**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT VĨNH LONG**

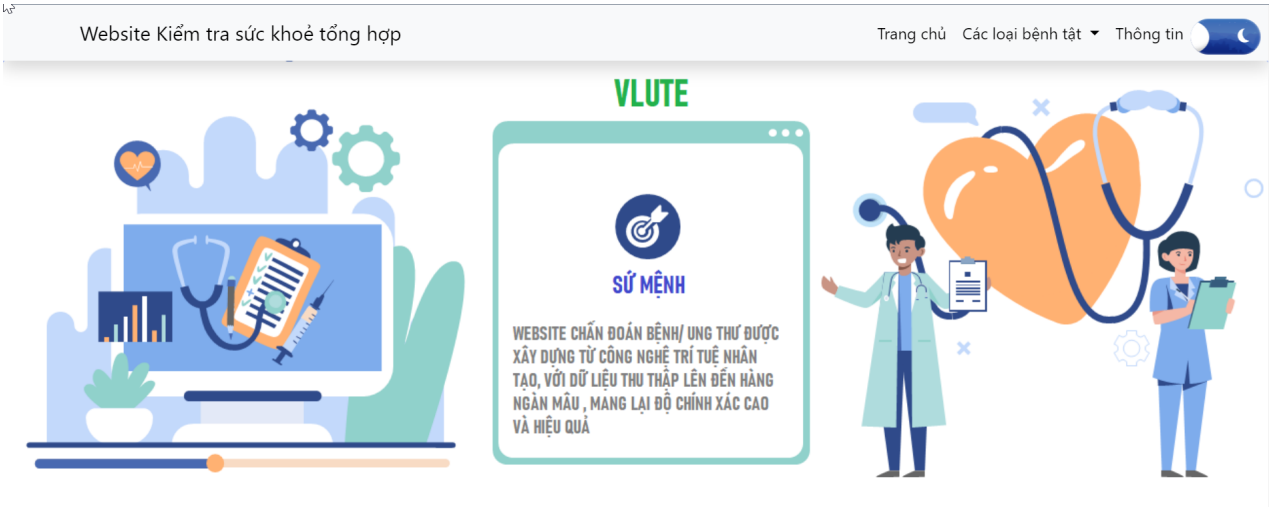
**KHOA CÔNG NGHỆ THÔNG TIN**

————————o0o————————

# HỌC PHẦN TƯƠNG TÁC NGƯỜI MÁY

# WEBSITE ÁP DỤNG MACHINE LEARNING ĐỂ CHẨN ĐOÁN CÁC VẤN ĐỀ VỀ Y SINH

*Giảng viên hướng dẫn:* **ThS. Trần Hồ Đạt**





Sinh viên: **18004115 - Nguyễn Duy Tân**

**18004134 - Lê Thị Mỹ Tiên**

**18004135 - Nguyễn Mạnh Tiến**

Lớp: **ĐH. CNTT 2018**


**VĨNH LONG, 10/2021**

# MỞ ĐẦU

Tác giả: *Nguyễn Duy Tân, Lê Thị Mỹ Tiên , Nguyễn Mạnh Tiến* [1]

**Abstracts:** Trí tuệ nhân tạo là một ngành khoa học máy tính mới mẽ và phát triển mạnh với khả năng ứng dụng với nhiều lĩnh vực. Trong đó, đặc biệt với ngành y học - sinh học. Từ thực tiễn đó, nhóm đã lên ý tưởng về việc xây dựng một website được ứng dụng công nghệ A.I thông qua các giải thuật được xây dựng với độ chính xác cao bằng một lượng lớn dữ liệu thu thập được để từ đó có thể tính toán - chẩn đoán các vấn đề bệnh lý ở người. Dự kiến kết quả thu được là chẩn đoán các bệnh lý về người với độ chính xác cao, và mở rộng quy mô phần mềm

**Keyword:** Machine Learning - Biomedical - Malaria - Diabetes - Fetal Health

# TỔNG QUAN

Dựa theo các bài nghiên cứu khoa học hàng đầu thế giới như NCBI, CARE, ACA hay WHO, nhóm đã lấy đó làm nền tảng cơ sở để nghiên cứu và phát triển phần mềm trở nên trực quan, sinh động và tối ưu hơn. Trong đó các bài báo khoa học nhóm dùng để thực hiện cho việc nghiên cứu đó là tiểu đường và sốt rét, cụ thể như sau:

- **Tiểu đường - đái tháo đường**: có lẽ là một trong những căn bệnh lâu đời nhất được biết đến với con người. Lần đầu tiên nó được báo cáo trong bản thảo Ai Cập về 3000 năm trước. Nhưng các vấn đề với nó vẫn còn tồn tại và dai dẳng đến ngày nay, theo thống kê của WHO, số người mắc bệnh tiểu đường tăng từ 108 triệu người năm 1980 lên 422 triệu người trong năm 2014, tức là tỉ lệ tăng lên 400% trong vòng 34 năm. Từ năm 2000 đến năm 2016, tỉ lệ tử vong vì bệnh tiểu đường tăng lên 5%, đặc biệt trong năm 2019, 1.5 triệu người đã ra đi trực tiếp vì tiểu đường và 2.2 triệu người đã ra đi vì lượng đường trong máu tăng cao.

  Dưới sức ảnh hưởng nghiêm trọng và tàn phá nặng nề như vậy, nhóm đã tìm các tài liệu về việc chẩn đoán tiểu đường bằng trí tuệ nhân tạo, như nhận biết bằng Faster-RCNN [1], nhận biết bằng phương pháp Random Forest Classifier [2], nhận biết bằng phương pháp hồi quy Logistics [3]

- **Sốt rét - Malaria:** khoảng một nửa dân số thế giới vẫn có nguy cơ sốt rét. Bệnh sốt rét hay gặp ở châu Phi, Ấn Độ và các khu vực khác của Nam Á, Đông Nam Á, Bắc và Nam Triều Tiên, Mexico, Trung Mỹ, Haiti, Cộng hòa Dominican, Nam Mỹ (bao gồm các phần phía bắc Argentina), Trung Đông (bao gồm Thổ Nhĩ Kỳ, Syria, Iran, và Iraq), và Trung Á. CDC cung cấp thông tin về các quốc gia cụ thể nơi truyền bệnh sốt rét (Yellow Fever and Malaria Information [4]), các loại sốt rét, các mô hình đề kháng và đề phòng dự phòng.

  Năm 2015, có khoảng 214 triệu ca sốt rét trên toàn thế giới, với 438.000 ca tử vong, chủ yếu ở trẻ em < 5 tuổi ở Phi Châu. Kể từ năm 2000, tử vong do sốt rét đã giảm

60% nhờ nỗ lực của Roll Back Malaria Program, trong đó có trên 500 đối tác (bao gồm các quốc gia lưu hành và các tổ chức, tổ chức khác nhau). Sốt rét đã 1 lần thành dịch ở Mỹ. Hiện tại có khoảng 1500 trường hợp xảy ra ở Mỹ mỗi năm. Gần như tất cả đều nhiễm ở nước ngoài, nhưng một số nhỏ do truyền máu hoặc hiếm khi lây truyền bởi muỗi địa phương đốt người nhập cư hoặc những người đi du lịch bị nhiễm bệnh.

Với sự phát triển mạnh mẽ của trí tuệ nhân tạo, ta có thể xác định sốt rét thông qua các tế bào hồng cầu, với ý tưởng đó, nhóm đã tham khảo các nguồn tài liệu như: nhận biết sốt rét bằng phương pháp mạng nơ-ron tích hợp (CNN)[5], nhận biết bằng phương pháp ResNet-5[6], nhận biết bằng phương pháp VGG-19 [7]

- **Sức khoẻ thai nhi:** Mang thai và sự phát triển của thai nhi là một quá trình sinh học cực kỳ phức tạp, trong quá trình mang thai có thể là bào thai khoẻ mạnh, cũng có thể là có nguy cơ tiềm ẩn sức khoẻ không tốt, hoặc thật sự bào thai có vấn đề. Và một trong những phương pháp để xác định thai nhi có phát triển theo đúng mong đợi hay không là chụp tim mạch. Mục đích của kỹ thuật chẩn đoán này là đo nhịp tim của thai nhi và các cơn co thắt tử cung của mẹ, thường là trong ba tháng cuối của thai kỳ khi tim của thai nhi hoạt động đầy đủ. Kết quả đầu ra của điện tâm đồ thường được hiểu là thuộc một trong ba trạng thái: bình thường, nghi ngờ (tiềm ẩn bệnh) và mắc bệnh lý. Xác định mục tiêu như vậy, nhóm đã lần lượt tìm hiểu và phân tích các phương pháp khác nhau, cụ thể: chẩn đoán sức khoẻ thai nhi bằng cách sử dụng phương pháp Random Forest Classification (RFC) [8], chẩn đoán bằng phương pháp cây quyết định (Decision Tree Classifier)[9]

# PHÂN TÍCH ƯU & NHƯỢC ĐIỂM

## Theo dõi sức khỏe thai nhi

*Ưu điểm*

- Sử dụng máy chụp tim (CTG) [10] trong việc theo dõi thai nhi bằng cách tiếp cận dựa trên bằng chứng.

- Ghi hình CTG là để xác định khi có lo ngại về tình trạng sức khỏe của thai nhi để cho phép tiến hành các biện pháp can thiệp trước khi thai nhi bị tổn hại.

-Phân tích tỷ lệ mắc bệnh do thiếu oxy máu cho thấy sử dụng CTG làm giảm tỷ lệ co giật ở thời kỳ sơ sinh và tỷ lệ bại não.

*Nhược điểm*

- CTG chỉ có thể được đánh giá nếu tỷ lệ tín hiệu thất bại dưới 15%

- Các kết quả đọc CTG phải luôn được đánh giá bởi một nữ hộ sinh hoặc bác sĩ và các kết quả đọc phải được ký tắt bằng chữ ký có thể nhận dạng được, đối với bệnh nhân hoặc thân nhân thì yêu cầu có số liệu bệnh án mới có thể thực hiện thử nghiệm lâm sàng.

- Chịu tác động từ nhiều yếu tố bên ngoài khi thu thập tín hiệu

## Bệnh tiểu đường - đái tháo đường

*Ưu điểm* - Triển khai thuật toán AI hoạt động tốt trong chẩn đoán GDM trong một bối cảnh đòi hỏi ít nhân viên và thiết bị y tế hơn và thiết lập một ứng dụng dựa trên thuật toán AI.

- Cho thấy kết quả có hiệu suất tương đương với các bác sĩ lâm sàng

- Thuật toán chẩn đoán tự động có thể cung cấp chẩn đoán chính xác và thời gian thực với ít nguồn lực y tế hơn, đòi hỏi ít thiết bị hơn và nhân viên y tế chuyên nghiệp

- Giải quyết các vấn đề y tế ở những khu vực thiếu tài nguyên và ứng dụng này có thể sẽ giúp giải quyết tình trạng thiếu nguồn lực y tế

*Nhược điểm*

- Tập dữ liệu của tương đối nhỏ chỉ đến từ một bệnh viện.

- Bộ dữ liệu liên quan đến dân số Quảng Đông và Hồng Kông, cả về đặc điểm bệnh nhân và hệ thống nên khả năng áp dụng cho các nhóm dân số khác còn hạn chế

- Không có thuật toán nào hoạt động tốt về độ nhạy so với các chuyên gia về con người, minh họa rằng khả năng chẩn đoán GDM của AI đòi hỏi phải cải thiện.

# Mục lục


**1 CƠ SỞ LÝ THUYẾT** **1**

1.1 TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO . . . . . 1

1.1.1 KHÁI NIỆM . . . . . 1

1.1.2 QUÁ TRÌNH . . . . . 1

1.2 TỔNG QUAN VỀ TƯƠNG TÁC NGƯỜI MÁY . . . . . 1

1.2.1 KHÁI NIỆM . . . . . 1

1.2.2 QUÁ TRÌNH . . . . . 2

**2 CÔNG CỤ THỰC HIỆN** **3**

2.1 PAGESPEED INSIGHTS GOOGLE . . . . . 3

2.1.1 Tổng quan về Pagespeed Insights Google . . . . . 3

2.2 PAGESPEED INSIGHTS GOOGLE . . . . . 3

2.2.1 Tổng quan về Pagespeed Insights Google . . . . . 3

2.3 PAGESPEED INSIGHTS GOOGLE . . . . . 4

2.3.1 Tổng quan về Pagespeed Insights Google . . . . . 4

**3 NỘI DUNG NGHIÊN CỨU** **5**

3.1 Tiêu chuẩn CMMI . . . . . 5

3.1.1 CMMI là gì? . . . . . 5

3.2 Tiêu chuẩn CMMI . . . . . 5

3.2.1 CMMI là gì? . . . . . 5

3.3 Tiêu chuẩn CMMI . . . . . 6

3.3.1 CMMI là gì? . . . . . 6

**4 KẾT QUẢ THỰC HIỆN** **7**

4.1 Tập dữ liệu được sử dụng để huấn luyện . . . . . 7

4.2 Thực nghiệm . . . . . 8

4.2.1 Xử lí dữ liệu . . . . . 8

**5 KẾT LUẬN** **9**

5.1 Kết quả đạt được . . . . . 9

5.2 Hạn chế khó khăn . . . . . 9

5.3 Hướng phát triển . . . . . 9

# Danh sách hình vẽ

# Danh sách bảng

# Danh sách thuật ngữ

| Tiếng Anh | Tiếng Việt |
|---|---|
| accuracy | độ chính xác |
| anomalous | bất thường, dị thường |
| artificial intelligence | trí tuệ nhân tạo, trí thông minh nhân tạo |
| attribute | thuộc tính |
| binary decision tree | cây quyết định nhị phân |
| child node | nút con |
| classification | phân loại |
| conditional probability | xác suất có điều kiện |
| continuous | liên tục |
| cross-validation score | độ chính xác khi dự đoán trên tập kiểm thử |
| cross validation | một kĩ thuật để đánh giá độ chính xác của mô hình |
| decision tree | cây quyết định |
| discrete | rời rạc |
| entropy | độ hỗn loạn thông tin |
| gini index | chỉ số đo lường độ không sạch của dữ liệu |
| indicator variable | biến nhị phân |
| information gain | độ lợi thông tin |
| joint probobability | xác xuất hợp |
| label | nhãn |
| leaf node | nút lá, nút không có con |
| leave-one-out | còn lại một |
| machine learning | máy học, học máy |
| non-leaf node | nút trong, nút có con |
| normal | bình thường |
| outcome | đầu ra của dữ liệu |
| overfitting | quá khớp |
| random variable | biến ngẫu nhiên |
| regression | hồi quy |
| request body | phần nội dung trong gói tin HTTP request |
| request line | chỉ dòng đầu tiên trong gói tin HTTP request |
| root node | nút gốc |

| Tiếng Anh | Tiếng Việt |
|---|---|
| sibling node | các nút có cùng nút cha |
| supervised learning | học có giám sát |
| test error | mất mát trên dữ liệu kiểm tra |
| test data | dữ liệu kiểm tra |
| traffic | lưu lượng truy cập mạng |
| train error | mất mát trên dữ liệu huấn luyện |
| training score | độ chính xác khi huấn luyện |
| training data | dữ liệu huấn luyện |
| underfitting | không khớp |
| unsupervised learning | học không giám sát |
| validation | một kĩ thuật để đánh giá độ chính xác của mô hình |
| validation set | tập dữ liệu đánh giá |
| validation error | mất mát trên tập đánh giá |

# Chương 1

# CƠ SỞ LÝ THUYẾT

## 1.1 TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO

GHI ĐẠI Ý

### 1.1.1 KHÁI NIỆM

J. McCarthy là người đầu tiên đưa cụm từ "Trí tuệ nhân tạo" (artificial intelligence-AI) trở thành một khái niệm khoa học. Trong [27], J. McCarthy và cộng sự cho rằng nghiên cứu TTNT nhằm mô tả chính xác các khía cạnh của xử lý trí tuệ và học (để có được tri thức) và tạo ra được các hệ thống, máy mô phỏng hoạt động học và xử lý trí tuệ. Ở giai đoạn đầu, TTNT hướng tới xây dựng các hệ thống, máy có khả năng sử dụng ngôn ngữ tự nhiên, trừu tượng hóa -hình thức hóa các khái niệm và giải quyết vấn đề dựa trên tiếp cận lô gic, ra quyết định trong điều kiện thiếu thông tin. TTNT là lĩnh vực liên ngành của Triết học, Tâm lý học, Khoa học thần kinh, Toán học, Điều khiển học, Khoa học máy tính, Ngôn ngữ học, Kinh tế

### 1.1.2 QUÁ TRÌNH

GHI ĐẠI Ý

## 1.2 TỔNG QUAN VỀ TƯƠNG TÁC NGƯỜI MÁY

### 1.2.1 KHÁI NIỆM

J. McCarthy là người đầu tiên đưa cụm từ "Trí tuệ nhân tạo" (artificial intelligence-AI) trở thành một khái niệm khoa học. Trong [27], J. McCarthy và cộng sự cho rằng nghiên cứu TTNT nhằm mô tả chính xác các khía cạnh của xử lý trí tuệ và học (để có được tri thức) và tạo ra được các hệ thống, máy mô phỏng hoạt động học và xử lý trí tuệ. Ở giai đoạn đầu, TTNT hướng tới xây dựng các hệ thống, máy có khả năng sử dụng ngôn ngữ tự nhiên, trừu tượng hóa -hình thức hóa các khái niệm và giải quyết vấn đề dựa trên tiếp cận lô gic, ra quyết định trong điều kiện thiếu thông tin. TTNT là lĩnh vực liên ngành của Triết học, Tâm lý học, Khoa học thần kinh,

Toán học, Điều khiển học, Khoa học máy tính, Ngôn ngữ học, Kinh tế

### 1.2.2 QUÁ TRÌNH

GHI ĐẠI Ý

# Chương 2

# CÔNG CỤ THỰC HIỆN

## 2.1 PAGESPEED INSIGHTS GOOGLE

### 2.1.1 Tổng quan về Pagespeed Insights Google

Pagespeed Insights là công cụ tối ưu hóa hiệu suất website, cũng như đưa ra đánh giá chi tiết cho website của bạn. Một trang web chất lượng và được tối ưu hóa tốt sẽ có vai trò đặc biệt quan trọng trong việc xây dựng thương hiệu cũng như tăng khả năng tiếp cận khách hàng mục tiêu. Pagespeed Insights chính là công cụ hữu hiệu trong việc hỗ trợ người dùng phân tích và đánh giá trang web, từ đó đưa ra những đề xuất chỉnh sửa sao cho hoàn thiện nhất. Pagespeed Insights là một công cụ do Google phát triển. Công cụ này được nhiều chuyên gia trang web lựa chọn nhằm tối ưu hiệu suất. Đồng thời cũng đánh giá chất lượng cho trang web của mình, dựa trên những tiêu chuẩn đánh giá của Google. Khi dùng Pagespeed Insight, người dùng có thể nhận báo cáo về hiệu suất Website trên cả máy tính và di động. Ngoài ra, công cụ này cũng sẽ cung cấp cho người dùng những đề xuất nhằm tối ưu trang web thông qua báo cáo UX của Chrome.

## 2.2 PAGESPEED INSIGHTS GOOGLE

### 2.2.1 Tổng quan về Pagespeed Insights Google

Pagespeed Insights là công cụ tối ưu hóa hiệu suất website, cũng như đưa ra đánh giá chi tiết cho website của bạn. Một trang web chất lượng và được tối ưu hóa tốt sẽ có vai trò đặc biệt quan trọng trong việc xây dựng thương hiệu cũng như tăng khả năng tiếp cận khách hàng mục tiêu. Pagespeed Insights chính là công cụ hữu hiệu trong việc hỗ trợ người dùng phân tích và đánh giá trang web, từ đó đưa ra những đề xuất chỉnh sửa sao cho hoàn thiện nhất. Pagespeed Insights là một công cụ do Google phát triển. Công cụ này được nhiều chuyên gia trang web lựa chọn nhằm tối ưu hiệu suất. Đồng thời cũng đánh giá chất lượng cho trang web của mình, dựa trên những tiêu chuẩn đánh giá của Google. Khi dùng Pagespeed Insight, người dùng có thể nhận báo cáo về hiệu suất Website trên cả máy tính và di động. Ngoài ra, công cụ này cũng sẽ cung cấp cho người dùng những đề xuất nhằm tối ưu trang web thông qua báo cáo UX của Chrome.

## 2.3 PAGESPEED INSIGHTS GOOGLE

### 2.3.1 Tổng quan về Pagespeed Insights Google

Pagespeed Insights là công cụ tối ưu hóa hiệu suất website, cũng như đưa ra đánh giá chi tiết cho website của bạn. Một trang web chất lượng và được tối ưu hóa tốt sẽ có vai trò đặc biệt quan trọng trong việc xây dựng thương hiệu cũng như tăng khả năng tiếp cận khách hàng mục tiêu. Pagespeed Insights chính là công cụ hữu hiệu trong việc hỗ trợ người dùng phân tích và đánh giá trang web, từ đó đưa ra những đề xuất chỉnh sửa sao cho hoàn thiện nhất. Pagespeed Insights là một công cụ do Google phát triển. Công cụ này được nhiều chuyên gia trang web lựa chọn nhằm tối ưu hiệu suất. Đồng thời cũng đánh giá chất lượng cho trang web của mình, dựa trên những tiêu chuẩn đánh giá của Google. Khi dùng Pagespeed Insight, người dùng có thể nhận báo cáo về hiệu suất Website trên cả máy tính và di động. Ngoài ra, công cụ này cũng sẽ cung cấp cho người dùng những đề xuất nhằm tối ưu trang web thông qua báo cáo UX của Chrome.

# Chương 3

# NỘI DUNG NGHIÊN CỨU

## 3.1 Tiêu chuẩn CMMI

### 3.1.1 CMMI là gì?

CMMI được viết tắc của (Capability Maturity Model Integration) được phát triển tại Viện Kỹ Nghệ Phần Mềm của Mỹ (Viện SEI – nay đổi thành Viện CMMI) tại trường Đại học Carnegie Mellon ở Pittsburgh. CMMI là mô hình năng lực trưởng thành tích hợp cung cấp một định nghĩa rõ ràng về những hành động cần được doanh nghiệp xúc tiến để nâng cao năng suất hoạt động. Với năm "Mức trưởng thành" hoặc ba "Mức năng lực", CMMI xác định những yếu tố quan trọng nhất để xây dựng nên hệ thống có thể sản xuất ra những sản phẩm tốt, hoặc cung cấp dịch vụ tốt. Là một tập các phương thức và giải pháp nhằm tối ưu hóa quy trình phát triển phần mềm. Trọng tâm chính của CMMI là tập trung xây dựng các công cụ hỗ trợ việc cải thiện các quy trình dùng để phát triển và ổn định các hệ thống và sản phẩm. Kết quả của CMMI là một bộ các sản phẩm cung cấp một phương pháp tiếp cận tích hợp trên toàn doanh nghiệp.

## 3.2 Tiêu chuẩn CMMI

### 3.2.1 CMMI là gì?

CMMI được viết tắc của (Capability Maturity Model Integration) được phát triển tại Viện Kỹ Nghệ Phần Mềm của Mỹ (Viện SEI – nay đổi thành Viện CMMI) tại trường Đại học Carnegie Mellon ở Pittsburgh. CMMI là mô hình năng lực trưởng thành tích hợp cung cấp một định nghĩa rõ ràng về những hành động cần được doanh nghiệp xúc tiến để nâng cao năng suất hoạt động. Với năm "Mức trưởng thành" hoặc ba "Mức năng lực", CMMI xác định những yếu tố quan trọng nhất để xây dựng nên hệ thống có thể sản xuất ra những sản phẩm tốt, hoặc cung cấp dịch vụ tốt. Là một tập các phương thức và giải pháp nhằm tối ưu hóa quy trình phát triển phần mềm. Trọng tâm chính của CMMI là tập trung xây dựng các công cụ hỗ trợ việc cải thiện các quy trình dùng để phát triển và ổn định các hệ thống và sản phẩm. Kết quả của CMMI là một bộ các sản phẩm cung cấp một phương pháp tiếp cận tích hợp trên toàn doanh nghiệp.

## 3.3 Tiêu chuẩn CMMI

### 3.3.1 CMMI là gì?

CMMI được viết tắc của (Capability Maturity Model Integration) được phát triển tại Viện Kỹ Nghệ Phần Mềm của Mỹ (Viện SEI – nay đổi thành Viện CMMI) tại trường Đại học Carnegie Mellon ở Pittsburgh. CMMI là mô hình năng lực trưởng thành tích hợp cung cấp một định nghĩa rõ ràng về những hành động cần được doanh nghiệp xúc tiến để nâng cao năng suất hoạt động. Với năm "Mức trưởng thành" hoặc ba "Mức năng lực", CMMI xác định những yếu tố quan trọng nhất để xây dựng nên hệ thống có thể sản xuất ra những sản phẩm tốt, hoặc cung cấp dịch vụ tốt. Là một tập các phương thức và giải pháp nhằm tối ưu hóa quy trình phát triển phần mềm. Trọng tâm chính của CMMI là tập trung xây dựng các công cụ hỗ trợ việc cải thiện các quy trình dùng để phát triển và ổn định các hệ thống và sản phẩm. Kết quả của CMMI là một bộ các sản phẩm cung cấp một phương pháp tiếp cận tích hợp trên toàn doanh nghiệp.

# Chương 4

# KẾT QUẢ THỰC HIỆN

## 4.1 Tập dữ liệu được sử dụng để huấn luyện

Dữ liệu dùng cho giai đoạn xây dựng cây là tập dữ liệu CSIC 2010 [1].

Tập dữ liệu **HTTP CSIC 2010** chứa những traffic nhắm đến những ứng dụng web thương mại điện tử phát triển tại CSIC. Tập dữ liệu này được tạo tự động và chứa khoảng 36.000 những request bình thường và hơn 25.000 request bất thường đã được gán nhãn. Tập dữ liệu bao gồm những dạng tấn công như: *SQL injection, Bufer overflow, information gathering, files disclosure, CRLF injection, XSS, server side include, parameter tampering, ....*

Lưu lượng web này được tạo ra bằng các bước sau:

- Đầu tiên, dữ liệu thật được thu thập với tất cả các tham số của ứng dụng web. Tất cả các dữ liệu (như: *tên, họ, địa chỉ,...*) được lấy chính xác từ cơ sở dữ liệu thực tế. Những giá trị này được lưu trữ trong hai cơ sở dữ liệu: **normal** (*bình thường*) và **anomalous** (*bất bình thường*). Ngoài ra, tất cả các trang của ứng dụng web cũng được liệt kê.
- Kế đó, những *requests normal* và *anomalous* được tạo cho mỗi trang của web. Trong trường hợp requests normal, các tham số được lắp đầy với dữ liệu được lấy từ *cơ sơ dữ liệu normal* một cách ngẫu nhiên. Quá trình xử lý tương tự với requests anomalous, các tham số được lấy từ cơ sở dữ liệu anomalous.

Có ba loại **requests anomalous** được quan tâm:

- **Static attacks:** cố gắng truy cập vào các tài nguyên bị ẩn. Những requests này bao gồm: những tập tin ít dùng, *Session ID* trong URL rewrite, những tập tin cấu hình, những tập tin mặc định, ...
- **Dynamic attacks:** chỉnh lại những tham số hợp lệ của request để thực hiện các cuộc tấn công *SQL injection, CRLF injection, XSS, buffer overflows*, ...

[1] **CISC** (viết tắt của *Consejo Superior de Investigaciones Científicas* theo tiếng Tây Ban Nha) là tổ chức cộng đồng lớn nhất dành cho nghiên cứu ở Tây Ban Nha, và lớn thứ 3 ở Châu Âu. Tổ chức này tạo ra 20% trong tổng số bài báo khoa học trong nước.

- **Unintentional illegel requests**: những requests này không cố ý chứa những thứ độc hại, tuy nhiên họ không tuân theo những hành vi bình thường của ứng dụng web và không có cấu trúc như những tham số bình thường. Ví dụ trường (field) nhập số điện thoại có kiểu là *số* nhưng người dùng lại nhập vào đó là *ký tự*.

Tập dữ liệu này được chia thành ba phần khác nhau. Một phần cho giai đoạn *huấn luyện*, chỉ chứa những traffic bình thường. Và hai phần còn lại được dùng cho giai đoạn *kiểm tra*, một với những traffic bình thường, một với những traffic malicious (lưu lượng độc hại).

## 4.2 Thực nghiệm

Chọn **Python 3** để hiện thực thuật toán [2].

### 4.2.1 Xử lí dữ liệu

Từ tập dữ liệu CSIC 2010 trình bày ở trên, tiến hành xử lý để đưa các HTTP request thành những vector đặc trưng phục vụ cho quá trình huấn luyện.

Trong các HTTP requests, không phải tất cả các tham số đều có giá trị sử dụng (tham số có ảnh hưởng đến quyết định đầu ra). Dựa vào các lỗ hổng phổ biến chúng tôi quyết định tập trung vào phần **Request Line** và **Request Message Body** trong cấu trúc của gói tin HTTP request được mô tả ở hình **??** để khai thác.

Cụ thể các đặc trưng được chọn để chuyển đổi thành vector như sau:

[2]Toàn bộ source code được kèm theo báo cáo này

# Chương 5

# KẾT LUẬN

## 5.1 Kết quả đạt được

Sau khi thực hiện các đánh giá đảm bảo chất lượng phần mềm từ quy trình kiểm thử chất lượng phần mềm. Các kết quả mang lại được là: Nhận thức và đánh giá được các tiêu chuẩn mà Việt Nam Lược Sử cần phải đạt được, hoạch định được quy trình và kế hoạch cần thiết mà một website cần có. Đồng thời hiểu rõ và sâu sắc về quy tắc khi xây dựng một trang web cho người dùng, đồng thời tiếp nhận được kiến thức về lĩnh vực đảm bảo hệ thống nói chung và nói riêng ở đây là kiến thức hoạt động mà một website cần có. Nắm được các cấu trúc cũng như quy tắc hoạt động của từng ngôn ngữ hay các công cụ hỗ trợ trong quá trình sử dụng và tìm hiểu được áp dụng cho trang website. Đảm bảo được đầy đủ các chức năng phù hợp với nhu cầu người xem đẩy mạnh lượt truy cập trang website từ đó trang website sẽ được phổ biến hơn.

## 5.2 Hạn chế khó khăn

Tốc độ tải trang chỉ tương đối. Yếu tố này phụ thuộc vào nên tảng, chỉ mới thực hiện trên các hosting miễn phí. Các báo cáo thì khẳng định tốc độ tải trang ảnh hưởng đến trải nghiệm người dùng và cảm nhận của họ đối với dịch vụ được cung cấp. Chưa đánh giá được phần bảo mật cao vì website chỉ mới sử dụng hosing free nên phần bảo mật còn chưa tối ưu. Chưa hoàn toàn tối ưu được mọi nền tảng, đặc biệt là ở các nền tảng màn hình nhỏ như Mobile.

## 5.3 Hướng phát triển

Sử dụng công nghệ Responsive cho một website hoàn hảo. Công nghệ này mang lại lợi ích cho tất cả các website trong mọi lĩnh vực, đặc biệt là website du lịch, website bán vé máy bay, website bán hàng, website tin tức,... Tạo được hiệu ứng sử dụng tin tức khi đưa lên trang website cho người xem lẫn người quản trị hệ thống của trang website. Cập nhật thông tin và nhu cầu người xem hơn

# TÀI LIỆU THAM KHẢO

# Tài liệu tham khảo


[1] Saleh Albahli, Tahira Nazir, Aun Irtaza and Ali Javed *Recognition and Detection of Diabetic Retinopathy UsingDensenet-65 Based Faster-RCNN*, CMC, 2021, vol.67, no.2

[2] Xuchun Wang, MengMeng Zhai, ZepingRen, HaoRen, *Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier*, BMC Medical Informatics and Decision Making, Article: 105 (2021)

[3] Ram D. Joshi and Chandra K. Dhakal *Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches*, Environmental Research and Public Health, Published: 9 July 2021

[4] Mark D. Gershman, Emily S. Jentes, Rhett J. Stoney (Yellow Fever) Kathrine R. Tan, Paul M. Arguin (Malaria) *Yellow Fever Vaccine & Malaria Prophylaxis Information*, by Country.

[5] W. David Pan, Yuhang Dong and Dongsheng Wu*Classification of Malaria-Infected Cells Using Deep Convolutional Neural Networks*, Published: September 19th 2018, DOI: 10.5772/intechopen.72426

[6] A. Sai Bharadwaj Reddy; D. Sujitha Juliet *Transfer Learning with ResNet-50 for Malaria Cell-Image Classification*, 4-6 April 2019, INSPEC Accession Number: 18619266

[7] Rishika Kapoor*TMalaria Detection using Deep Convolutional Neural Network*, June 2017

[8] T. Peterek, P. Gajdos *Human Fetus Health Classification on Cardiotocographic Data Using Random Forests* DOI:10.1007/978-3-319-07773-4 19Corpus ID: 49407443

[9] M. Ramla; S. Sangeetha; S. Nickolas *Fetal Health State Monitoring Using Decision Tree Classifier from Cardiotocography Measurements* DOI: 10.1109/ICCONS.2018.8663047 11 March 2019

[10] Dr. Hayley Willacy, *Cardiotocography*, 28 Jun 2021