**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KĨ THUẬT THÔNG TIN**



**NHẬP MÔN BẢO ĐẢM VÀ AN NINH THÔNG TIN**

**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

|  |  |
| --- | --- |
| **GVHD:** | TS. Nguyễn Tấn Cầm |
|  |  |

**Lớp:** IE105.O21

**Sinh viên thực hiện:** Trần Trung Hiếu – 22520446

Võ Quang Nhật Hoàng – 22520482

🙡🙢 Tp. Hồ Chí Minh, 02/2024 🙠🙣

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……., ngày……...tháng……năm 2024*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên****)***

**MỤC LỤC**

***I. Explainable AI (XAI) 6***

***II. Đặc điểm chung của SHAP và LIME 8***

***III. Phương pháp giải thích SHAP 9***

***IV. Phương pháp giải thích LIME 12***

***V. Sự khác biệt về nguyên lý làm việc 14***

***VI. So sánh hiệu năng SHAP và LIME 15***

***VII. Kết luận và tương lai của Explainable AI 15***

**MỤC LỤC HÌNH ẢNH**

***Hình 1: Biểu đồ thác nước thể hiện giải thích của dự đoán 11***

***Hình 2:******Thể hiện mức độ và vị trí ảnh hưởng của giá trị SHAP 11***

***Hình 3: Hình ảnh gốc ta cung cấp cho LIME 13***

***Hình 4: Những điểm ảnh mà LIME nghĩ là chó Labrador 13***

**MỤC LỤC BẢNG**

***Bảng 1:*** ***Bảng so sánh nguyên lý làm việc của SHAP và LIME 14***

***Bảng 2: Bảng so sánh hiệu năng SHAP và LIME 15***

**NỘI DUNG BÀI LÀM**

**I. Explainable AI (XAI)**

***1. Định nghĩa của Explainable AI***

*- Explainable Artificial Intelligence (XAI)* là một tập hợp quy trình và phương pháp cho phép người dùng hiểu và tin tưởng vào kết quả và đầu ra được tạo ra bởi các thuật toán học máy.

- *Explainable AI* được sử dụng để mô tả một mô hình trí tuệ nhân tạo (AI) giúp xác định độ chính xác của mô hình, tính minh bạch, công bằng và kết quả trong quá trình ra quyết định dựa trên AI. XAI giúp xây dựng niềm tin và sự tự tin trong việc sử dụng AI và cũng giúp đảm bảo việc phát triển AI được thực hiện một cách có trách nhiệm.

***2. Tầm quan trọng của Explainable AI***

- Cốt lõi của việc sử dụng Explainable AI là giải quyết vấn đề về niềm tin. Nếu chúng ta không thể tin tưởng được một hệ thống thì nó sẽ trở thành một thách thức trong việc sử dụng nó một cách tối ưu. Nên khi đến với những mô hình học máy, niềm tin là vô cùng quan trọng vì nó đảm bảo rằng mô hình đưa ra nhưng quyết định đúng đắn.

- Hơn nữa, nó giúp chúng ta biết được vấn đề khi một mô hình thất bại, đưa cho chúng ta kiểm soát về hiệu năng của mô hình tốt hơn và chúng ta có thể đưa ra lựa chọn rằng nên cải tiến mô hình dựa trên kiến thức đó hay không sử dụng mô hình đó trong những tình huống nhất định.

***3. Phân loại Explainable AI***

XAI được phân loại theo phương pháp diễn giải 1 mô hình học máy. Cấu trúc của XAI có thể được xác định và phân loại bằng phương pháp khác nhau, tùy thuộc vào mục đích, thời điểm, hay tính ứng dụng. Nhưng chung quy lại đều được phân chia theo các loại sau:

a. Intrinsic Explainations (giải thích nội tại) và Post-hoc Explainations (giải thích hậu tại)

- Phương pháp nội tại đề cập đến việc giới hạn độ phức tạp của mô hình. Khi cấu trúc của các mô hình đã đủ đơn giản, chúng ta có thể hiểu cách mô hình đưa ra dự đoán bằng việc nhìn vào chính mô hình.

* Một số mô hình nội tại phổ biến bao gồm: Linear Regression, Logistic Regression, GLM, GAM, RuleFit, KNN, Naive Bayes,…

- Trong khi đó, các phương pháp hậu tại phân tích mô hình sau khi đã được đào tạo, ví dụ như việc hoán vị các đặc trưng quan trọng (permutation feature importance). Các phương pháp này gồm 2 nhánh nhỏ là mô hình bất khả tri (model-agnostic) và mô hình cụ thể (model-specific).

b. Model-agnostic (mô hình bất khả tri) và Model-specific (mô hình cụ thể)

- Đối với mô hình bất khả tri, ta có thể áp dụng nó cho bất kì mô hình học máy nào, trong khi đó, đối với mô hình cụ thể ta chỉ có thể áp dụng cho 1 số loại mô hình nhất định.

- Ví dụ, một phương pháp chỉ hoạt động với mạng neuron (Neural Network) thì ta có thể coi phương pháp đó chỉ dành cho riêng mô hình đó. Ngược lại, nếu 1 phương pháp có thể dùng để giải thích cho mọi mô hình đã được đào tạo là hộp đen thì nó sẽ được coi là mô hình bất khả tri.

c. Local (cục bộ) và Global (toàn cục)

- Phương pháp giải thích cục bộ nhằm giải thích 1 điểm dữ liệu hoặc 1 dự đoán duy nhất, còn phương pháp toàn cục cố gắng cung cấp lời giải thích toàn diện về cách mô hình hoạt động dựa trên toàn bộ dữ liệu và dự đoán. Các phương pháp cục bộ có thể biến thành kỹ thuật toàn cục thông qua việc tổng hợp các kết quả cục bộ.

* Một số ví dụ cho phương pháp cục bộ như: Individual Conditional Expectation Curves, LIME, Counterfactual Explanations, Shapley values, SHAP,...
* Một số phương pháp toàn cục phổ biến bao gồm Partial dependence plot, Accumulated local effect plots, Feature interaction (H-statistics), Functional decomposition, Permutation feature importance…

**II**. **Đặc điểm chung của SHAP và LIME**

- SHAP và LIME đều thuộc loại giải thích hậu tại (post-hoc explanation), nghĩa là chúng không ảnh hưởng đến quá trình huấn luyện của mô hình mà chỉ can thiệp sau khi mô hình đã được xây dựng xong.

- SHAP và LIME đều là các phương pháp giải thích Black Box Model. Nghĩa là chúng được thiết kế để làm rõ cách thức hoạt động của các mô hình phức tạp, vốn thường được gọi là mô hình đen hộp vì tính phức tạp và khó hiểu của chúng. Cả hai phương pháp này đều cố gắng giải thích các quyết định của mô hình bằng cách đưa ra các yếu tố đầu vào (features) và đánh giá tầm quan trọng của chúng đối với kết quả đầu ra của mô hình. Qua đó, người dùng có thể hiểu rõ hơn về cách mà mô hình đưa ra các quyết định dựa trên dữ liệu.

- SHAP và LIME đều có khả năng giải thích bất kỳ loại mô hình nào, từ các mô hình tuyến tính đơn giản đến các mô hình phi tuyến phức tạp như cây quyết định, mạng nơ-ron, và máy học sâu. Đây là đặc điểm làm cho cả hai phương pháp trở nên rất linh hoạt và dễ áp dụng trong nhiều tình huống khác nhau. Điều này đặc biệt quan trọng trong các ứng dụng thực tế, nơi mà dữ liệu và mô hình thường rất đa dạng và phức tạp.

- Một điểm giống nhau khác là cả hai phương pháp đều hướng đến việc tạo ra các giải thích cục bộ (local explanations). Điều này có nghĩa là SHAP và LIME không cố gắng giải thích toàn bộ mô hình một cách tổng quát, mà thay vào đó, chúng tập trung vào việc giải thích các quyết định cụ thể của mô hình cho từng điểm dữ liệu cụ thể. Cách tiếp cận này giúp các giải thích trở nên chi tiết và cụ thể hơn, từ đó giúp người dùng hiểu rõ hơn về cách mô hình hoạt động trong các tình huống cụ thể.

- SHAP và LIME sử dụng các mẫu ngẫu nhiên để tạo ra các giải thích. LIME sử dụng các mẫu dữ liệu giả lập để xây dựng mô hình đơn giản quanh một điểm dữ liệu cụ thể, trong khi SHAP sử dụng các mẫu ngẫu nhiên của các hoán vị các yếu tố đầu vào để tính toán giá trị Shapley. Vì vậy, cả hai phương pháp có thể áp dụng linh hoạt trên nhiều loại dữ liệu và mô hình khác nhau mà không cần thay đổi cấu trúc cơ bản của chúng.

**III. Phương pháp giải thích SHAP**

***1. Định nghĩa***

- SHAP (SHapley Additive exPlanations) là một AI khá phổ biến được sử dụng trong học máy. Sử dụng Lý thuyết game để giải thích cho kết quả của các mô hình học máy. Nó kết nối việc phân bổ tín dụng tối ưu với giải thích cục bộ bằng cách sử dụng các giá trị Shapley cổ điển có từ lý thuyết game và các tiện ích mở rộng liên quan. Từ đó SHAP sẽ đưa ra kết quả cho dự đoán.

***2. Phương pháp thực hiện***

- SHAP sẽ tính toán các giá trị Shapley cho từng đặc trưng đầu vào. Sau đó nó sẽ tính toán mức độ ảnh hưởng của từng đặc trưng đó đến kết quả dự đoán của mô hình.

- Để tính toán giá trị Shapley, SHAP thực hiện các bước sau:

**1. Tạo các tập dữ liệu con:**

- Cho một điểm dữ liệu cụ thể mà bạn muốn giải thích, sau đó SHAP sẽ tạo ra một tập các tập dữ liệu con.

- Mỗi tập dữ liệu con bao gồm các đặc điểm dữ liệu ban đầu và một số tính năng được chọn ngẫu nhiên.

**2. Tính toán dự đoán cho mỗi tập dữ liệu con:** Sau khi đã tạo các tập dữ liệu con. Dựa vào model mà ta đã sử dụng để train, SHAP sẽ tính toán các dự đoán cho từng tập dữ liệu con. Đối với mỗi tập dữ liệu con, giá trị của từng dự đoán sẽ được tính toán dựa trên các tính năng có mặt trong tập dữ liệu con đó.

**-** Quá trình này lặp lại cho tất cả các tập dữ liệu đã tạo ở bước 1.

**3. Tính toán mức độ đóng góp của mỗi features:** Sử dụng kết quả dự đoán từ các tập dữ liệu con.

→ SHAP sẽ tính toán mức độ đóng góp của mỗi features cho dự đoán của mô hình cho dữ liệu ban đầu.

**4. Tổng hợp mức độ đóng góp:** Mức độ đóng góp của mỗi features sẽ được tổng hợp lại để đưa ra kết quả dự đoán của mô hình.

**→** Qua các bước này, SHAP cung cấp một giải thích giải thích đầy đủ và toàn diện về cách mà mô hình AI đưa ra các dự đoán, giúp người dùng hiểu rõ hơn về cách mà mỗi đặc trưng ảnh hưởng đến kết quả của dự đoán.

***3. Công thức SHAP***

a. Công Thức SHAP

- Giả sử ta có một mô hình *f*  và một bộ đặc trưng *x* gồm *N* đặc trưng {*x*1, *x*2, …, *xN*}.

b. Giá trị Shapley cho một đặc trưng *i* được tính bằng công thức:

A black and white math equation

Description automatically generatedTrong đó:

* + *S* là một tập con của tất cả các đặc trưng trừ đặc trưng *i*.
  + *F(S)* là giá trị dự đoán của mô hình khi chỉ có các đặc trưng trong tập *S*.
  + ∣*S*∣ là số lượng đặc trưng trong tập *S*.
  + ∣*N* ∣ là tổng số đặc trưng.

c. Giải Thích Công Thức

* *ϕ*i(*f*, *x*) là giá trị Shapley của đặc trưng *i*, đại diện cho sự đóng góp của *i* vào dự đoán của mô hình.
* **∑**S⊆N∖{*i*}là tổng hợp tất cả các tập con *S* không chứa *i*.
* là trọng số cho mỗi tập con *S*. Trọng số này đảm bảo rằng tất cả các cách sắp xếp của các đặc trưng được xem xét công bằng.
* *f* (*S* ∪ {*i*}) – *f* (*S*) là sự thay đổi trong giá trị dự đoán khi thêm đặc trưng *i* vào tập con *S*.

A graph with numbers and text

Description automatically generated

*Hình 1:* *Biểu đồ thác nước thể hiện giải thích của dự đoán*

A collage of birds

Description automatically generated

*Hình 2:* *Thể hiện mức độ và vị trí ảnh hưởng của giá trị SHAP*

**IV. Phương pháp giải thích LIME (Local Interpretable Model-agnostic Explaination)**

- Giải thích dựa trên việc xây dựng một mô hình đơn giản và dễ hiểu hơn để mô phỏng cách mà một mô hình (thường là một mô hình tuyến tính đơn giản như hồi quy tuyến tính) AI đưa ra dự đoán cho một điểm dữ liệu cụ thể. Thay vì giải thích toàn bộ mô hình AI, LIME tập trung vào việc xây dựng một mô hình giải thích đơn giản hơn, dễ hiểu hơn bằng cách tạo ra các điểm dữ liệu giả định xung quanh điểm dữ liệu đã chọn và sử dụng chúng để huấn luyện một mô hình giải thích.

- Công thức chính của LIME không phải là một công thức toán học duy nhất, mà là một quy trình bao gồm nhiều bước. Dưới đây là quy trình cơ bản của LIME:

*1. Chọn điểm dữ liệu cần giải thích:*

- Giả sử ta có một điểm dữ liệu *x* cần được giải thích.

*2. Tạo ra dữ liệu giả định (perturbations):*

- Tạo ra nhiều biến thể của *x* bằng cách thêm nhiễu vào *x* . Những điểm dữ liệu này tạo thành tập dữ liệu *Z* .

*3. Dự đoán với mô hình phức tạp:*

- Sử dụng mô hình phức tạp ban đầu (black-box model) để dự đoán các nhãn hoặc xác suất cho các điểm trong tập dữ liệu *Z* . Kết quả này được gọi là *f(z)* với z \in *Z* .

*4. Tính trọng số (weights):*

- Tính trọng số cho mỗi điểm dữ liệu trong *Z* dựa trên độ giống nhau của chúng với điểm gốc *x* . Trọng số có thể được tính bằng cách sử dụng một hàm kernel như hàm Gaussian:

A math equation with black text

Description automatically generated

Trong đó, D(*x*, *z*) là khoảng cách giữa *x* và *z* , và *σ* là một tham số điều chỉnh.

*5. Huấn luyện mô hình đơn giản (interpretable model):*

- Sử dụng các điểm dữ liệu trong Z cùng với các trọng số \pi\_x(z) và các dự đoán *f(z)* để huấn luyện một mô hình đơn giản, như mô hình hồi quy tuyến tính hoặc cây quyết định, để dự đoán *f(z)* .

*6. Giải thích kết quả:*

- Mô hình đơn giản này được sử dụng để giải thích dự đoán của mô hình phức tạp tại điểm *x*. Các tham số của mô hình đơn giản cho thấy tầm quan trọng của các đặc trưng của *x* đối với dự đoán của mô hình phức tạp.

Tóm lại, LIME sử dụng một quy trình bao gồm việc tạo dữ liệu giả định, tính trọng số cho các điểm dữ liệu giả định, và huấn luyện một mô hình đơn giản để giải thích các dự đoán của mô hình phức tạp. Điều này giúp người dùng hiểu rõ hơn về cách mà mô hình phức tạp đưa ra dự đoán tại một điểm dữ liệu cụ thể.

A dog and cat touching noses

Description automatically generated

*Hình 3:* *Hình ảnh gốc ta cung cấp cho LIME*

A dog's face on a black background

Description automatically generated

*Hình 4: Những điểm ảnh mà LIME nghĩ là chó Labrador*

**V. Sự khác biệt về nguyên lý làm việc**

- Việc lựa chọn giữa hai phương pháp này thường phụ thuộc vào mục tiêu cụ thể của người dùng và độ phức tạp của mô hình đang làm việc.

|  |  |  |
| --- | --- | --- |
|  | **SHAP** | **LIME** |
| **Mục tiêu chính** | Tập trung vào việc tính toán giá trị Shapley, để phân bổ đóng góp của mỗi đặc trưng vào dự đoán của mô hình. | Tập trung vào việc xây dựng các mô hình đơn giản cục bộ để giải thích dự đoán của mô hình AI tại một điểm dữ liệu cụ thể. |
| **Phương pháp tính toán** | Dựa trên một phương pháp tính toán được xác định cụ thể bằng các giá trị Shapley hoặc các phương pháp ước lượng SHAP khác như TreeSHAP. | Sử dụng các mô hình đơn giản như hồi quy tuyến tính để xấp xỉ mô hình AI và giải thích dự đoán. |
| **Mô hình giải thích** | Không tạo ra một mô hình giải thích cụ thể. Thay vào đó, nó cung cấp một giá trị Shapley cho mỗi đặc trưng. | Xây dựng một mô hình giải thích cục bộ để mô phỏng cách mà Mô hình AI đưa ra dự đoán tại một điểm dữ liệu cụ thể. |
| **Điểm mạnh** | Tính toán chính xác giá trị Shapley, một phép đo đối với các đặc trưng, cung cấp một cái nhìn tổng quan về cách mà mỗi đặc trưng đóng góp vào dự đoán.  Có thể áp dụng cho mọi loại mô hình, từ tuyến tính đến mạng nơ-ron. | Tạo ra các mô hình đơn giản cục bộ, dễ hiểu và dễ diễn giải.  Tính linh hoạt cao, có thể được áp dụng cho các mô hình hộp đen mà không cần biết cấu trúc của chúng. |
| **Điểm yếu** | Tính toán SHAP có thể phức tạp và tốn tài nguyên, đặc biệt là cho các mô hình phức tạp. | Khả năng chính xác của mô hình giải thích có thể phụ thuộc nhiều vào việc chọn lựa và xử lý dữ liệu con. |

*Bảng 1:* *Bảng so sánh nguyên lý làm việc của SHAP và LIME*

**VI. So sánh hiệu năng SHAP và LIME**

|  |  |  |
| --- | --- | --- |
|  | **SHAP** | **LIME** |
| **Độ tin cậy** | - SHAP sử dụng lý thuyết Shapley để cung cấp các giải thích có tính chất toàn vẹn, tức là đảm bảo tính chắc chắn và khả năng tái sản xuất.  → Thường đáng tin cậy hơn so với LIME. | - Tính chắc chắn của LIME có thể bị ảnh hưởng bởi việc lựa chọn dữ liệu giả định.  → Cũng cung cấp giải thích khá tin cậy, nhưng không đảm bảo tính toàn vẹn như SHAP. |
| **Tính ổn định** | - Sử dụng lý thuyết Shapley, đảm bảo tính ổn định trong việc cung cấp giải thích dữ liệu.  →Thường ổn định hơn LIME. | - Phụ thuộc vào dữ liệu giả định và mô hình tuyến tính được sử dụng.  → Không ổn định bằng SHAP |
| **Tính linh hoạt** | - Thường linh hoạt hơn trong việc xử lý các loại mô hình và dữ liệu khác nhau, bao gồm cả mô hình tuyến tính và mô hình học sâu. | - Khá linh hoạt, tuy nhiên, có thể gặp khó khăn khi áp dụng cho một số loại mô hình phức tạp hoặc dữ liệu có cấu trúc phức tạp. |
| **Độ phức tạp** | - Sử dụng lý thuyết Shapley và cần phải tính toán các giá trị Shapley cho mỗi điểm dữ liệu.  → Đòi hỏi nhiều tài nguyên tính toán hơn. | - Thường đơn giản hơn và nhanh chóng hơn trong việc tính toán giải thích, đặc biệt là khi sử dụng mô hình tuyến tính. |
| **Hiệu suất** | - Hiệu suất tính toán có thể thấp hơn so với LIME do tính phức tạp của thuật toán. | - Thường có hiệu suất tính toán tốt hơn do sử dụng phương pháp đơn giản hơn và ít phụ thuộc hơn vào dữ liệu giả định. |

*Bảng 2: Bảng so sánh hiệu năng SHAP và LIME*

**KẾT LUẬN**

**VII. Kết luận và tương lai của Explainable AI**

- Cả SHAP và LIME đều là những công cụ quan trọng trong lĩnh vực giải thích trí tuệ nhân tạo (XAI), và có thể được cải tiến để tăng cường hiệu suất và tính ứng dụng của chúng trong tương lai. Sau đây là một số hướng phát triển tiềm năng cho cả hai phương pháp:

**1. Tăng cường tính ổn định và tin cậy:**

- Nghiên cứu về cách tăng cường tính ổn định và tin cậy của cả SHAP và LIME sẽ giúp cải thiện khả năng sử dụng của các phương pháp này trong các ứng dụng thực tế.

- Phát triển các phương pháp mới để đánh giá và đảm bảo tính chắc chắn và ổn định của các giải thích được tạo ra bởi SHAP và LIME.

**2. Tối ưu hóa tính hiệu suất:**

- Nghiên cứu cách tối ưu hóa hiệu suất tính toán của SHAP và LIME để giảm thiểu thời gian và tài nguyên tính toán cần thiết cho quá trình giải thích mô hình học máy.

- Phát triển các phương pháp và công nghệ mới để tăng cường hiệu suất tính toán của SHAP và LIME trên các nền tảng phần cứng khác nhau.

**3. Mở rộng khả năng áp dụng:**

- Nghiên cứu và phát triển các phương pháp mở rộng khả năng áp dụng của SHAP và LIME cho nhiều loại mô hình học máy và tập dữ liệu khác nhau.

- Tích hợp SHAP và LIME vào các công cụ và framework học máy phổ biến để tạo ra các công cụ mạnh mẽ hơn cho việc giải thích mô hình.

**4. Tích hợp thêm thông tin ngữ cảnh:**

- Nghiên cứu và phát triển cách tích hợp thêm thông tin ngữ cảnh vào quá trình giải thích của SHAP và LIME, giúp cải thiện sự hiểu biết và tin cậy của các giải thích.

- Sử dụng các phương pháp học máy khác, chẳng hạn như mạng nơ-ron hồi quy, để hỗ trợ việc tích hợp thông tin ngữ cảnh vào các giải thích.

**5. Khám phá các ứng dụng mới:**

- Nghiên cứu và phát triển các ứng dụng mới cho SHAP và LIME trong các lĩnh vực mới, bao gồm y - sinh học, tài chính, và hợp tác xã hội.

- Khám phá các cách sử dụng SHAP và LIME để hỗ trợ đưa ra quyết định trong các tình huống phức tạp và quan trọng.

**→** Bằng cách thúc đẩy các hướng phát triển này, SHAP và LIME có thể ngày càng trở thành những công cụ quan trọng hơn trong việc hiểu và giải thích các mô hình trí tuệ nhân tạo.

**TÀI LIỆU THAM KHẢO**

+ [Explainable AI Methods – A Brief Overview](https://link.springer.com/chapter/10.1007/978-3-031-04083-2_2?utm_medium=referral&utm_source=slink&utm_content=RM&utm_term=null&utm_campaign=HSCR_BOOKS_AWA1_GL_MPAS_005KU_LNCS50-AA) – *Tác giả: Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. – 2022*

+ [Explainable AI - Giải Thích AI Bằng Cách Nào](https://codelearn.io/sharing/explainable-ai-giai-thich-bang-cach-nao) – *Tác giả: khangnvt1* – *09/12/2020*

*+* [XAI—Explainable artificial intelligence](https://sci-hub.se/https:/www.science.org/doi/abs/10.1126/scirobotics.aay7120) – *Tác giả: David Gunning, Mark Stefik, Jaesik Choi , Timothy Miller, Simone Stumpf, Guang-Zhong Yang – 18/12/2019*

[+ Comparison of Contextual Importance and Utility with LIME and Shapley Values](https://sci-hub.st/https:/link.springer.com/chapter/10.1007/978-3-030-82017-6_3) *– Tác giả: Kary Främling, Marcus Westberg, Martin Jullum, Manik Madhikermi & Avleen Malhi – 2021*

[+ SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk](https://www.frontiersin.org/articles/10.3389/frai.2021.752558/full) *– Tác giả: Alex Gramegna, Paulo Giudici – 2021*

+ [Extracting spatial effects from machine learning model using local interpretation method](https://www.sciencedirect.com/science/article/pii/S0198971522000898) – *Tác giả: Ziqi Li* – *18/06/2022*

+ [Failure mode and effects analysis of RC members based on machinelearning-based SHapley Additive exPlanations (SHAP) approach](https://sci-hub.se/https:/www.sciencedirect.com/science/article/abs/pii/S0141029620307513) – *Tác giả: Sujith Mangalathu , Seong-Hoon Hwang, Jong-Su Jeon – 12/06/2020*

+ [Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis](https://sci-hub.se/https:/www.sciencedirect.com/science/article/abs/pii/S0001457519311790) – *Tác giả: Amir Bahador Parsaa, Ali Movahedia, Homa Taghipoura, Sybil Derribleb, Abolfazl (Kouros) Mohammadian – 15/12/2019*

+ [What Does LIME Really See in Images?](http://proceedings.mlr.press/v139/garreau21a/garreau21a.pdf) *– Tác giả: Damien Garreau, Dina Mardaoui – 2021*

+ [XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions](https://arxiv.org/pdf/2206.03585) *– Tác giả: Gautam Srivastava, Rutvij H Jhaveri, Sweta Bhattacharya, Sharnil Pandya, Rajeswari, Praveen Kumar Reddy Maddikunta, Gokul Yenduri, Jon G. Hall, Mamoun Alazab, Thippa Reddy Gadekallu – 3/6/2022*

[+ Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP](https://www.nature.com/articles/s41598-023-35795-0) *-Tác giả: Rasheed Omobolaji Alabi, Mohammed Elmusrati, Ilmo Leivo, Alhadi Almangush & Antti A. Mäkitie – 2/2023*