉ nhóm một số lượng đường viền nhỏ hơn đáng kể so với loại phương pháp tạo nhóm khác. Trong một nghiên cứu khác, tác giả Vijini Mallawaarachchi et al [32] đề xuất MetaCoAG, một công cụ tạo nhóm metagenomic sử dụng các biểu đồ lắp ráp với thông tin về thành phần và phạm vi bao phủ của các đường viền. MetaCoAG ước tính số lượng thùng ban đầu bằng cách sử dụng các gen đánh dấu một bản sao đơn, lặp đi lặp lại gán các đường viền vào các thùng và điều chỉnh linh hoạt số lượng thùng trong suốt quá trình tạo thùng. MetaCoAG vượt trội đáng kể so với các công cụ tạo nhóm hiện đại bằng cách tạo ra các thùng tương tự chẳng hạn như cải thiện tính hoàn chỉnh của các thùng trong khi vẫn duy trì mức độ tinh khiết cao và tạo ra nhiều thùng chất lượng cao hơn. Từ kết quả thử nghiệm cho thấy MetaCoAG đạt được hiệu suất tạo nhóm tốt nhất cho cả tập dữ liệu mô phỏng và tập dữ liệu thực khi so sánh với các công cụ tiên tiến, đặc biệt là về chất lượng nhóm. Tuy nhiên, các vấn đề trong quá trình lắp ráp như lắp ráp sai có thể khó xử lý và cần được điều tra thêm.

Cách đây khá lâu, tác giả Yunan Luo et al [33] đã đề xuất bài báo đã trình bày Opal, một phương pháp dựa trên thành phần mới để tạo nhóm metagenomic. Bằng cách rút ra ý tưởng từ các mã low-density parity check (LDPC) của Gallager từ lý thuyết mã hóa, bài viết đã thiết kế một nhóm hàm LSH hiệu quả và có tính phân biệt để xây dựng các đặc điểm cấu thành nhằm nắm bắt các mối phụ thuộc tầm xa trong các đoạn metagenomic. OPAL nhanh và chính xác hơn nhiều so với các phương pháp căn chỉnh, vượt trội hơn các phương pháp tạo nhóm metagenomic hiện có tại thời điểm đó. Opal cũng hoạt động đặc biệt tốt trong việc phát hiện dòng dõi mới. Ngoài ra, Opal có hiệu quả ở cấp độ phân loài, mặc dù nó cần được đào tạo bổ sung để có hiệu quả do sự tương đồng trong bộ gen của các loài liên quan.

Tác giả Joachim Johansen et al [34] đã giới thiệu Phages from Metagenomics Binning (PHAMB) như một cách tiếp cận cho phép binning hàng nghìn bộ gen virus trực tiếp từ dữ liệu metagenomics số lượng lớn, đồng thời cho phép phân cụm các bộ gen virus thành các quần thể virus được phân loại chính xác. Sử dụng thuật toán học sâu được phát triển gần đây để tạo nhóm metagenomic (VAMB) dựa trên việc tạo nhóm toàn bộ tập dữ liệu của các đường viền được lắp ráp. Khi áp dụng trên bộ dữ liệu Human Microbiome Project 2 (HMP2), PHAMB đã phục hồi 6.077 bộ gen chất lượng cao từ 1.024 quần thể vi rút và xác định các tương tác giữa vi r rút và vật chủ. Bộ phân loại có thể phân loại các thùng không phải thể thực khuẩn từ bất kỳ tập dữ liệu nào với độ chính xác rất cao (93–99%). Ngoài ra, bài báo cho thấy sự gia tăng lên tới 210% bộ gen virus HQ được trích xuất bằng cách kết hợp các nhánh vào thùng chứa virus.

Bộ tổng hợp metagenomic BinChill trong nghiên cứu của tác giả Oliver S. Bak et al [35], là một bộ tổng hợp sử dụng phương pháp tiếp cận sự xuất hiện của đối tượng dựa trên ACE, đồng thời triển khai thông tin miền vào phép đo độ tương tự, dưới dạng SCG để tập hợp và tinh chỉnh các thùng từ nhiều phân vùng. Hơn nữa, bài viết cũng triển khai một phiên bản độc lập của BinChill, trong đó thay vì dựa vào các phân vùng từ các phương pháp khác, nó tạo ra các phân vùng riêng để có thể khám phá các kỹ thuật khác nhau trong khuôn khổ. Kết quả cho thấy điểm mạnh của nhiều phân vùng có thể được kết hợp để tạo ra phân vùng có chất lượng cao hơn. Tuy nhiên, cải tiến đạt được bằng cách tiếp cận tổng hợp phải trả giá bằng thời gian chạy cao hơn, cho thấy khả năng mở rộng của phương pháp có thể được cải thiện.

Việc tích hợp thông tin về cấu trúc nhiễm sắc thể trong mô hình học tập tự giám sát cũng giúp cải thiện quá trình binning metagenome. Như bài nghiên cứu của tác giả Harrison Ho et al [36] đã đề xuất một chiến lược mới để tích hợp thông tin Hi-C vào khung học tập tự giám sát. MetaBAT-LR, phần mềm được sử dụng để xây dựng mô hình rừng ngẫu nhiên từ một tập hợp con của kết quả tạo thùng ban đầu. Sau đó, nó dự đoán khả năng kết nối giữa tất cả các cặp giàn giáo trong tập dữ liệu và sử dụng dự đoán để hợp nhất các thùng chứa bộ gen chưa hoàn chỉnh hoặc tuyển dụng các giàn giáo chưa được phân loại. MetaBAT-LR có thể tận dụng khả năng đọc Hi-C để cải thiện tính hoàn chỉnh của việc tạo thùng và khám phá các thùng mới bị các công cụ thay thế bỏ sót. MetaBAT-LR có khả năng cải thiện các thử nghiệm tạo thùng của tất cả các công cụ tạo thùng metagenome vì nó không phụ thuộc vào đầu ra của bất kỳ công cụ tạo thùng cụ thể nào. Một hạn chế trong phương pháp này là sự phụ thuộc của metaBAT-LR vào khả năng kết nối cặp đọc Hi-C tốt.

## Hướng tiếp cận giải quyết của đề tài

Một trong những vấn đề với K-mean là một thuật toán phân cụm cứng. Theo điều này, mỗi điểm chỉ được liên kết với một cụm. Do đó, không có xác suất có thể cho biết có bao nhiêu điểm dữ liệu được liên kết với một cụm cụ thể. Do đó, chúng tôi sử dụng phương pháp phân cụm mềm. Gaussian Mixture Models là một ứng cử viên hoàn hảo cho việc này. Mô hình hỗn hợp Gaussian sử dụng nhiều phân bố Gaussian để phù hợp với dữ liệu có hình dạng tùy ý.

Mô hình hỗn hợp Gaussian (GMM) là mô hình xác suất giả định tất cả các điểm dữ liệu được tạo ra từ hỗn hợp của một số phân phối Gaussian hữu hạn với các tham số chưa biết. Nhiều bộ dữ liệu có thể được mô hình hóa dễ dàng với sự trợ giúp của Gaussian Distribution. Do đó, người ta có thể giả định rằng các cụm có các Phân phối Gaussian khác nhau. Ý tưởng cốt lõi của mô hình là dữ liệu được mô hình hóa với một số hỗn hợp của các phân phối Gaussian [37]

Có hai loại giá trị tham số hóa Gaussian Mixture Model – trọng số thành phần và phương sai / hiệp phương sai. Gaussian Mixture Model với K thành phần, μk là giá trị trung bình của thành phần thứ k. Hơn nữa, trường hợp đơn biến sẽ có phương sai là σk trong khi trường hợp đa biến sẽ có ma trận hiệp phương sai là Σk. Φk là định nghĩa của trọng lượng thành phần hỗn hợp dành cho mọi thành phần Ck. Điều này có một ràng buộc rằng ∑Ki = 1 ϕi = 1 sao cho tổng xác suất được chuẩn hóa thành 1.

\* Ưu điểm:

Về tốc độ: GMM được coi là thuật toán nhanh nhất trong việc học mô hình hỗn hợp.

Khách quan: Bởi vì thuật toán này tối đa hóa chỉ xác suất hợp lý, nó không làm cho các giá trị trung bình tiến gần về không, cũng không làm cho kích thước của các cụm có các cấu trúc cụ thể mà có thể hoặc không áp dụng.

\* Nhược điểm:

Sự đơn điệu: Khi có quá ít điểm cho mỗi hỗn hợp, việc ước lượng ma trận hiệp phương sai trở nên khó khăn, và thuật toán có thể không ổn định và tìm ra các giải pháp có xác suất vô hạn trừ khi ta điều chỉnh nhân ký của hiệp phương sai một cách nhân tạo.

Số lượng thành phần: Thuật toán này luôn sử dụng tất cả các thành phần có sẵn, cần phải sử dụng dữ liệu được giữ lại hoặc các tiêu chí thông tin lý để quyết định số lượng thành phần cần sử dụng trong trường hợp không có dấu hiệu bên ngoài.

# THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP

## Kiến trúc tổng quát hệ thống

Trong bài niên luận này, mục tiêu của chúng tôi về việc phân cụm metagenomics là phân loại và gom nhóm dữ liệu metagenomic để hiểu cấu trúc, sự đa dạng và tổ chức của các cộng đồng vi sinh vật trong môi trường mẫu. Thông qua việc áp dụng các thuật toán phân cụm, mục tiêu này nhằm tìm ra các mẫu có đặc điểm tương đồng nhau hoặc có mối quan hệ gần gũi với nhau, từ đó giúp nhìn nhận được sự phân bố, cấu trúc và mối quan hệ giữa các loại vi sinh vật trong môi trường metagenomic đã nghiên cứu.

Quy trình và các bước quan trọng trong quá trình phân cụm bao gồm tiền xử lý dữ liệu, mô hình hóa mô hình Gaussian Mixture Models (GMM), xác định nhãn và tính toán bins\_break, huấn luyện, phân cụm và đánh giá mô hình. Một vài thư viện được sử dụng như scikit-learn, numpy, pandas, math, GaussianMixture.

Ở đây dữ liệu được chúng tôi thu thập từ trang web trực tuyến Kaggle, gồm tổng hợp sử dụng bao gồm 3302 cột và 3611 dòng. Bộ dữ liệu này được tạo bởi nhóm Edoardo Pasolli và các cộng sự, họ đã xuất bản một bài báo nghiên cứu vào tháng 7 năm 2016 và tạo ra MetAML - Phân tích dự đoán Metagenomic dựa trên Học máy [38]. Ban đầu, chúng tôi đã xem xét tổng cộng 3610 mẫu metagenomic có sẵn công khai (từ mười tám nghiên cứu/bộ dữ liệu chính). Sau khi xử lý trước đã giảm xuống còn 1644 mẫu và 6 bệnh khác nhau nằm trong 6 nghiên cứu được dành cho đặc điểm của hệ vi sinh vật đường ruột ở người khi mắc các bệnh khác nhau. Ngoài ra, các nghiên cứu còn lại tập trung vào các đối tượng khỏe mạnh và không liên quan chặt chẽ đến hệ vi sinh vật đường ruột đã được bỏ qua. Đa số các mẫu được thu thập từ mẫu phân của các bệnh nhân.

## Xây dựng các mô hình

Phân cụm metagenomics là quá trình nhằm phân loại hoặc gom nhóm các dữ liệu metagenomic để hiểu về cấu trúc và sự đa dạng của các cộng đồng vi sinh vật trong môi trường mẫu. Khi sử dụng thuật toán Gaussian Mixture Models (GMM) trong ngữ cảnh này, chúng tôi tạo ra một mô hình xác suất để phân loại hoặc gom nhóm các mẫu metagenomic dựa trên sự phân phối của chúng.

Đầu tiên, chúng tôi tiền xử lý dữ liệu metagenomic được thu thập và chuẩn bị để có thể sử dụng trong mô hình GMM. Chúng tôi đã xử lý thống nhất dữ liệu hệ vi sinh vật metagenomic cho 1644 mẫu từ 6 nghiên cứu về 6 loại bệnh. Xơ gan bao gồm 118 bệnh nhân bị xơ gan và 114 người khỏe mạnh. Kết quả bệnh ung thư đại trực tràng bao gồm tổng cộng 134 mẫu, 48 mẫu trong số đó bị bệnh ung thư đại trực tràng. IBD đại diện cho bộ dữ liệu metagenomic gồm 2 loại bệnh ibd ulcerative colitis và ibd crohn's disease với tổng cộng 173 mẫu bệnh, và 319 mẫu khỏe mạnh. Béo phì bao gồm 164 người béo phì và 89 người bị gầy và 25 mẫu khỏe mạnh. Thay vào đó, hai nghiên cứu riêng biệt liên quan đến sự thay đổi hệ vi sinh vật ở những đối tượng mắc bệnh tiểu đường loại 2 (T2D). Trong bộ dữ liệu T2D, có 170 bệnh nhân T2D và 193 người đối chứng không mắc bệnh tiểu đường. WT2D bao gồm 53 bệnh nhân T2D, 49 người bị suy giảm dung nạp glucose và 43 người dung nạp glucose bình thường.

*Bảng 1 Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | tổng số mẫu |  |  |
| Chatelier\_gut\_obesity | 278 | n | 25 |
|  |  | leaness | 89 |
|  |  | obesity | 164 |
| metahit | 492 | n | 319 |
|  |  | ibd\_ulcerative\_colitis + ibd\_crohn\_disease | 173 |
| Quin\_gut\_liver\_cirrhosis | 232 | n | 114 |
|  |  | cirrhosis | 118 |
| T2D | 363 | t2d | 170 |
|  |  | n | 193 |
| WT2D | 145 | n | 43 |
|  |  | impaired\_glucose\_tolerance | 49 |
|  |  | t2d | 53 |
| Zeller\_fecal\_colorectal\_cancer | 134 | n | 47 |
|  |  | cancer | 48 |
|  |  | large\_adenoma | 13 |
|  |  | small\_adenoma | 26 |

Bước 1: Xử lý tập dữ liệu và gắn nhãn các tập dữ liệu cần tạo thùng:

* Ở đây chúng tôi sử dụng 6 bộ dữ liệu, mỗi bộ chia thành 2 tập x và y.
* Dữ liệu ban đầu được làm phẳng thành một vector để chuẩn bị cho việc phân nhóm.
* Dữ liệu được biến đổi thành một cột duy nhất với mỗi phần tử của nó là một hàng
* Giá trị không mong muốn hoặc giá trị 0 được loại bỏ khỏi dữ liệu để làm sạch và cải thiện chất lượng dữ liệu.

Bước 2: Sử dụng mô hình Gaussian Mixture Model (GMM) được tạo và huấn luyện trên dữ liệu đã được xử lý.

* Mô hình phân phối Gaussian được áp dụng lên dữ liệu đã được xử lý để phân cụm nó thành các cụm có thể được coi là tương đồng với nhau dựa trên phân phối Gaussian.
* Cụ thể, n\_components là số lượng các cụm cần tạo, fit dùng để học mô hình trên dữ liệu, và predict để dự đoán nhãn của các mẫu, phân cụm dữ liệu với số cụm k = 10.
* Số lượng cụm được chọn là num\_bin -1.

Bước 3: Tính toán giá trị của các bins\_break (ngưỡng cho từng nhóm): Sử dụng hàm get\_bin\_break để tính toán bins\_break dựa trên dữ liệu đã xử lý

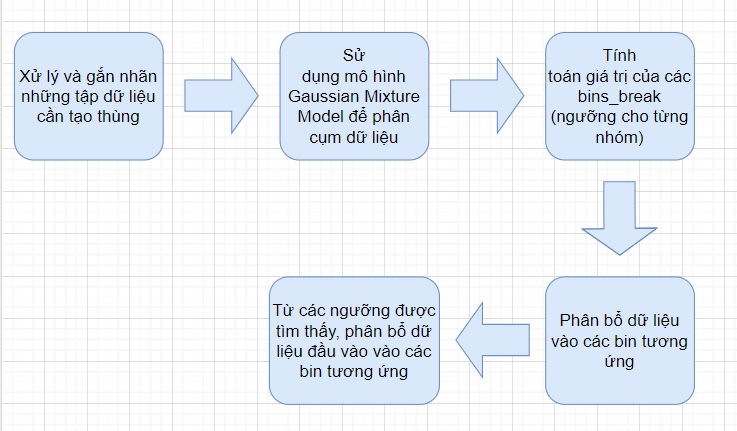
* Sắp xếp dữ liệu theo nhãn của cụm và giá trị của mỗi điểm.
* Từ các nhóm dữ liệu đã được phân, chúng ta xác định giá trị lớn nhất và giá trị nhỏ nhất của mỗi nhóm.
* Để tạo ngưỡng cho các bin tiếp theo, chúng ta tính giá trị trung bình của hai giá trị đầu mút của hai nhóm liền kề. Kết quả này sẽ là ngưỡng cho bin tiếp theo trong quá trình chia dữ liệu thành các bin.

Bước 4: Phân bổ dữ liệu vào các bin tương ứng:

* Dựa trên các ngưỡng được xác định từ bước trước, chúng ta phân bổ dữ liệu đầu vào vào các bin tương ứng. Các bin được tạo ra dựa trên các ngưỡng đã tính toán, giúp chia dữ liệu thành các phạm vi tương ứng với mỗi bin.

Bước 5: Từ các ngưỡng được tìm thấy, phân bổ dữ liệu đầu vào vào các bin tương ứng.

* Sử dụng các ngưỡng đã tính toán từ bước trước, chúng ta phân bổ dữ liệu đầu vào vào các bin tương ứng. Điều này giúp tạo ra các bin và phân loại dữ liệu vào các phạm vi tương ứng với mỗi bin.



*Hình 1. Sơ đồ các bước tạo nhóm bằng GMM*

## Giải pháp cài đặt

Trong bài niên luận này chúng tôi sử dụng chủ yếu ngôn ngữ lập trình python.

Thư viện sử dụng: scikit-learn cho GMM.

Môi trường phát triển: Anaconda (hoặc các môi trường Python khác)

Phiên bản Python: Python 3.x

* Chuẩn bị dữ liệu:
  + Chuẩn bị dữ liệu metagenomic theo định dạng phù hợp, ở đây tôi sử dụng CSV.
* Xây dựng và huấn luyện mô hình GMM:
  + Sử dụng GaussianMixture từ thư viện scikit-learn để xây dựng và huấn luyện mô hình GMM dựa trên dữ liệu metagenomic cho việc phân loại dữ liệu metagenomic vào các nhóm khác nhau.

# KIỂM THỬ VÀ ĐÁNH GIÁ

## Kịch bản kiểm thử

Trong bài niên luận này, mục tiêu của chúng tôi là kiểm tra chức năng chuyển đổi dữ liệu sang dạng nhị phân sử dụng GMM của cả 6 dữ liệu nằm trong tập tin data. Vì dữ liệu đã được chuyển đổi thử công từ trước nên chúng tôi sẽ không chuyển vị dữ liệu. Sau đó chọn cột dữ liệu thứ 2 và sử dụng phương pháp GMM để chuyển đổi dữ liệu sang dạng nhị phân. Sau khi nhúng nhị phân cho dữ liệu, chúng tôi chọn số lượng nhóm là 10 và số lần lặp là 500 (esochs) để huấn luyện dữ liệu. Bằng những hiển thị từ kết qủa cho thấy những thùng này có thể giúp cải thiện hiệu quả thực hiện các nhiệm vụ phân loại ở một vài dữ liệu.

Chúng tôi cũng thử nghiệm nhúng dữ liệu dạng raw, nhúng nhị phân và phương pháp GMM kết hợp cũng như sử dụng một mô hình RF và SVM với số lượng nhóm và số lần lặp tương tự để so sánh.

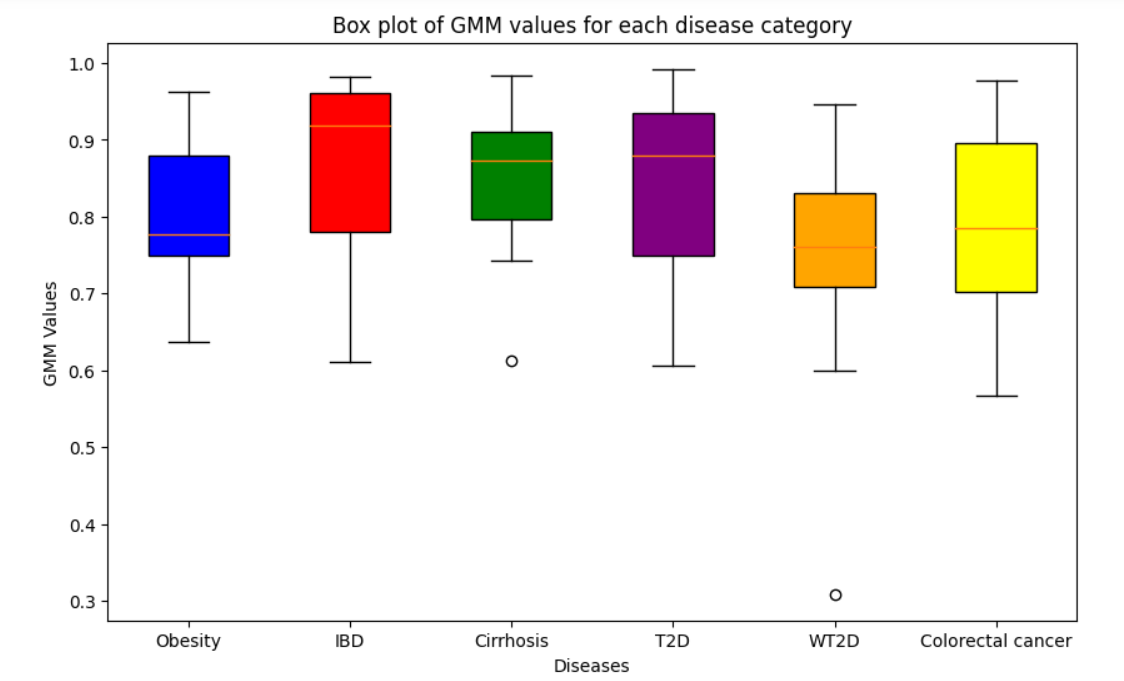
## Kết quả kiểm thử

Dưới đây là các kết quả thu được của niên luận:

*Bảng 2. So sánh độ đo AUC và ACC bài báo của* [39]*,* [38]*, và GMM*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | EQW [39] | | MetAML[38] | GMM | |
| AUC | ACC | AUC | AUC | ACC |
| Chatelier\_gut\_obesity | 0.67300 | 0.64400 | 0.65500 | **0.76600** | 0.59300 |
| metahitIBD | 0.85100 | 0.80800 | 0.89000 | **0.92100** | **0.84600** |
| Quin\_gut  \_liver\_cirrhosis | **0.95500** | **0.87700** | 0.94500 | 0.92100 | 0.85800 |
| T2D | **0.78000** | 0.66400 | 0.74400 | 0.73600 | **0.66600** |
| WT2D | **0.81200** | 0.70500 | 0.76200 | 0.61000 | 0.43500 |
| Zeller\_fecal\_colorectal\_cancer | 0.83700 | **0.80500** | **0.87300** | 0.74300 | 0.52300 |

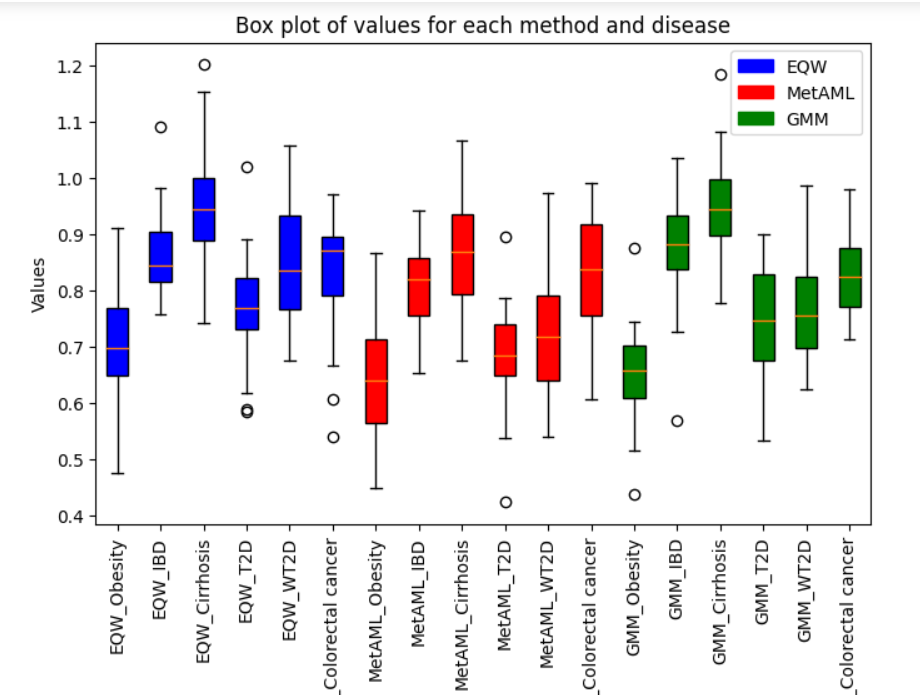
So sánh độ chính xác của bài báo gốc, kết quả của chúng tôi cho thấy trong mọi trường hợp kết quả có cả cao hơn và thấp hơn. Cụ thể, độ chính xác khác nhau đáng kể đối với hai tập dữ liệu: AUC tăng từ 0.655 lên 0.766 đối với bệnh béo phì, từ 0.89 lên 0.921 đối với IBD. Trong bốn trường hợp còn lại, từ 0.945 giảm 0.921 đối với IBD, AUC giảm từ 0.744 lên 0.736 đối với bệnh T2D, từ 0.762 giảm còn 0.61 đối với WT2D và từ 0.873 giảm còn 0.743 đối với ung thư đại trực tràng.

So sánh độ chính xác của EQW, kết quả của chúng tôi cho thấy trong mọi trường hợp kết quả tương tự và kém hơn. Tương tự như trên, độ chính xác khác nhau đáng kể đối với hai tập dữ liệu: AUC tăng từ 0.673 lên 0.766 đối với bệnh béo phì, từ 0.851 lên 0.921 đối với IBD. Trong bốn trường hợp còn lại, từ 0.955 giảm 0.921 đối với IBD, AUC giảm từ 0.78 lên 0.736 đối với bệnh T2D, từ 0.812 giảm còn 0.61 đối với WT2D và từ 0.837 giảm còn 0.743 đối với ung thư đại trực tràng.

*Hình 2. Biểu đồ box plots thể hiện từng loại bệnh của GMM*

Kết quả thể hiện của GMM với độ đo AUC là:

Kết quả thể hiện của GMM với độ đo ACC là:



*Hình 3. Biểu đồ box plots thể hiện từng loại bệnh của GMM, MetAML, EQW*

*Bảng 3. So sánh độ đo AUC và ACC của GMM, binning GMM với RF và SVM*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | GMM | | (GMM) RF | | (GMM) SVM | |
| ACC | AUC | ACC | AUC | ACC | AUC |
| Chatelier\_gut\_obesity | **0.59300** | **0.76600** | 0.48600 | 0.65900 | 0.48100 | 0.52400 |
| metahitIBD | 0.84600 | **0.92100** | **0.86000** | 0.86100 | 0.82700 | 0.81100 |
| Quin\_gut\_liver\_cirrhosis | 0.85800 | **0.92100** | **0.87500** | 0.87900 | 0.80200 | 0.80300 |
| T2D | 0.66600 | **0.73600** | **0.71100** | 0.71400 | 0.57900 | 0.57700 |
| WT2D | **0.43500** | **0.61000** | 0.04100 | 0.51400 | 0.37200 | 0.5700 |
| Zeller\_fecal\_colorectal\_cancer | **0.52300** | **0.74300** | 0.17200 | 0.57400 | 0.39600 | 0.59100 |

So sánh độ chính xác của kết quả chúng tôi với của bộ phân loại SVM và RF.

Đối với RF và SVM cho thấy trong mọi trường hợp kết quả của chúng tôi tốt hơn. Cụ thể, độ chính xác khác nhau đáng kể đối với ba tập dữ liệu: AUC tăng từ 0.524 đối với SVM và từ 0.659 của RF lên 0.766 đối với béo phì, từ 0.811 đối với SVM và từ 0.861 của RF lên 0.921 đối với IBD và từ 0.803 đối với SVM và từ 0.879 của RF lên 0.921 đối với CIR, từ 0.577 đối với SVM và từ 0.714 của RF lên 0.736 đối với T2D, từ 0.57 đối với SVM và từ 0.514 của RF lên 0.61 đối với WT2D, và từ 0.591 đối với SVM và từ 0.574 của RF lên 0.743 đối với WT2D.

So với hai bộ phân loại phổ biến SVM và RF, kết quả của chúng tôi có xu hướng tốt hơn. GMM đã thể hiện hiệu suất vượt trội đối với ba tập dữ liệu với độ chính xác cao hơn so với cả SVM và RF. Do đó, về mặt phương pháp, kết quả của thuật toán phân cụm chúng tôi đã đề xuất tốt hơn so với 2 mô hình RF và SVM.

*Bảng 4. So sánh ACC và AUC dữ liệu gốc và GMM*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | GMM | | raw | |
| ACC | AUC | ACC | AUC |
| Chatelier\_gut\_obesity | **0.59300** | **0.76600** | 0.55700 | 0.73400 |
| metahitIBD | **0.84600** | **0.92100** | 0.84600 | 0.91800 |
| Quin\_gut\_liver\_cirrhosis | **0.85800** | **0.92100** | 0.77600 | 0.86700 |
| T2D | **0.66600** | 0.73600 | 0.64200 | **0.74800** |
| WT2D | 0.43500 | 0.61000 | **0.60700** | **0.80900** |
| Zeller\_fecal\_colorectal\_cancer | **0.52300** | **0.74300** | 0.45000 | 0.67400 |

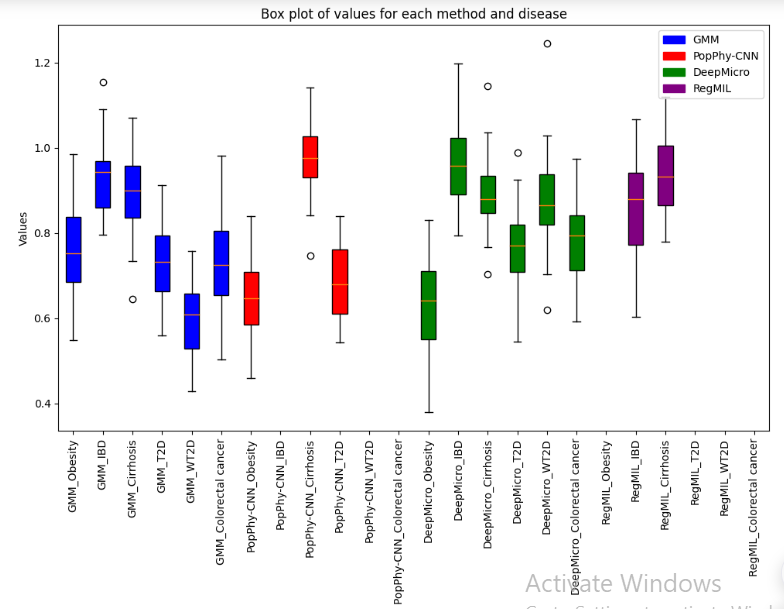
So sánh độ chính xác với dữ liệu gốc raw, kết quả của chúng tôi cho thấy trong mọi trường hợp kết quả có cả cao hơn. Cụ thể, độ chính xác khác nhau đáng kể đối: AUC tăng từ 0.734 lên 0.766 đối với bệnh béo phì, từ 0.918 lên 0.921 đối với IBD. Đối với CIR, AUC tăng từ 0.867 lên 0.921 và AUC tăng từ 0.671 lên 0.743 đối với ung thư đại trực tràng. AUC giảm từ 0.748 lên 0.736 đối với đối với bệnh T2D, AUC giảm từ 0.809 lên 0.61 đối với WT2D.

*Bảng 5. So sánh độ đo AUC của các bài nghiên cứu về dữ liệu metagenomics đường ruột*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bins | GMM | PopPhy-CNN [40] | DeepMicro [41] | RegMIL [42] |
| Chatelier\_gut\_obesity | **0.76600** | 0.66600 | 0.65900 | - |
| metahitIBD | 0.92100 | - | **0.95500** | 0.84420 |
| Quin\_gut\_liver\_cirrhosis | 0.92100 | **0.94600** | 0.94000 | 0.92720 |
| T2D | 0.73600 | 0.69000 | **0.76300** | - |
| WT2D | 0.61000 | - | **0.89900** | - |
| Zeller\_fecal\_colorectal\_cancer | 0.74300 | - | **0.80300** | - |

Từ bảng so sánh trên có thể thấy GMM có kết quả AUC tốt nhất cho Obesity với giá trị là 0.766. Ở bệnh IBD cũng cho ra kết quả khả quan là 0.921 chênh lệch thấp hơn so với DeepMicro là 0.955. Ở bệnh CIR, GMM cho ra kết quả thấp nhất, tuy nhiên không quá chênh lệch so với các nghiên cứu còn lại. Với dữ liệu bệnh T2d, AUC của GMM cao hơn của PopPhy nhưng thấp hơn DeepMicro. Đối với bệnh ung thư đại trực tràng thì độ đo AUC của GMM thấp hơn hẳn.

Mỗi bài nghiên cứu sử dụng một mô hình khác nhau và đôi khi có kết quả khác nhau cho từng loại bệnh lý. Tuy nhiên, GMM xuất sắc cho nhiều loại bệnh lý khác nhau như Obesity. DeepMicro cũng đạt kết quả tốt cho WT2D, CRC, T2D, IBD.

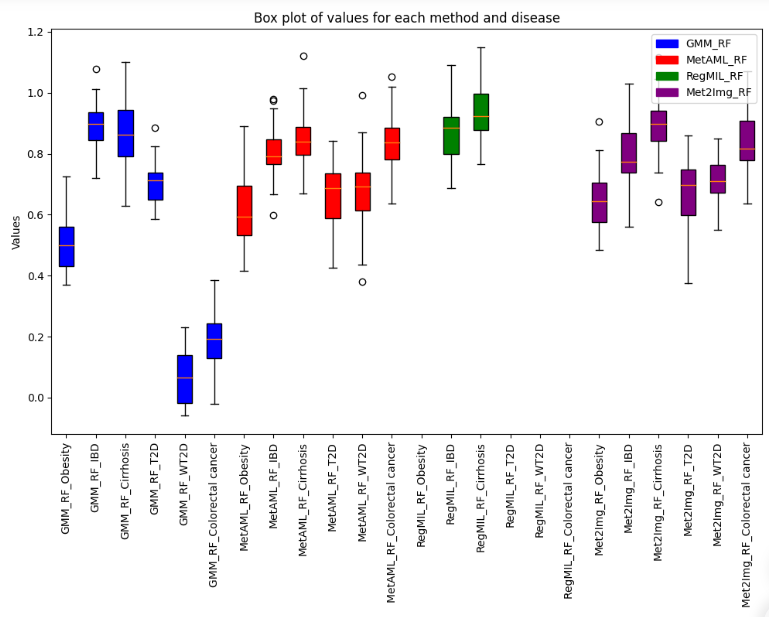


*Hình 4. Biểu đồ box plots thể hiện từng loại bệnh của GMM, PopPhy-CNN, DeepMicro, RegMIL*

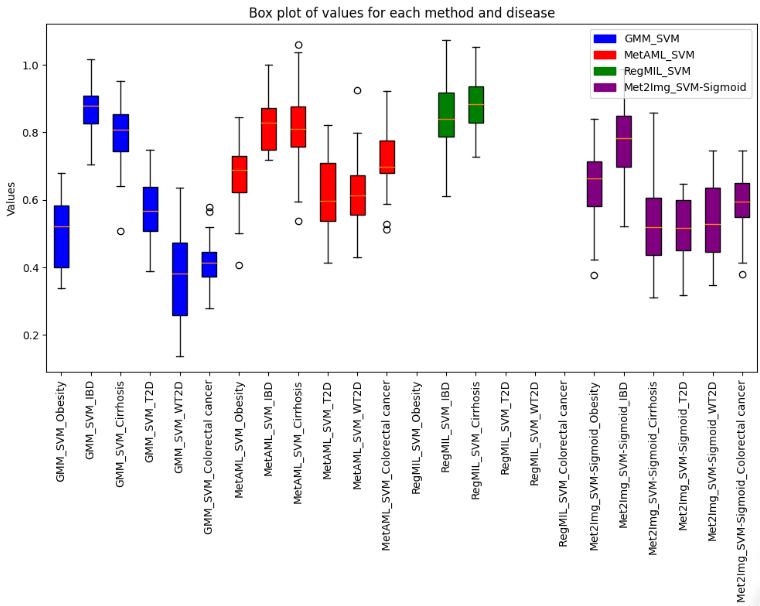
*Bảng 6. So sánh độ đo ACC của các bài nghiên cứu về dữ liệu metagenomics đường ruột*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bins | GMM | | MetAML [38] | | RegMIL [42] | | Met2Img [43] | |
| Model | RF | SVM | RF | SVM | RF | SVM | RF | SVM-Sigmoid |
| Chatelier\_gut\_obesity | 0.48600 | 0.48100 | 0.64400 | 0.63600 | - | - | **0.64500** | **0.64800** |
| metahitIBD | **0.86000** | **0.82700** | 0.80900 | 0.80900 | 0.84660 | 0.82370 | 0.80800 | 0.75500 |
| Quin\_gut\_liver\_cirrhosis | 0.87500 | 0.80200 | 0.87700 | 0.83400 | **0.92810** | **0.89500** | 0.87700 | 0.50900 |
| T2D | **0.71100** | 0.57900 | 0.66400 | **0.61300** | - | - | 0.67200 | 0.51500 |
| WT2D | 0.04100 | 0.37200 | **0.70300** | **0.59600** | - | - | **0.70300** | 0.55300 |
| Zeller\_fecal\_colorectal\_cancer | 0.17200 | 0.39600 | 0.80500 | **0.74300** | - | - | **0.81200** | 0.60300 |

Với dữ liệu bệnh béo phì, RF và SVM của GMM cho kết quả ACC thấp hơn đáng kể so với các nghiên cứu còn lại. Tuy nhiên dữ liệu IBD, GMM lại cho ra kết quả cao nhất cả RF lẫn SVM với giá trị là 0.86 và 0.827. Trong khi bệnh CIR, GMM của RF là 0.875 chỉ thấp hơn giá trị RF của RegMI và SVM cao hơn SVM-Sigmoid của Met2Img. Kết quả của T2D có độ đo Acc cao nhất về cả RF và SVM so với tất cả các nghiên cứu còn lại. Còn với bệnh WT2D và CRC, GMM của RF và SVM đều cho ra kết quả không khả quan.



*Hình 5. Biểu đồ box plots thể hiện từng loại bệnh theo RF của GMM, MetAML, RegMIL, Met2Img*



*Hình 6. Biểu đồ box plots thể hiện từng loại bệnh theo SVM của GMM, MetAML, RegMIL, Met2Img*

Các kết quả ACC khác nhau cho từng mô hình và loại bệnh lý. GMM của RF và SVM cho ra kết quả khả quan nhất đối với các bệnh IBD, T2D, CIR. Một số mô hình như RF, SVM của MetAML và RegMIL thường cho kết quả cao hơn so với các mô hình khác trong một số trường hợp. Điều này cho thấy sự đa dạng và sự khác biệt trong hiệu suất của các mô hình khi áp dụng vào các bộ dữ liệu khác nhau.

Nhìn chung thuật toán Gaussian Mixture Models cho ra kết quả khả quan, cải thiện đáng kể độ chính xác. Kết quả của chúng tôi cho thấy sự biến đổi đáng kể trong độ chính xác so với bài báo gốc. Trong một số trường hợp, chúng tôi đã ghi nhận độ chính xác cao hơn, đặc biệt là đối với bệnh béo phì và IBD. Tuy nhiên, cũng có những trường hợp mà kết quả của chúng tôi có sự giảm so với bài báo gốc, đặc biệt là đối với CRC và bệnh WT2D.

So với hai bộ phân loại phổ biến SVM và RF, kết quả của chúng tôi có xu hướng tốt hơn. GMM đã thể hiện hiệu suất vượt trội đối với ba tập dữ liệu với độ chính xác cao hơn so với cả SVM và RF.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết luận

Triển khai việc phân cụm metagenomics sử dụng Gaussian Mixture Models (GMM) trong ngữ cảnh cụ thể của nghiên có ý nghĩa quan trọng. GMM cho phép phân tích và hiểu rõ hơn về đa dạng sinh học của cộng đồng vi sinh vật trong mẫu metagenomics. Việc này có thể giúp phát hiện và phân loại các loại vi khuẩn, vi rút, hoặc vi sinh vật khác tồn tại trong mẫu. Bằng cách phân cụm dữ liệu, GMM giúp xác định cấu trúc và tổ chức của cộng đồng vi sinh vật trong mẫu metagenomic. Điều này có thể tiết lộ mối quan hệ tương tác giữa các thành phần các virus gây bệnh. Triển khai GMM trong phân cụm metagenomics có thể ứng dụng rộng rãi trong lĩnh vực y học. Việc hiểu rõ cấu trúc và sự đa dạng của cộng đồng vi sinh vật có thể cung cấp thông tin hữu ích để xác định các bệnh lý, đánh giá sức khỏe. Hơn thế nữa phân cụm metagenomics không chỉ mang lại thông tin cần thiết cho ứng dụng cụ thể mà còn tạo ra nền tảng nghiên cứu cho các nghiên cứu tiếp theo về cấu trúc và sự đa dạng của cộng đồng vi sinh vật trong mẫu metagenomics.

Dựa trên các kết quả thực nghiệm và so sánh độ chính xác giữa các phương pháp phân loại khác nhau, kết quả của nghiên cứu cho thấy sự ứng dụng của thuật toán phân cụm Gaussian Mixture Models (GMM) trong việc xử lý và dự đoán các bệnh lý dựa trên dữ liệu metagenomics đạt được kết quả khả quan. So với các phương pháp khác như Random Forest (RF) và Support Vector Machine (SVM), GMM đã thể hiện hiệu suất tốt hơn trong nhiều trường hợp.

Khi so sánh với kết quả từ bài báo gốc của Edoardo Pasolli và các cộng sự, chúng tôi đã quan sát thấy sự tương đồng và thậm chí có sự cải thiện về độ chính xác trong một số tập dữ liệu nhất định, như béo phì và bệnh đại trực tràng. Mặc dù trong một số trường hợp khác, kết quả của chúng tôi có thể thấp hơn so với bài báo gốc, nhưng vẫn cho thấy sự tiến triển và khả năng áp dụng rộng rãi của phương pháp GMM. Ngoài ra, so sánh với hai mô hình RF và SVM cũng cho thấy hiệu suất của GMM vượt trội trong việc dự đoán và phân loại các bệnh lý dựa trên các chỉ số đánh giá như Area Under Curve (AUC). Điều này cho thấy tiềm năng và hiệu quả của phương pháp phân cụm GMM trong việc xử lý dữ liệu metagenomic và dự đoán các bệnh lý. Cuối cùng, việc so sánh kết quả khi sử dụng dữ liệu gốc (raw) và kết quả được xử lý bởi GMM đã thể hiện sự cải thiện rõ rệt trong độ chính xác. Việc áp dụng phương pháp GMM đã giúp tăng cường hiệu suất và chính xác của dự đoán, đặc biệt là trong việc phân loại các bệnh lý từ dữ liệu metagenomics.

Tổng quan lại, thuật toán GMM đã chứng minh được khả năng hiệu quả và tiềm năng trong việc xử lý dữ liệu metagenomic và dự đoán các bệnh lý, đồng thời mở ra triển vọng trong việc áp dụng rộng rãi trong lĩnh vực y học dựa trên nguyên liệu gen.

Tuy nhiên vẫn còn một số trường hợp kết quả của chúng tôi không thể đạt được độ chính xác như kết quả từ bài báo gốc. Điều này có thể phần nào được giải thích qua sự khác biệt trong cách tiếp cận phân tích và cách tiền xử lý dữ liệu khác nhau.

## Hướng phát triển

Nghiên cứu đã thể hiện tiềm năng và hiệu quả của thuật toán GMM trong việc xử lý và phân loại dữ liệu genomic để dự đoán bệnh lý. Mặc dù còn tồn tại một số hạn chế, nhưng kết quả tích cực đã cung cấp cơ sở để tiếp tục nghiên cứu và phát triển trong lĩnh vực này.

Trong tương lai, chúng tôi dự tính sẽnghiên cứu cách kết hợp GMM với các mô hình khác như Neural Networks hoặc Deep Learning để tạo ra mô hình kết hợp mạnh mẽ hơn. Song song tiến hành các thử nghiệm thực nghiệm và đánh giá kết quả với các bộ dữ liệu lớn và đa dạng hơn kết hợp sử dụng thêm dữ liệu bằng hình ảnh. Nghiên cứu nguyên nhân và cách sửa chữa các sai số trong quá trình dự đoán. Viết báo cáo hoặc bài báo để chia sẻ kết quả nghiên cứu với cộng đồng khoa học và y học. Những hướng phát triển của chúng tôi có thể giúp tăng cường sức mạnh và ứng dụng thực tế của thuật toán GMM trong việc phân loại dựa trên dữ liệu metagenomics.

**CHƯƠNG 5. TÀI LIỆU THAM KHẢO**

[1] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, “Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products,” *Chem. Biol.*, vol. 5, no. 10, pp. R245–R249, Oct. 1998, doi: 10.1016/S1074-5521(98)90108-9.

[2] K. Chen and L. Pachter, “Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities,” *PLoS Comput. Biol.*, vol. 1, no. 2, p. e24, 2005, doi: 10.1371/journal.pcbi.0010024.

[3] C. Simon and R. Daniel, “Metagenomic Analyses: Past and Future Trends,” *Appl. Environ. Microbiol.*, vol. 77, no. 4, pp. 1153–1161, Feb. 2011, doi: 10.1128/AEM.02345-10.

[4] A. K. Wani *et al.*, “Discovering untapped microbial communities through metagenomics for microplastic remediation: recent advances, challenges, and way forward,” *Environ. Sci. Pollut. Res.*, vol. 30, no. 34, pp. 81450–81473, Jan. 2023, doi: 10.1007/s11356-023-25192-5.

[5] *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, D.C.: National Academies Press, 2007, p. 11902. doi: 10.17226/11902.

[6] C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand, “From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1722–1734, 2020, doi: 10.1016/j.csbj.2020.06.028.

[7] K. Sartaj, A. Patel, L. Matsakas, and R. Prasad, “Unravelling Metagenomics Approach for Microbial Biofuel Production,” *Genes*, vol. 13, no. 11, p. 1942, Oct. 2022, doi: 10.3390/genes13111942.

[8] J. Chakraborty, K. Palit, and S. Das, “Metagenomic approaches to study the culture-independent bacterial diversity of a polluted environment—a case study on north-eastern coast of Bay of Bengal, India,” in *Microbial Biodegradation and Bioremediation*, Elsevier, 2022, pp. 81–107. doi: 10.1016/B978-0-323-85455-9.00014-X.

[9] Y. Liang *et al.*, “Prediction of BMI traits in the Chinese population based on the gut metagenome,” In Review, preprint, Sep. 2023. doi: 10.21203/rs.3.rs-3337879/v1.

[10] J. Zhang, W. Liu, H. Simayijiang, P. Hu, and J. Yan, “Application of Microbiome in Forensics,” *Genomics Proteomics Bioinformatics*, vol. 21, no. 1, pp. 97–107, Feb. 2023, doi: 10.1016/j.gpb.2022.07.007.

[11] K. T. T. Kwok, D. F. Nieuwenhuijse, M. V. T. Phan, and M. P. G. Koopmans, “Virus Metagenomics in Farm Animals: A Systematic Review,” *Viruses*, vol. 12, no. 1, p. 107, Jan. 2020, doi: 10.3390/v12010107.

[12] G. Marais, D. Hardie, and A. Brink, “A case for investment in clinical metagenomics in low-income and middle-income countries,” *Lancet Microbe*, vol. 4, no. 3, pp. e192–e199, Mar. 2023, doi: 10.1016/S2666-5247(22)00328-7.

[13] X. Xu, A. Wu, X. Zhang, M. Su, T. Jiang, and Z.-M. Yuan, “MetaDP: a comprehensive web server for disease prediction of 16S rRNA metagenomic datasets,” *Biophys. Rep.*, vol. 2, no. 5–6, pp. 106–115, Dec. 2016, doi: 10.1007/s41048-016-0033-4.

[14] F. Yang and Q. Zou, “DisBalance: a platform to automatically build balance-based disease prediction models and discover microbial biomarkers from microbiome data,” *Brief. Bioinform.*, vol. 22, no. 5, p. bbab094, Sep. 2021, doi: 10.1093/bib/bbab094.

[15] Y. Long and J. Luo, “WMGHMDA: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network,” *BMC Bioinformatics*, vol. 20, no. 1, p. 541, Dec. 2019, doi: 10.1186/s12859-019-3066-0.

[16] J. Zhu *et al.*, “Statistical modeling of gut microbiota for personalized health status monitoring,” *Microbiome*, vol. 11, no. 1, p. 184, Aug. 2023, doi: 10.1186/s40168-023-01614-x.

[17] M. Zhao, R. Yue, X. Wu, Z. Gao, M. He, and L. Pan, “The diagnostic value of metagenomic next-generation sequencing for identifying Pneumocystis jirovecii infection in non-HIV immunocompromised patients,” *Front. Cell. Infect. Microbiol.*, vol. 12, p. 1026739, Oct. 2022, doi: 10.3389/fcimb.2022.1026739.

[18] X. Zhang *et al.*, “Diagnosis of Non-Tuberculous Mycobacterial Pulmonary Disease by Metagenomic Next-Generation Sequencing on Bronchoalveolar Lavage Fluid,” *Infect. Drug Resist.*, vol. Volume 16, pp. 4137–4145, Jun. 2023, doi: 10.2147/IDR.S417088.

[19] J. Li *et al.*, “Diagnostic Value of Metagenomic Next-Generation Sequencing for Pneumonia in Immunocompromised Patients,” *Can. J. Infect. Dis. Med. Microbiol.*, vol. 2022, pp. 1–10, Dec. 2022, doi: 10.1155/2022/5884568.

[20] M. Acharya, S. Yadav, and K. K. Mohbey, “How can we create a recommender system for tourism? A location centric spatial binning-based methodology using social networks,” *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 1, p. 100161, Apr. 2023, doi: 10.1016/j.jjimei.2023.100161.

[21] Z. Ma *et al.*, “An Unsupervised Crop Classification Method Based on Principal Components Isometric Binning,” *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 11, p. 648, Oct. 2020, doi: 10.3390/ijgi9110648.

[22] K. E. Hartley and R. W. Wilson, “Optimized temporal binning of comparison star measurements for differential photometry,” *Mon. Not. R. Astron. Soc.*, vol. 526, no. 3, pp. 3482–3494, Oct. 2023, doi: 10.1093/mnras/stad2964.

[23] Q. Zhu *et al.*, “Impacts of Binning Methods on High‐Latitude Electrodynamic Forcing: Static Versus Boundary‐Oriented Binning Methods,” *J. Geophys. Res. Space Phys.*, vol. 125, no. 1, p. e2019JA027270, Jan. 2020, doi: 10.1029/2019JA027270.

[24] C. D. Dean, A. A. Chiarenza, and S. C. R. Maidment, “Formation binning: a new method for increased temporal resolution in regional studies, applied to the Late Cretaceous dinosaur fossil record of North America,” *Palaeontology*, vol. 63, no. 6, pp. 881–901, Nov. 2020, doi: 10.1111/pala.12492.

[25] A. Fyrdahl, A. Ullvin, J. G. Ramos, N. Seiberlich, M. Ugander, and A. Sigfridsson, “Three‐dimensional sector‐wise golden angle–improved k‐space uniformity after electrocardiogram binning,” *Magn. Reson. Med.*, vol. 90, no. 3, pp. 1041–1052, Sep. 2023, doi: 10.1002/mrm.29698.

[26] I. Moskowitz, E. Gawiser, A. Bault, A. Broussard, J. A. Newman, and J. Zuntz, “Improved Tomographic Binning of 3 × 2 pt Lens Samples: Neural Network Classifiers and Optimal Bin Assignments,” *Astrophys. J.*, vol. 950, no. 1, p. 49, Jun. 2023, doi: 10.3847/1538-4357/accc88.

[27] D. Weckmüller and A. Dunkel, “AN APPLICATION-ORIENTED IMPLEMENTATION OF HEXAGONAL ON-THE-FLY BINNING METRICS FOR CITY-SCALE GEOREFERENCED SOCIAL MEDIA DATA,” *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLVIII-4/W7-2023, pp. 253–260, Jun. 2023, doi: 10.5194/isprs-archives-XLVIII-4-W7-2023-253-2023.

[28] J. N. Nissen *et al.*, “Improved metagenome binning and assembly using deep variational autoencoders,” *Nat. Biotechnol.*, vol. 39, no. 5, pp. 555–560, May 2021, doi: 10.1038/s41587-020-00777-4.

[29] G. P. Schmartz *et al.*, “BusyBee Web: towards comprehensive and differential composition-based metagenomic binning,” *Nucleic Acids Res.*, vol. 50, no. W1, pp. W132–W137, Jul. 2022, doi: 10.1093/nar/gkac298.

[30] S. Kieser, J. Brown, E. M. Zdobnov, M. Trajkovski, and L. A. McCue, “ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data,” *BMC Bioinformatics*, vol. 21, no. 1, p. 257, Dec. 2020, doi: 10.1186/s12859-020-03585-4.

[31] Y. Du and F. Sun, “HiFine: integrating Hi-C-based and shotgun-based methods to refine binning of metagenomic contigs,” *Bioinformatics*, vol. 38, no. 11, pp. 2973–2979, May 2022, doi: 10.1093/bioinformatics/btac295.

[32] V. Mallawaarachchi and Y. Lin, “Accurate Binning of Metagenomic Contigs Using Composition, Coverage, and Assembly Graphs,” *J. Comput. Biol.*, vol. 29, no. 12, pp. 1357–1376, Dec. 2022, doi: 10.1089/cmb.2022.0262.

[33] Y. Luo, Y. W. Yu, J. Zeng, B. Berger, and J. Peng, “Metagenomic binning through low-density hashing,” *Bioinformatics*, vol. 35, no. 2, pp. 219–226, Jan. 2019, doi: 10.1093/bioinformatics/bty611.

[34] J. Johansen *et al.*, “Genome binning of viral entities from bulk metagenomics data,” *Nat. Commun.*, vol. 13, no. 1, p. 965, Feb. 2022, doi: 10.1038/s41467-022-28581-5.

[35] O. S. Bak, M. D. Jensen, F. M. Trudslev, A. Windfeld, and A. Lamurias, “BinChill: A Metagenomic Binning Ensemble Method,” *IEEE Access*, vol. 11, pp. 49561–49577, 2023, doi: 10.1109/ACCESS.2023.3277755.

[36] H. Ho *et al.*, “Integrating chromatin conformation information in a self-supervised learning model improves metagenome binning,” *PeerJ*, vol. 11, p. e16129, Sep. 2023, doi: 10.7717/peerj.16129.

[37] DỤC ĐOÀN TRÌNH, “Gaussian Mixture Model trong machine learning,” Dec. 2021.

[38] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights,” *PLOS Comput. Biol.*, vol. 12, no. 7, p. e1004977, Jul. 2016, doi: 10.1371/journal.pcbi.1004977.

[39] T. H. Nguyen and J.-D. Zucker, “Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches,” in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, Da Nang, Vietnam: IEEE, Oct. 2019, pp. 1–5. doi: 10.1109/KSE.2019.8919295.

[40] D. Reiman, A. A. Metwally, J. Sun, and Y. Dai, “PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype From Metagenomic Data,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2993–3001, Oct. 2020, doi: 10.1109/JBHI.2020.2993761.

[41] M. Oh and L. Zhang, “DeepMicro: deep representation learning for disease prediction based on microbiome data,” *Sci. Rep.*, vol. 10, no. 1, p. 6026, Apr. 2020, doi: 10.1038/s41598-020-63159-5.

[42] M. A. Rahman and H. Rangwala, “RegMIL: Phenotype Classification from Metagenomic Data,” in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Washington DC USA: ACM, Aug. 2018, pp. 145–154. doi: 10.1145/3233547.3233585.

[43] T. H. Nguyen, E. Prifti, Y. Chevaleyre, N. Sokolovska, and J.-D. Zucker, “Disease Classification in Metagenomics with 2D Embeddings and Deep Learning,” 2018, doi: 10.48550/ARXIV.1806.09046.