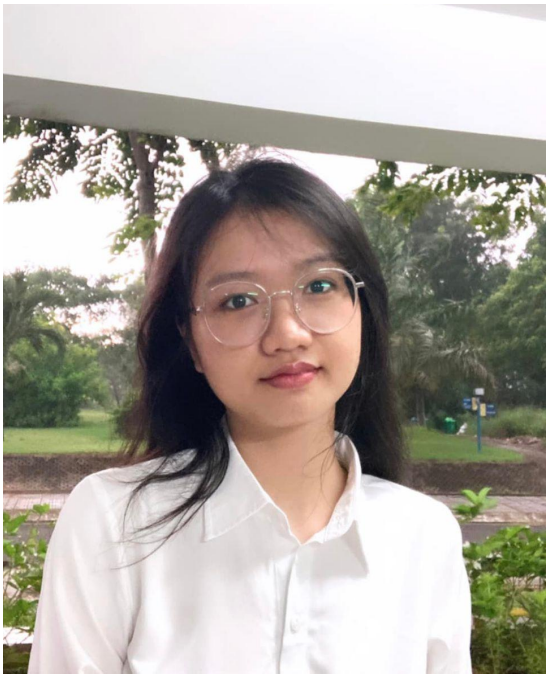
THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/6A2D_7dynYs*
- Link slides (dạng .pdf đặt trên Github): *https://github.com/Trinhtruc1831/CS2205.CH170/blob/main/Tru%CC%81c%20Tri%CC%A3nh%20Thi%CC%A3%20Thanh%20-%20xCS2205.DeCuong.FinalReport.Template.Slide.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Trịnh Thị Thanh Trúc • MSSV: 19521059 	• Lớp: CS2205.CH1702 - APR2023 • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 2 • Số câu hỏi QT cá nhân: 7 • Số câu hỏi QT của cả nhóm: 2 • Link Github: https://github.com/Trinhtruc1831/CS2205.CH170/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

XÂY DỰNG MÔ HÌNH DỰ BÁO THỜI ĐIỂM BÙNG PHÁT BỆNH NHIỆT ĐỚI BỊ LÃNG QUÊN

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

CONSTRUCT THE PREDICTION MODEL FOR NEGLECTED TROPICAL DISEASE OUTBREAKS

TÓM TẮT *(Tối đa 400 từ)*

Tiêu chảy - một căn bệnh cực kỳ nhạy cảm với nguồn nước, là nguyên nhân hàng đầu gây tử vong và tình trạng bệnh nhiễm trùng ở trẻ em và phổ biến trong các nước đang phát triển. Trong khi đó, biến đổi khí hậu cùng với hiện tượng thời tiết cực đoan được coi là một trong những nguyên nhân chính gây ô nhiễm nước trong những báo cáo gần đây. Là một quốc gia đang phát triển và nằm trong vùng khí hậu nhiệt đới, Việt Nam có nguy cơ cao về bùng phát tiêu chảy, dẫn đến nhu cầu cấp bách về hệ thống cảnh báo sớm cho căn bệnh này. Tuy nhiên, các nghiên cứu trước đây chỉ thực hiện hồi quy tỉ lệ ca mắc tiêu chảy. Trong nghiên cứu này, chúng tôi tập trung dự báo các sự kiện bùng phát tiêu chảy theo hướng tiếp cận phân lớp chuỗi thời gian các đặc trưng khí hậu với mục tiêu dự báo các điểm bùng nổ hoặc không bùng nổ bệnh trong tương lai. Thực nghiệm 22 thuật toán máy học bao gồm các thuật toán máy học thống kê và 4 thuật toán học sâu tiên tiến: CNN, LSTM, LSTM với cơ chế Attention và Transformer nhằm so sánh hiệu quả của hai loại mô hình trong miền dữ liệu hiện có. Ngoài ra, chúng tôi tiến hành so sánh kết quả dự báo bùng nổ tính toán trên ca nhiễm hồi quy tiêu chảy tốt nhất từ nghiên cứu [9] tương tự [10] và so sánh với kết quả phân lớp các điểm bùng nổ trực tiếp của chúng tôi. Tiếp đến, chúng tôi đề xuất một phương pháp lọc dự đoán kết hợp tri thức từ các mô hình khả thi dựa trên thuật toán Apriori để có thể loại bỏ được các dự báo giả được đưa ra. Sau cùng, Ensemble Voting được thực hiện cho TOP-k mô hình với kỳ vọng sẽ lọc nhiều hơn nữa các dự báo giả khi không nhận được sự đồng thuận từ đa số.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Bệnh nhiệt đới lãng quên là một nhóm các loại bệnh phổ biến chủ yếu ở vùng nhiệt đới, đặc biệt là các quốc gia đang phát triển. Mặc dù hầu hết các bệnh trong danh sách này đều có thể được phòng ngừa và chữa trị trong điều kiện y học phát triển như hiện nay. Tuy nhiên, việc chủ quan trong khâu nhận biết sớm và điều trị có thể gây hậu quả nghiêm trọng cho sức khoẻ, kinh tế và xã hội. Trong danh sách được liệt kê bởi WHO, các bệnh dẫn đến triệu chứng tiêu chảy thường là mối quan tâm số 1 của các nhà chức trách. Bởi, tiêu chảy xảy ra dễ dàng và hàng loạt chính là nguyên nhân phổ biến thứ hai gây tử vong ở trẻ em dưới 5 tuổi

và là nguyên nhân hàng đầu gây suy dinh dưỡng ở trẻ nhỏ [1]. Tiêu chảy thường là triệu chứng của nhiễm trùng đường ruột với nguyên nhân chủ yếu là do vi khuẩn, vi rút, ký sinh trùng lây lan qua nguồn nước ô nhiễm kém vệ sinh. Biến đổi khí hậu đi kèm những điều kiện thời tiết cực đoan chính là một trong những yếu tố góp phần làm ô nhiễm môi trường sống đặc biệt là nguồn nước. Nhiều nghiên cứu liên quan cũng đã dấy sự chú ý của mối liên hệ giữa về khí hậu và các ca Tiêu chảy được ghi nhận [2]-[9]. Ngoài ra, yếu tố thời gian cũng được đề cập nhiều trong các nghiên cứu trên, khi các phân tích cho thấy thời điểm các sự kiện thời tiết xảy ra cũng tạo nên mức độ khác biệt trong số lượng các nhiễm được ghi nhận.
Trong khi đó, Việt Nam – một quốc gia đang phát triển, với vị trí địa lý nằm hoàn toàn trong vành đai khí hậu nhiệt đới thì việc chịu ảnh hưởng nặng nề bởi biến đổi khí hậu dẫn đến các nguy cơ cao gây mắc và bùng phát hàng loạt các ca tiêu chảy đang dấy lên mối quan tâm và lo ngại hàng đầu cho các cơ quan kiểm soát bệnh tật tại Việt Nam. Từ đây, nhu cầu về xây dùng một hệ thống có khả năng dự báo và phát hiện nguy cơ bùng phát dịch bệnh là đang vô cùng cấp thiết.
Nhiều nghiên cứu trước đây cũng đã khai thác sức mạnh máy học trong việc dự báo tỉ lệ ca nhiễm và mức độ bùng phát tiêu chảy trong tương lai dựa vào các đặc trưng khí hậu. Mặt khác, các mô hình học sâu cũng được áp dùng rất mạnh mẽ với kỳ vọng khai thác sức mạnh của các kiến trúc này trong việc truy xuất thông tin từ dữ liệu hiện có. Tuy nhiên, vẫn còn khá ít nghiên cứu tập trung vào mục tiêu dự báo bùng phát dịch bệnh. Thay vào đó là các dự báo ca nhiễm được đưa ra hướng tới những đối tượng có kiến thức chuyên môn đủ để xác định được mức độ nguy hiểm của các ca nhiễm được dự báo. Vì thế nghiên cứu này tập trung phát hiện trực tiếp các mốc thời gian xảy bùng phát dịch bệnh trong khu vực thay vì hậu xứ lý với kết quả dự báo ca nhiễm như các nghiên cứu trước đây. Các tiếp cận bài toán được chúng tôi thực hiện là phân lớp chuỗi thời gian các đặc trưng khí hậu để đưa ra dự báo cho các điểm bùng nổ trong tương lai. Thực nghiệm được xây dựng với các mô hình dự báo học máy thống kê lẫn các mô hình học sâu khác nhau với kỳ vọng sẽ khai thác được ưu điểm của từng loại mô hình này trên miền dữ liệu hiện tại, sau cùng đưa ra được kết quả phát hiện hiện đúng đắn, làm cơ sở để xây dựng được một hệ thống cảnh báo và nhận biết sớm dịch bệnh hiệu quả.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

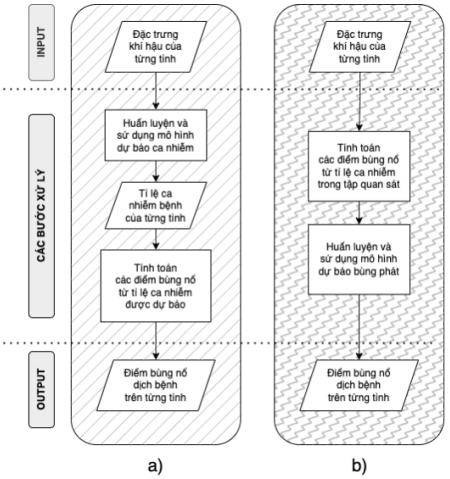
1. Khảo sát mối liên hệ giữa các đặc trưng khí hậu và mức độ bùng phát dịch bệnh tiêu chảy cũng như các phương pháp tiếp cận cho đến thời điểm hiện tại.
2. Đề xuất một phương pháp lọc hai pha giúp loại bỏ các dự báo giả theo các mức độ khác nhau.
3. Tóm tắt kết quả nghiên cứu về: (1) Hiệu quả của hai hướng tiếp cận máy học thống kê và học sâu tiên tiến; (2) Hiệu quả của hướng phân lớp trực tiếp các điểm bùng nổ với hậu xử lý bùng nổ từ tỉ lệ ca nhiễm được dự báo hồi quy; (3) Hiệu quả của phương pháp lọc hai giai đoạn.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

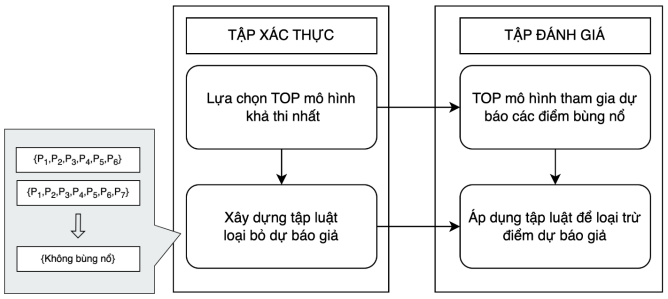
1. Nội dung

- **Đối tượng và phạm vi:**
  - Đối tượng nghiên cứu: Bệnh Sốt Xuất Huyết (Dengue Fever - DF), Bệnh Tiêu Chảy (Diarrhea - DH). Trên các tỉnh thành tại Việt Nam từ 1997 – 2016.
  - Các đặc trưng khí hậu sử dụng: độ ẩm, lượng bốc hơi, lượng mưa, nhiệt độ, số giờ nắng.
  - Nguồn dữ liệu được lấy từ: Viện Vệ sinh dịch tễ Trung ương và Viện Khoa học Khí tượng Thủy văn và Biến đổi khí hậu
- **Phát biểu bài toán:**
  - Đầu vào:
    - Đặc trưng khí hậu Việt Nam
  - Đầu ra:
    - Điểm bùng nổ dịch bệnh cho từng thời điểm trong từng tỉnh



Hình 1: Hai hướng tiếp cận dự báo bùng nổ được thực nghiệm so sánh

- Lọc và loại bỏ điểm loại báo giả:
  - Việc lọc và loại bỏ điểm dự báo giả được thực hiện theo hai pha. Trong đó pha đầu tiên sẽ khai thác tập luật phổ biến đưa ra dự báo giả các TOP các mô hình khả thi, sau đó, áp dụng các tập luật này vào các điểm dự báo trong tập đánh giá (Hình 4). Pha thứ hai sẽ tiến hành Ensemble voting TOP-k các mô hình để tìm ra các dự báo bùng nổ có mức độ tin cậy cao nhất.



Hình 4: Minh hoạ luồng xử lý khai thác tập luật để loại bỏ dự báo giả được thực hiện ở pha lọc đầu tiên

2. Phương pháp

- Tìm hiểu cơ bản về các căn bệnh nhiệt đới, các thống kê về nguyên nhân và mức độ nguy hiểm của loại bệnh này.
- Tìm hiểu tổng quan về các thuật toán máy học hướng thống kê truyền thống và hướng học sâu.
- Tìm hiểu về đặc trưng dữ liệu chuỗi thời gian và các kỹ thuật tiền xử lý dữ liệu.
- Xây dựng một bộ dữ liệu về các đặc trưng khí hậu và tỷ lệ các ca nhiễm bệnh nhiệt đới theo thời gian và theo vùng địa lý tại Việt Nam.
- Cài đặt và thực nghiệm các thuật toán dự báo đã khảo sát trên bộ dữ liệu được xây dựng.
- Phân tích kết quả thực nghiệm và so sánh hiệu quả các phương pháp cũng như so sánh với các công trình đi trước.
- Tổng hợp, đánh giá kết quả và viết báo cáo.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Kỳ vọng kết quả khi tiếp cận phân lớp trực tiếp chuỗi đặc trưng khí hậu để dự báo bùng nổ dịch bệnh tiêu chảy sẽ tốt hơn với phương pháp hậu xử lý tỉ lệ ca nhiễm được hồi quy trước đây.
- Kỳ vọng phương pháp lọc dự báo giả hai pha sẽ giúp loại bỏ được các dự báo giả, trong đó: ở pha đầu tiên sẽ hạn chế tối đa các dự báo đúng bị loại bỏ trong quá trình lọc; ở pha thứ hai sẽ loại trên cơ sở của pha thứ nhất, tuy nhiên mạnh mẽ hơn và các dự báo bùng nổ còn lại sẽ có mức độ tin cậy rất cao.
- Chúng tôi dự kiến công bố:
    - 01 bài báo hội nghị quốc tế thuộc danh mục SCOPUS.

- Một báo cáo chi tiết về kết quả nghiên cứu.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] W. H. Organization, “Diarrhoeal disease,” 2017, https://www.gso.gov. vn/en/population/ [Accessed: (July 10, 2023)].

[2] D. Onozuka and M. Hashizume, “Weather variability and paediatric infectious gastroenteritis,” Epidemiology & Infection, vol. 139, no. 9, pp. 1369–1378, 2011.

[3] D. Phung et al., “Association between climate factors and diarrhoea in a mekong delta area,” International journal of biometeorology, vol. 59, pp. 1321–1331, 2015.

[4] D. PHUNG et al., “Temporal and spatial patterns of diarrhoea in the mekong delta area, vietnam,” Epidemiology amp; Infection, vol. 143, no. 16, p. 3488–3497, 2015.

[5] C. N. Thompson et al., “The impact of environmental and climatic variation on the spatiotemporal trends of hospitalized pediatric diarrhea in ho chi minh city, vietnam,” Health & place, vol. 35, pp. 147–154, 2015.

[6] D. Phung et al., “Heavy rainfall and risk of infectious intestinal diseases in the most populous city in vietnam,” Science of The Total Environment, vol. 580, pp. 805–812, 2017.

[7] K. Wangdi and A. C. Clements, “Spatial and temporal patterns of diarrhoea in bhutan 2003–2013,” BMC infectious diseases, vol. 17, no. 1, pp. 1–9, 2017.

[8] R. D’souza et al., “Climatic factors associated with hospitalizations for rotavirus diarrhoea in children under 5 years of age,” Epidemiology & Infection, vol. 136, no. 1, pp. 56–64, 2008.

[9] T. D. Do et al., “Diarrhoea incidence prediction using climate data: Machine learning approaches,” in 2022 RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE, 2022, pp. 1–6.

[10] V.-H. Nguyen et al., “Deep learning models for forecasting dengue fever based on climate data in vietnam,” PLoS Neglected Tropical Diseases, vol. 16, no. 6, p. e0010509, 2022.