

# BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

## (HƯỚNG NGHIÊN CỨU)

**Tên đề tài :** Vấn đề bất đồng trong giải thích của XAI khi ứng dụng trên dữ liệu ảnh bằng Saliency Map

**Sinh viên thực hiện :** 19120298 – Mai Duy Nam

19120460 – Nguyễn Hữu Bình

**Giảng viên phản biện:** ThS. Lê Ngọc Thành

### 1. Chủ đề và ý tưởng nghiên cứu:

Trong thế giới ngày càng phụ thuộc vào trí tuệ nhân tạo và học máy, việc hiểu được lý do mà các mô hình dự đoán như thế nào đã trở thành một yêu cầu quan trọng. Đối với các hệ thống học máy mạnh mẽ như mạng nơ-ron sâu, giải thích các quyết định của chúng là một vấn đề thách thức. Điều này dẫn đến xu hướng sử dụng các kỹ thuật Giải thích Trí tuệ nhân tạo (XAI) để giúp hiểu cách mà các mô hình đưa ra các quyết định dự đoán. Tuy nhiên, một trong những thách thức đối với XAI là vấn đề bất đồng. Một mô hình có thể có nhiều giải thích khác nhau cho cùng một dự đoán, hoặc hai mô hình khác nhau có thể đưa ra các giải thích khác nhau cho cùng một trường hợp dữ liệu.

Để hiểu hơn độ bất đồng này, nhóm sinh viên thực hiện các nghiên cứu các phương pháp giải thích khác nhau trên dữ liệu ảnh chụp phổi liên quan đến hai bệnh viêm phổi và tràn dịch màng phổi. Sau đó thực hiện so sánh kết quả giải thích trên các độ đo. Dựa trên kết quả so sánh, nhóm phân tích và đưa ra nhận định các yếu tố gây ra các điểm bất đồng này.

### 2. Phương pháp nghiên cứu:

Nhóm sinh viên thực hiện nghiên cứu một số công trình của các tác giả như Krishna, Neely, Brughmans, và Roy. Sau đó chọn ra 8 phương pháp giải thích khác nhau gồm DeepLift, GradientShap, GuidedBackprop, GuidedGradCam, IntgegratedGraidents, LRP, Occlusion, Saliency. Hai mô hình blackbox được chọn là InceptionV3 và RestNet101 do kiến trúc nhỏ gọn, thời gian huấn luyện nhanh. Nhóm sinh viên chạy thực nghiệm trên 2 tập dữ liệu và so sánh các kết quả trên các chỉ số SSIM, Feature Agreement, Sign agreement, Rank Correlation. Các nhận xét được rút ra dựa trên các kết quả phân tích này.

### **3. Đóng góp Khoa học và thực tiễn:**

Về lý thuyết, nhóm sinh viên đã tìm hiểu về chủ đề bất đồng trong giải thích và trình bày các phương pháp giải thích khác nhau. Tuy nhiên, các mô tả cần chi tiết thêm và có các minh họa để thể hiện sự khác nhau của phương pháp giải thích. Đối với mô hình black box, nhóm đã trình bày tóm tắt đặc điểm cơ bản của hai mô hình InceptionV3 và ResNet101. Về thực nghiệm, nhóm sinh viên chủ yếu chạy lại các phương pháp giải thích được tích hợp sẵn trong thư viện Captum. Các tương quan được thể hiện chủ yếu trên bản đồ nhiệt. Chưa đổi chiều nhiều đối với các kết quả đã được công bố trong các công trình liên quan. Nhóm sinh viên đã rút ra được một số nhận xét, chú trọng vào tính chất gradient. Tuy nhiên, các nhận xét chưa có cơ sở hay chứng minh rõ ràng để thuyết phục.

### **4. Báo cáo viết:**

Báo cáo trình bày bằng tiếng Anh gồm 6 chương. Chương 4 cần xem xét lại tiêu đề phù hợp hơn. Phần mô tả Dataset (4.3.3) nên chuyển sang chương thực nghiệm (Chương 5).

### **5. Trình bày trước hội đồng:**

Trình bày tốt trước hội đồng.

### **6. Công bố khoa học/ ứng dụng thực tế:**

Chưa có công bố khoa học.

**Đánh giá xếp loại:** Đáp ứng yêu cầu của khóa luận cử nhân CNTT.

TP.HCM, ngày 01 tháng 08 năm 2023

**Giảng viên phản biện**

(Ký và ghi rõ họ tên)



**Lê Ngọc Thành**