

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA KỸ THUẬT MÁY TÍNH

NGUYỄN DUY TÀI  
VŨ VĂN MẠNH

KHÓA LUẬN TỐT NGHIỆP  
XÁC ĐỊNH CHẤT LƯỢNG TRÚNG DỰA TRÊN CẢM  
BIẾN QUANG VÀ AI  
DETERMINATION OF EGGS QUALITY USING  
OPTICAL SENSOR AND AI

KỸ SƯ NGÀNH KỸ THUẬT MÁY TÍNH

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA KỸ THUẬT MÁY TÍNH

NGUYỄN DUY TÀI – 19522152

VŨ VĂN MẠNH – 19521831

KHÓA LUẬN TỐT NGHIỆP  
XÁC ĐỊNH CHẤT LƯỢNG TRỨNG DỰA TRÊN CẢM  
BIÊN QUANG VÀ AI

DETERMINATION OF EGGS QUALITY USING  
OPTICAL SENSOR AND AI

KỸ SƯ NGÀNH KỸ THUẬT MÁY TÍNH

GIẢNG VIÊN HƯỚNG DẪN  
PHẠM QUỐC HÙNG

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

## **THÔNG TIN HỘI ĐỒNG CHẤM KHÓA LUẬN TỐT NGHIỆP**

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số 11/QĐ-ĐHCNTT, ngày 05 tháng 01 năm 2024 của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

## LỜI CẢM ƠN

Lời nói đầu tiên, nhóm em muốn bày tỏ lời cảm ơn chân thành đến tất cả các giảng viên, các thầy cô khoa Kỹ thuật máy tính của trường Đại học Công nghệ thông tin Đại học Quốc gia Thành phố Hồ Chí Minh đã tạo điều kiện và cơ hội cho nhóm em thực hiện luận văn tốt nghiệp này. Tất cả các thầy cô đều luôn hỗ trợ cho nhóm em bằng rất nhiều các khía cạnh, qua việc truyền đạt kiến thức và kinh nghiệm. Đó là những giá trị vô cùng quý báu mà nhóm chúng em vô cùng cảm kích.

Đặc biệt, nhóm em xin gửi lời cảm ơn chân thành và sự tri ân đối với thầy hướng dẫn của nhóm là thầy Ts. Phạm Quốc Hùng. Cảm ơn thầy, đã dành ra những thời gian quý báu để hướng dẫn nhóm em, chỉ bảo nhóm em tận tình trong khoảng thời gian làm luận văn và cả trong khoảng thời gian dài gần ba năm nay.

Nhóm chúng em xin chân thành cảm ơn và muôn thể hiện lòng biết ơn của nhóm em đối với gia đình và bạn bè, những người luôn khuyên khích chúng tôi đạt được mục tiêu của mình và cung cấp cho nhóm em sự hỗ trợ vô điều kiện trong suốt chặng đường sinh viên qua.

Trong quá trình làm luận văn tốt nghiệp, nhóm đã học thêm được nhiều kinh nghiệm quý báu. Tuy nhiên, trong quá trình làm cũng khó tránh khỏi sai sót, rất mong các thầy cô bỏ qua.

Chúng em xin chân thành cảm ơn!

Sinh viên thực hiện

**Vũ Văn Mạnh**

Khoa Kỹ thuật máy tính – Lớp MTCL2019.2

Sinh viên thực hiện

**Nguyễn Duy Tài**

Khoa Kỹ thuật máy tính – Lớp MTCL2019.3

## MỤC LỤC

<b>TÓM TẮT KHÓA LUẬN.....</b>	8
<b>CHƯƠNG 1. GIỚI THIỆU VỀ ĐỀ TÀI .....</b>	12
1.1. Tổng quan đề tài.....	12
1.2. Lý do chọn đề tài .....	14
1.3. Mục tiêu đề tài .....	15
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT .....</b>	16
2.1. Các phương pháp đo chất lượng trứng gà.....	16
2.1.1. Phương pháp đo truyền thống.....	16
2.1.2. Phương pháp đo xâm lấn .....	17
2.1.3. Phương pháp đo không xâm lấn.....	18
2.2. Ứng dụng quang phổ trong việc đánh giá chất lượng và độ tươi của trứng .....	18
2.2.1. Quang phổ .....	18
2.2.2. Quang phổ cận hồng ngoại NIR .....	19
2.2.3. Các ưu điểm của phổ cận hồng ngoại NIR .....	20
2.2.4. Phân loại quang phổ cận hồng ngoại .....	20
2.2.5. Cơ sở lý thuyết về dải bước sóng của trứng gà .....	21
2.3. Machine Learning .....	23
2.3.1. Giới thiệu về Machine Learning.....	23
2.3.2. Học có giám sát .....	23
2.3.3. Học không giám sát.....	23
2.4. Phương pháp đo .....	24
<b>CHƯƠNG 3. THIẾT KẾ HỆ THỐNG NHÚNG.....</b>	27
3.1. Tổng quan hệ thống nhúng .....	27
3.2. Thành phần hệ thống .....	27
3.2.1. Bóng đèn Halogen Philips W5W T10.....	27
3.2.2. Cảm Biến AS7263 .....	28
3.2.3. Vi điều khiển Raspberry Pi 4B .....	31
3.2.4. Màn hình Oled .....	32
3.2.5. Module relay 5v.....	33

3.2.6. Nguồn tần số UP30DAC.....	34
3.3. Thiết bị và phần mềm hỗ trợ.....	35
3.3.1. Thiết bị hỗ trợ cân tiểu ly điện tự 500g và thước đo 20cm .....	35
3.3.2. Các ứng dụng hỗ trợ .....	36
3.4. Mô hình hệ thống .....	41
3.4.1. Mô tả kết nối hệ thống .....	41
3.4.2. Vị trí đặt mẫu vật.....	42
3.4.3. Lưu đồ giải thuật việc thu thập dữ liệu thô từ mẫu vật trứng gà .....	43
3.4.4. Lưu đồ giải thuật mô hình dự đoán chất lượng, độ tươi trứng gà .....	44
<b>CHƯƠNG 4. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN CHẤT LƯỢNG ĐỘ TUƠI CỦA TRỨNG GÀ CÔNG NGHIỆP MÀU NÂU .....</b>	<b>45</b>
4.1. Thu thập tập dữ liệu thô .....	45
4.2. Tiền xử lý dữ liệu.....	46
4.2.1. Standard Scaler .....	47
4.2.2. Standard Normal Variate_SNV và Multiplicative Scatter Correction_MSC .....	49
4.3. Mô hình máy học thực nghiệm .....	51
4.3.1. Mô hình Multiple Linear Regression .....	51
4.3.2. Mô hình Support Vector Regression (SVM) .....	52
4.3.3. Decision Tree .....	55
4.4. Đánh giá mô hình máy học .....	55
4.4.1. Đánh giá mô hình máy học dựa trên độ chính xác phân tích $R^2$ .....	56
4.4.2. Đánh giá mô hình dựa trên độ lệch gốc căn bậc hai của mức trung bình bình phương (RMSE) .....	57
<b>CHƯƠNG 5. KẾT QUẢ THỰC NGHIỆM.....</b>	<b>58</b>
5.1. Hiện thực phần cứng .....	58
5.2. Tập dữ liệu thô thu được .....	60
5.2.1. Mô tả chi tiết tập dữ liệu .....	60
5.2.2. Đánh giá cảm quan trứng gà qua các ngày .....	64
5.2.3. Đánh giá sự thay đổi của trứng gà qua các ngày thông Haugh Units.....	67
5.2.4. Dải bước sóng NIR qua các ngày .....	68

<b>5.3. Tiết xử lý dữ liệu.....</b>	69
<b>5.4. Thực nghiệm mô hình Multiple Linear Regression.....</b>	72
<b>5.5. Thực hiện mô hình Decision Tree .....</b>	77
<b>5.6. Thực hiện mô hình Support Vector Regression .....</b>	80
<b>5.7. Lưu và tải mô hình máy học.....</b>	85
<b>5.8. Kiểm thử mô hình dự đoán .....</b>	86
<b>CHƯƠNG 6. KẾT LUẬN .....</b>	88
<b>6.1. Kết quả đạt được .....</b>	88
<b>6.2. Những điều đạt được khi thực hiện đề tài .....</b>	89
<b>6.3. Những khó khăn trong quá trình thực hiện đề tài.....</b>	89
<b>6.4. Hướng phát triển chính cho đề tài.....</b>	90
<b>TÀI LIỆU THAM KHẢO.....</b>	93

## **DANH MỤC HÌNH**

Hình 1.1: Sự khác biệt chất lượng trứng gà vào ngày bảo quản số 0 và 28.

Hình 2.1: Phương pháp đánh giá chất lượng trứng bằng cách thả trứng vào nước.

Hình 2.2: Đo chiều cao của lòng trắng trứng.

Hình 2.3 Ví dụ trung bình dữ liệu quang phổ thô của trứng theo thời gian lưu trữ.

Hình 2.4 Hình vẽ biểu diễn hàm tuyế́n tính với đơn biế́n.

Hình 2.5: Cấu hình khoa học phân nhánh và cấu hình quang học  $0^\circ/45^\circ$ .

Hình 2.6: Phản xạ gương.

Hình 1.1: Dải bước sóng của bóng đèn Halogen Philips W5W T10.

Hình 1.2: Bóng đèn Halogen Philips W5W T10.

Hình 3.3 SparkFun AS7263.

Hình 3.4 Dải 6 bước sóng đỉnh thu được của cảm biến AS7263.

Hình 3.5: Raspberry Pi 4B.

Hình 3.6 Màn hình OLED.

Hình 3.7: Module relay 5v.

Hình 3.8: Nguồn tổ ong UP30DAC.

Hình 3.9: Cân tiểu ly điện tử và thước đo.

Hình 3.10: Advanced Ip Scanner tìm kiếm địa chỉ của Raspberry.

Hình 3.11: Bật VNC Server thông qua Putty.

Hình 3.12: VNC hoạt động dựa trên mô hình client/server.

Hình 3.13: Sử dụng VNC Viewer lập trình trên Raspberry từ xa.

Hình 3.14: Sử dụng Win SCP chuyển file qua lại với Raspberry.

Hình 3.15 Sơ đồ kết nối các thành phần của hệ thống.

Hình 3.16: Mô phỏng các vị trí đặt mẫu vật trứng gà để thu thập data.

Hình 3.17: Lưu đồ giải thuật thu thập dữ liệu thô từ trứng gà.

Hình 3.18: Lưu đồ giải thuật của toàn bộ hệ thống.

Hình 4.1: Ví dụ về SVM.

Hình 5.1: Mặt bên trên, mặt sau của phần cứng sau khi hiện thực.

Hình 5.2 Mặt trước của phần cứng sau khi hiện thực.

Hình 5.3: Raspberry bị lỗi “won’t boot fix”.

Hình 5.4: Các vị trí đặt trứng thực nghiệm.

Hình 5.5: Mẫu vật trứng gà nhận vào 22/11/2023 được để trong vỉ không bị bẹp hay vỡ.

Hình 5.6: Đo độ cao lòng trắng trứng thủ công.

Hình 5.7: Tập dữ liệu thô ở vị trí ở giữa quả trứng (mid).

Hình 5.8: Lòng đỏ và lòng trắng của trứng gà tại ngày số 4.

Hình 5.9: Lòng đỏ và lòng trắng của trứng gà tại ngày số 9.

Hình 5.10: Trứng xuất hiện những nấm li ti trên khắp bề mặt vỏ trứng (ngày 16).

Hình 5.11: Lòng trắng và lòng đỏ trứng gà có vỏ màu sẫm ngày 18.

Hình 5.12: Biểu đồ sự thay đổi trung bình của Haugh Units qua số ngày tuổi của trứng.

Hình 5.13 Sự thay đổi của bước sóng qua các ngày.

Hình 5.14: chương trình của 2 phương pháp tiền xử lý Standard Normal Variate\_SNV và Multiplicative Scatter Correction\_MSC.

Hình 5.15: SNV với tập dữ liệu thô mid + bot.

Hình 5.16: MSC với tập dữ liệu thô mid + bot.

Hình 5.17 Dữ liệu thô sau khi tiền xử lý standard scaler.

Hình 5.18 Kiểm tra giá trị trung bình và độ lệch chuẩn.

Hình 5.19 Chia tập dữ liệu tiền xử lý thành tập huấn luyện và tập kiểm định.

Hình 5.20 Kết quả train mô hình Multiple Linear Regression.

Hình 5.21: Kết quả train mô hình Decision Tree bằng Google Colab.

Hình 5.22 Ví dụ về Margin trong SVM.

Hình 5.23: Kết quả huấn luyện mô hình Support Vector Regression bằng Google Colab.

Hình 5.24 Lưu mô hình máy học.

## **DANH MỤC BẢNG**

- Bảng 1.1: Kết quả một số nghiên cứu đánh giá chất lượng trứng sử dụng NIR.
- Bảng 1.1: So sánh quang phổ cận hồng ngoại bước sóng dài với bước sóng ngắn.
- Bảng 2.2: Sự khác nhau giữa cấu hình quang học phân nhánh và cấu hình quang học  $0^{\circ}/45^{\circ}$ .
- Bảng 1.1: Thông số kỹ thuật của bóng đèn Halogen Philips W5W T10.
- Bảng 1.2: Thông số kỹ thuật cảm biến AS7263.
- Bảng 3.3 Đặc tính quang học của AS7263.
- Bảng 1.4: Thông số kỹ thuật của Raspberry Pi 4B.
- Bảng 1.5: Thông số kỹ thuật của màn hình Oled 0.96 Inch.
- Bảng 1.6: Thông số kỹ thuật của module relay 5v.
- Bảng 1.7: Thông số kỹ thuật của nguồn tần số UP30DAC.
- Bảng 1.8: Thông số kỹ thuật của cân tiểu ly điện tử có định lượng 500 gam.
- Bảng 4.1 Sự khác nhau giữa MinMax Scaling và Standard Scaler.
- Bảng 5.1: Mô tả chi tiết tập dữ liệu thô mid+bot.
- Bảng 5.2: Kết quả mô hình học máy khi sử dụng Standard Scaler.
- Bảng 5.3: Kết quả mô hình học máy khi sử dụng SNV.
- Bảng 5.4: Kết quả mô hình học máy khi sử dụng MSC.
- Bảng 5.5: So sánh giữa các giá trị  $R^2$ .
- Bảng 5.6: Kết quả mô hình học máy Decision Tree khi sử dụng Standard Scaler.
- Bảng 5.7: Kết quả mô hình học máy Decision Tree khi sử dụng SNV.
- Bảng 5.8: Kết quả mô hình học máy Decision Tree khi sử dụng MSC.
- Bảng 5.9: Kết quả mô hình Support Vector Regression khi sử dụng Standard Scaler.
- Bảng 5.10: Kết quả mô hình Support Vector Regression khi sử dụng SNV.
- Bảng 5.11: Kết quả mô hình Support Vector Regression khi sử dụng MSC.
- Bảng 5.12 Input mới để kiểm thử.
- Bảng 5.13 Kiểm tra kết quả input mới.

## DANH MỤC CÁC TỪ VIẾT TẮT

NIR	Near Infrared Reflectance
NIRS	Near Infrared Spectroscopy
UT	Ultrasonic Testing
RT	Radiographic Testing
PT	Liquid Penetrant Testing
TCVN	Tiêu chuẩn Việt Nam
HU	Haugh Units
I2C	Inter-IC
SPI	Serial Peripheral Interface
GPIO	General-purpose input/output
VNC	Virtual Network Computing
MSC	Multiplicative Scatter Correction
SNV	Standard Normal Variate
ESS	Residual Sum of Squares
TSS	Total Sum of Squares
RMSE	Root mean squared error
SVM	Support Vector Machine
SVR	Support Vector Regression

## TÓM TẮT KHÓA LUẬN

Trứng gia cầm như trứng gà hay trứng vịt là một loại thực phẩm phổ biến, được sử dụng rộng rãi trên thị trường với hàm lượng dinh dưỡng cao, giá rẻ và dễ chế biến. Vì thế trứng được phân bố hầu hết ở mọi nơi trên thị trường như chợ, siêu thị và các cửa hàng để đáp ứng cho người tiêu dùng với số lượng lớn. Trứng gia cầm có chứa hàm lượng dinh dưỡng cao nhưng rất khó bảo quản vì trên vỏ trứng có nhiều lỗ khí nên để lâu dài khả năng nước bị bốc hơi rất cao, vào mùa hè chỉ sau 5-7 ngày trứng gia cầm đã xuất hiện hiện tượng hư hỏng. Đặc biệt trứng gia cầm được bán ở chợ, cửa hàng thường chưa có biện pháp xử lý và bảo quản hợp lý nên chất lượng trứng có thể bị giảm là điều không tránh khỏi.

Hiện tại, việc xác định chất lượng trứng trên thị trường tiêu dùng hầu như đều sử dụng các phương pháp truyền thống, nhưng mang lại kết quả không chính xác. Trong công nghiệp, việc đánh giá chất lượng trứng ở Việt Nam hay trên thế giới được phân hạng dựa trên: bên ngoài quả trứng (hình dạng, màu sắc, trạng thái, mùi và nấm mốc), ở bên trong quả trứng (trạng thái, mùi, màu sắc và nấm mốc). Những các tiêu chuẩn đây chỉ mang tính chất tương đối một phần nhưng không hoàn toàn chính xác toàn bộ. Bên cạnh đó, sau thời gian bảo quản và phân phối, chất lượng trứng khi đến tay người tiêu dùng sẽ ít nhiều bị thay đổi.

Đề tài này nhóm tập trung tìm hiểu, nghiên cứu và thiết kế hệ thống nhúng có thể xác định chất lượng trứng gà công nghiệp (vỏ màu nâu sẫm) một cách nhanh chóng, chính xác và không gây xâm lấn bằng quang phổ cận hồng ngoại NIR nhằm mục đích giúp người tiêu dùng có thể xác định chất lượng trứng dựa trên số ngày tuổi của trứng và Haugh units, qua đó lựa chọn hoặc loại bỏ trứng kém chất lượng nhằm đảm bảo an toàn khi sử dụng.

## CHƯƠNG 1. GIỚI THIỆU VỀ ĐỀ TÀI

### 1.1. Tổng quan đề tài

Nhiều người thấy trứng là một nguồn dinh dưỡng hợp lý. Protein trong trứng chứa nhiều các axit amin (tryptophan, methionin, cysteine, arginin, photpho, protein) không thể thay thế mà tỷ lệ rất hợp lý, cân đối. Không chỉ có vậy bên trong trứng còn chứa nhiều vitamin như vitamin A, D, B1, hàm lượng khoáng cao như sắt, photpho, đặc biệt bên trong lòng đỏ trứng chứa một hàm lượng lipit cao ở dạng nhũ hóa dễ tiêu hóa.

Tuy nhiên, độ tươi và chất lượng của trứng bị ảnh hưởng nhiều bởi thời gian và điều kiện bảo quản. Sự thay đổi về độ tươi có thể bị người tiêu dùng coi là thiếu chất lượng. Ngoài ra, sự suy thoái có thể đạt đến điểm mà quả trứng không còn thích hợp để sử dụng. Những thay đổi này bao gồm sự mỏng đi của albumen, sự suy yếu của màng vi Telline và sự gia tăng hàm lượng nước trong lòng đỏ. Thời gian bảo quản, nhiệt độ, độ ẩm, chất lượng không khí và cách xử lý là các yếu tố bên ngoài có thể góp phần vào sự suy thoái của trứng.

Đặc biệt là thời gian bảo quản theo Akter và cộng sự (2014) [1] đã chứng minh rằng khối lượng trứng, nồng độ pH, quá trình oxy hóa và Haugh Units [2] trong trứng bị ảnh hưởng bất lợi khi thời gian bảo quản của trứng kéo dài. Ví dụ như hình 1.1 bên dưới có thể thấy từ ngày số 0 đến ngày 28 hàm lượng lòng trắng giảm đáng kể có thể quan sát được, điều này cũng chứng minh sau một khoảng thời gian trứng cũng mất đi một hàm lượng dinh dưỡng đáng kể. Vì lý do này việc phát triển các phương pháp giám sát thời gian bảo quản trứng cũng là một tiêu chí đánh giá chất lượng của trứng.



Hình 1.1: Sự khác biệt chất lượng trứng gà vào ngày bảo quản số 0 và 28.

Vấn đề xác định chất lượng trứng và độ tươi thông qua thời gian bảo quản cũng được nhiều người quan tâm. Trong những năm gần đây, các kỹ thuật không phá hủy để đánh giá độ tươi và thời gian bảo quản ở nhiệt độ phòng đã xuất hiện. Các kỹ thuật này bao gồm mũi điện tử (Yongwei và cộng sự, 2009) [3], siêu âm (Aboonajmi và cộng sự, 2014) [4]. Đặc biệt việc sử dụng quang phổ NIR đã được sử dụng trong một thời gian dài vì đây là một kỹ thuật phân tích chất lượng chính xác, nhanh chóng và không phá hủy chất lượng. Một vài dự án tiêu biểu như: Alessandro và cộng sự 2008 [5], Zhao và cộng sự 2010 [6], Coronel Reyes và cộng sự 2018 [7], Suktanarak và cộng sự 2017 [8], Douglas và cộng sự 2021 [9] cho thấy được tiềm năng của phương pháp quang phổ NIR trong việc đánh giá chất lượng và độ tươi của trứng.

Bảng 1.1: Kết quả một số nghiên cứu đánh giá chất lượng trứng sử dụng NIR

<b>Tên đề tài</b>	Determination of egg storage time at room temperature using a low-cost NIR spectrometer and machine learning techniques	On-line monitoring of egg freshness using a portable NIR spectrometer in tandem with machine learning	Non-destructive freshness assessment of shell eggs using FT-NIR spectroscopy	Identification of egg's freshness using NIR and support vector data description	Non-destructive quality assessment of hens' eggs using hyperspectral images
<b>Tác giả</b>	Julian, Coronel-Reyes, Ivan Ramirez-Morales, Enrique Fernandez-Blanco, Daniel Rivero, Alejandro Pazos	J.P. Cruz-Tirado, Maria Lucimar da Silva Medeiros, Douglas Fernandes Barbin	Alessandro Giunchi, Annachiara Berardinelli, Luigi Ragni, Angelo Fabbri, Florina Aurelia Silaghi	Jiewen Zhao, Hao Lin, Quansheng Chen, Xingyi Huang, Zongbao Sun, Fang Zhou	Suktanarak Sineenart, Teerachaichayut Sontisuk
<b>Năm</b>	2018	2021	2008	2010	2017

Thiết bị	SCIOTM	DLPR NIRscan™ Nano	FT-NIR	Antaris II NIR analyzer	Specim, Spectral Imaging
Dải quang phổ	740 - 1070nm	900–1700nm	833–2500nm	1000-2500nm	900-1700 nm
Đầu ra	Ngày tuổi của trứng	Đơn vị Haugh	Đơn vị Haugh	Đơn vị Haugh	Đơn vị Haugh
Giá thiết bị		999\$	12000\$	9999\$	40520\$
$R^2$	$0,8319 \pm 0,0377$	0.93	0.676	93.3% (validation)	0.85
RMSE	1.97	3.31	9.1		6.29

Từ bảng 1.1 bên trên có thể thấy việc sử dụng quang phổ NIR để xác định chất lượng trứng được sử dụng rất phổ biến và đạt được nhiều kết quả cao. Nhưng đặc điểm chung của các dự án này chi phí thiết bị rất đắt đỏ không phù hợp với thị trường tiêu dùng chỉ phù hợp để nghiên cứu hay sử dụng trong công nghiệp.

## 1.2. Lý do chọn đề tài

Mặc dù trứng là một thực phẩm đem lại giá trị dinh dưỡng cao và có tính phổ biến nhưng để làm sao xác định được trứng già cầm tại cửa hàng, chợ, tại nhà có chất lượng tốt hay không, còn có thể sử dụng hay không là chưa có. Nhưng khi sử dụng những phương pháp truyền thống thì không mang lại tính chính xác cao. Để tránh mua phải trứng bị hư hoặc trứng để lâu ngày, việc lựa chọn thực phẩm sạch là cực kỳ quan trọng giúp giảm thiểu hoặc loại bỏ việc tiêu thụ và tránh những tác động xấu đến sức khỏe khi ăn phải trứng bị hư. Vì vậy đề sử dụng phai trứng bị hư có thể mang đến nhiều ảnh hưởng xấu đến sức khỏe người tiêu dùng như:

- Nhiễm khuẩn Salmonella: gây tiêu chảy, buồn nôn, đau bụng và trong một số trường hợp nặng hơn, có thể dẫn đến viêm ruột hoặc sốt thương hàn.
- Tăng cường tình trạng dị ứng: trứng hư có thể tạo điều kiện cho nấm và vi khuẩn phát triển, mang lại các vấn đề như dị ứng hoặc kích thích một số bệnh lý nếu được tiêu thụ.
- Tác động đến hệ tiêu hóa: trứng bị hư có thể chứa những chất độc hại hoặc vi khuẩn có thể gây tổn thương đến hệ tiêu hóa, gây ra các triệu chứng như đau bụng, buồn nôn, và tiêu chảy.

Bên cạnh đó, qua bảng 1.1 bên trên nhóm nhận thấy được tiềm năng của NIR trong việc đánh giá chất lượng độ tươi của trứng. Vì vậy nhóm muốn nghiên cứu đề xuất một thiết bị nhỏ gọn giá rẻ sử dụng quang phổ giá rẻ NIR áp dụng máy học để có thể xác định nhanh chóng, chính xác, không xâm lấn thời gian bảo quản của trứng gà công nghiệp (vỏ màu nâu), thông qua đó đánh giá chất lượng và độ tươi của trứng phù hợp với người tiêu dùng dựa trên những đề tài đánh giá chất lượng trứng dựa trên quang phổ NIR. Bên cạnh đó nhóm kết hợp thêm phương pháp Haugh Units (HU) [9] để có thể phân loại độ tươi của trứng gà: loại AA ( $HU \geq 72$ ), loại A ( $60 \leq HU < 72$ ), loại B ( $31 \leq HU < 60$ ) theo tiêu chuẩn quốc gia TCVN 1858:2018 về trứng gà.

### 1.3. Mục tiêu đề tài

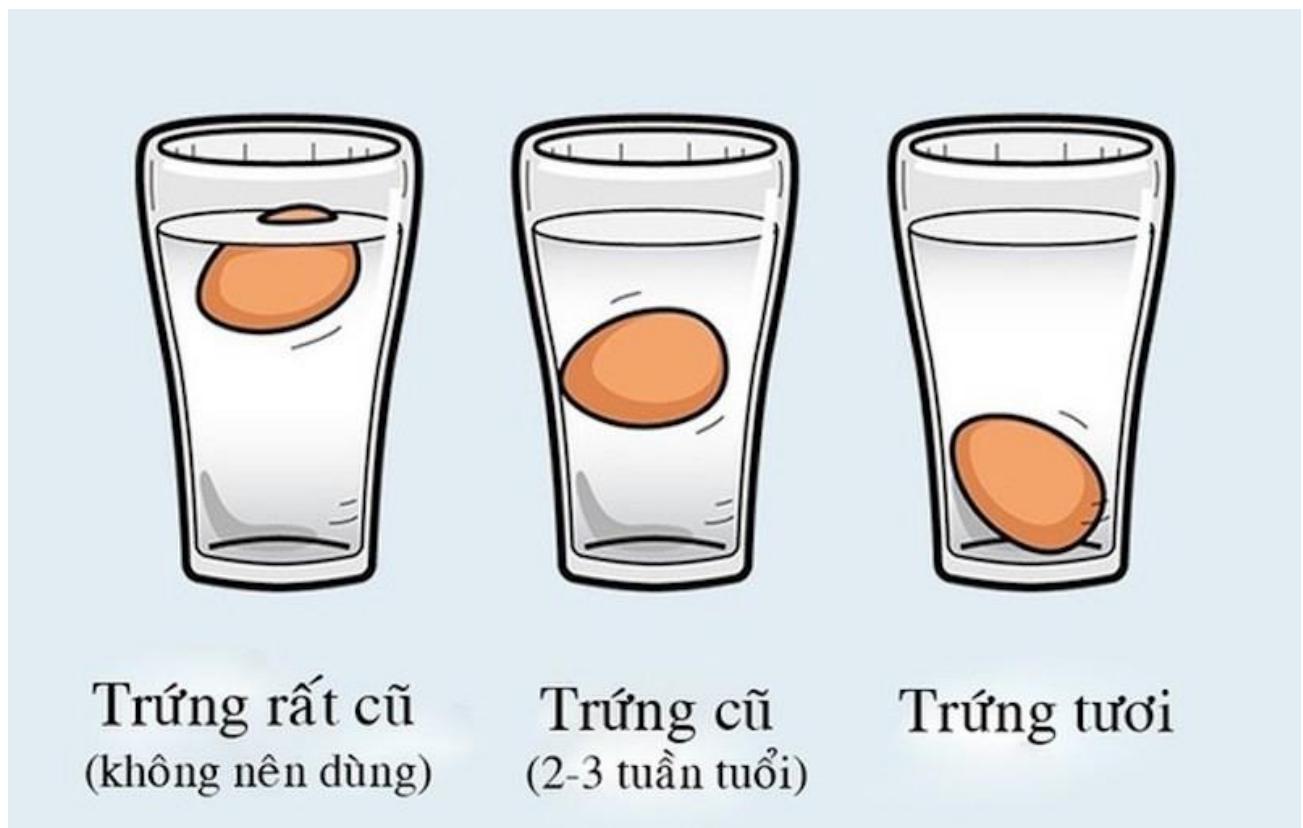
- Tìm hiểu lý thuyết về bước sóng quang phổ cận hồng ngoại NIR và ứng dụng của quang phổ cận hồng ngoại trong việc đánh giá chất lượng trứng không xâm lấn.
- Nghiên cứu và thiết kế thiết bị phần cứng đánh giá chất lượng và độ tươi của trứng sử dụng quang phổ cận hồng ngoại NIR giá rẻ.
- Nghiên cứu và xây dựng mô hình học máy đánh giá nhanh chóng chất lượng và độ tươi của trứng gà công nghiệp màu nâu sử dụng phổ bước sóng cận hồng ngoại dựa trên thời gian bảo quản và Haugh units.

## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

### 2.1. Các phương pháp đo chất lượng trứng gà

#### 2.1.1. Phương pháp đo truyền thống

Soi trứng: phương pháp này rất đơn giản, chỉ cần lấy đèn soi vào quả trứng và quan sát nếu trứng gà còn mới thì buồng khí nhỏ, lòng đỏ tròn không di động và nằm chính giữa, lòng trắng trong suốt có màu cam đỏ và hồng nhạt. Còn trứng cũ để lâu ngày bên trong thường có màu đỏ với nhiều đường vân và buồng khí khá lớn.



Hình 2.1: Phương pháp đánh giá chất lượng trứng bằng cách thả trứng vào nước.

Thả trứng vào nước là một phương pháp truyền thống thường được sử dụng nhiều. Thực hiện phương pháp thả trứng này bằng cách nhẹ nhàng đặt quả trứng vào bát hoặc cốc nước, ví dụ như hình 2.1. Có thể thấy nếu quả trứng chìm xuống bên dưới chứng tỏ quả trứng còn tươi, còn khi quả trứng nghiêng lên trên hoặc thậm chí nổi lên, thì trứng đã cũ. Điều này xuất hiện là do khi thời gian bảo quản trứng kéo dài một phần nước bên trong

trứng được thoát ra và được thay thế bằng không khí, túi khí dần to hơn khiến trứng có thể nổ. Điều này cũng được nhóm chứng minh rõ trong khi thực hiện do Haugh Units kết quả được trình bày ở mục 5.2.3: đánh giá sự thay đổi của trứng qua các ngày thông qua Haugh Units.

### 2.1.2. Phương pháp đo xâm lấn

Haugh unit (HU) hay đơn vị Haugh được Raymond Haugh [10] giới thiệu vào năm 1937 là thước đo quan trọng để đánh giá chất lượng và phân hạng cùng với các thước đo khác như độ dày vỏ, trạng thái, mùi nấm mốc và được sử dụng phổ biến trong công nghiệp để phân hạng trứng gà.



Hình 2.2: Đo chiều cao của lòng trắng trứng.

Haugh unit được tính bằng công thức:

$$HU = 100 \log (H + 7.57 - 1.7W^{0.37}) \quad (1)$$

- $H(\text{mm})$  là chiều cao của lòng trắng trứng bao quanh lòng đỏ đo bằng cách đập vỡ quả trứng lên 1 bề mặt phẳng sau đó đo chiều cao từ mặt phẳng đến điểm cao nhất của lòng trắng.
- $W(\text{g})$  là cân nặng của quả trứng.

Chỉ số Haugh unit càng cao thì chất lượng trứng càng tốt (trứng càng tươi, chất lượng trứng càng cao thì lòng trắng trứng càng dày): loại AA ( $HU \geq 72$ ), loại A ( $60 \leq HU < 72$ ), loại B ( $31 \leq HU < 60$ ) theo tiêu chuẩn quốc gia TCVN 1858:2018 về trứng gà.

### **2.1.3. Phương pháp đo không xâm lấn**

Qua mục 2.1.2 và 2.1.3 có thể thấy được phương pháp truyền thống có thể đánh giá được chất lượng trứng qua cảm quan một cách nhanh chóng nhưng kết quả thu được chỉ mang tính tương đối không chính xác. Còn phương pháp đo xâm lấn sử dụng Haugh Units mang lại kết quả chính xác nhanh chóng nhưng lại phá hủy vật mẫu gây xâm lấn, phương pháp Haugh Units có thể phù hợp cho việc đánh giá tổng quan lô hàng trứng trong công nghiệp, đặc biệt không phù hợp khi sử dụng ở siêu thị, chợ hay các cửa hàng.

Với sự phát triển của công nghệ cảm biến trong thời đại 4.0, một chiến lược hấp dẫn để không phá hủy xác định độ tươi và chất lượng của trứng được nhiều tác giả quan tâm. Đặc biệt là việc sử dụng quang phổ NIR đã và đang được sử dụng trong một khoảng thời gian dài vì đây là một kỹ thuật phân tích chất lượng chính xác, nhanh chóng và không phá hủy chất lượng của vật mẫu. Những nghiên cứu tiêu biểu liên quan đến việc sử dụng NIR được nêu ở bảng 1.1: Kết quả một số nghiên cứu đánh giá chất lượng trứng sử dụng NIR đã cho thấy được tiềm năng của phương pháp quang phổ NIR trong việc đánh giá chất lượng và độ tươi của trứng. Mà gần đây, một số thiết bị NIR giá rẻ đã xuất hiện trên thị trường, làm cho các ứng dụng NIR có giá cả phải chăng dễ tiếp cận và phát triển rộng rãi hơn.

## **2.2. Ứng dụng quang phổ trong việc đánh giá chất lượng và độ tươi của trứng**

### **2.2.1. Quang phổ**

Quang phổ là sự phân tích và phân loại ánh sáng thành các bước sóng (hoặc mức năng lượng) khác nhau. Ánh sáng được phân tích thành các bước sóng riêng lẻ hoặc nhóm các bước sóng, tùy thuộc vào tính chất của nguồn ánh sáng và phương pháp phân tích.

Ánh sáng là một dạng sóng điện từ, nó có thể lan truyền qua không gian hoặc các chất liệu khác nhau và có thể chứa nhiều bước sóng có các bước sóng khác nhau. Khi ánh sáng tiếp xúc với các vật chất, nó tương tác và tạo ra các hiện tượng như hấp thụ, phản xạ, phát xạ, chuyển đổi năng lượng, hay phản chiếu tùy thuộc vào tính chất của chất liệu đó.

Ứng dụng của quang phổ được sử dụng rộng rãi trong nhiều lĩnh vực khoa học và công nghệ, bao gồm:

- Quang hóa học: Sử dụng để xác định thành phần của các chất hóa học dựa trên phản ứng phổ hấp thụ và phát xạ của chúng với ánh sáng.
- Quang phổ cận hồng ngoại (NIR): Được sử dụng để phân tích thành phần của các mẫu ví dụ như đánh giá chất lượng trứng gà dựa trên thời gian bảo quản, đánh giá độ tươi của cá hồi Đại Tây Dương, táo, thịt lợn, đo đường huyết không xâm lấn và nhiều ứng dụng khác.
- Quang phổ hấp thụ nguyên tử: Dùng để xác định nồng độ của các nguyên tố hóa học trong mẫu.
- Quang phổ Raman: Được sử dụng để xác định thành phần phân tử và cấu trúc phân tử của các vật chất.
- Quang phổ điện tử: Sử dụng để nghiên cứu và xác định các vật chất trong vũ trụ.

Các phân tích quang phổ cung cấp thông tin về cấu trúc, tính chất và thành phần của các chất và chúng là một công cụ quan trọng trong nghiên cứu và ứng dụng trong nhiều lĩnh vực khoa học và công nghệ.

### **2.2.2. Quang phổ cận hồng ngoại NIR**

Theo William và cộng sự (1998) [11], khi có một nguồn ánh sáng chiếu qua các mẫu vật, vùng ánh sáng cận hồng ngoại (750 nm - 2500 nm) được hấp thụ bởi các liên kết phân tử C-H, N-H và O-H có trong các chất hữu cơ của vật mẫu sinh học. Đo cường độ ánh sáng phản xạ lại hoặc hấp thụ từ các mẫu vật sinh học sẽ đem lại được các thông tin về thành phần hóa học của mẫu vật đó.

Quang phổ cận hồng ngoại NIR là một phương pháp sử dụng vùng hồng ngoại gần của quang phổ điện tử khoảng 750 nm - 2500 nm. Bằng cách đo ánh sáng phân tán ra khỏi và xuyêng qua mẫu, phổ NIR có thể được sử dụng để nhanh chóng xác định các thành phần hóa học của thực phẩm (như protein và lipid), nồng độ các chất nhanh hơn nhiều so với xét

nghiệm sinh hóa thông thường. Từ đó có thể sử dụng kết quả để đánh giá chất lượng thực phẩm.

### **2.2.3. Các ưu điểm của phổ cận hồng ngoại NIR**

Phương pháp phổ cận hồng ngoại (NIR) không cần sử dụng thêm các dung môi hoặc thuốc thử độc hại, điều này giúp nó trở thành một kỹ thuật an toàn với môi trường và tiết kiệm tài nguyên. Đặc biệt thiết bị phổ NIR có thể được sử dụng bởi những người không có chuyên môn và kiến thức sâu về hóa học và phân tích, nhờ đó thời gian đào tạo nhân viên là rất ngắn.

Thêm vào đó, phương pháp NIR không làm tổn hại cấu trúc tế bào của mẫu vật, không cần tiến hành phá hủy mẫu vật bằng nhiệt độ hoặc hóa chất trong quá trình phân tích, điều này giúp mẫu có thể được tái sử dụng sau khi phân tích. Sự không phá hủy này giúp bảo tồn mẫu và giảm chi phí thử nghiệm.

Ngoài ra, phương pháp phổ NIR thực hiện phân tích nhanh chóng và đem lại kết quả chính xác chỉ trong ít phút. Điều này tăng tính hiệu quả và giúp quy trình kiểm tra và kiểm soát chất lượng trở nên dễ dàng và hiệu quả hơn.

### **2.2.4. Phân loại quang phổ cận hồng ngoại**

Quang phổ cận hồng ngoại (NIR) được phân loại dựa theo bước sóng, gồm hai loại chính: quang phổ bước sóng dài (Long-wave NIR Spectroscopy) với khoảng bước sóng từ 1300nm đến 2500nm và quang phổ bước sóng ngắn (Short-wave NIR Spectroscopy) với khoảng bước sóng từ 780nm đến 1300nm. Bảng 2.1 sẽ cung cấp các đặc điểm cũng như ưu điểm của từng loại.

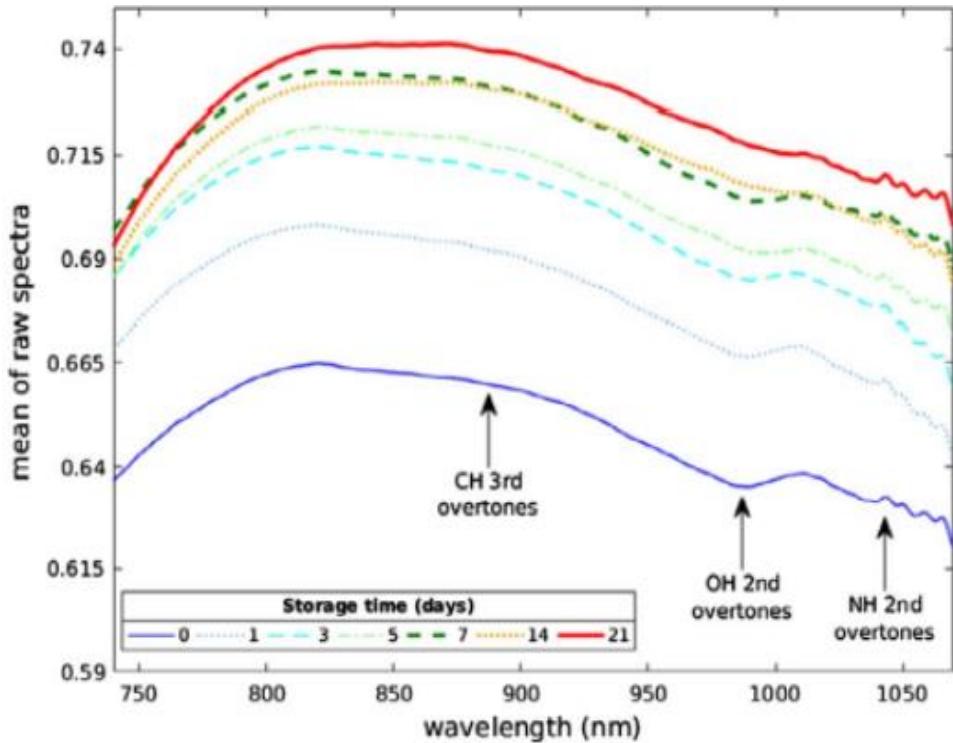
Việc sử dụng bước sóng dài có hạn chế là khó có khả năng phát hiện các phân tử đường bên dưới lớp vỏ trứng do có độ thâm nhập nông, đồng thời chi phí của các cảm biến ở vùng bước sóng này tương đối cao. Điều này làm cho vùng bước sóng ngắn trở nên lựa chọn phù hợp hơn cho đề tài. Với khoảng bước sóng ngắn hơn, phương pháp này sẽ có khả năng thâm thấu sâu hơn vào bên trong quả trứng và từ đó tiếp cận và phân tích một cách hiệu quả hơn.

Bảng 2.1: So sánh quang phổ cận hồng ngoại bước sóng dài với bước sóng ngắn.

Quang phổ cận hồng ngoại bước sóng dài	Quang phổ cận hồng ngoại bước sóng ngắn
Giúp tạo ra sự cộng hưởng giữa O-H và C-H cho âm bội đầu tiên. Thu được sự cộng hưởng giữa liên kết O-H và C-H cho âm bội thứ nhất cao hơn so với bước sóng ngắn.	Độ hấp thụ và độ phản xạ ánh sáng sắc nét hơn và mạnh hơn ở âm bội thứ nhất so với âm bội thứ hai và thứ ba (âm bội cao hơn).
Hạn chế về khả năng thâm nhập nông.	Sự rung động của phân tử C-H đã được quan sát ở 920 nm.
Độ hấp thụ và phản xạ ánh sáng tương đối kém.	Sự cộng hưởng của liên kết C-H <sub>2</sub> (âm bội cao thứ hai) cao hơn và rõ ràng hơn so với bước sóng dài.

### 2.2.5. Cơ sở lý thuyết về dải bước sóng của trứng gà

Kỹ thuật phân tích dựa trên phổ cận hồng ngoại (NIR) mang lại hiệu quả tốt hơn những kỹ thuật phân tích cấu trúc khác như: nhiễu xạ tia X và cộng hưởng từ điện tử. Quang phổ cận hồng ngoại (NIR) hoạt động theo nguyên lý đơn giản: Những hợp chất hóa học trong vật mẫu có khả năng lựa chọn hấp thụ bức xạ hồng ngoại. Khi các hóa chất trong vật mẫu hấp thụ những bức xạ hồng ngoại, các phân tử trong các hợp chất hóa học dao động với nhiều vận tốc dao động và xuất hiện dải phổ hấp thụ gọi là phổ hấp thụ bức xạ hồng ngoại. Các dao động phân tử tồn tại trong vùng NIR tổng thể được gọi là sóng hài hoặc âm bội. Rung động phân tử phụ thuộc vào rung động liên kết như vẩy, uốn cong, lắc lư, kéo căng và xoắn.



Hình 2.3 Ví dụ trung bình dữ liệu quang phổ thô của trứng theo thời gian lưu trữ.

Theo Lammertyn và cộng sự (2000) [12] độ sâu xuyên qua của ánh sáng phụ thuộc trực tiếp vào bước sóng. Ví dụ, một ánh sáng trong phạm vi từ 700 đến 900 nm có thể xuyên qua đến 4 mm, trong khi ánh sáng từ 900 đến 1900 nm đặt khả năng xuyên qua tối đa từ 2 đến 3 mm. Phạm vi quang phổ của dụng cụ đo, được sử dụng trong công việc này, cho phép độ xuyên thấu từ 3 mm đến 4 mm, đủ để đi qua vỏ trứng và tiếp cận bên trong. Do đó, những thay đổi trong quang phổ liên quan trực tiếp đến vỏ, lớp biểu bì và albumen của trứng mặc dù không có thông tin của lòng đỏ vì khả năng thâm nhập không đủ để cung cấp thông tin về điều này. Các dải đáp ứng trong vùng quang phổ NIR (700–1100 nm), chủ yếu là kết quả của các âm bội thứ 3(CH), âm bội thứ 2(OH - NH). Nhìn hình 2.3 quan sát được sự thay đổi của các dải này qua thời gian bảo quản trứng. Vì vậy vùng quang phổ (700nm–1100nm) phù hợp để thực hiện đề tài.

## 2.3. Machine Learning

### 2.3.1. Giới thiệu về Machine Learning

Machine Learning là một trong những lĩnh vực nghiên cứu chủ yếu của Trí Tuệ Nhân Tạo giúp tạo ra những hệ thống máy tính có khả năng học tập và giải quyết vấn đề giống con người mà không cần phải lập trình một cách cụ thể hay rõ ràng. Các mô hình máy học yêu cầu dữ liệu tương đối lớn để huấn luyện và cải thiện độ chính xác của mô hình.

Hiện nay có 2 loại thuật toán máy học chính là học có giám sát (Supervised learning) và học không có giám sát (Unsupervised learning).

### 2.3.2. Học có giám sát

Hay được gọi là Supervised learning: nhóm các thuật toán sử dụng những cặp dữ liệu cũ đã được dán nhãn để tiến hành tìm ra một hàm thể hiện mối quan hệ giữa đầu vào và đầu ra từ đó có thể dự đoán đầu ra ứng với một đầu vào dữ liệu mới. Có 2 nhóm bài toán cơ bản trong Supervised learning là Phân loại (Classification) và Hồi quy (Regression):

Bài toán phân loại: một bài toán được gọi là Classification khi các biến đầu vào của bài toán được phân chia thành từng lớp dựa trên các tham số khác nhau và nhiệm vụ của bài toán là tiến hành tìm ra hàm số để dự đoán các đầu vào ứng với các đầu ra rời rạc.

Regression (Bài toán hồi quy): ngược lại so với Classification, khi đầu vào của bài toán không được phân chia theo từng lớp và dữ liệu có tính liên tục thì đây là bài toán Regression. Nhiệm vụ của bài toán này là tiến hành tìm ra hàm số để dự đoán các đầu ra ứng với các đầu vào liên tục.

### 2.3.3. Học không giám sát

Unsupervised learning (Học không giám sát) là thuật toán mà các dữ liệu đầu vào không được dán nhãn và sẽ chỉ sử dụng cấu trúc hay các đặc tính về thông tin từ dữ liệu để đưa ra kết quả mong muốn. Học không giám sát được chia thành 2 loại: clustering (phân nhóm) và association (kết hợp).

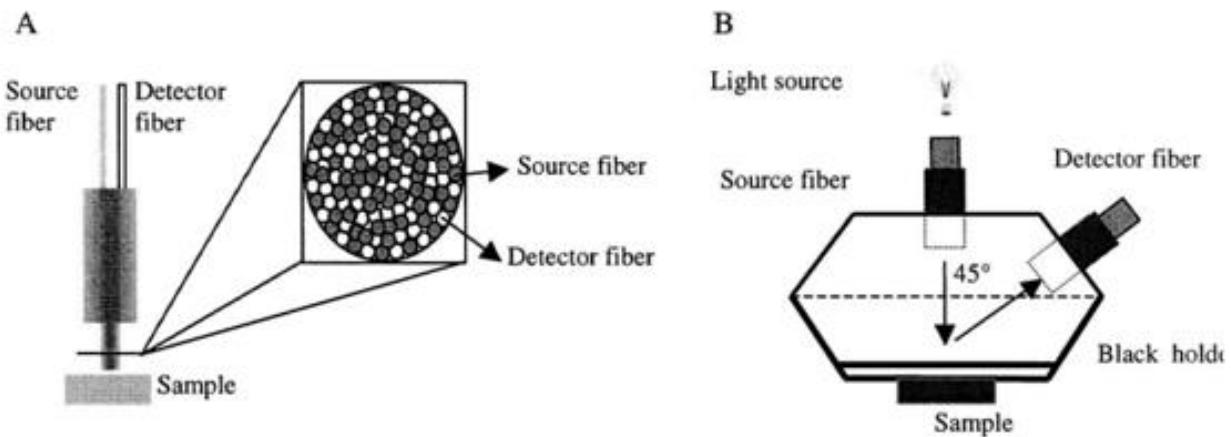
Bài toán phân nhóm (Clustering): giúp phân chia bộ dữ liệu thành các nhóm dữ liệu khác nhau và mỗi điểm dữ liệu trong từng nhóm có điểm giống tương tự với các điểm dữ liệu còn lại và khác so với các điểm dữ liệu của nhóm khác. Clustering có điểm chung so với classification tuy nhiên đối với bài toán phân loại có đầu ra cụ thể còn bài toán phân nhóm thì không có đầu ra.

Association (Bài toán kết hợp): bài toán mà người dùng muốn có đầu ra một cách mới hoàn toàn và tìm ra một quy luật mới dựa trên bộ dữ liệu có sẵn.

#### **2.4. Phương pháp đo**

Hiện nay có nhiều cách đo ánh sáng tán xạ từ vật mẫu được sử dụng phổ biến trong đo đặc kiểm tra chất lượng nông sản, phổ biến nhất gồm 3 cấu hình: Cấu hình quang học đo xuyên thấu, Cấu hình quang học phân nhánh, Cấu hình quang học  $0^0/45^0$ .

Nguồn sáng gồm có đèn halogen 12V/20W được sử dụng để quan sát được vùng ánh sáng nhìn thấy và vùng hồng ngoại gần. Ở mục 2.2.5: Cơ sở dải bước sóng của trứng gà có nhắc đến một ánh sáng có bước sóng ở 700 nm đến 900 nm có thể xuyên qua 4 mm, trong khi ánh sáng có bước sóng ở 900 nm đến 1900 nm chỉ xuyên qua từ 2 đến 3 mm. Mặc dù sử dụng cấu hình quang học xuyên thấu về lý thuyết sẽ cho phép chúng ta thu được những ánh sáng tán xạ chưa được những dữ liệu về đa số thành phần của quả trứng gồm: vỏ, lòng trắng và lòng đỏ, nhưng độ xuyên thấu của ánh sáng của quang phổ NIR không cho phép sử dụng cấu hình quang học này. Quang phổ NIR có bước sóng từ 750 nm đến 2500 nm vì vậy độ xuyên thấu của quang phổ này cao nhất đạt được khoảng 4 mm, như vậy sẽ không thể xuyên qua được trứng.



Hình 2.5: Cấu hình khoa học phân nhánh và cấu hình quang học  $0^\circ/45^\circ$ .

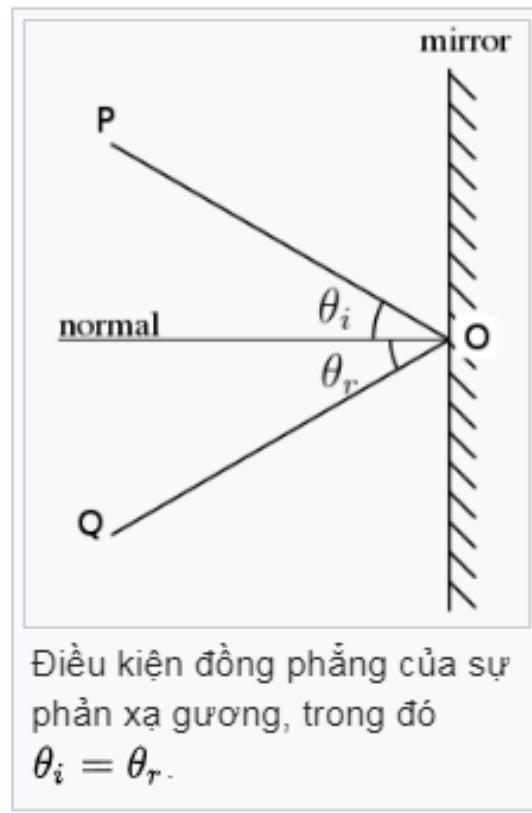
Cấu hình quang học phân nhánh sẽ dẫn ánh sáng đến vật mẫu bằng những sợi nguồn sáng và thu những nguồn sáng phản xạ tán xạ bằng các sợi cảm biến. Những sợi nguồn và sợi dò được đặt ngẫu nhiên trong đầu của dò.

Cấu hình bao gồm một hộp đèn chứa nguồn sáng và đầu dò (cảm biến) được đặt ở một góc  $45^\circ$  để tránh phản xạ gương (là sự phản xạ giống như gương của các sóng từ một bề mặt).

Bảng 2.2: Sự khác nhau giữa cấu hình quang học phân nhánh và cấu hình quang học  $0^\circ/45^\circ$ .

Cấu hình quang học phân nhánh	cấu hình quang học $0^\circ/45^\circ$ .
Hầu như được thiết kế sẵn	Phải thiết lập vị trí, môi trường đo
Bề mặt được chiếu sáng bé	Bề mặt chiếu sáng lớn hơn
Khoảng cách thu ánh sáng gần	Linh hoạt điều chỉnh từng thành phần của cấu hình
Cường độ ánh sáng tập trung vào vật mẫu cao, đầu cảm biến thu được ánh sáng có cường độ cao	Cường độ ánh sáng chiếu vào vật mẫu thấp hơn, cường độ ánh sáng thu được thấp hơn

Mặc dù có kết quả không hiệu quả bằng cách hình quang học phân nhánh (có thể chấp nhận chênh lệch 1% - 8%) nhưng nhằm hướng đến sản phẩm có giá thành phải chăng, dễ dàng tự xây dựng mô hình nhóm đã sử dụng cách hình quang học  $0^0/45^0$ . Thêm: Phản xạ chiếu hay phản xạ đều là gương giống sự phản chiếu của sóng, chẳng hạn như ánh sáng, từ một bề mặt. Định luật phản xạ phát biểu rằng một tia phản xạ ánh sáng phát ra từ bề mặt phản xạ cùng một góc so với bề mặt pháp tuyến là tia tới, nhưng nằm ở phía đối diện của bề mặt pháp tuyến trong mặt phẳng được tạo bởi tia tới và tia phản xạ.



Hình 2.6: Phản xạ gương.

## CHƯƠNG 3. THIẾT KẾ HỆ THỐNG NHÚNG

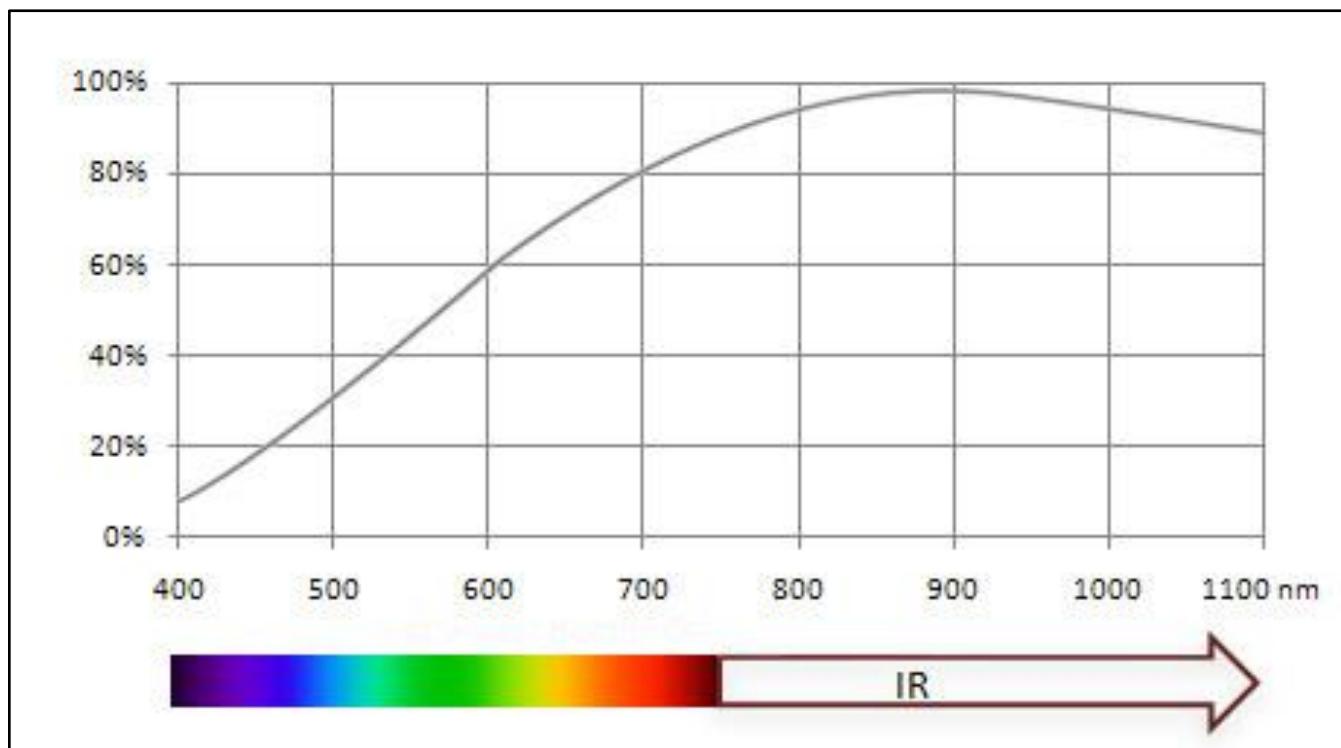
### 3.1. Tổng quan hệ thống nhúng

Hệ thống nhúng được nhóm xây dựng như sau: bước đầu hệ thống nhúng được nghiên cứu và xây dựng để thu thập dữ liệu thô. Các dữ liệu thô thu thập được từ hệ thống sẽ được lưu trữ cho bước huấn luyện mô hình. Sau khi mô hình hoàn tất quá trình huấn luyện sẽ được đưa lên Raspberry để dự đoán các kết quả về số ngày tuổi của trứng, Haugh Units và được in ra màn hình Oled 1 cách trực quan.

### 3.2. Thành phần hệ thống

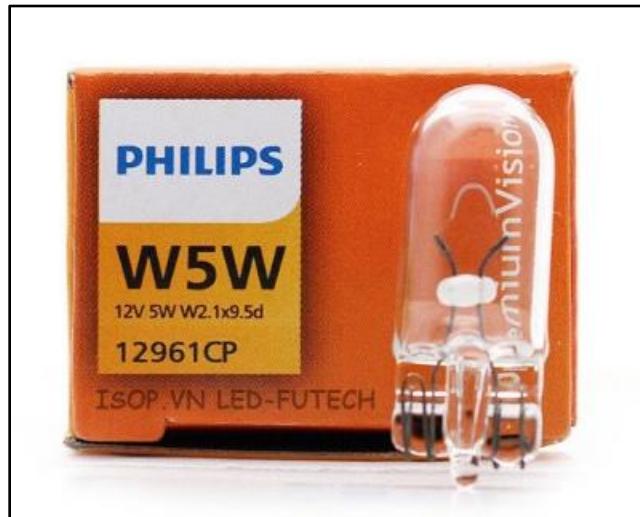
#### 3.2.1. Bóng đèn Halogen Philips W5W T10

Nhóm quyết định sử dụng nguồn phát từ bóng đèn Halogen vì dải bước sóng liên tục mà nó phát ra có thể tận dụng toàn bộ vùng bước sóng mà cảm biến hỗ trợ. Hình 3.1 minh họa rõ ràng dải bước sóng mà bóng đèn Halogen có thể phát ra.



Hình 3.1: Dải bước sóng của bóng đèn Halogen Philips W5W T10.

Hình 3.2 Mô tả bóng đèn mà nhóm sử dụng.



Hình 3.2: Bóng đèn Halogen Philips W5W T10.

Thông số kỹ thuật của bóng đèn Halogen Philips W5W T10 được nêu dưới bảng 3.1 bên dưới.

Bảng 3.1: Thông số kỹ thuật của bóng đèn Halogen Philips W5W T10.

Thông số kỹ thuật
<ul style="list-style-type: none"><li>Điện áp cấp 12V</li><li>Công suất 5W</li><li>Tuổi thọ ~400h</li></ul>

### 3.2.2. Cảm Biến AS7263

Cảm biến quang phổ cận hồng ngoại (NIR) SparkFun AS7263 đo lường và mô tả cách các vật liệu khác nhau hấp thụ và phản xạ các bước sóng ánh sáng khác nhau. Cảm biến quang phổ AS7263 phát hiện các bước sóng trong phạm vi khả biến ở 610nm, 680nm, 730nm, 760nm, 810nm và 860nm của ánh sáng, mỗi bước sóng có khả năng phát hiện tối đa một nửa chiều rộng đầy đủ là 20nm.



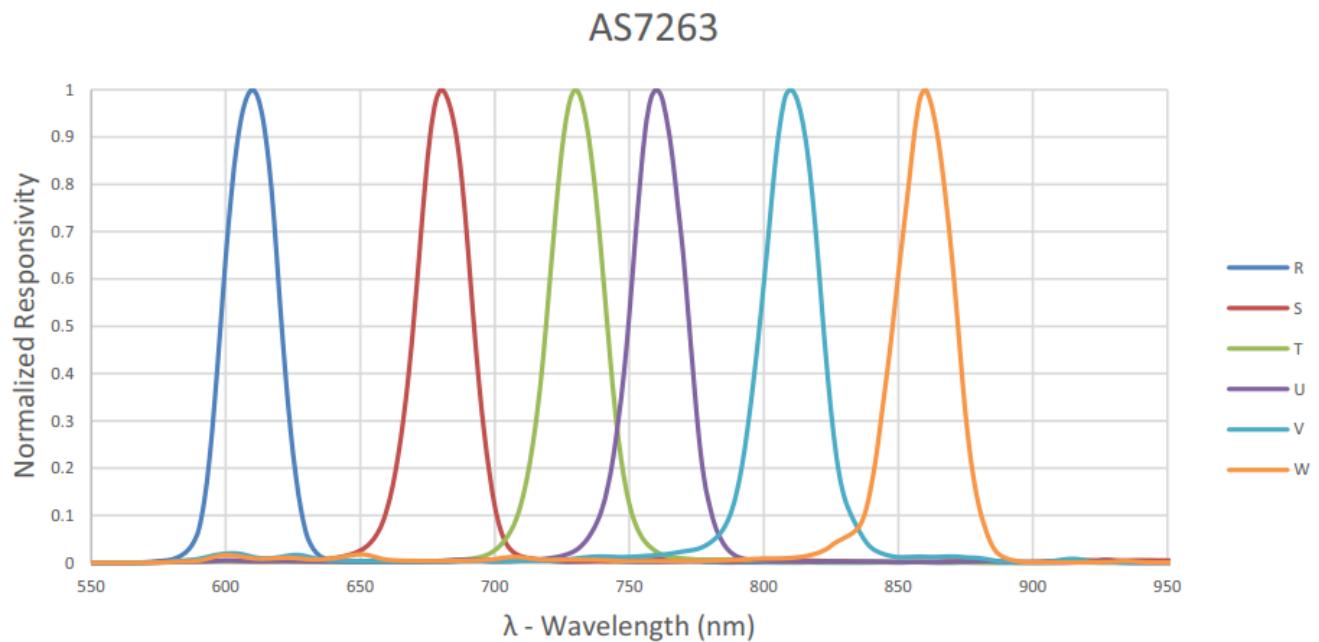
Hình 3.3: SparkFun AS7263.

Thông số kỹ thuật của AS7263 được nêu dưới bảng 3.2 bên dưới.

Bảng 3.2: Thông số kỹ thuật cảm biến AS7263.

Thông số kỹ thuật
<ul style="list-style-type: none"><li>6 kênh hồng ngoại gần: 610nm, 680nm, 730nm, 760nm, 810nm và 860nm, mỗi cái có 20nm FWHM.</li><li>Tích hợp giao tiếp I2C và UART.</li><li>Độ chính xác: +/- 12%.</li><li>Bộ lọc NIR được thực hiện bởi bộ lọc nhiễu silicon.</li><li>Hoạt động ở điện áp thấp: 2.7v đến 3.6v với I2C.</li><li>Nhiệt độ -40°C đến 85°C</li></ul>

Hình 3.4 bên dưới mô tả khả năng đáp ứng quang phổ với 6 bước sóng tương ứng với 6 kênh của cảm biến AS7263.



Hình 3.4: Dải 6 bước sóng đỉnh thu được của cảm biến AS7263.

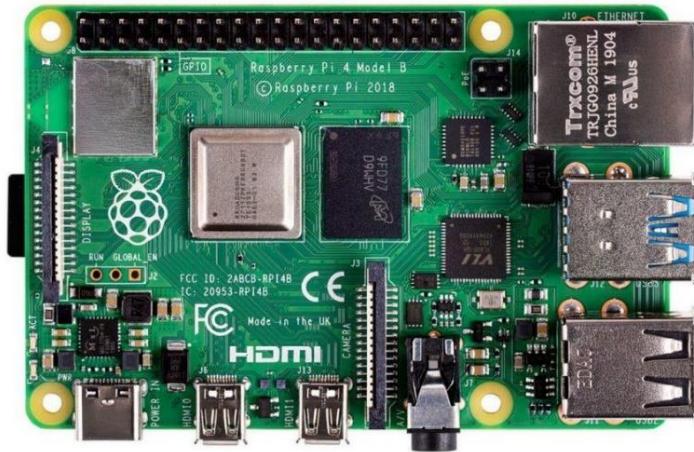
Bảng 3.3 bên dưới cho thấy các giá trị quang học của AS7263. Phần Test Condition có đề cập tới việc cảm biến này hoạt động tốt đối với môi trường của đèn sợi đốt, trong đề tài này nhóm chọn bóng đèn halogen phù hợp với tính chất này của cảm biến. Bên cạnh đó những phân tích về bước sóng áp dụng trong đề tài là từ 700nm đến 1100 nm nên 4 bước sóng ở kênh T, U, V, và W sẽ được chọn tương ứng 730nm, 760nm, 810nm và 860nm để thực hiện đề tài.

Bảng 3.3: Đặc tính quang học của AS7263.

Symbol	Parameter	Test Conditions	Channel (nm)	Min	Typ	Max	Unit
R	Channel R	Incandescent <a href="#">(2)</a> , <a href="#">(4)</a>	610		35 <a href="#">(3)</a> , <a href="#">(4)</a>		counts/ ( $\mu$ W/cm $^2$ )
S	Channel S	Incandescent <a href="#">(2)</a> , <a href="#">(4)</a>	680		35 <a href="#">(3)</a> , <a href="#">(4)</a>		counts/ ( $\mu$ W/cm $^2$ )
T	Channel T	Incandescent <a href="#">(2)</a> , <a href="#">(4)</a>	730		35 <a href="#">(3)</a> , <a href="#">(4)</a>		counts/ ( $\mu$ W/cm $^2$ )
U	Channel U	Incandescent <a href="#">(2)</a> , <a href="#">(4)</a>	760		35 <a href="#">(3)</a> , <a href="#">(4)</a>		counts/ ( $\mu$ W/cm $^2$ )
V	Channel V	Incandescent <a href="#">(2)</a> , <a href="#">(4)</a>	810		35 <a href="#">(3)</a> , <a href="#">(4)</a>		counts/ ( $\mu$ W/cm $^2$ )
W	Channel W	Incandescent <a href="#">(2)</a> , <a href="#">(4)</a>	860		35 <a href="#">(3)</a> , <a href="#">(4)</a>		counts/ ( $\mu$ W/cm $^2$ )
FWHM	Full Width Half Max		20		20		nm
Wacc	Wavelength Accuracy				$\pm 5$		nm
dark	Dark Channel Counts	GAIN=64, $T_{AMB}=25^{\circ}C$				5	counts
f	Angle of Incidence	On the sensors			$\pm 20.0$		deg

### 3.2.3. Vị điều khiển Raspberry Pi 4B

Raspberry Pi là một máy tính nhúng có kích thước tương đối nhỏ gọn, dễ sử dụng và được phát triển bởi Raspberry Pi Foundation. Trong đồ án này Raspberry Pi 4B sẽ giúp nhóm liên kết cảm biến và bộ phát để thu thập các thông số cho dataset, đồng thời thực hiện các mô hình máy học phức tạp trên chính board Raspberry.



Hình 3.5: Raspberry Pi 4B.

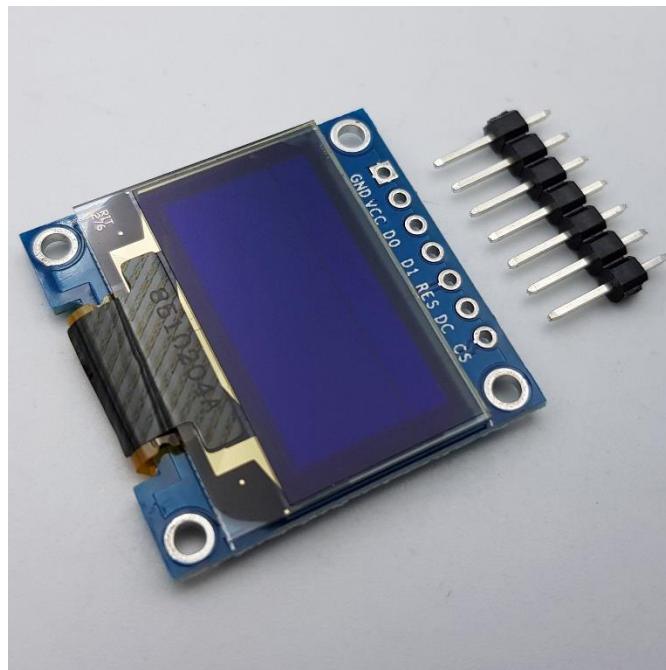
Một số thông số kỹ thuật chính của Raspberry Pi 4B được nêu dưới bảng 3.4 bên dưới.

Bảng 3.4: Thông số kỹ thuật của Raspberry Pi 4B.

Thông số kỹ thuật
<ul style="list-style-type: none"><li>Vi xử lý; Quad core 64-bit ARM-Cortex A72 (1.5GHz).</li><li>Ram: 1, 2 and 4 (GB), RAM LPDDR4.</li><li>Tích hợp 28 GPIO: UART, I2C, SPI, DPI, PCM, PWM, CPCLK.</li><li>Nguồn điện sử dụng: 5V/2.5A.</li><li>Màn hình HDMI lên đến 4Kp60.</li></ul>

### 3.2.4. Màn hình Oled

Để hiển thị kết quả sau khi dự đoán số ngày trung đã trải qua, nhóm chọn màn hình OLED có kích thước 0.96 Inch. Màn hình hiển thị rõ vào ban ngày, đồng thời tiết kiệm điện năng đi kèm với giá cả phải chăng.



Hình 3.6: Màn hình OLED.

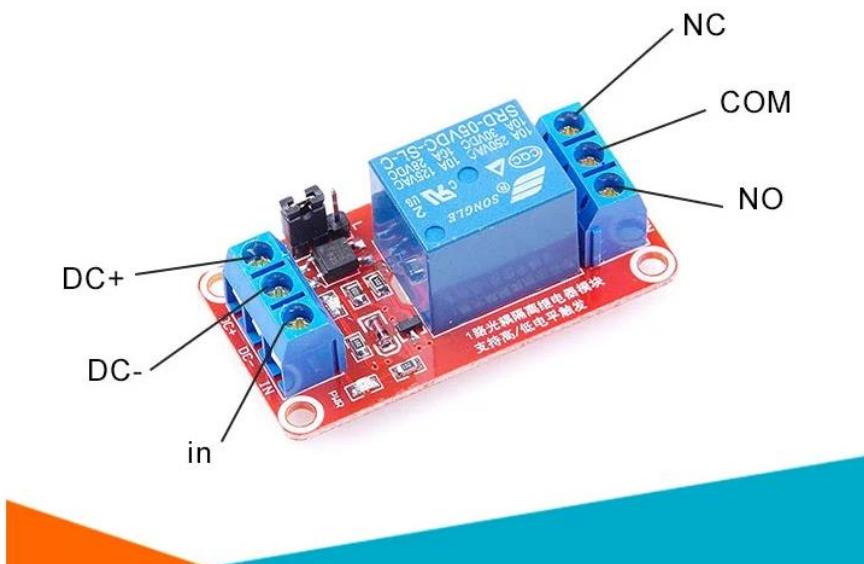
Thông số kỹ thuật của màn hình Oled được nêu dưới bảng 3.5 bên dưới.

Bảng 3.5: Thông số kỹ thuật của màn hình Oled 0.96 Inch.

Thông số kỹ thuật
<ul style="list-style-type: none"><li>Nguồn đầu vào cung cấp: 2.2V - 5.5V</li><li>Công suất màn hình Oled: 0.04w</li><li>Số điểm ảnh: 128x64</li><li>Độ lớn khung hình của màn hình Oled: 0.96 inch</li><li>Màu hiển thị: Trắng / Xanh Dương</li><li>Giao thức: SPI</li></ul>

### 3.2.5. Module relay 5v

Module relay 5VDC được dùng trong nhiều thiết kế để điều khiển đóng mở, dùng điện áp nhỏ hơn để kích mở điện áp lớn. Mục đích nhóm sử dụng module relay 5VDC là để đóng mở nguồn 12v cấp nguồn cho bóng đèn Halogen được điều khiển thông qua nút nhấn được kết nối với Raspberry bằng giao tiếp GPIO vì Raspberry không có nguồn đầu ra là 12v.



Hình 3.7: Module relay 5v.

Thông số kỹ thuật của module relay 5v được nêu dưới bảng 3.6 bên dưới.

Bảng 3.6: Thông số kỹ thuật của module relay 5v.

Thông số kỹ thuật
<ul style="list-style-type: none"><li>• Nguồn điện cung cấp: 5 VDC</li><li>• Có 2 mức thiết lập : Thấp (L) hoặc Cao (H)</li><li>• COM: Tiếp điểm relay 220V 10A</li><li>• NO: chân thường mở</li><li>• NC: chân thường đóng</li></ul>

### 3.2.6. Nguồn tổ ong UP30DAC

Vì phần cứng sử dụng 2 thành phần bóng đèn Halogen Philips có nguồn đầu vào 12V và vi điều khiển Raspberry Pi 4B có nguồn đầu vào là 5V 3A nên cần một nguồn đầu vào ổn định có thể cấp nguồn đồng thời cho cả 2 thành phần. Nguồn tổ ong UP30ADC là lựa chọn phù hợp cho phần cứng của đè tài với 2 đầu ra 12V 1.2A, 5V 3A.



Hình 3.8: Nguồn tổ ong UP30DAC.

Thông số kỹ thuật của nguồn tổ ong UP30DAC được nêu dưới bảng 3.7 bên dưới.

Bảng 3.7: Thông số kỹ thuật của nguồn tổ ong UP30DAC.

Thông số kỹ thuật
<ul style="list-style-type: none"><li>Nguồn đầu vào : 85V-264V ~1.5A</li><li>Công suất thực: 30W</li><li>Đầu ra :5V 3A, 12V 1.2A</li></ul>

### 3.3. Thiết bị và phần mềm hỗ trợ

#### 3.3.1. Thiết bị hỗ trợ cân tiểu ly điện tử 500g và thước đo 20cm

Để hỗ trợ cho việc tính toán Haugh Units cần sử dụng cân tiểu ly với độ sai lệch thấp để đo khối lượng. Bên cạnh đó cần thêm thước đo có độ phân giải 1 milimet nhằm mục tiêu đo độ cao của lòng trắng trứng.



Hình 3.9: Cân tiêu ly điện tử và thước đo.

Một số thông số kỹ thuật chính của cân tiêu ly điện tử có định lượng 500g được nêu dưới bảng 3.8 bên dưới.

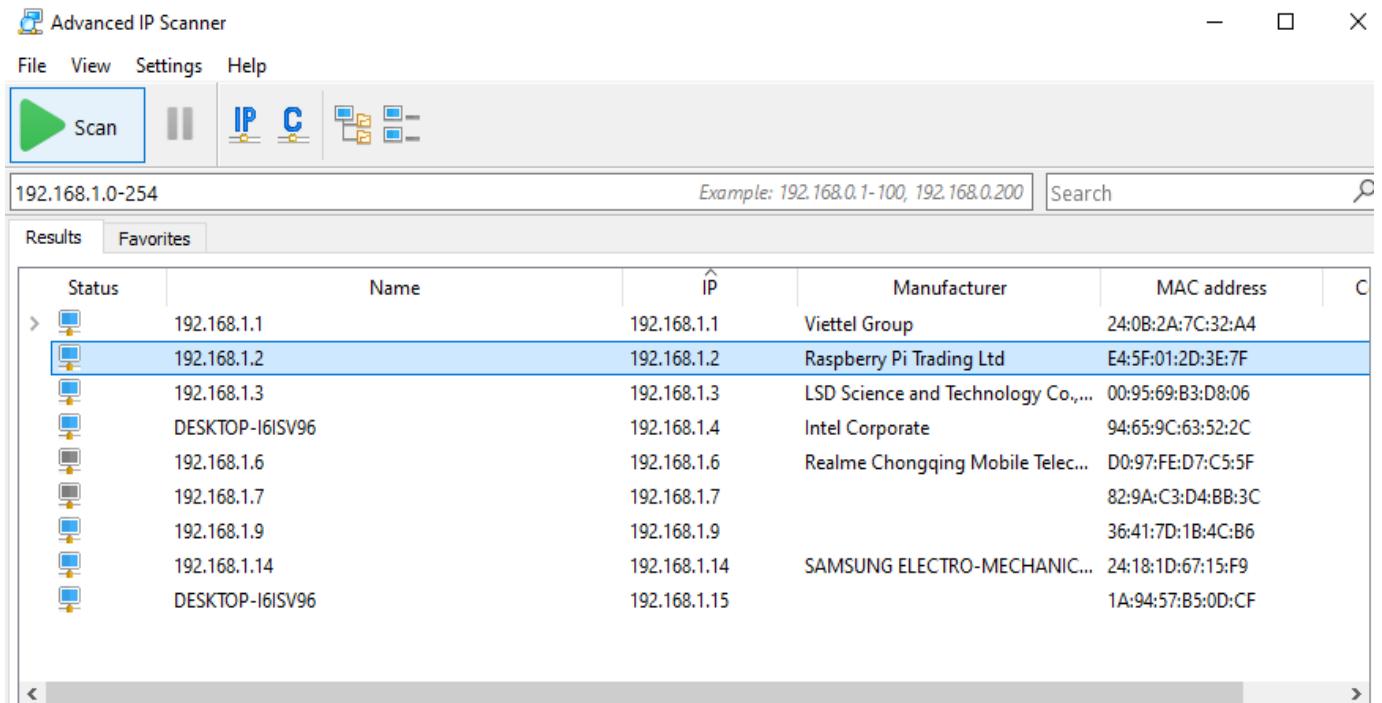
Bảng 3.8: Thông số kỹ thuật của cân tiêu ly điện tử có định lượng 500 gam.

<b>Thông số kỹ thuật</b>
<ul style="list-style-type: none"> <li>• Cân Tiêu Ly Điện Tử có định lượng 500gam</li> <li>• Độ chia / sai số : 0.01g hoặc 0.1g</li> <li>• Sử dụng pin AAA x2</li> <li>• Kích thước sản phẩm : 12.8 x 7.4 cm</li> <li>• Đơn vị cân : g ( Gam)-Oz ( Ounce)-Ct( Carat)- Gn( Grain)</li> <li>• Màn Hình Hiển Thị : LCD</li> </ul>

### 3.3.2. Các ứng dụng hỗ trợ

- Advanced Ip Scanner

Đây là ứng dụng quét mạng miễn phí và tin cậy, nó giúp hiển thị tất cả IP của các thiết bị mạng trong cùng một router, cho phép truy cập các thư mục chia sẻ, cung cấp khả năng điều khiển máy tính từ xa.



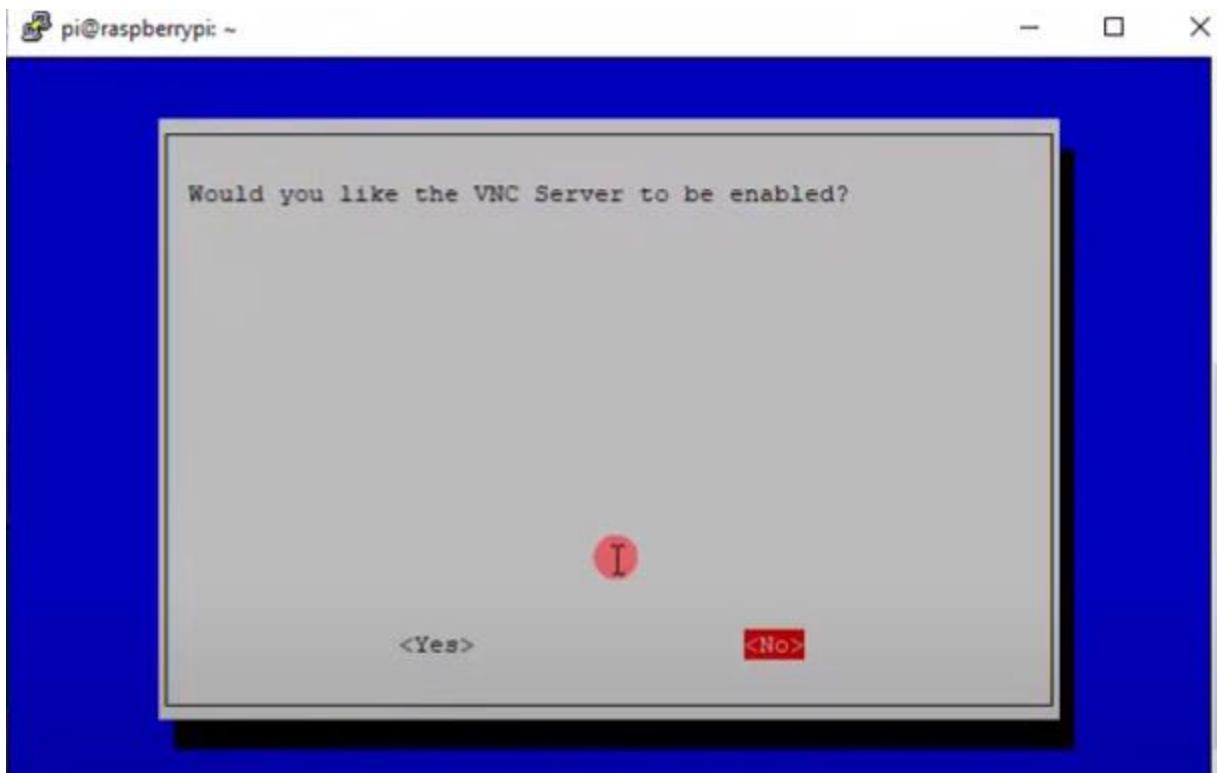
Hình 3.10: Advanced Ip Scanner tìm kiếm địa chỉ của Raspberry.

Trong đề tài mục đích dùng ứng dụng nhằm tìm kiếm địa chỉ Ip của Raspberry. Sau đó sử dụng Ip để kết nối với các ứng dụng hỗ trợ Putty, VNC Viewer, Win SCP với Raspberry từ xa.

- Ứng dụng PuTTY:

Đây là phần mềm miễn phí giúp người dùng thực hiện thao tác kết nối, quản lý và điều khiển các máy chủ thông qua mạng. PuTTY có nhiều tính năng và chức năng nổi bật như: thao tác kết nối dễ dàng, phù hợp cho hệ điều hành Windows 32 bit và 64 bit, dễ sử dụng, giao diện đơn giản, truy cập điều khiển máy tính thông qua giao thức SSH và nhiều tính năng khác.

Putty sẽ được kết nối với Raspberry thông qua Ip đã tìm được khi sử dụng Advanced IP Scanner và giao thức SSH. Từ đó truy cập vào cài đặt của Raspberry bật VNC Server giúp VNC Viewer có thể kết nối được và điều khiển Raspberry từ xa.

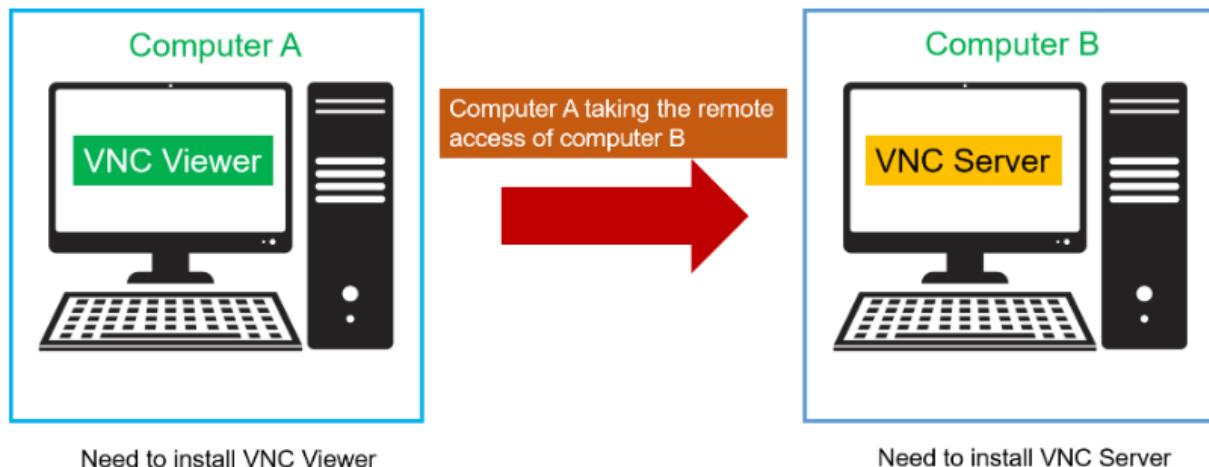


Hình 3.11: Bật VNC Server thông qua Putty.

- **VNC Viewer**

VNC\_Virtual Network Computing là một hệ thống để chia sẻ màn hình giữa các thiết bị khác nhau với mục đích điều khiển ở khoảng cách xa. Người dùng từ xa có thể xem và điều khiển một máy tính khác như thể họ đang ngồi trước máy tính đó.

VNC hoạt động dựa trên client/server. Máy tính cần được cài đặt thành một máy chủ VNC, trong khi máy tính khác muốn điều khiển từ xa cần cài đặt một trình xem VNC, hoặc còn gọi là client. Khi hai thành phần này được kết nối, máy chủ VNC sẽ chuyển gửi hình ảnh màn hình từ xa đến trình xem VNC.



Hình 3.12: VNC hoạt động dựa trên mô hình client/server.

Việc sử dụng VNC Viewer trong đề tài nhằm mục đích sử dụng Ip đã tìm được thông qua Advanced Ip Scanner và VNC Server đã bật trước đó trên Raspberry thông qua Putty, từ đó có thể điều khiển, lập trình trên Raspberry từ xa thông qua máy tính không cần màn hình riêng.

```

66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86

```

```

start=time.time()
GPIO.output(relay, True)
a0 = sensor.green
a1 = sensor.yellow
a2 = sensor.orange
a3 = sensor.red

with open('1.txt', 'w') as f:
    f.write("{} {} {} {}".format(a0,a1,a2,a3))
with open('1.txt', 'r') as f:
    w = f.readline()
    w = w.split()

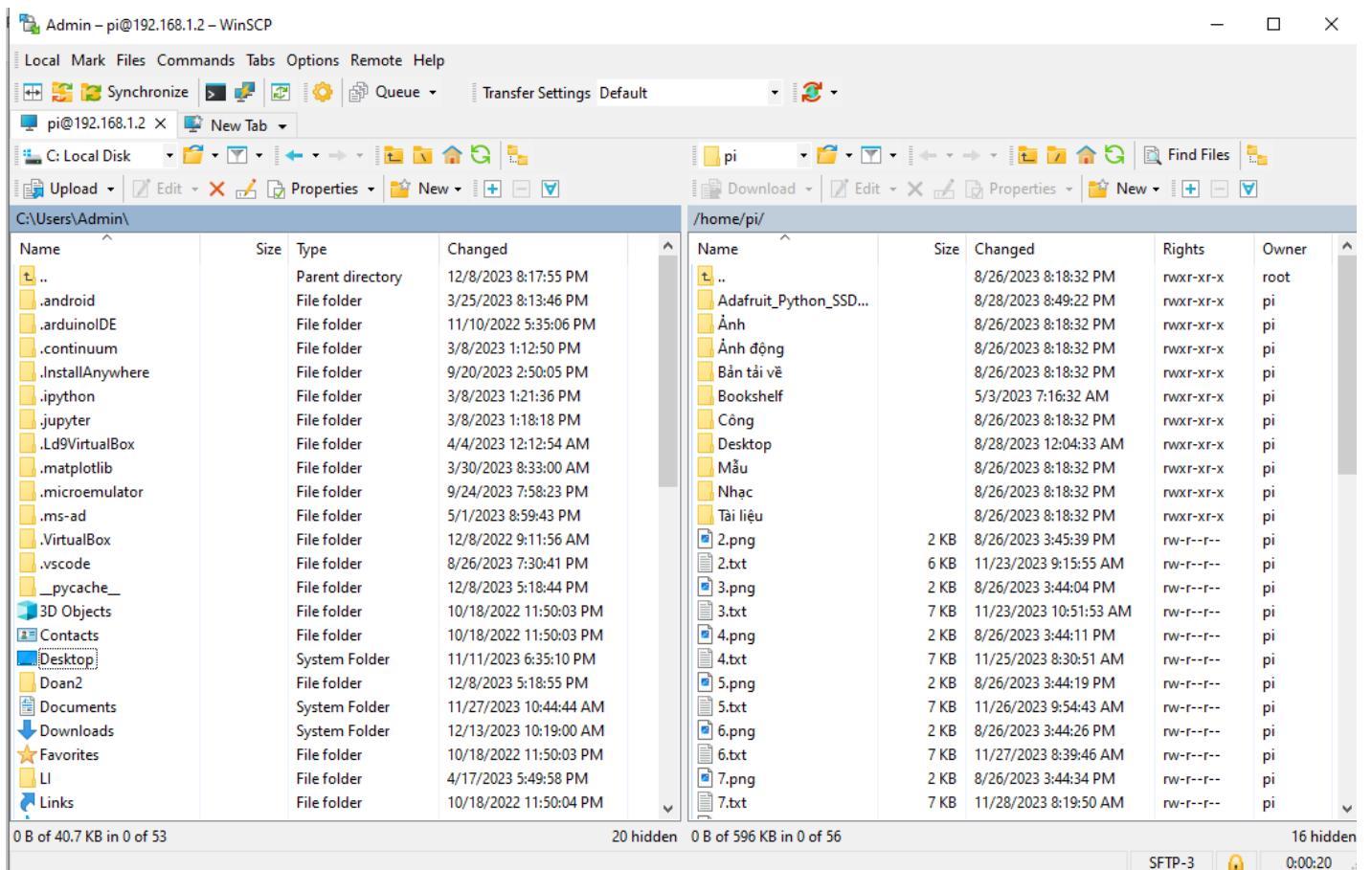
ML_Model = 'model.sav'
loaded_model = pickle.load(open(ML_Model, 'rb'))
data = {'1': [float(w[0])], '2': [float(w[1])], '3': [float(w[2])], '4': [float(w[3])]}
df=pd.DataFrame(data)
dk=np.array(df)
arr=scaler.transform(dk)
re = loaded_model.predict(arr)
print("Kq %s", re);

```

Hình 3.13: Sử dụng VNC Viewer lập trình trên Raspberry từ xa.

- Win SCP

WinSCP\_Windows Secure Copy là một phần mềm miễn phí có mã nguồn mở dành cho Windows, phần mềm này hỗ trợ người dùng kết nối với một máy tính từ xa để có thể chuyển file an toàn giữa các máy tính trong cùng một mạng.



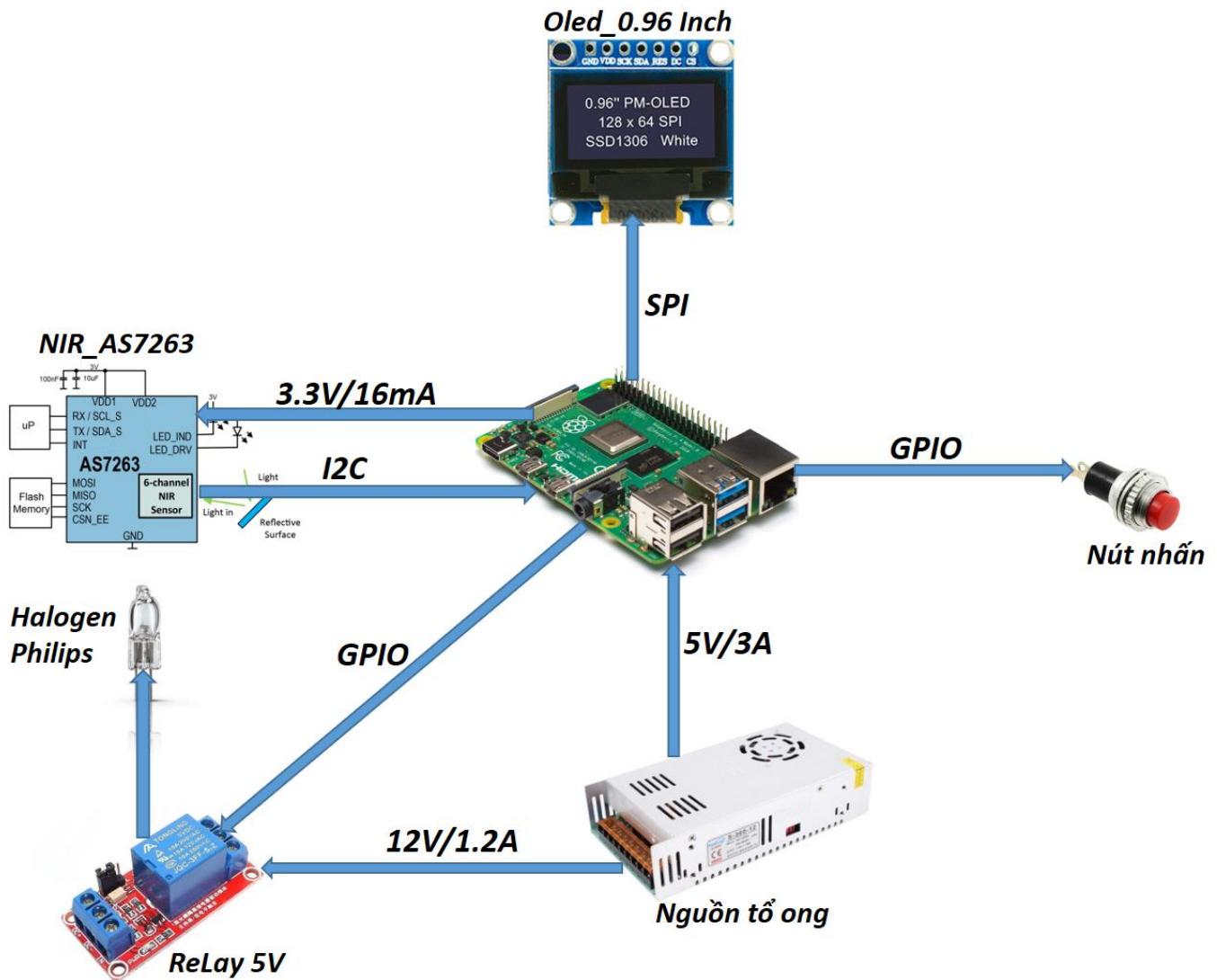
Hình 3.14: Sử dụng Win SCP chuyển file qua lại với Raspberry.

Việc sử dụng Win SCP nhằm mục đích chuyển một số file cần thiết như file dữ liệu nhóm thu thập trên Raspberry qua máy tính để tổng hợp và xử lý phổ thông thu được và đưa lên Google Colab để tiến hành xử lý và chạy chương trình máy học.

### 3.4. Mô hình hệ thống

#### 3.4.1. Mô tả kết nối hệ thống

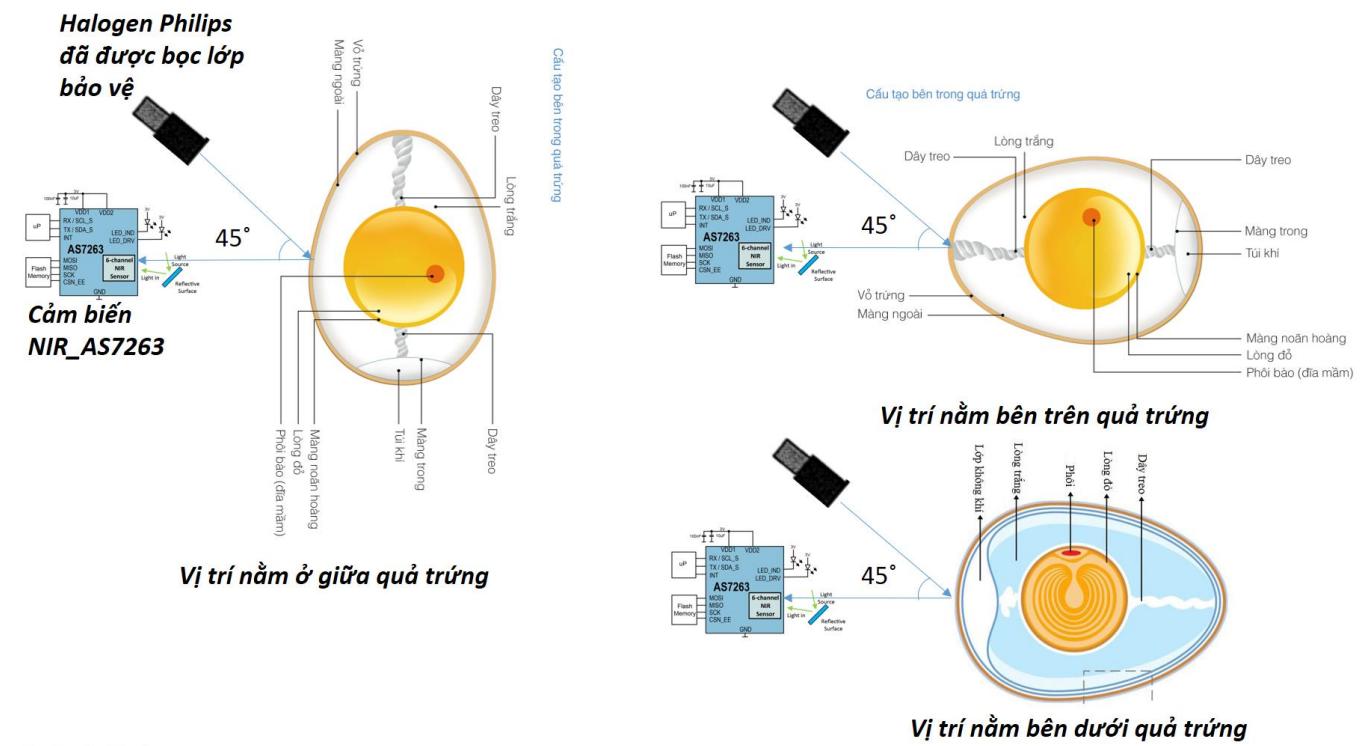
Sau khi tìm hiểu lý thuyết, các thiết bị chính và ứng dụng hỗ trợ, nhóm thực hiện xây dựng mô hình hệ thống nhúng cho việc thu thập dữ liệu phổ thô và lập trình hệ thống nhúng dự đoán ngày bảo quản trứng, Haugh Units. Hình 3.15 bên dưới cho thấy cách kết nối và các giao thức giữa các thành phần hệ thống.



Hình 3.15 Sơ đồ kết nối các thành phần của hệ thống.

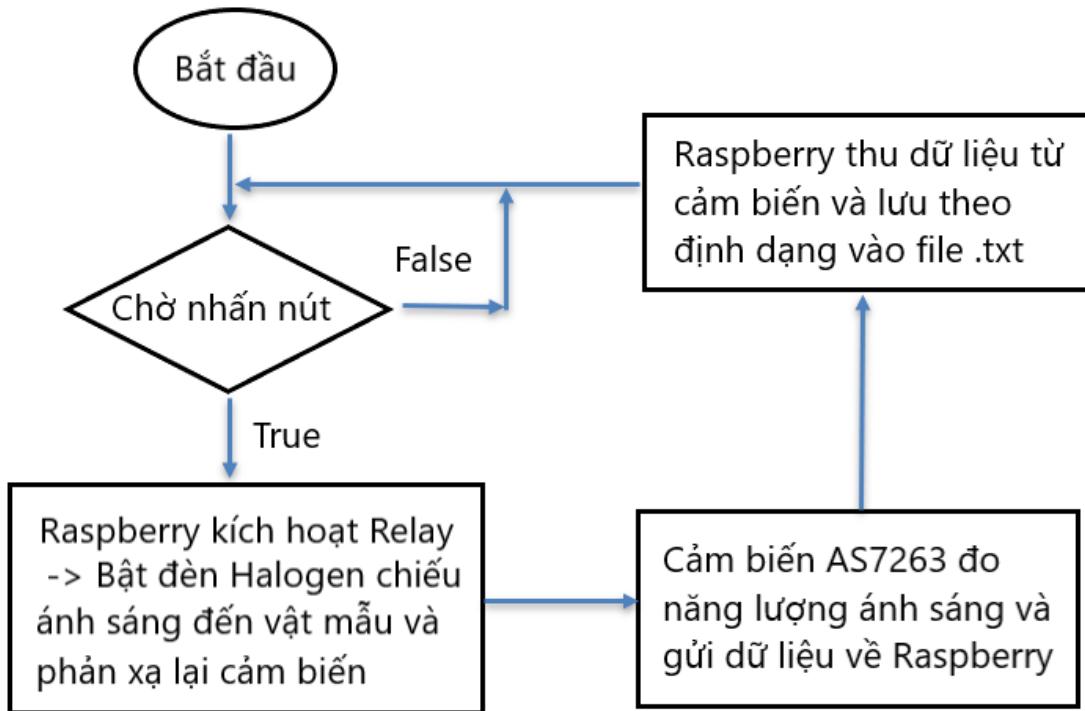
### 3.4.2. Vị trí đặt mẫu vật

Mô hình được xây dựng theo cơ sở lý thuyết trưng đặt càng gần cảm biến càng tốt và góc phản xạ ánh sáng là  $\sim 45^\circ$  điều này được chứng minh ở mục 2.5 Phương pháp đo. Bóng đèn Halogen Philips được bọc xung quanh bằng một vật liệu có mặt bên trong trắng và bên ngoài màu đen mục đích nhằm tạo một luồng ánh sáng chiếu thẳng đến trứng và không có ánh sáng tràn ra ảnh hưởng đến kết quả thu thập được của cảm biến.



Hình 3.16: Mô phỏng các vị trí đặt mẫu vật trứng gà để thu thập data.

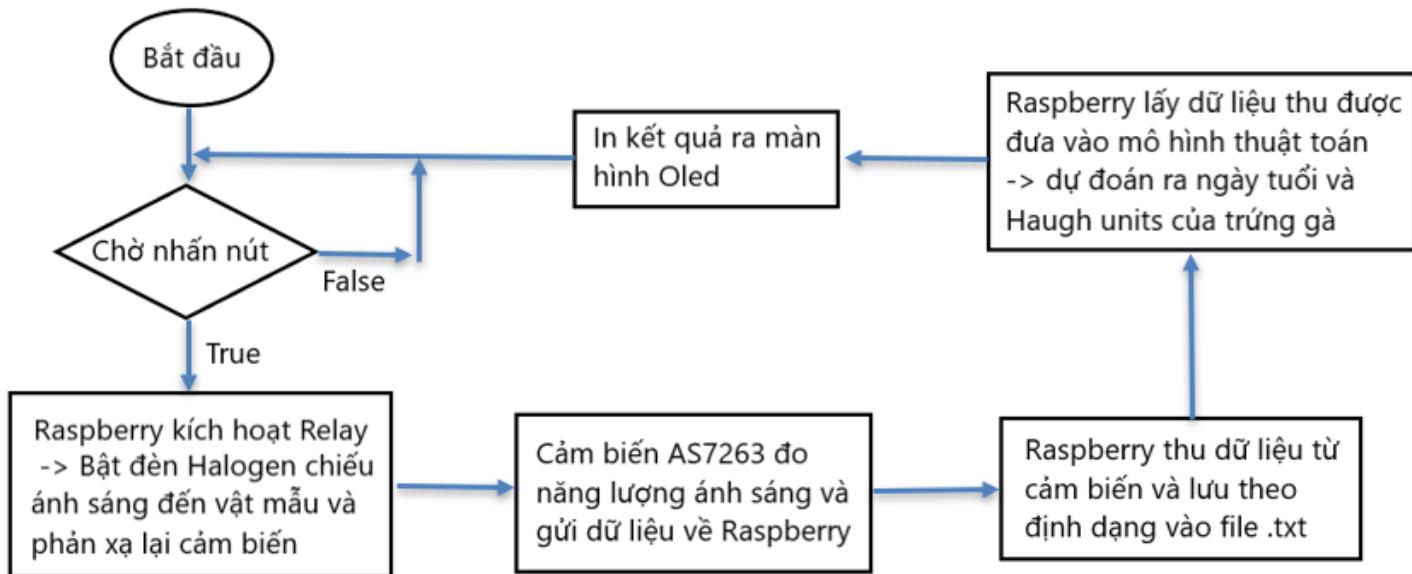
### 3.4.3. Lưu đồ giải thuật việc thu thập dữ liệu thô từ mẫu vật trứng gà



Hình 3.17: Lưu đồ giải thuật thu thập dữ liệu thô từ trứng gà.

Cách vận hành của hệ thống để thu thập dữ liệu phô thô: bắt đầu chương trình hệ thống sẽ chờ đợi tín hiệu khi nút bấm được nhấn Raspberry sẽ nhận được tín hiệu kích hoạt Module Relay cấp nguồn cho đèn Halogen sáng và chiếu tới mẫu vật trứng gà và phản xạ lại cảm biến. Cảm biến AS7263 đo năng lượng ánh sáng phản xạ lại qua 6 kênh R, S, T, U, V, W sau đó gửi lại Raspberry. Sau khi nhận được Raspberry sẽ ghi những dữ liệu lại vô file txt. Sau đó chuyển những dữ liệu thô qua máy tính để tổng hợp, xử lý, xây dựng mô hình máy học. Bên cạnh đó sau khi thu thập xong quang phổ nhóm tiến hành đo khối lượng và độ cao lòng trắng trứng để hoàn thành tập dữ liệu, việc này sẽ được trình bày ở mục 4.1 Thu thập tập dữ liệu thô.

### 3.4.4. Lưu đồ giải thuật mô hình dự đoán chất lượng, độ tươi trứng gà



Hình 3.18: Lưu đồ giải thuật của toàn bộ hệ thống.

Cách vận hành của thiết bị: bắt đầu chương trình hệ thống sẽ chờ đợi tín hiệu khi nút bấm được nhấn Raspberry sẽ nhận được tín hiệu kích hoạt Module Relay cấp nguồn cho đèn Halogen sáng và chiếu tới mẫu vật trứng gà và phản xạ lại cảm biến. Cảm biến AS7263 đo năng lượng ánh sáng phản xạ lại qua 6 kênh R, S, T, U, V, W sau đó gửi lại Raspberry. Sau khi nhận được Raspberry sẽ ghi những dữ liệu lại vào file txt. Sau đó raspberry đọc dữ liệu từ file để làm đầu vào cho mô hình máy học, kết quả dự đoán mô hình cho ra gồm ngày tuổi của trứng, Haugh Units sẽ được in ra màn hình Oled.

Mô hình máy học tạo ra bằng những bộ dữ liệu thu được trước đó, đã qua tiền xử lý dữ liệu, training và test trên Google Colab. Điều này sẽ được trình bày rõ ở hai Chương 4: Xây dựng mô hình máy học, Chương 5: Kết quả thực nghiệm.

## CHƯƠNG 4. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN CHẤT LƯỢNG ĐỘ TUƠI CỦA TRÚNG GÀ CÔNG NGHIỆP MÀU NÂU

Để xây dựng được một mô hình phù hợp cho việc dự đoán chất lượng và độ tươi của trứng gà công nghiệp cần trải qua 4 bước:

- Bước 1: Tiến hành thu thập và xây dựng tập dữ liệu.
- Bước 2: Tiền xử lý những dữ liệu thô.
- Bước 3: Huấn luyện mô hình.
- Bước 4: Chạy thử nghiệm.

### 4.1. Thu thập tập dữ liệu thô

Thu thập tập dữ liệu thô bằng cách sử dụng thiết bị đã xây dựng được từ trước: dữ liệu thô được thu thập bằng việc sử dụng cảm biến quang phổ NIR AS7263 đo năng lượng ánh sáng phản xạ lại của đèn halogen khi chiếu ánh sáng vào trứng gà sẽ thu được 6 tín hiệu thông qua 6 kênh kênh R, S, T, U, V, W của cảm biến (tương ứng với các dải bước sóng 610nm, 680nm, 730nm, 760nm, 810nm và 860nm) những tín hiệu này sẽ được hệ thống tiếp nhận và lưu vào file txt được định sẵn. Với mỗi 6 giá trị thu được sẽ tương ứng với số ngày bảo quản trứng hiện tại. Mỗi lần đo trứng gà được đặt ở 3 vị trí bên trên, bên dưới, ở giữa như hình 3.16: Vị trí đặt mẫu vật.

Sau khi thực hiện việc lấy giá trị quang phổ bước tiếp theo tiến hành đo khối lượng (W) của trứng bằng cân tiêu ly điện tử có định lượng 500g và độ sai lệch 0.01g. Sau đó đập nhẹ quả trứng ra một mặt phẳng gỗ bằng phẳng không có chỗ lồi hay lõm và dùng thước Inox độ phân giải 1 milimet để đo độ cao (H) của lòng trắng trứng. Kết quả có được sẽ được cho vô công thức (1) để tính Haugh Units.

Kết quả thu được là tập dữ liệu gồm 11 cột bao gồm 6 cột đầu là giá trị 6 tín hiệu thông qua 6 kênh kênh R, S, T, U, V, W của cảm biến (tương ứng với các dải bước sóng 610nm, 680nm, 730nm, 760nm, 810nm và 860nm), 5 cột sau là số ngày tuổi của trứng gà, trọng lượng của trứng, Haugh Units, chiều cao lòng trắng trứng và vị trí đặt mẫu vật.

## 4.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là quá trình chuẩn bị và biến đổi dữ liệu gốc trước khi phân tích hoặc mô hình hóa. Nó bao gồm làm sạch dữ liệu, xử lý dữ liệu thiếu, loại bỏ nhiễu và giá trị ngoại lai, chọn lọc đặc trưng và biến đổi dữ liệu. Mục tiêu là đảm bảo được sự phù hợp, tính chính xác và đáng tin cậy của bộ dữ liệu trước khi tiếp tục các bước xử lý dữ liệu tiếp theo. Để thực hiện được mục tiêu này, nhóm đã thực thi các bước tiền xử lý dữ liệu một cách tỉ mỉ để đảm bảo được tính chính xác và hiệu quả của mô hình. Dưới đây là chi tiết về các quy trình tiền xử lý dữ liệu đã được thực hiện.

Bước đầu nhóm sẽ xử lý dữ liệu thiếu vì dữ liệu chiều cao lòng trắng, khối lượng của trứng được nhóm nhập thủ công nên có thể xảy ra các trường hợp mẫu bị sai sót hoặc thiếu khi nhập liệu. Để khắc phục, nhóm ghi các kết quả đo được về hai giá trị này trên giấy trước, sau khi nhập liệu xong nhóm tiến hành xem xét lại dữ liệu để kiểm tra sự nhất quán và sửa chữa những giá trị thiếu hoặc sai sót dựa trên thông tin này. Các trường hợp không thể xác định chính xác giá trị, nhóm quyết định loại bỏ những mẫu này để đảm bảo tính tin cậy và chính xác của dữ liệu.

Tiếp theo là quá trình chọn lựa đặc trưng. Ở trong đề tài này nhóm quan tâm đến mối quan hệ giữa các giá trị đo từ cảm biến với số ngày bảo quản trứng và Haugh Units. Do đó, nhóm chỉ sử dụng các đặc trưng liên quan đến các giá trị đo từ cảm biến.

Để lựa chọn kỹ thuật tiền xử lý phù hợp nhất là một quá trình khá khó khăn. Vì vậy nhóm đề xuất ba kỹ thuật tiền xử lý để khảo sát và tìm ra chọn ra phương pháp có kết quả tốt nhất:

- Multiplicative Scatter Correction
- Standard Scaler
- Standard Normal Variate

#### 4.2.1. Standard Scaler

Trong tập dữ liệu phổ thông các giá trị quang phổ NIR nhóm thu được ở dải bước sóng R\_610nm = ~6000 (counts/ ( $\mu$ W/cm<sup>2</sup>)), còn ở dải bước sóng W\_860nm = ~32000 (counts/ ( $\mu$ W/cm<sup>2</sup>)) có chênh lệch quá lớn nên cần phải được chuẩn hóa trước khi thực thi huấn luyện mô hình máy học. Chuẩn hóa giá trị của đầu vào (giá trị bước sóng) và đầu ra (ngày tuổi của trứng, Haugh Units) của bộ dữ liệu trước khi huấn luyện là một bước thực hiện quan trọng, đồng thời việc chuẩn hóa giúp quá trình huấn luyện nhanh hơn. Nếu đầu vào (giá trị bước sóng) không được chuẩn hóa có thể dẫn đến quá trình huấn luyện không được ổn định dẫn đến kết quả không tốt.

Chuẩn hóa có thể tạo ra sự khác biệt lớn giữa một mô hình kém và một mô hình tốt. Bước tiền xử lý dữ liệu liên quan đến hai kỹ thuật normalization và standardization để rescale lại đầu vào và đầu ra trước khi huấn luyện mô hình.

Bảng 4.1: Sự khác nhau giữa MinMax Scaling và Standard Scaler.

	MinMax Scaling	Standard Scaler
Nguyên lý	Là sự biến đổi của các đặc trưng dựa vào giá trị nhỏ nhất và lớn nhất.	Là sự biến đổi của các đặc trưng dựa vào giá trị trung bình và độ lệch chuẩn.
Giới hạn	Thường là [0;1] hoặc [-1;1] hoặc có thể lựa chọn khoảng min-max khác.	Không bị giới hạn trong 1 phạm vi nhất định.
Nhiều	Bị ảnh hưởng bởi các giá trị ngoại lai (outlier).	Ít bị ảnh hưởng bởi các giá trị ngoại lai (outlier).
Dữ liệu	Có thể xử lý tốt dữ liệu một chiều.	Có thể xử lý tốt dữ liệu đa chiều.

Với tập dữ liệu đa chiều cùng với việc các giá trị ngoại lai xuất hiện do khi lấy mẫu có thể quả trứng bị ung (bên trong trứng màu vàng và có mùi trứng thối vì protein có trong thành phần của trứng bị phân hủy và tạo thành sulfua hydro H<sub>2</sub>S – hấp thụ ánh sáng nhiều hơn) nên các dữ liệu thu được sẽ nhỏ hơn các giá trị trứng không bị hỏng. Từ đây cho thấy kỹ thuật Standard Scaler cũng khá phù hợp sử dụng cho đề tài.

Kỹ thuật Standard Scaler sẽ giúp chuẩn hóa giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1 với mỗi đặc trưng (ứng với mỗi cột bước sóng ở tập dữ liệu). Sau việc chuẩn hóa hoàn tất, các thuật toán như Logistic Regression, Multiple Linear Regression, Linear Regression, tập dữ liệu được cải thiện.

Công thức tính toán:

$$x' = \frac{x - \text{mean}(x)}{\text{std}(x)} \quad (2)$$

Trong đó:

- x': vector đặc trưng sau khi chuẩn hóa.
- x: vector đặc trưng ban đầu.
- mean(x): giá trị trung bình của vector đặc trưng.
- std(x): độ lệch chuẩn của vector đặc trưng.

#### 4.2.2. Standard Normal Variate\_SNV và Multiplicative Scatter Correction\_MSC

Các tác nhân ảnh hưởng đến hình dạng của mỗi quang phổ cận hồng ngoại(NIR) là:

- Các bước sóng khác nhau sẽ có sự hấp thụ khác nhau đối với mẫu vật, do bản chất hóa học của chính mẫu vật đó. Trong hầu như các trường hợp, đây là tín hiệu muốn đo và nó liên quan đến chất phân tích quan tâm.
- Sự khác biệt về kích thước hạt trong vật liệu sẽ khiến ánh sáng bị lệch ở nhiều góc khác nhau tùy thuộc vào bước sóng của nó. Hiệu ứng tán xạ, cùng với sự khác biệt về độ dài đường truyền là nguyên nhân chính gây ra sự biến đổi trong phổ NIR.
- Sự khác biệt về độ dài đường dẫn đến mẫu do sự thay đổi vị trí hoặc sự không đồng đều trên bề mặt mẫu.

Ý tưởng đằng sau việc hiệu chỉnh tán xạ là loại bỏ tất cả các hiệu ứng xấu có tác động vào bản chất hóa học của mẫu. Vì vậy, ý tưởng là nếu chúng ta có thể loại bỏ trước những tác động không mong muốn này, thì chúng ta sẽ có được một mô hình tốt hơn. Đặc trưng trong việc loại bỏ những hiệu ứng không liên quan này có hai phương pháp là Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV).

Kỹ thuật SNV được giới thiệu bởi Barnes vào năm 1989 [13] thực hiện chuẩn hóa phổ bao gồm việc trừ từng phổ theo giá trị trung bình của chính nó và chia nó cho độ lệch chuẩn của chính nó. Sau SNV, mỗi phổ sẽ có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. SNV có gắng làm cho tất cả quang phổ có thể so sánh được về cường độ (hoặc mức độ hấp thụ). Có thể hữu ích khi hiệu chỉnh quang phổ đối với những thay đổi về độ dài đường quang và sự tán xạ ánh sáng. Theo Rinnan và công sự, 2009 [14] công thức tính SNV:

$$X_{corr} = \frac{X_{org} - a_0}{a_1} \quad (3)$$

Trong đó:

- $x_{corr}$ : quang phổ đã hiệu chỉnh.

- $a_0$ : giá trị trung bình của phô mẫu cần hiệu chỉnh.
- $x_{org}$ : dữ liệu phô thu bằng cảm biến NIR.
- $a_1$ : độ lệch chuẩn của phô mẫu.

Multiplicative Scatter Correction\_MSC lần đầu tiên được Martens giới thiệu vào năm 1983 [15] thường sử dụng để bù cho hiệu ứng tán xạ ánh sáng và thay đổi về độ dài đường đi. Thật vậy, sự tán xạ ánh sáng từ các mẫu rắn và nhũ tương có thể gây ra những sai lệch cấp số nhân phụ thuộc vào bước sóng. MSC giảm thiểu những sai lệch này bằng cách khớp mô hình tuyến tính giữa phô tham chiếu và phô khác của tập dữ liệu bằng phương pháp bình phương tối thiểu tuyến tính. Phô tham chiếu thường được chọn làm giá trị trung bình của tất cả các phô trong tập dữ liệu. Các hệ số mô hình được sử dụng để tính toán phô lý tưởng. Sau khi áp dụng MSC, tất cả các quang phô dường như có cùng mức độ hấp thụ. MSC và SNV thường dẫn đến kết quả tương tự. Theo Rinnan và công sự, 2009 [14] MSC bao gồm 2 bước:

- Ước tính hệ số hiệu chỉnh:

$$x_{org} = b_0 + b_{ref,1} * x_{ref} + e \quad (4)$$

- Hiệu chỉnh quang phô:

$$x_{corr} = \frac{x_{org} - b_0}{b_{ref,1}} = x_{ref} + \frac{e}{b_{ref,1}} \quad (5)$$

Trong đó:

- $x_{org}$ : dữ liệu phô thu bằng cảm biến NIR.
- $e$ : phần chưa được mô hình hóa của  $x_{org}$ .
- $x_{ref}$ : phô tham chiếu được sử dụng để tiền xử lý (phô trung bình của bộ dữ liệu được sử dụng làm phô tham chiếu).
- $x_{corr}$ : quang phô đã hiệu chỉnh.
- $b_0$  và  $b_{ref,1}$ : các tham số vô hướng, khác nhau đối với mỗi mẫu.

### 4.3. Mô hình máy học thực nghiệm

Cơ sở lý thuyết của đề tài dựa vào việc sử dụng tín hiệu thu được từ cảm biến để tìm ra hàm thể hiện mối quan hệ của dữ liệu đầu vào là các tín hiệu của cảm biến và kết quả đầu ra là giá trị tham chiếu số ngày trứng trải qua và Haugh Units. Từ đó có thể dự đoán được thời gian bảo quản trứng, HU với một đầu vào dữ liệu mới mà cảm biến thu được. Điều này phù hợp với nhóm thuật toán học có giám sát. Dựa vào các đặc tính của phân tử trong trứng có thể thay đổi liên tục tại mỗi thời điểm bất kỳ nên việc tìm một hàm số để dự đoán đầu ra ứng với các đầu vào liên tục của cảm biến trở nên phù hợp với đề tài. Các đặc tính trên phù hợp với bài toán hồi quy.

Việc lựa chọn được một mô hình máy học thích hợp cho bài toán tương đối khó khăn, nên trong đề tài này nhóm khảo sát áp dụng ba mô hình học máy với bài toán hồi quy để huấn luyện, sau đó so sánh độ chính xác để có thể tìm ra được mô hình phù hợp nhất. Các phương pháp mà nhóm sẽ sử dụng bao gồm: Multiple Linear Regression, Support Vector Regression và Decision Tree. Bằng cách khám phá sự đa dạng này, nhóm hy vọng tìm ra được mô hình có kết quả tốt nhất trong việc dự đoán số ngày tuổi của trứng và Hu.

#### 4.3.1. Mô hình Multiple Linear Regression

Multiple Linear Regression hay hồi quy đa biến là mô hình được sử dụng nhằm phân tích và ước tính mối quan hệ giữa hai hay nhiều biến độc lập và một biến phụ thuộc. Mô hình tìm ra phương trình tổng quát về những mối quan hệ chặt chẽ của biến phụ thuộc và các biến độc lập, ngoài ra mô hình có thể xác định được giá trị của biến phụ thuộc tại một giá trị cố định cụ thể của biến độc lập.

Công thức chung của mô hình Multiple Linear Regression là:

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n \quad (6)$$

Trong đó:

- $y$  là biến phụ thuộc (đầu ra).
- $X_1, X_2, X_3, \dots, X_n$ : các biến độc lập.

- $\beta_1, \beta_2, \beta_3, \dots, \beta_n$  lần lượt là các hệ số của mô hình, quyết định độ nghiêng của đường thẳng.
- $\beta_0$  là hằng số quyết định sự dịch chuyển của đường thẳng so với gốc tọa độ.

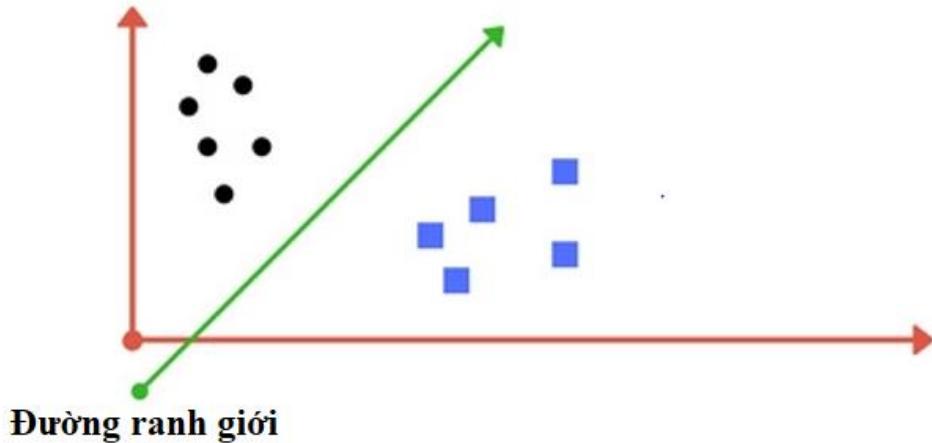
Trong quá trình huấn luyện, các hệ số  $\beta$  sẽ được mô hình tính toán ra, sau đó có thể dùng bộ số này để xử lý các đầu vào mới liên tục để tính toán các giá trị đầu ra. Mục tiêu chính của phân tích hồi quy là ước lượng các giá trị hệ số hồi quy dựa trên dữ liệu mẫu.

Dựa vào các dữ liệu phô thu được từ cảm biến AS7263 ở 6 bước sóng (610nm, 680 nm, 730nm, 760nm, 810nm, 860nm). Nhưng do ngưỡng bước sóng lựa chọn để tài là từ 700nm – 1100nm nên sẽ loại bỏ 2 dải bước sóng đầu là 610 nm và 680 nm. Lúc này bài toán cần được xử lý với nhiều biến độc lập nên mô hình Multiple Linear Regression là một lựa chọn để sử dụng nhằm đạt được mục tiêu là dự đoán biến phụ thuộc số ngày trứng đã trải qua và Haugh Units.

Áp dụng để tài cho mô hình Multiple Linear Regression, biến đầu ra y sẽ là ngày tuổi của trứng hoặc HU, biến đầu vào gồm  $X_1, X_2, X_3, X_4$  tương ứng là các giá trị bước sóng 730 nm – 760 nm - 810nm - 860nm thu được vào cảm biến. Mục đích từ đầu ra và đầu vào mô hình là tìm ra các hệ số  $\beta_1, \beta_2, \beta_3, \beta_4$  từ đây có thể dùng đầu vào mới để tính toán dự đoán được đầu ra y.

#### 4.3.2. Mô hình Support Vector Regression (SVM)

Mô hình Support Vector Regression (SVM) là một phương pháp máy học thường được sử dụng để dự đoán và phân loại các điểm dữ liệu. SVM tạo ra một ranh giới (tuyến tính hoặc phi tuyến tính) nhằm mục đích phân chia các tập dữ liệu có tính chất khác nhau. Nó cố gắng tìm ranh giới tối ưu nhất giữa những điểm dữ liệu gần nhau nhất của từng tập dữ liệu, ranh giới này được gọi là support vectors.



Hình 4.1: Ví dụ về SVM.

Support Vector Regression (SVR) là một phương pháp máy học thường được sử dụng trong việc phân tích hồi quy và SVR được phát triển dựa trên SVM. SVR sẽ cố gắng tìm ra hàm gần đúng về mối liên hệ giữa các biến đầu vào và biến đầu ra liên tục, đồng thời giảm thiểu sai số dự đoán. Không giống như SVM thường được sử dụng cho các bài toán về phân loại, SVR cố gắng tìm ra một hyperplane hợp lý nhất của các điểm dữ liệu trong một không gian liên tục. Điều này đạt được bằng cách mô hình sẽ chiếu các biến đầu vào lên một không gian đặc trưng có nhiều chiều và tìm siêu phẳng giúp tối đa hóa ranh giới giữa hyperplane và các điểm dữ liệu gần nhất của những lớp dữ liệu khác nhau, đồng thời giảm thiểu lỗi dự đoán. Hyperplane là khoảng ranh giới chia không gian đầu vào thành hai phần hoặc nhiều vùng khác nhau, mỗi một phần sẽ tương ứng với một loại đầu ra khác nhau. Hyperplane trong không gian hai chiều là một đường thẳng chia không gian thành hai nửa. Tuy nhiên, trong không gian ba chiều, hyperplane là một mặt phẳng chia không gian thành hai nửa.

Để có thể xử lý được các mối quan hệ phi tuyến tính của các biến đầu vào và biến mục tiêu, SVR đã dùng hàm kernel để chiếu dữ liệu tới một không gian có nhiều chiều hơn. Vì vậy SVR là một công cụ mạnh mẽ để giải quyết các tác vụ hồi quy có thể chứa các mối quan hệ phức tạp của các biến đầu vào và biến đầu ra. Có một số hàm kernel phổ biến mà SVM sử dụng như là:

- Linear Kernel: Đây là hàm kernel đơn giản nhất, thực hiện ánh xạ dữ liệu vào không gian đặc trưng tuyến tính từ không gian ban đầu. Nó được sử dụng trong SVM tuyến tính.
- Polynomial Kernel: Hàm kernel đa thức ánh xạ dữ liệu vào không gian đặc trưng bằng cách sử dụng các hàm mũ. Tham số mũ được có tác dụng là điều chỉnh mức độ phức tạp của hàm kernel. Đa thức kernel cho phép SVM phân loại các lớp dữ liệu phi tuyến tính.
- Radial Basis Function (RBF) Kernel: Đây là hàm kernel phổ biến nhất trong SVM. Nó ánh xạ dữ liệu vào không gian đặc trưng sử dụng hàm cơ sở đồng tâm. RBF kernel cho phép SVM phân loại các lớp dữ liệu phi tuyến tính và cũng có thể mô hình hóa các mối quan hệ phức tạp hơn so với các hàm kernel khác.
- Sigmoid Kernel: Hàm kernel sigmoid ánh xạ dữ liệu vào không gian đặc trưng bằng cách sử dụng hàm sigmoid. Nó cho phép SVM phân loại các lớp dữ liệu phi tuyến tính và có tính chất gần giống với hàm sigmoid của mạng neuron nhân tạo.

SVR tìm ra hàm gần đúng cho mối quan hệ giữa các biến đầu vào và biến đầu ra liên tục, đồng thời giảm thiểu sai số dự đoán, cho thấy mô hình SVR thích hợp cho việc tìm ra hàm tổng quát về mối quan hệ của các bước sóng thu được từ cảm biến AS7263 và các biến đầu ra là HU và ngày tuổi của trứng. Từ hàm thu được có thể sử dụng dự đoán HU và ngày tuổi của trứng thông qua đầu vào mới.

Công thức SVR trong bài toán dự đoán HU và ngày tuổi của trứng từ dữ liệu cảm biến AS7263 có dạng:

$$f(x) = sign(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{18} x_{18}) \quad (7)$$

Trong đó:

- $f(x)$  là hàm quyết định, nó trả về nhãn dự đoán của Hu hoặc ngày tuổi của trứng (+1 hoặc -1).
- $sign(x)$  là hàm signum, trả về +1 nếu  $x \geq 0$  và -1 nếu  $x < 0$ .
- $\beta_0, \beta_1, \beta_2 \dots \beta_{18}$  là các tham số học được trong quá trình huấn luyện SVR.

### 4.3.3. Decision Tree

Mô hình Decision Tree (cây quyết định) là một phương pháp máy học được sử dụng để dự đoán và phân loại dữ liệu. Giải pháp chính của mô hình này là xây dựng được cây quyết định, trong đó mỗi nút được thể hiện cho một thuộc tính và mỗi nhánh được thể hiện cho một giá trị của thuộc tính. Mô hình được dùng để tạo ra các quy tắc phân loại dựa trên các thuộc tính của dữ liệu đầu vào.

Trong bài toán dự đoán Hu và ngày tuổi của trúng từ dữ liệu cảm biến AS7263, để sử dụng mô hình Decision Tree để xây dựng một cây quyết định dựa trên giá trị đo từ cảm biến AS7263. Mô hình Decision Tree giúp chúng ta tìm ra các quy tắc dựa trên giá trị đo từ cảm biến AS7263 để phân loại các mẫu vào các nhãn HU và ngày tuổi của trúng khác nhau. Công thức của Decision Tree không có dạng toán học như các mô hình Multiple Linear Regression hay Support Vector Machine. Thay vào đó, mô hình này được xây dựng theo quy tắc rẽ nhánh (split rule) để phân loại dữ liệu. Mỗi nút của cây quyết định thể hiện cho một thuộc tính và mỗi nhánh thể hiện cho một giá trị của thuộc tính. Quy tắc rẽ nhánh được sử dụng để xác định điều kiện phân loại tại mỗi nút. Các quy tắc này được phát triển theo các thuộc tính và giá trị của chúng trong tập dữ liệu huấn luyện.

## 4.4. Đánh giá mô hình máy học

Việc lựa chọn mô hình thực nghiệm hoàn tất với ba mô hình cần được khảo sát MLR, SVR và Decision Tree. Bước tiếp theo cần phải có phương pháp đánh giá mô hình máy học xem có thích hợp với bài toán mà nhóm đã đặt ra hay không nhóm sử dụng hai phương pháp để đánh giá mô hình máy học:

- Đánh giá mô hình dựa trên độ lệch gốc căn bậc hai của mức trung bình bình phương \_RMSE.
- Đánh giá mô hình máy học dựa trên độ chính xác phân tích\_  $R^2$ .

#### 4.4.1. Đánh giá mô hình máy học dựa trên độ chính xác phân tích $R^2$

Mô hình trên Google Colab được đánh giá dựa trên hệ số hiệu chỉnh R bình phương ( $R^2$  Score):

- $R$  bình phương được dùng mô hình hồi quy tuyến tính (Regression Model) để thể hiện mức độ phù hợp của mô hình đối với các biến đầu vào của mô hình. Giới hạn trong khoảng từ 0 – 1.  $R$  bình phương càng lớn càng thể hiện mức độ chính xác và sự phù hợp của mô hình ứng với bài toán và ngược lại.
- Công thức tính hệ số  $R$  bình phương:

$$R^2 = 1 - \frac{ESS}{TSS} \quad (8)$$

Trong đó:

- ESS (Residual Sum of Squares): tổng các độ lệch bình phương của phần dư.
- TSS (Total Sum of Squares): tổng độ lệch bình phương của toàn bộ các nhân tố nghiên cứu.

Đi kèm với  $R^2$ , ta còn có  $R^2$  hiệu chỉnh được tính theo công thức:

$$R_{hc}^2 = 1 - \frac{ESS}{TSS} \div \frac{n - k}{n - 1} \quad (9)$$

Sau khi biến đổi:

$$R_{hc}^2 = 1 - \frac{n - 1}{n - k} (1 - R^2) \quad (10)$$

Trong đó:

- $n$  là số lượng mẫu thực hiện.
- $k$  là tham số của mô hình.

#### 4.4.2. Đánh giá mô hình dựa trên độ lệch gốc căn bậc hai của mức trung bình bình phương (RMSE)

RMSE là một thước đo thống kê phổ biến được sử dụng trong việc đánh giá mức độ chính xác của mô hình dự đoán. Nó đo lường sự khác biệt trung bình giữa giá trị dự đoán và giá trị thực tế, dựa trên bình phương sai số.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}} \quad (11)$$

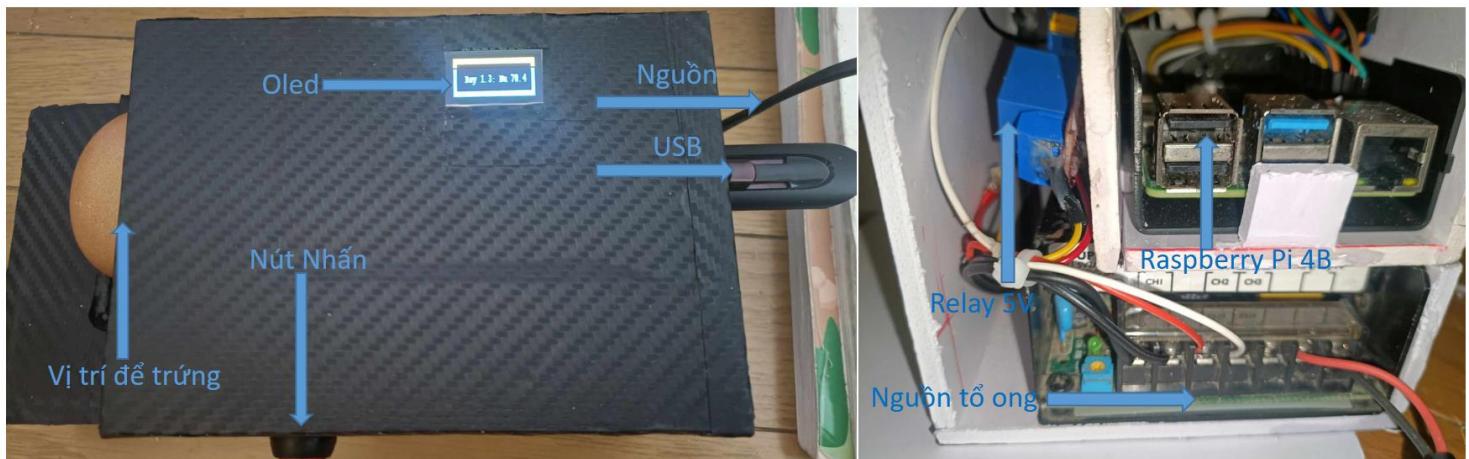
Trong đó:

- $RMSE$ : Giá trị căn bậc hai của trung bình sai số bình phương
- $\hat{y}_i$ : Dữ liệu dự đoán thứ  $i$ .
- $y_i$ : Dữ liệu thực thứ  $i$ .
- $n$ : Tổng số dữ liệu.

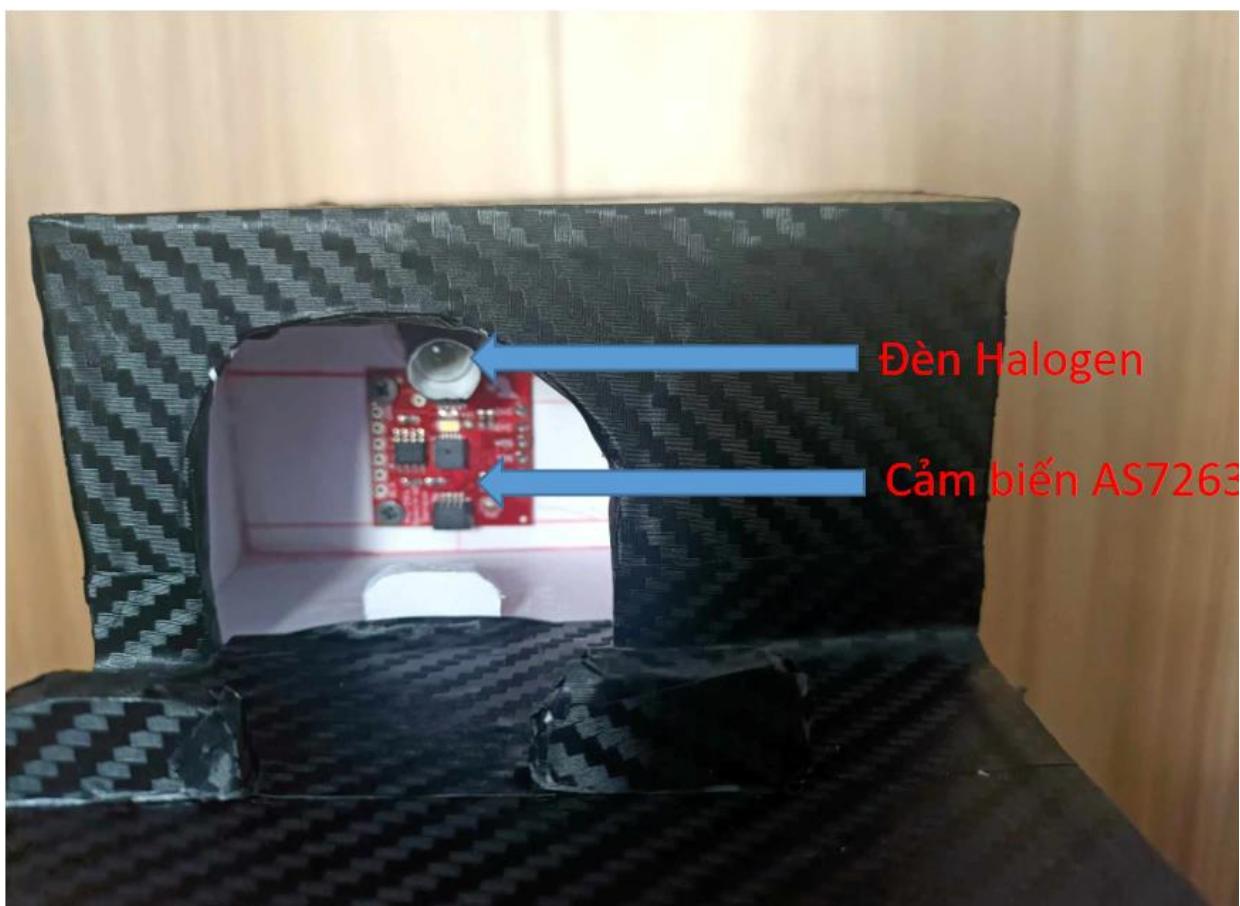
RMSE tính toán bình phương sai số giữa giá trị dự đoán và giá trị thực tế, sau đó lấy căn bậc hai của trung bình các bình phương sai số này. Giá trị RMSE càng nhỏ, thể hiện mức độ chính xác cao hơn của mô hình, tức là mô hình dự đoán gần giá trị thực tế.

## CHƯƠNG 5. KẾT QUẢ THỰC NGHIỆM

### 5.1. Hiện thực phần cứng



Hình 5.1: Mặt bên trên, mặt sau của phần cứng sau khi hiện thực.



Hình 5.2: Mặt trước của phần cứng sau khi hiện thực.

Hình 5.1 và 5.2 là hình ảnh phần cứng được nhóm hiện thực hóa từ những thành phần hệ thống đã được tìm hiểu ở mục 3.2: Thành phần hệ thống. Bóng đèn Halogen và cảm biến AS7263 được nhóm thiết kế được đặt trong 1 hộp kín 5 mặt (tránh ánh sáng ngoài môi trường ảnh hưởng tới việc đo đặc ánh sáng) và tạo thành một góc  $\sim 45^0$ . Một mặt được cắt hở 1 khoảng giúp khi đặt mẫu vật trưng bày. Khi đặt quả trưng nằm dọc, bên ngoài miệng trống của hộp, có 2 thanh nhựa nằm 2 bên giúp cố định quả trưng và giúp quả trưng đặt vừa phần định hoặc phần đáy của quả trưng.

```

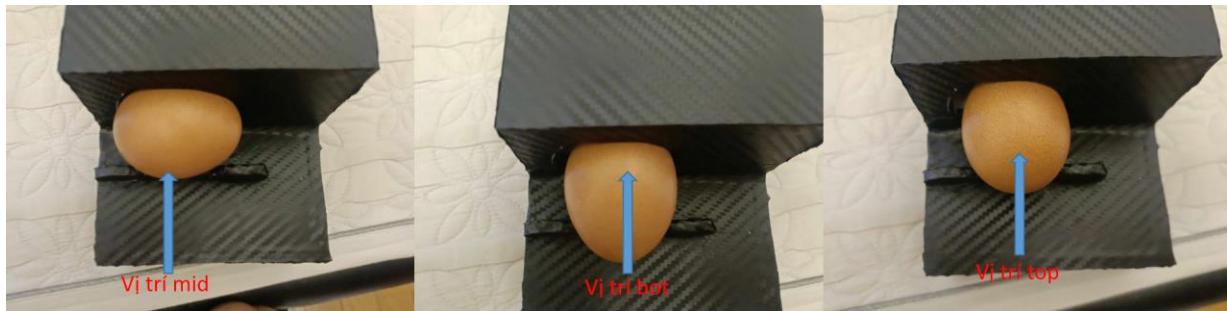
Raspberry Pi 4 Model B - 2GB
bootloader: d1be7b5b Jan 16 2021
update-ts: 1611216292
00:22
board: b03111 f483a565 dc:a6:32:01:62:69
boot: mode USB-MSD 4 order f41 retry 1/128 restart 0/-1
SD: card not detected
part: 0 mbr [0x0b:00000002 0x00:00000000 0x00:00000000 0x00:00000000]
fw: start.elf fixup.dat
net: down ip: 0.0.0.0 sn: 0.0.0.0 gw: 0.0.0.0
tftp: 0.0.0.0 00:00:00:00:00:00

MSD [02:00] 3.00 000000:02 register MSD
MSD [02:00] 3.00 000000:02 LUN 0
MSD INQUIRY [02:00] 3.00 000000:02
HUB [01:00] 2.16 000000:01 init port 3 speed 1
MSD [02:00] 3.00 000000:02 lun 0 block-count 1465149168 block-size 512
Trying partition: 0
lba: 2 oem: 'BSD 4.4' volume: 'SDCARD'
rsc 32 fat-sectors 178808 c-count 22887367 c-size 64 r-dir 2 r-sec 0
Trying partition: 0
lba: 2 oem: 'BSD 4.4' volume: 'SDCARD'
rsc 32 fat-sectors 178808 c-count 22887367 c-size 64 r-dir 2 r-sec 0
Firmware not found

```

Hình 5.3: Raspberry bị lỗi “won’t boot fix”.

Thiết bị USB lắp thêm do trong lúc sử dụng và thao tác trên Raspberry bị xuất hiện lỗi “won’t boot fix” nên không thể sử dụng hệ điều hành thông qua thẻ nhớ cắm trực tiếp vào raspberry mà phải cần một usb hỗ trợ để chạy hệ điều hành.



Hình 5.4: Các vị trí đặt mẫu vật trứng thực nghiệm.

Quan sát hình 5.4 là các vị trí đặt mẫu vật trứng mid, top, bot trên phần cứng dùng để thu thập dữ liệu thử.

## 5.2. Tập dữ liệu thử thu được

### 5.2.1. Mô tả chi tiết tập dữ liệu

Mẫu vật trứng gà sử dụng nghiên cứu gồm 420 quả trứng gà màu nâu còn nguyên vỏ với trọng lượng từ 55g đến 65g (cỡ M) được mua tại trại gà Hoàng Lan địa chỉ: B184, Bình Thuận, Thuận An, Bình Dương, Việt Nam. Do khối lượng trứng gà khá lớn và địa điểm khá xa nên phải nhờ trang trại đóng gói và vận chuyển bằng xe chuyên dụng, sau khi nhận được trứng nhóm tiến hành kiểm tra lại xem có mẫu vật nào bị nứt hay vỡ trong quá trình vận chuyển hay không và tiến hành loại bỏ, những mẫu vật trứng gà nứt thường bị hư rát nhanh chóng chỉ khoảng 5 ngày đến 7 ngày đều này sẽ ảnh hưởng đến kết quả đo được.

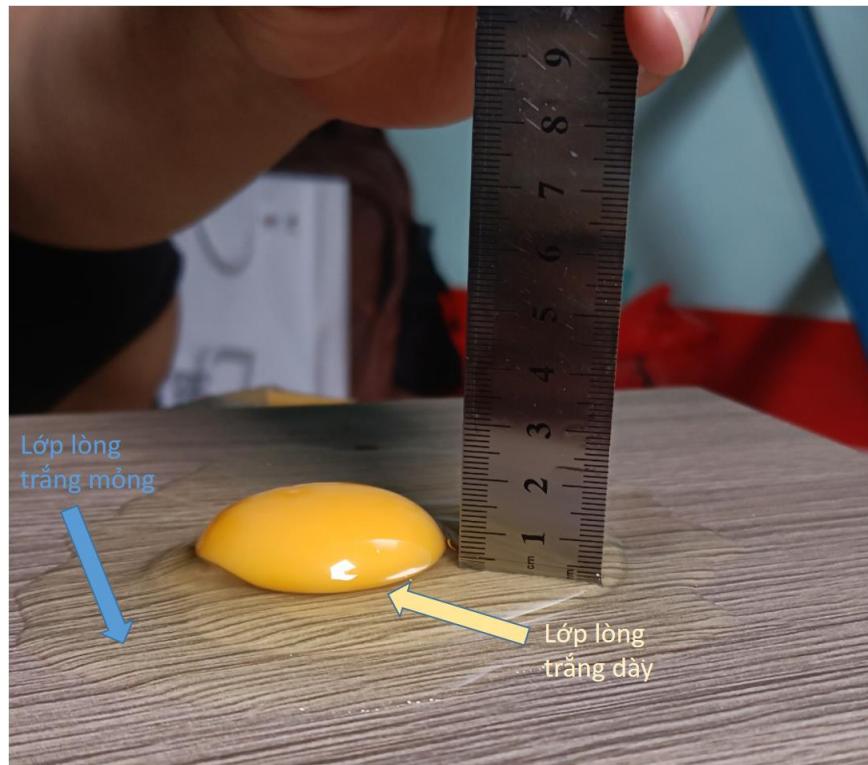


Hình 5.5: Mẫu vật trứng gà nhận vào 22/11/2023 được để trong vỉ không bị bẹp hay vỡ.

Vì việc mua và vận chuyển mất khá nhiều thời gian nên ngày tiếp nhận mẫu vật trứng gà là vào ngày 22/11/2023 nhưng ngày trứng đẻ vào ngày 21/11/2023. Nên các phép đo bỏ qua ngày tuổi số một và thực hiện bắt đầu từ khi ngày tuổi của trứng là 2 đến khi trứng được 21 ngày tuổi (12/11/2023) khoảng cách giữa các lần đo là 24 tiếng. Mẫu vật trứng gà được bảo quản ở nơi khô ráo trên cao, ở nhiệt độ phòng ( $24^0 - 30^0$ ).

Sau khi tiếp nhận 420 mẫu vật trứng gà nhóm tiên hành chia đều cho 20 ngày mỗi ngày đo 20 quả trứng với thiết bị phần cứng được thiết kế trước đó, bằng cách đo lần lượt điểm giữa của mỗi quả trứng (mid), điểm trên đầu mỗi quả trứng (top) và điểm đáy của từng quả trứng (bot) như vậy ta mỗi ngày thu được 60 đường dữ liệu gồm 20 đường dữ liệu giữa mỗi quả trứng (mid), 20 đường dữ liệu đỉnh mỗi quả trứng (top) và 20 đường dữ liệu đáy mỗi quả trứng (bot), với mỗi 1 đường dữ liệu chứa 6 giá trị ánh sáng gồm các bước sóng: 610nm, 680nm, 730nm, 760nm, 810nm và 860nm. Còn lại 20 quả trứng là để thay thế cho việc khi đo chiều cao lòng trắng trứng có thể bị vỡ hay hư hỏng trong quá trình thu thập dữ liệu.

Sau khi đo xong quang phổ của từng quả trứng, ta tiến hành đo đặc đơn vị Haugh của từng quả trứng bằng cách đo khối lượng của từng quả trứng bằng cân tiêu ly. Sau khi đo xong khối lượng, ta tiến hành đập từng quả trứng đặt lên 1 mặt phẳng để đo chiều cao lòng trắng của từng quả trứng. Tại bước này ta cần tiến hành cẩn thận tránh làm vỡ lòng đỏ trứng, cũng như khi đo lòng trắng ta cần đặt thước đo trong vùng lòng trắng nhằm tránh tình trạng vỡ lòng tránh và tránh đo phần lòng trắng có màu đục hơn này gần lòng đỏ (phần này sẽ cao hơn những vùng lòng trắng khác), những điều này sẽ làm giảm độ chính xác của đơn vị Haugh. Khi đo lòng trắng, ta tiến hành đo 3 lần lòng trắng xung quanh lòng đỏ, sau đó tính là lấy giá trị trung bình của 3 lần đo.



Hình 5.6: Đo độ cao lòng trắng trứng thủ công.

Sau khi việc thu thập dữ liệu hoàn tất nhóm sẽ có 3 tập dữ liệu tương ứng với 3 vị trí ở giữa(mid), bên trên(top) và bên dưới(bot). Sau đó nhóm tiến hành tính trung bình mid+top, mid+bot, bot+top, top+mid+bot, mục đích để so sánh giá trị trung bình tốt nhất. Cuối cùng tạo thành 7 tập dữ liệu thô với mỗi bộ dữ liệu thô sẽ có 400 hàng 11 cột nhưng do dải bước sóng nhóm chọn nằm trong khoảng từ (700nm - 1100nm) nên nhóm sẽ loại bỏ hai bước sóng đầu 610nm - 680nm nên dữ liệu sẽ còn là 400 hàng và 9 cột.

Hình 5.7 bên dưới mô tả một số tập dữ liệu thô đã thu thập được. Phần còn nhóm sẽ để trong link google drive:

[https://drive.google.com/drive/folders/196WTPwrhDfrkCOKrSdJxNyNNyfNtZH5x?usp=drive\\_link](https://drive.google.com/drive/folders/196WTPwrhDfrkCOKrSdJxNyNNyfNtZH5x?usp=drive_link)

```

      T_730nm      U_760nm      V_810nm      W_860nm  DATE location \
0  13118.49414  15181.95801  30960.63281  30443.13867  2    mid
1  13357.27441  15381.11230  35937.54297  34033.54297  2    mid
2  11890.62988  13640.63770  34465.28516  32531.52148  2    mid
3  13066.00391  15148.92285  36151.74609  33882.97266  2    mid
4  13078.35449  14880.86719  32264.94727  30228.91406  2    mid
..   ...
395 12603.88184  14660.94824  33726.14063  32041.86523  21   mid
396 12830.31152  14752.50293  35243.65234  33316.19531  21   mid
397 12325.99121  14333.42969  36186.94141  34389.76563  21   mid
398 12213.80566  14190.90723  36983.41016  34907.57813  21   mid
399 13868.79883  16108.82715  35134.03906  34106.98828  21   mid

      weight  height      HU
0    62.38   0.70  82.7666
1    61.62   0.80  88.9867
2    60.00   0.80  89.4133
3    60.00   0.90  94.6290
4    61.19   0.96  97.2217
..   ...
395  59.02   0.42  61.1051
396  57.92   0.40  59.5165
397  58.34   0.37  55.8296
398  60.00   0.38  56.0707
399  58.77   0.32  49.0764

```

[400 rows x 9 columns]

Hình 5.7: Tập dữ liệu thô ở vị trí ở giữa quả trứng (mid).

Bảng 5.1 bên dưới mô tả chi tiết về tập dữ liệu thô mid + bot. Phần còn lại nhóm sẽ để link google drive:

[https://docs.google.com/spreadsheets/d/1kQm9XjVKqCXj\\_6JM1cy0V1kT5VUNwsGQ/edit?usp=sharing&ouid=112513283677115077136&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1kQm9XjVKqCXj_6JM1cy0V1kT5VUNwsGQ/edit?usp=sharing&ouid=112513283677115077136&rtpof=true&sd=true)

Bảng 5.1: Mô tả chi tiết tập dữ liệu thô mid+bot.

Đặc trưng	T_730nm	U_760nm	V_810nm	W_860nm	DATE	weight	height	HU
<b>mean</b>	10353.45	12380.91	33091.70	31044.11	11.50	59.26	0.56	72.33
<b>std</b>	725.23	909.31	1932.67	1785.55	5.77	1.62	0.16	11.54
<b>min</b>	8435.52	10249.36	27152.27	25465.79	2.00	54.64	0.30	44.60
<b>25%</b>	9802.08	11687.80	31818.69	29937.88	6.75	58.12	0.45	64.39
<b>50%</b>	10343.71	12350.86	33250.98	31084.59	11.50	59.17	0.50	69.37
<b>75%</b>	10865.26	12979.23	34527.89	32332.60	16.25	60.17	0.66	80.62
<b>max</b>	12239.54	14918.62	37285.10	35863.63	21.00	64.19	0.99	98.85

### 5.2.2. Đánh giá cảm quan trứng gà qua các ngày

Theo quan sát thực tế của nhóm tại những ngày tuổi đầu, chất lượng trứng gà đang ở trạng thái tốt nhất:

- Vỏ trứng trơn, bóng: khi bóc trứng vỏ trứng không bị vỡ thành nhiều mảnh nhỏ, lớp màng sau vỏ dai.
- Lòng trắng trứng có màu trong, khá đặc lớp lòng trắng phủ quanh lòng đỏ dễ dàng nhận biết được.



Hình 5.8: Lòng đỏ và lòng trắng của trứng gà tại ngày số 4.

Từ hình 5.8 có thể quan sát thấy trứng ở ngày số 4 khi trứng còn mới và tươi có lòng trắng trong suốt và lòng đỏ màu vàng. Đặc biệt lớp lòng trắng dày bao bọc lòng đỏ trứng khá bền không dễ bị vỡ khi trứng bị đập vỡ. Bên cạnh đó ở những ngày đầu khi đo chiều cao của lòng trắng trứng và khối lượng nhóm tiến hành tính toán Haugh Units cho kết quả tương đối cao trung bình ngày số 2 HU = 94.32, trung bình ngày số 3 HU = 92.93, trung bình ngày số 4 HU = 87.76. Nếu theo bảng 1.1 tiêu chuẩn quốc gia TCVN 1858:2018 về trứng gà thì ở những ngày đầu tiên trứng đang được xếp loại ở loại AA là trứng còn tươi và mới. Điều này cho thấy rằng ở những ngày đầu tiên trứng vẫn đảm bảo được chất lượng và độ tươi.



Hình 5.9: Lòng đỏ và lòng trắng của trứng gà tại ngày số 9.

Quan sát hình 5.9 cho thấy từ ngày 9 theo quan sát của nhóm đã xuất hiện một vài quả trứng có hiện tượng màng chứa lòng trắng có hiện tượng mỏng đi, một vài quả trứng bị vỡ màng chứa lòng trắng khi bị bô ra. Điều này cũng chứng minh trứng có dấu hiệu giảm về chất lượng và độ tươi theo thời gian. Những quả trứng bị vỡ lòng trắng nhóm sẽ bỏ qua và thay thế bằng một quả trứng khác không bị vỡ lớp lòng trắng dày bên trong việc này để đảm bảo dữ liệu thu được chính xác.



Hình 5.10: Trứng xuất hiện những nấm li ti trên khắp bề mặt vỏ trứng (ngày 16).

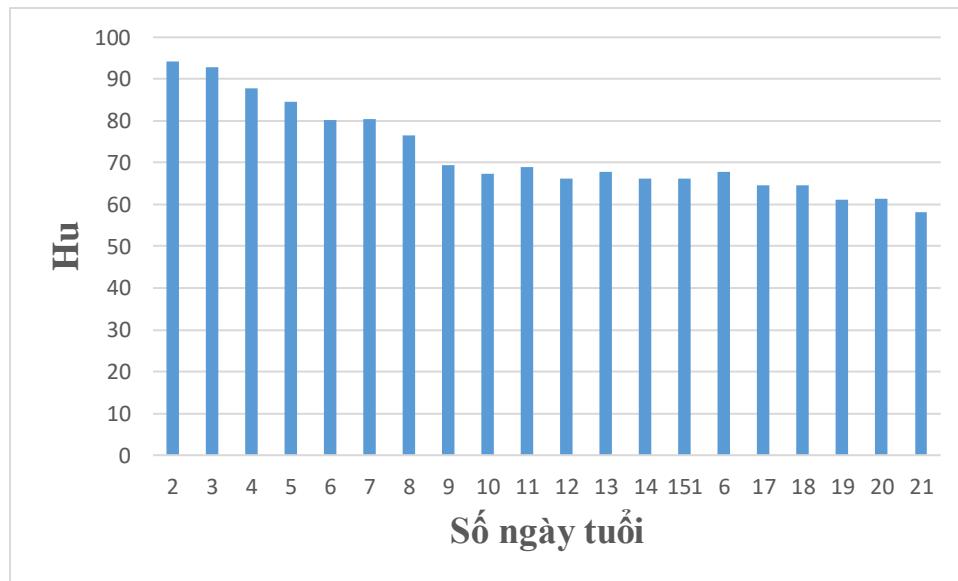
Quan sát hình 5.10 cho thấy ngày 16 vỏ trứng xuất hiện nấm mốc trăng ti li phủ quanh quả trứng. Càng để lâu màu của lòng trắng trứng chuyển từ màu trong suốt sang ngả vàng, vỏ trứng giòn hơn khi để lâu. Càng về sau việc lấy mẫu càng khó khăn việc đập trứng cần sự nhẹ nhàng tỉ mỉ để tránh vỡ lòng trắng.



Hình 5.11: Lòng trắng và lòng đỏ trứng gà có vỏ màu sẫm ngày 18.

Sau khi quan sát một thời gian nhóm nhận thấy rằng những quả trứng có màu sáng hơn vỏ trứng sẽ giòn, dễ vỡ hơn những quả trứng có vỏ màu thẫm. Lớp màng bao lòng trắng của những quả trứng có màu sáng thường mỏng và thường xuyên vỡ màng bao khi bóc trứng. Có vẻ như những quả trứng có màu thẫm sẽ có chất lượng trứng tốt hơn những quả trứng có vỏ sáng màu với cùng thời gian bảo quản.

### 5.2.3. Đánh giá sự thay đổi của trứng gà qua các ngày thông Haugh Units



Hình 5.12: Biểu đồ sự thay đổi trung bình của Haugh Units qua số ngày tuổi của trứng.

Quan sát biểu đồ hình 5.12 có thể thấy từ ngày số 2 đến ngày số 8 Haugh Units  $> 72$  tức trứng đang thuộc loại trứng gà AA là loại trứng gà tốt, tươi vẫn còn sử dụng được. Và ở những ngày đầu tiên thì lòng trắng trứng trong, dai, bao bọc lòng đỏ tốt, khó bị vỡ khi đập trứng điều này được thể hiện ở hình 5.8.

Bắt đầu từ ngày 9 trứng bắt đầu xuất hiện hiện tượng lòng trắng không còn được dai như trước có một số quả đập ra không cẩn thận dẫn đến bị vỡ lòng trắng bao bọc lòng đỏ điều này được thể hiện ở hình 5.9. Và bắt đầu từ ngày 9 có thể quan sát thấy  $HU < 72$  tức chất lượng trứng đã giảm xuống loại A lúc này trứng vẫn còn tốt và tươi, nhưng chất lượng cũng có giảm đáng kể tính từ ngày 2.

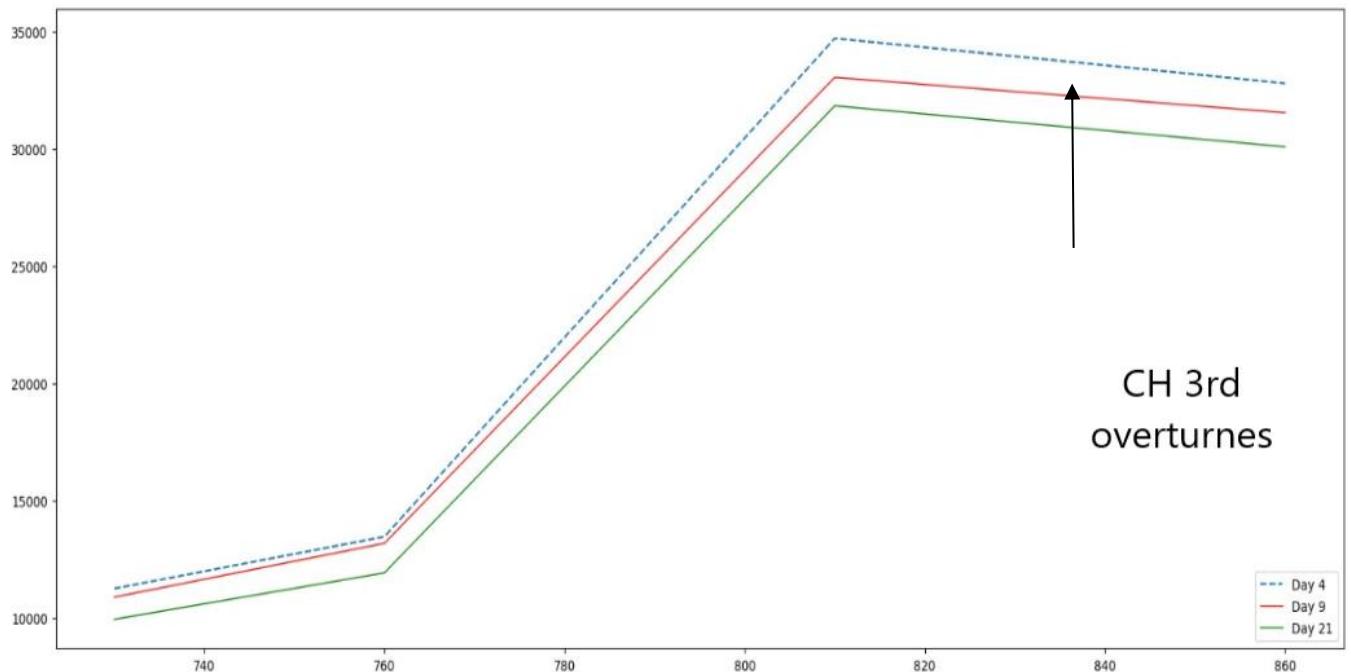
Từ ngày 17 đến ngày 21, phần lớn màng bao lòng trắng đã mỏng hơn rất nhiều dễ bị vỡ khi trứng bị bỗn nêng việc đập trứng cần sự tỉ mỉ, thận trọng, chiều cao lòng trắng thu được rất thấp hơn những ngày trước đó khá nhiều. Điều này có thể thấy được  $HU$  từ ngày 17 đến ngày 20 gần chạm mốc 60 tức mốc thấp nhất của loại A, đến ngày số 21  $HU$  đã giảm xuống dưới 60 tức trứng lúc này đã giảm xuống loại B chất lượng kém hơn ban đầu rất nhiều.

Qua việc quan sát trực quan và đánh giá Hu có thể thấy được chất lượng trứng giảm dần theo thời gian từ trứng loại AA ở ngày 2 xuống loại B ở ngày 21. Qua đây có thể thấy được tầm quan trọng việc dự đoán thời gian bảo quản, Haugh Units để đánh giá chất lượng và độ tươi của trứng gà.

#### 5.2.4. Dải bước sóng NIR qua các ngày

Theo Lammertyn và cộng sự, 2000 [12], ánh sáng trong phạm vi từ 700nm đến 900 nm có thể xuyên qua đến 4 mm, đủ để đi qua vỏ trứng và tiếp cận lớp biểu bì và albumen của trứng mặc dù không có thông tin của lòng đỏ vì không đủ khả năng thâm nhập.

Các dải đáp ứng trong vùng phổ NIR (730nm - 860 nm) chủ yếu là kết quả của các âm bội CH<sub>3</sub> trên tông màu của các liên kết phân tử cơ bản trong vùng hồng ngoại giữa (Mid Infrared Region). Trong hình 5.13, giá trị trung bình của dữ liệu phổ thô được trình bày theo thời gian lưu trữ và các âm bội của các liên kết phân tử cơ bản được đánh dấu. Tại bước sóng 850 nm, các âm bội CH giảm giá trị theo thời gian có thể là hậu quả do sự thay đổi, giảm giá trị thành phần protein của lớp biểu bì và lớp albumen.



Hình 5.13: Sự thay đổi của bước sóng qua các ngày.

Sự giảm giá trị của lớp albumen bao gồm sự biến đổi hóa học của lớp albumen thành những chất khác và tăng lượng nước có trong lòng trắng trứng. Đồng thời lớp vỏ trứng trong quá trình bảo quản lâu dài dần xuất hiện nhiều lỗ nhỏ trên bề mặt, nước từ lòng trắng trứng sẽ thoát ra ngoài môi trường từ từ thông qua những lỗ nhỏ tì li trên bề mặt vỏ trứng. Điều này dẫn tới trứng càng để lâu thành phần của lòng trắng chứa càng nhiều nước (chiều cao lòng trắng giảm), khối lượng quả trứng ngày càng giảm và chất lượng vỏ trứng suy giảm (giòn, dễ vỡ) sau một thời gian bảo quản. Những điều này đã được thể hiện trong quá trình thực nghiệm đo đơn vị Haugh ở mục 5.2.3.

### 5.3. Tiền xử lý dữ liệu

Hình 5.14 bên dưới là chương trình cho hai phương pháp: Standard Normal Variate(SNV) và Multiplicative Scatter Correction(MSC).

```
[3] def msc(input_data, reference=None):
    # mean centre correction
    for i in range(input_data.shape[0]):
        input_data[i,:] -= input_data[i,:].mean()

    # Get the reference spectrum. If not given, estimate it from the mean
    if reference is None:
        # Calculate mean
        ref = np.mean(input_data, axis=0)
    else:
        ref = reference

    # Define a new array and populate it with the corrected data
    data_msc = np.zeros_like(input_data)
    for i in range(input_data.shape[0]):
        # Run regression
        fit = np.polyfit(ref, input_data[i,:], 1, full=True)
        # Apply correction
        data_msc[i,:] = (input_data[i,:] - fit[0][1]) / fit[0][0]

    return (data_msc, ref)

[4] def snv(input_data):

    # Define a new array and populate it with the corrected data
    output_data = np.zeros_like(input_data)
    for i in range(input_data.shape[0]):

        # Apply correction
        output_data[i,:] = (input_data[i,:] - np.mean(input_data[i,:])) / np.std(input_data[i,:])

    return output_data
```

Hình 5.14: chương trình của 2 phương pháp tiền xử lý Standard Normal Variate\_SNV và Multiplicative Scatter Correction\_MSC.

Hình 5.15 bên dưới mô tả bộ dữ liệu thô mid + bot sau khi trải qua phương pháp tiền xử lý SNV, dữ liệu được xử lý ở bộ tập dữ liệu thô gồm các giá trị ở 4 đặc trưng ứng với 4 cột bước sóng.

---

```
[[ -1.13467781  1.07284465 -0.85214598  0.91397913]
 [ -1.11345086  1.13833475 -0.86804453  0.84316064]
 [ -1.12661158  1.0783512  -0.86095187  0.90921225]
 ...
 [-1.09425688  1.07943728 -0.89738499  0.91220459]
 [-1.10353501  1.04900043 -0.88933557  0.94387014]
 [-1.10287475  1.07169325 -0.88847084  0.91965234]]
```

Hình 5.15: SNV với tập dữ liệu thô mid + bot.

Hình 5.16 bên dưới mô tả tập dữ liệu mid + bot sau khi chạy phương pháp tiền xử lý MSC, dữ liệu trong tập dữ liệu thô sẽ được xử lý gồm các giá trị ở 4 đặc trưng ứng với 4 cột bước sóng.

```
[[ -11807.82737121  11164.37139265  -8867.70902169  9511.16500025]
 [-11589.2700517  11848.27220307  -9034.97660806  8775.97445669]
 [-11721.42762618  11219.31979652  -8957.46618006  9459.57400972]
 ...
 [-11381.31745076  11227.17938137  -9333.66154967  9487.79961906]
 [-11483.49979773  10916.00734351  -9254.51819785  9822.01065207]
 [-11472.03082285  11147.68286483  -9241.81539284  9566.16335085]]
```

Hình 5.16: MSC với tập dữ liệu thô mid + bot.

Hình 5.17 bên dưới mô tả tập dữ liệu mid + bot sau khi chạy phương pháp tiền xử lý standard scaler, những dữ liệu sẽ được xử lý gồm các giá trị ở 4 đặc trưng ứng với 4 cột bước sóng.

2.079911753612824	1.8366396578124633	1.3653557650763528	-0.25574393614172786
1.659749863937633	1.4320713548932982	1.1333904884455293	-0.34237184748982324
1.7194977017263977	1.485107992328748	1.3149743095065094	-0.26935790908117074
1.613493488139322	1.2826047973978678	0.9814582266712497	-1.0551813054719985
1.67179571417567	1.3505441004861405	1.069364076560294	-0.8633654616545948
1.924278527321183	1.6358897176642682	1.4396166586500059	-0.08434806449832245
2.0461830986366896	1.7643172177674782	1.1533335089330885	-0.39001581660204326
2.135322756144339	1.8646921878415756	1.1158091146138873	-0.7668395305319632
2.2861377337773092	1.9308782361471635	1.1126605921623862	-0.9537053772269328
2.0061905832494666	1.767823727332848	1.7552894154781506	0.48676550457505224

Hình 5.17: Dữ liệu thô sau khi tiến xử lý standard scaler mid+bot.

Dữ liệu được đảm bảo tính đúng đắn với ý nghĩa của thuật toán bằng cách kiểm tra giá trị trung bình cũng như độ lệch chuẩn sau khi xử lý, trong đó:

- Giá trị trung bình  $\approx 0$ .
- Giá trị độ lệch  $\approx 1$ .

**mean** 0.000000 -1.421085e-16 1.894781e-16 -7.223851e-16

**std** 1.000834 1.000834e+00 1.000834e+00 1.000834e+00

Hình 5.18: Kiểm tra giá trị trung bình và độ lệch chuẩn.

Sau khi dữ liệu được xử lý, giá trị trung bình ở các cột gần bằng 0 và độ lệch chuẩn gần bằng 1, điều này được thể hiện ở hình 5.18 đảm bảo tính đúng đắn với thuật toán.

Sau khi tiến hành tiền xử lý bảy tập dữ liệu thô hoàn tất khi sử dụng ba phương pháp tiền xử lý: standard scaler, multiplicative scatter correction (MSC) và standard normal variate (SNV), nhóm đã thu được 21 bộ dữ liệu cho việc huấn luyện để so sánh và đánh giá ba mô hình học máy: MLR, SVR và Decision Tree. Vì mô hình nhóm hướng đến dự đoán hai đặc trưng ngày tuổi của trứng và Haugh Units, hai đặc trưng khác nhau nên việc huấn luyện mô hình cùng lúc hai đặc trưng này sẽ không mô tả hết được những đặc điểm riêng của ngày tuổi của trứng và Haugh Units liên quan đến bước sóng. Nhóm sẽ tiến hành tách riêng số ngày tuổi của trứng và Haugh Units để training trên các mô hình khác nhau. Việc này sẽ khiến mô hình có thể bao quát tốt hơn liên hệ giữa bước sóng NIR thu được với số ngày tuổi của trứng và Haugh Units.

#### 5.4. Thực nghiệm mô hình Multiple Linear Regression

Việc training mô hình Multiple Linear Regression được nhóm thực hiện trên Google colab. Trước khi bước vào training mô hình nhóm tiến hành phân chia 21 tập dữ liệu tiền xử lý với mỗi tập dữ liệu chia thành 2 tập nhỏ ngẫu nhiên bao gồm: Tập huấn luyện (80% dữ liệu) và tập kiểm định (20% dữ liệu). Hình 5.19 bên dưới thể hiện việc phân chia dữ liệu. Trong đó:

- 80% tập huấn luyện được chia thành 2 phần là:
  - x\_train: đây là giá trị của 4 đặc trưng (tương ứng với 4 cột bước sóng mà cảm biến hỗ trợ), dùng để huấn luyện.
  - y\_train\_: đây là giá trị của 1 đặc trưng tương ứng với thời gian bảo quản trứng hoặc Haugh Units dùng để huấn luyện.
- 20% tập kiểm định được chia thành 2 phần là:
  - x\_test\_: đây là giá trị của 4 đặc trưng (tương ứng với 4 cột bước sóng mà cảm biến hỗ trợ), dùng để kiểm định.
  - y\_test: đây là giá trị của 1 đặc trưng tương ứng với thời gian bảo quản trứng hoặc Haugh Units dùng để kiểm định.

```
[▶] from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_std,y_original,test_size = 0.2)
print(x_test)
print(y_test)
```

Hình 5.19: Chia tập dữ liệu tiền xử lý thành tập huấn luyện và tập kiểm định.

Khi việc phân chia tập dữ liệu xong bắt đầu tiến hành huấn luyện mô hình và đánh giá kết quả trên 21 tập dữ liệu tiền xử lý. Hình 5.20 bên dưới thể hiện việc huấn luyện và đánh giá 1 tập dữ liệu trong 21 tập dữ liệu với mô hình Multiple Linear Regression.

## Training model

```
✓ 0 [137] from sklearn.linear_model import LinearRegression  
giây      ml = LinearRegression()  
      ml.fit(x_train,y_train)  
      print(ml.coef_)  
      print(ml.intercept_)
```

```
➡ [ 1.62108401  2.63864652  5.69545716 -5.248466  ]  
72.32243734202353
```

```
✓ 0 [138] y_pred = ml.predict(x_test)  
giây
```

```
✓ 0 [139] from sklearn.metrics import r2_score  
giây      r2_score(y_test, y_pred)
```

```
0.3790524040226628
```

```
✓ 0 [140] mse = mean_squared_error(y_test,y_pred)  
giây      Rmse=math.sqrt(mse)  
      print(Rmse)
```

```
8.858600104497338
```

Hình 5.20: Kết quả train mô hình Multiple Linear Regression.

Bằng ba phương pháp tiền xử lý và một mô hình máy học, nhóm đã có 21 đánh giá  $R^2$  và Rmse đã được tính toán cho mô hình được thể hiện ở 3 bảng 5.2, 5.3, 5.4 bên dưới:

Bảng 5.2: Kết quả mô hình học máy Multiple Linear Regression khi sử dụng Standard Scaler.

Standar Scaler	R <sup>2</sup> _Hu	R <sup>2</sup> _Date	RMSE_Hu	RMSE_Date(ngày)
Mid	31.56%	29.85%	9.06	4.59
Top	39.86%	40.01%	8.13	4.63
Bot	<b>43.17%</b>	<b>48.85%</b>	<b>13.09</b>	<b>6.45</b>
Mid top	22.42%	24.95%	10.85	5.09
Mid bot	37.90%	44.86%	8.85	4.93
Bot top	41.31%	33.15%	8.75	4.80
Mid top bot	40.94%	34.41%	8.86	4.62

Bảng 5.3: Kết quả mô hình học máy Multiple Linear Regression khi sử dụng MSC.

SNV	R <sup>2</sup> _Hu	R <sup>2</sup> _Date	RMSE_Hu	RMSE_Date(ngày)
Mid	22.07%	18.57%	10.02	5.29
Top	44.93%	<b>41.82%</b>	8.93	<b>3.98</b>
Bot	<b>47.66%</b>	29.48%	<b>8.83</b>	4.65
Mid top	22.83%	28.22%	10.18	5.17
Mid bot	39.65%	25.67%	8.03	5.01
Bot top	41.57%	38.52%	9.47	4.43
Mid top bot	37.48%	28.80%	8.64	4.57

Bảng 5.4: Kết quả mô hình học máy Multiple Linear Regression khi sử dụng SNV.

MSC	R <sup>2</sup> _Hu	R <sup>2</sup> _Date	RMSE_Hu	RMSE_Date(ngày)
Mid	24.58%	17.01%	8.93	5.47
Top	36.51%	<b>38.05%</b>	7.78	<b>4.47</b>
Bot	38.07%	34.91%	9.08	4.86
Mid top	23.69%	21.50%	10.18	4.94
Mid bot	41.33%	28.10%	8.59	5.08
Bot top	<b>43.46%</b>	36.51%	<b>9.70</b>	4.63
Mid top bot	34.66%	31.24%	9.24	5.14

Với phương pháp tiền xử lý Standard Scaler và mô hình học máy Multiple Linear Regression kết quả được thể hiện bảng 5.2 thấy rằng hầu hết các sai số dự đoán trung bình khoảng hơn 5.01 ngày, HU khoảng 9.65, sai số dự đoán thấp nhất là 4.59 ngày, HU là 8.13, sai số dự đoán cao nhất lên đến 6.45 ngày, HU là 13.09. Đối với phương pháp tiền xử lý MSC ở bảng 5.3 hầu như các sai số dự đoán trung bình khoảng 4.72 ngày, HU khoảng 9.15, sai số dự đoán thấp nhất 3.98 ngày, HU là 8.03, sai số dự đoán lớn nhất là 5.29 ngày, HU là 10.18. Đối với phương pháp tiền xử lý SNV ở bảng 5.4 hầu như các sai số dự đoán trung bình khoảng 4.79 ngày, HU khoảng 9.07, sai số dự đoán thấp nhất 4.47 ngày, HU là 7.78, sai số dự đoán lớn nhất là 5.47 ngày, HU là 10.18.

RMSE là một chỉ tiêu sử dụng để đánh giá độ chính xác của mô hình với so với dữ liệu thực tế, giá trị của RMSE thể hiện sự khác biệt trung bình giữa giá trị dự báo và giá trị thực tế, giá trị RMSE càng thấp, mô hình dự báo càng chính xác. Từ đây cũng có thể thấy rằng Multiple Linear Regression được training bằng tập dữ liệu sử dụng phương pháp tiền xử lý MSC có độ chính xác cao nhất, còn hai phương pháp Standard Scaler, SNV độ chính xác khá thấp.

Nhưng nhìn lại một cách tổng quan các kết quả từ các phương pháp có được thấy rằng trung bình sai số dự đoán (RMSE) thấp nhất về dự đoán ngày tuổi của trứng cũng chỉ khoảng 3.98(ngày) và HU là 7.78, RMSE cao nhất lên đến 6.45 ngày và Hu là 13.09. Kết quả này nhóm nhận thấy rằng chưa khả quan vẫn cần cải thiện và đánh giá thêm.

Bảng 5.5: So sánh giữa các giá trị  $R^2$ .

Standar Scaler	$R^2$ _Hu	$R^2$ _Date	MSC	$R^2$ _Hu	$R^2$ _Date	SNV	$R^2$ _Hu	$R^2$ _Date
Mid	31.56%	29.85%	Mid	22.07%	18.57%	Mid	24.58%	17.01%
Top	39.86%	40.01%	Top	44.93%	41.82%	Top	36.51%	38.05%
Bot	43.17%	<b>48.85%</b>	Bot	<b>47.66%</b>	29.48%	Bot	38.07%	34.91%
Mid top	22.42%	24.95%	Mid top	22.83%	28.22%	Mid top	23.69%	21.50%
Mid bot	37.90%	44.86%	Mid bot	39.65%	25.67%	Mid bot	41.33%	28.10%
Bot top	41.31%	33.15%	Bot top	41.57%	38.52%	Bot top	43.46%	36.51%
Mid top bot	40.94%	34.41%	Mid top bot	37.48%	28.80%	Mid top bot	34.66%	31.24%
<b>Max</b>	43.17%	<b>48.85%</b>	<b>Max</b>	<b>47.66%</b>	41.82%	<b>Max</b>	43.46%	38.05%

Nhìn vào bảng 5.5 bên trên ta có thể thấy giá trị  $R^2$  khi áp dụng 3 kỹ thuật tiền xử lý là Standard Scaler, SNV và MSC trên bảy tập dữ liệu khác nhau. Với mô hình học máy dành cho “ngày bảo quản trứng”,  $R$  bình phương sẽ đạt giá trị tốt nhất là 48.85% khi nhóm sử dụng kỹ thuật tiền xử lý Standard Scaler với tập dữ liệu Bot.

Với mô hình học máy cho Haugh Units,  $R$  bình phương cao nhất là 47.66% khi nhóm sử dụng kỹ thuật tiền xử lý MSC với tập dữ liệu Bot. Như vậy việc sử dụng tập dữ liệu Bot để làm dữ liệu đầu vào cho bài toán học máy là tập dữ liệu phù hợp nhất với mô hình học máy.

$R$  bình phương được dùng trong mô hình hồi quy tuyến tính để thể hiện mức độ phù hợp của mô hình đối với các biến đầu vào. Từ giá trị  $R^2$  thấy được các kết quả thu được rất thấp, điều này chứng minh các đầu vào nhóm thu được có mức độ phù hợp với Multiple

Linear Regression không được cao nên nhóm vẫn cần phải tìm hiểu thêm mô hình hồi quy phù hợp hơn.

Mặc dù phương pháp đạt được chưa được khả quan nhưng để phải chọn ra mô hình sử dụng đối với việc dự đoán ngày tuổi của trứng thì tập dữ liệu top với phương pháp tiền xử lý MSC với  $R^2 = 41,82\%$ , RMSE = 3.98 (ngày) có thể áp dụng, còn với tập dữ liệu bot với kỹ thuật tiền xử lý Standard Scaler mặc dù được  $R^2 = 48.85\%$  nhưng RMSE lại khá cao lên đến 6.45 ngày. Đối với dự đoán Haugh Units thì tập dữ liệu ở vị trí bot với phương pháp tiền xử lý MSC có được kết quả tốt nhất  $R^2 = 47.66\%$ , RMSE = 8.83. Từ đây có thể thấy được phương pháp tiền xử lý Multiplicative Scatter Correction\_MSC là phương pháp tốt nhất trong 3 phương pháp nhóm đã đề cập. Vị trí đặt trứng tốt nhất để dự đoán ngày tuổi của trứng là tại điểm bên trên (top), để dự đoán Haugh Units là điểm bên dưới (bot).

### 5.5. Thực hiện mô hình Decision Tree

Tương tự Multiple Linear Regression trước khi bước vào training mô hình Decision Tree nhóm tiến hành phân chia 21 tập dữ liệu tiền xử lý với mỗi tập dữ liệu chia thành hai tập nhỏ ngẫu nhiên bao gồm: tập huấn luyện (80% dữ liệu) và tập kiểm định (20% dữ liệu). Khi việc phân chia tập dữ liệu xong bắt đầu tiến hành huấn luyện mô hình và đánh giá kết quả trên 21 tập dữ liệu tiền xử lý. Hình 5.21 thể hiện việc huấn luyện và đánh giá một tập dữ liệu trong 21 tập dữ liệu với mô hình Decision Tree. Nhưng đối với mô hình này chúng ta cần quan tâm đến một thông số `max_depth` là kích thước của cây hay còn gọi là số lớp (layers) hay độ sâu (depth). Nếu `max_depth` quá nông có thể dẫn đến underfitting, vì mô hình chỉ học được rất ít chi tiết từ dữ liệu. Ngược lại, cây quyết định quá sâu (deep) thì mô hình lại học quá nhiều chi tiết từ dữ liệu sẽ dẫn đến overfitting. Vì vậy trong quá trình training nhóm thử các giá giá `max_depth = 4,5,10,20,25` để có thể tìm được độ sâu mô hình cần thiết để có kết quả tốt nhất.

```

from sklearn import tree
clf = tree.DecisionTreeRegressor(random_state=42,max_depth=4)
clf = clf.fit(X_train, y_train)
prd=clf.predict(X_test)
from sklearn.metrics import r2_score
c=r2_score(y_test, prd)
mse = mean_squared_error(y_test,prd)
Rmse=math.sqrt(mse)
print(c)
print(Rmse)

```

0.3689796345157289

9.389255299223626

Hình 5.21: Kết quả train mô hình Decision Tree bằng Google Colab.

Bằng ba phương pháp tiền xử lý và một mô hình máy học, nhóm đã có 21 đánh giá  $R^2$  và RMSE đã được tính toán cho mô hình Decision Tree được thể hiện ở 3 bảng 5.6, 5.7, 5.8 bên dưới:

Bảng 5.6: Kết quả mô hình học máy Decision Tree khi sử dụng Standard Scaler.

Decision Tree và Standard Scaler						
Standard Scaler	$R^2$ _Hu	RMSE_Hu	Maxdepth_Hu	$R^2$ _Date	RMSE_Date (ngày)	Max_depth_Date
Mid	5.98%	12.29	5	20.01%	5.23	5
Top	44.05%	8.41	4	46.25%	4.56	4
Bot	<b>53.92%</b>	<b>7.92</b>	<b>4</b>	<b>57.74%</b>	<b>3.81</b>	<b>4</b>
Mid top	34.93%	10.25	4	42.38%	4.28	4
Mid bot	33.24%	10.40	4	30.93%	4.55	4
Bot top	46.01%	9.74	4	30.04%	4.74	4
Mid top bot	43.84%	8.24	5	41.01%	4.81	5

Bảng 5.7: Kết quả mô hình học máy Decision Tree khi sử dụng SNV.

Decision Tree và SNV						
SNV	R <sup>2</sup> _Hu	RMSE_Hu	Maxdepth_Hu	R <sup>2</sup> _Date	RMSE_Date (ngày)	Max_depth_Date
Mid	25.71%	10.58	4	26.07%	5.06	4
Top	<b>55,76%</b>	<b>8.99</b>	<b>4</b>	53.08%	3.97	4
Bot	39.27%	8.93	4	44.39%	4.44	4
Mid top	48.35%	8.73	4	39.43%	4.70	4
Mid bot	36.60%	10.07	4	40.30%	4.55	4
Bot top	48.84%	8.83	4	<b>59.23%</b>	<b>4.10</b>	<b>4</b>
Mid top bot	44.85%	9.10	4	54.07%	3.90	4

Bảng 5.8: Kết quả mô hình học máy Decision Tree khi sử dụng MSC.

Decision Tree và MSC						
MSC	R <sup>2</sup> _Hu	RMSE_Hu	Maxdepth_Hu	R <sup>2</sup> _Date	RMSE_Date (ngày)	Max_depth_Date
Mid	31.73%	9.08	4	21.07%	5.02	4
Top	45.40%	8.62	4	<b>58.57%</b>	<b>3.93</b>	<b>4</b>
Bot	40.66%	9.04	4	49.53%	4.15	4
Mid top	37.89%	8.52	4	32.39%	5.17	4
Mid bot	32.60%	9.75	4	37.62%	4.83	4
Bot top	46.93%	8.71	4	47.15%	4.11	4
Mid top bot	<b>53.17%</b>	<b>7.50</b>	<b>4</b>	51.03%	4.13	4

Thông qua ba bảng giá trị đánh giá với phương pháp Decision Tree và ba phương pháp tiền xử nhóp thấy được kết quả dự đoán Haugh Units có R<sup>2</sup> tốt nhất là 53.17% và RMSE là 7.50 với tập dữ liệu Mid+Top+Bot còn với tập dữ liệu top mặc dù có R<sup>2</sup>=55.76% nhưng kết quả RMSE = 8.99 là tương đối cao. So sánh với kết quả dự đoán Haugh Units

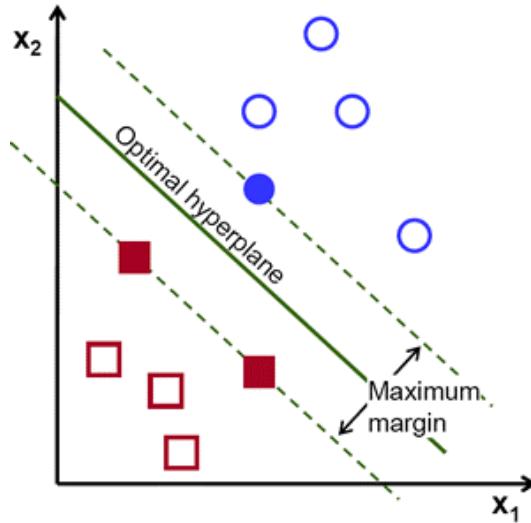
cao nhất có được từ phương pháp Multiple Linear Regression cho thấy kết quả mô hình Decision Tree cao hơn một chút với  $R^2$  là khoảng  $52.17\% - 47.66\% = 4.51\%$ , còn RMSE là khoảng  $8.83 - 7.50 = 1.33$ .

Còn với kết quả trong việc dự đoán ngày tuổi của trứng  $R^2$  tốt nhất là 58.37% và RMSE là 3.93 với tập dữ liệu Top còn với tập dữ liệu Bot+Top mặc dù có  $R^2=59.23\%$  nhưng kết quả RMSE = 4.10 là tương đối cao. Kết quả này nhóm mang so sánh với kết quả đánh giá trong việc dự đoán số ngày tuổi của trứng cao nhất có được từ phương pháp Multiple Linear Regression cho thấy kết quả mô hình Decision Tree cao hơn một chút với  $R^2$  là khoảng  $58.37\% - 41.82\% = 16.55\%$ , còn RMSE là khoảng  $3.98 - 3.93 = 0.05$ .

Qua đây cũng có thể thấy rằng mặc dù kết quả vẫn chưa đạt mong đợi nhưng có thể thấy được phương pháp học máy Decision Tree có được kết quả cao hơn. Và đầu vào nhóm thu được phù hợp với Decision Tree hơn so với mô hình Multiple Linear Regression. Bên cạnh đó qua ba bảng kết quả cho thấy với sự thử nghiệm  $\text{max\_depth} = 4, 5, 10, 20, 25$  hay chiều sâu của cây nhóm nhận thấy rằng với đầu vào này  $\text{max\_depth} = 4$  là đạt hiệu quả tốt nhất.

### **5.6. Thực hiện mô hình Support Vector Regression**

Tương tự Multiple Linear Regression trước khi bước vào huấn luyện mô hình Support Vector Regression, nhóm tiến hành phân chia 21 tập dữ liệu đã được trải qua bước tiền xử lý, mỗi tập dữ liệu được chia thành hai tập ngẫu nhiên bao gồm: Tập huấn luyện (80% dữ liệu) và tập kiểm định (20% dữ liệu). Sau khi phân chia xong hai tập dữ liệu, nhóm tiến hành huấn luyện mô hình và đánh giá kết quả của 21 tập dữ liệu. Hình 5.23 thể hiện việc huấn luyện và đánh giá một tập dữ liệu trong 21 tập dữ liệu với mô hình Support Vector Regression. Nhưng với mô hình này nhóm quan tâm đến hai thông số C và epsilon.



Hình 5.22: Ví dụ về Margin trong SVM.

Margin là khoảng cách nằm giữa từ Hyperplane đến hai điểm dữ liệu nằm gần Hyperplane nhất của các tập hợp dữ liệu khác nhau. Cách để SVM tối ưu thuật toán là SVM sẽ cố gắng tìm giá trị margin lớn nhất, từ đó SVM sẽ tìm ra Hyperplane tốt nhất để phân tách 2 lớp dữ liệu. Để tránh overfitting và có margin cao, ta có thể chấp nhận việc một số dữ liệu không được chia chính xác. Những dữ liệu không được chia chính xác này sẽ được gọi là nhiễu và margin trong trường hợp này là Soft Margin. Ngược lại, Hard Margin là thuật ngữ được gọi khi Margin không chứa những dữ liệu không được chia chính xác hoặc mô hình không có chứa những dữ liệu không được chia chính xác này. Với các bài toán thực tế, việc tìm được Hard Margin đôi lúc là điều không thể thực hiện được, vì thế việc chấp nhận mức độ sai lệch ở một nhường có thể chấp nhận được là một điều cần thiết.

Trong quá trình thực hiện SVM, cần quan tâm tới biến  $C$  với những quy ước sau:

- $C = \infty$ : Hard Margin (không có sai lệch xuất hiện).
- $c$ : cho phép mức độ sai lệch một phần nhỏ, đạt được Margin nhỏ.
- $C$ : cho phép mức độ sai lệch có thể lớn, đạt được Margin lớn.

Như vậy, tùy vào bài toán để ta điều chỉnh biến  $C$  cho việc thử nghiệm, từ đó được kết quả tốt nhất. Vì vậy trong khi thực hiện mô hình Support Vector Regression nhóm sẽ thử các tham số  $C = 1, 10, 100, 1000$  để tìm ra kết quả tốt nhất.

Giá trị của epsilon xác định mức độ chính xác của hàm gần đúng. Nó hoàn toàn dựa vào các giá trị đích trong tập huấn luyện. Nếu epsilon lớn hơn phạm vi giá trị mục tiêu thì kết quả thu được có thể không tốt. Việc chọn epsilon cũng chỉ có một độ chính xác nhất định và chỉ đảm bảo độ chính xác đó trên tập huấn luyện, thông thường để đạt được độ chính xác nhất định về tổng thể, chúng ta cần chọn epsilon nhỏ. Vì vậy bên cạnh việc lựa chọn thêm tham số C nhóm còn lựa chọn thêm Epsilon = 1, 0.1, 0.01, 0.001 để kết quả của mô hình có thể tốt hơn.

```
regr = svm.SVR(C=1000.0,epsilon=0.001)
regr.fit(X_train, y_train)
y_predict = regr.predict(X_test)
from sklearn.metrics import r2_score
b=r2_score(y_test, y_predict)
mse = mean_squared_error(y_test,y_predict)
Rmse=math.sqrt(mse)
print(b)
print(Rmse)
```

-0.1885771381335224

6.477275780720846

Hình 5.23: Kết quả huấn luyện mô hình Support Vector Regression bằng Google Colab.

Bằng ba phương pháp tiền xử lý và một mô hình máy học, nhóm đã có 21 nhóm kết quả để so sánh và đánh giá về  $R^2$  và RMSE đã được tính toán cho mô hình Support Vector Regression được thể hiện ở 3 bảng 5.9, 5.10, 5.11 bên dưới:

Bảng 5.9: Kết quả mô hình Support Vector Regression khi sử dụng Standard Scaler.

Support Vector Regression và Standard Scaler						
Standar Scaler	R <sup>2</sup> _Hu	RMSE_Hu	C, epsilon_Hu	R <sup>2</sup> _Date	RMSE_Date (ngày)	C, epsilon_Date
Mid	25.54%	10.89	C=10,epsilon=1	33.43%	4.51	C=10,epsilon=1
Top	58.41%	7.69	C=50,epsilon=1	<b>64.51%</b>	<b>3.52</b>	<b>C=10,epsilon=1</b>
Bot	<b>67.14%</b>	<b>6.92</b>	<b>C=10,epsilon=0.1</b>	64.70%	3.72	C=10,epsilon=0.1
Mid top	52.17%	8.17	C=100,epsilon=0.1	61.01%	3.57	C=10,epsilon=1
Mid bot	40.32%	9.08	C=10,epsilon=1	45.49%	4.48	C=10,epsilon=1
Bot top	54.36%	7.45	C=10,epsilon=1	61.38%	3.62	C=100,epsilon=0.01
Mid top bot	48.46%	9.09	C=100,epsilon=0.01	52.92%	3.87	C=100,epsilon=0.01

Bảng 5.10: Kết quả mô hình Support Vector Regression khi sử dụng SNV.

Support Vector Regression và SNV						
SNV	R <sup>2</sup> _Hu	RMSE_Hu	C, epsilon_Hu	R <sup>2</sup> _Date	RMSE_Date (ngày)	C, epsilon_Date
Mid	23.90%	9.99	C=1000,epsilon=0.001	26.60%	4.96	C=1000,epsilon=0.001
Top	36.77%	8.85	C=1000,epsilon=0.001	<b>38.35%</b>	<b>5.02</b>	<b>C=1000,epsilon=0.001</b>
Bot	<b>44.51%</b>	<b>9.48</b>	<b>C=1000,epsilon=0.001</b>	31.57%	5.04	C=1000,epsilon=0.001
Mid top	36.76%	9.25	C=1000,epsilon=0.001	31.67%	4.92	C=1000,epsilon=0.001
Mid bot	14.29%	10.43	C=1000,epsilon=0.001	30.08%	4.83	C=1000,epsilon=0.001
Bot top	20.34%	9.13	C=1000,epsilon=0.001	30.90%	4.89	C=1000,epsilon=0.001
Mid top bot	22.05%	9.97	C=1000,epsilon=0.001	25.17%	5.21	C=1000,epsilon=0.001

Bảng 5.11: Kết quả mô hình Support Vector Regression khi sử dụng MSC.

Support Vector Regression và MSC						
MSC	R <sup>2</sup> _Hu	RMSE_Hu	C, epsilon_Hu	R <sup>2</sup> _Date	RMSE_Date (ngày)	C, epsilon_Date
Mid	24.57%	9.42	C=1000,epsilon=0.001	19.14%	5.20	C=1000,epsilon=0.001
Top	30.12%	9.8	C=1000,epsilon=0.001	<b>36.70%</b>	<b>4.55</b>	<b>C=1000,epsilon=0.001</b>
Bot	35.04%	9.13	C=1000,epsilon=0.001	31.12%	4.82	C=1000,epsilon=0.001
Mid top	32.93%	9.92	C=1000,epsilon=0.001	32.74%	4.76	C=1000,epsilon=0.001
Mid bot	13.43%	10.30	C=1000,epsilon=0.001	30.83%	5.04	C=1000,epsilon=0.001
Bot top	<b>35.70%</b>	<b>9.18</b>	<b>C=1000,epsilon=0.001</b>	35.68%	4.66	C=1000,epsilon=0.001
Mid top bot	30.64%	9.66	C=1000,epsilon=0.001	31.70%	4.53	C=1000,epsilon=0.001

Thông qua ba bảng kết quả với mô hình Support Vector Regression và ba phương pháp tiền xử, trong việc dự đoán giá trị Haugh Units đã đạt được R<sup>2</sup> tốt nhất là 67.14% và RMSE là 6.92 khi sử dụng tập dữ liệu Bot để huấn luyện mô hình. Sau khi so sánh những kết quả dự đoán Haugh Units cao nhất khi sử dụng phương pháp Multiple Linear Regression, kết quả cho thấy mô hình Support Vector Regression cao hơn rất nhiều với R<sup>2</sup> là khoảng 67.14% - 47.66% = 19.48%, còn RMSE là khoảng 8.83 - 6.92 = 1.91 còn với mô hình Decision Tree thì cao hơn một ít với R<sup>2</sup> là khoảng 67.14% - 52.17% = 14.97%, còn RMSE là khoảng 7.5 - 6.92 = 0.58.

Còn với kết quả trong việc dự đoán ngày tuổi của trứng R<sup>2</sup> tốt nhất là 64.51% và RMSE là 3.52 với tập dữ liệu Top còn với tập dữ liệu Mid mặc dù có R<sup>2</sup>=64.70% nhưng kết quả RMSE = 3.72 là tương đối cao. So sánh với kết quả dự đoán số ngày tuổi của trứng cao nhất có được từ phương pháp Multiple Linear Regression cho thấy kết quả mô hình Support Vector Regression cao hơn rất nhiều với R<sup>2</sup> là khoảng 64.51% - 41.82% = 25.59%, còn

RMSE là khoảng  $3.98 - 3.52 = 0.46$  còn với mô hình Decision Tree thì cao hơn một ít với  $R^2$  là khoảng  $64.51\% - 58.37\% = 6.14\%$ , còn RMSE là khoảng  $3.93 - 3.52 = 0.41$ .

Qua mô hình Support Vector Regression cũng có thể thấy rằng mặc dù kết quả vẫn chưa đạt mong đợi nhưng có thể thấy được mô hình học máy Support Vector Regression đưa ra kết quả mô phỏng tốt hơn khi được so sánh với hai mô hình Multiple Linear Regression và mô hình Decision Tree. Qua đó, nhóm nhận thấy rằng  $C = 10$  và  $\text{epsilon} = 1$  là giá trị lựa chọn tốt nhất với mô hình.

Dựa vào thu thập và so sánh các kết quả đạt được từ ba phương pháp tiền xử lý và ba phương pháp máy học, kết quả đánh thu được đã cho thấy mô hình Support Vector Regression và phương pháp tiền xử lý Standard Scaler cho được kết quả tốt nhất với việc dự đoán HU là tập dữ liệu Bot và dự đoán ngày tuổi của trứng với tập dữ liệu Top. Sau khi so sánh và đánh giá, nhóm quyết định sử dụng hai phương pháp: tiền xử lý Standard Scaler và máy học Support Vector Regression để giải quyết vấn đề cho đẻ tài.

Kết quả  $R^2$  và RMSE nhóm thu được sau khi tiến hành mô phỏng được so sánh với kết quả của những bài báo khác với phương pháp tương tự ở bảng 1.1 cho thấy với việc dự đoán ngày kết quả các tác giả khác họ đạt được khá cao với RMSE = 1.96 ngày và  $R^2 = \sim 83\%$ , còn dự đoán HU cao nhất là RMSE = 3.31 và  $R^2 = \sim 93\%$  cao hơn nhiều so với kết quả của nhóm đạt được là dự đoán ngày  $R^2 = 64.51, 82\%$ , RMSE = 3.52(ngày), dự đoán Haugh Units với  $R^2 = 67.14\%$ , RMSE = 7.5. Nhưng chi phí cho thiết bị và phát triển đẻ tài của nhóm lại rẻ hơn rất nhiều so với các đẻ tài ở bảng 1.1 giá của cảm biến AS7263 nhóm sử dụng chỉ với 27.95\$ còn các thiết bị ở các bài báo trên lên đến gần 1000\$ thậm chí đến hơn 40000\$. Với chi phí thiết bị đắt đỏ này sẽ rất khó để các tác giả phát hành và ứng dụng rộng rãi thiết bị trên thị trường, việc sử dụng thiết bị giá rẻ như nhóm đang phát triển mang lại tiềm năng trong tương lai về một thiết bị có thể đánh giá được chất lượng và độ tươi của trứng với giá cả phù hợp hơn với người dùng.

## 5.7. Lưu và tải mô hình máy học

Sau khi có kết quả huấn luyện và dự đoán của mô hình, nhóm sử dụng module Pickle để lưu model lại:

```

[41] ML_Model = 'model.sav'
    pickle.dump(ml, open(ML_Model, 'wb'))

[42] loaded_model = pickle.load(open(ML_Model, 'rb'))

```

Hình 5.24: Lưu mô hình máy học.

Sau khi sử dụng lệnh pickle.dump để tiến hành lưu model, ở Google Colab sẽ tạo ra một file có tên “model.sav” và có thể tải về. Từ đó chuyển file sang Raspberry Pi, sử dụng lệnh pickle.load để load file “model.sav” và tiến hành dự đoán với đầu vào mới.

### 5.8. Kiểm thử mô hình dự đoán

Tiến hành dự đoán với input mới của 3 quả trứng ở ngày số 2, 6, 9, 12. Ta có giá trị như sau ở bảng 5.12.

Buộc sóng Số ngày	730 nm	760 nm	810 nm	860nm
Ngày số 2(quả trứng số 1)	11732.87621	14241.53316	31313.7779	29923.18594
Ngày số 2(quả trứng số 2)	11380.87219	13889.87335	30396.44211	28561.8518
Ngày số 2(quả trứng số 3)	11554.81957	14288.12249	33292.14944	32939.42797
Ngày số 6(quả trứng số 1)	12180.57883	14465.36631	35814.85256	33869.72484
Ngày số 6(quả trứng số 2)	11279.85549	13831.24885	33635.17031	31976.8793
Ngày số 6(quả trứng số 3)	9772.48562	11250.93655	33249.14874	30477.62309
Ngày số 9(quả trứng số 1)	10958.84746	12976.82363	32797.92445	31124.67547
Ngày số 9(quả trứng số 2)	10471.46414	12816.61459	31729.76117	30873.75219
Ngày số 9(quả trứng số 3)	10729.70137	12822.4623	31686.88617	30710.4768
Ngày số 12(quả trứng số 1)	9393.89748	11228.78529	33697.97531	31466.14663
Ngày số 12(quả trứng số 2)	9577.86191	11288.91137	32900.96313	29287.85375
Ngày số 12(quả trứng số 3)	9795.19961	11784.81738	33739.56516	30612.24819

Bảng 5.12: Input mới để kiểm thử.

Kết quả đo thực nghiệm so với số ngày thực của trứng đã trải qua với 3 mẫu vật trứng ở ngày số 2, 6, 9, 12 được thể hiện qua bảng 5.13.

Bảng 5.13: Kiểm tra kết quả input mới.

STT	Giá trị dự đoán ngày tuổi của trứng	Thực tế số ngày tuổi của trứng)	Giá trị dự đoán Haugh Units	Hu thực tế	Sai số giữa hệ thống và số ngày thực tế	Sai số giữa HU được dự đoán Và HU thực nghiệm	Thời gian thực hiện
1	5.04	2	87.91	94.41	3.04	6.5	1.573
2	6.71	2	88.38	89.53	4.71	1.15	1.574
3	5.53	2	81.55	89.21	3.53	7.66	1.574
4	9.57	6	80.50	86.09	3.87	5.59	1.573
5	8.38	6	81.00	85.81	2.38	4.81	1.573
6	14.28	6	67.81	82.71	8.28	14.9	1.573
7	11.55	9	76.79	68.71	2.55	8.08	1.574
8	10.24	9	72.36	70.92	1.24	1.44	1.573
9	10.56	9	74.81	76.20	1.56	1.39	1.574
10	11.20	12	69.34	72.60	0.8	3.26	1.574
11	15.43	12	67.91	69.31	3.43	1.4	1.573
12	15.63	12	67.90	63.69	3.63	4.21	1.573

Sau khi thực hiện kiểm thử đầu vào mới thực tế nhóm nhận thấy thời gian thực hiện chương trình tương đối nhanh khoảng ~1.573s. Kết quả không được tốt sai số dự đoán ngày tuổi của trứng ~ 3.25/ngày) và sai số dự đoán Hu ~ 5.03 kết quả này sai lệch vẫn quá cao không thể áp dụng cho thực tế một phần là do mô hình dự đoán chưa đạt độ chính xác cao, cần được nghiên cứu và cải thiện thêm. Nhưng về vấn đề thời gian và giá cả đã đạt được mong đợi của nhóm thời gian dự đoán tương đối nhanh chỉ mất tầm khoảng 1.573s đã có kết quả, giá thiết bị rẻ AS7263 có 27.95\$.

## CHƯƠNG 6. KẾT LUẬN

### 6.1. Kết quả đạt được

Trải qua một học kỳ nhóm đã tìm hiểu, nghiên cứu kiến thức về quang phổ cận hồng ngoại NIR và đã áp dụng vào việc đánh giá chất lượng và độ tươi của trứng dựa trên thời gian bảo quản và Haugh Units, nhóm đã thành công tạo được một mô hình thiết bị nhúng giá cả tương đối rẻ có thể dự đoán thời gian bảo quản và Haugh Units một cách nhanh chóng và không gây xâm lấn.

Nhóm cũng xây dựng được tập dữ liệu phổ thô tương đối lớn gồm 7 tập dữ liệu để có thể áp dụng vào mô hình máy học MLR, SVR, Decision Tree để tìm ra mô hình dự đoán tốt nhất. Cùng với sự khảo sát ba phương pháp tiền xử lý lý Standard Scaler, Multiplicative Scatter Correction\_MSC, Standard Normal Variate\_SNV kết quả có được sau khi training mô hình máy học đạt được tương đối nhiều 63 kết quả đánh giá ứng với 21 tập dữ liệu tiền xử lý và 3 phương pháp học máy áp dụng vào dự đoán ngày tuổi của trứng và 63 kết quả đánh giá ứng với 21 tập dữ liệu tiền xử lý và 3 phương pháp học máy áp dụng vào việc dự đoán Haugh Units. Nhưng sau khi đánh giá các mô hình dự đoán qua  $R^2$ , RMSE kết quả nhận được tương đối thấp. Kết quả cao nhất trong việc dự đoán ngày tuổi của trứng là phương pháp tiền xử lý Standard Scaler cùng với mô hình máy học Support Vector Regression với  $R^2 = 64.51,82\%$ , RMSE = 3.52(ngày), dự đoán Haugh Units với  $R^2 = 67.14\%$ , RMSE = 7.5. Kết quả này đã không đạt được kỳ vọng ban đầu mà nhóm đề ra, không thể áp dụng thực tế cho người dùng vì vậy vẫn cần cải thiện và đánh giá thử trên các phương pháp học máy hồi quy khác để có thể tìm ra mô hình phù hợp với tập dữ liệu nhóm đã xây dựng được.

Tuy nhiên trong khi tiến hành nghiên cứu nhóm gặp phải một số khó khăn khi tìm mẫu vật thu thập dữ liệu phổ thô nên vẫn chưa thể nghiên cứu thêm nhiều phương pháp để có thể nâng cao kết quả đạt được. Giai đoạn đầu khi bắt đầu nghiên cứu nhóm có sử dụng trứng gà ở siêu thị để làm mẫu vật và có được kết quả cao  $R^2 = \sim 82\%$  và sai số thu được tương đối thấp RMSE = 1,22 ngày nhưng nhóm nhận ra rằng trứng ở siêu thị không đảm bảo được tính chính xác dữ liệu thu được vì có thể trong một hộp trứng có nhiều quả

trứng khác nhau ở ngày tuổi nhóm nhận ra lỗi sai này do trong lúc nghiên cứu nhóm nhận thấy một số quả trứng có dấu hiệu hỏng sớm hơn bình thường. Vì vậy nhóm quyết định loại bỏ tập dữ liệu này và tìm một trang trại cung cấp mẫu vật tốt hơn, nhưng do ở trong thành phố không có trang trại gà đẻ trứng nên việc liên hệ, vận chuyển trứng từ một nơi khác mất tương đối nhiều thời gian dẫn đến việc khi thu thập lại dữ liệu xong không còn quá nhiều thời gian.

Bên cạnh đó, nhóm đã khảo sát được phương pháp tiền xử lý tốt nhất trong 3 phương pháp nhóm đã đề cập là phương pháp Standard Scaler và phương pháp máy học tốt nhất trong 3 phương pháp nhóm đã đề cập là Support Vector Regression.

Nhận xét chung: kết quả đạt của nhóm chưa đạt được như mong đợi có thể do thiết kế phần cứng chưa tối ưu dẫn đến dữ liệu thu được chưa ổn định do bị nhiễu. Cũng có thể do lượng mẫu thu còn ít hay mô hình máy học áp dụng chưa phù hợp với đề tài. Vì vậy, cần phải nghiên cứu và xem lại lại dữ liệu và phương pháp nghiên cứu để tìm ra hướng phát triển tốt hơn.

## **6.2. Những điều đạt được khi thực hiện đề tài**

Đã xây dựng được một hệ thống nhúng có thể dự đoán được ngày tuổi của trứng và Haugh Units với thời gian đáp ứng nhanh chóng, thao tác sử dụng dễ dàng và không xâm lấn.

Hiểu biết hơn về những công nghệ, phương pháp liên quan về đánh giá chất lượng trứng gia cầm. Ứng dụng được những lý thuyết nhóm tìm hiểu được vào đề tài.

Hiểu biết hơn về phương pháp học máy, nắm rõ quy trình của một bài toán học máy. Tìm hiểu được nhiều thuật toán giải quyết vấn đề của từng bước trong bài toán học máy. Tìm hiểu được nhiều loại vi xử lý, vi điều khiển, máy tính nhúng cũng như các loại phần cứng, linh kiện điện tử công nghệ cao có thể tìm thấy trên thị trường hiện nay.

## **6.3. Những khó khăn trong quá trình thực hiện đề tài**

Sự tăng cao nhiệt độ của mô hình trong quá trình thu đầu dữ liệu đầu vào: đèn halogen tỏa nhiều nhiệt khi sử dụng lâu dài, do phải tập trung ánh sáng vào trứng nhóm

đã bọc xung quanh đèn lại dẫn đèn không thể tỏa nhiệt ra xung quanh. Mô hình máy học còn đơn giản, chưa phù hợp với dữ liệu đầu vào. Tìm hiểu thêm về những mô hình học máy khác để tìm được phương pháp tốt nhất là điều cần thiết cho việc phát triển đề tài.

Thời gian thu được bộ dữ liệu thô còn dài ngày: để thu thập đủ dữ liệu quang phổ thô cho bài toán nhóm cần đo các vật mẫu trong khoảng thời gian 20 ngày. Vì vậy để cải thiện sản phẩm, nhóm cần thời gian để phân tích dữ liệu, chỉnh sửa phần cứng và đo lại dữ liệu để cải thiện kết quả cho đề tài này, những điều này cần thực hiện trong một khoảng thời gian dài với chi phí vốn kén và được lặp lại nhiều lần để cải thiện chất lượng mô hình.

Một hạn chế quan trọng là thiết bị chỉ đánh giá được trứng gà công nghiệp màu nâu và ở nhiệt độ phòng. Nhưng trong thực tế có nhiều loại môi trường bảo quản trứng khác nhau, màu sắc trứng khác nhau, nhiều loại trứng khác nhau nên đây một đề tài khá lớn cần thêm một lượng thời gian đủ lớn để cải thiện, và phát triển thêm.

#### **6.4. Hướng phát triển chính cho đề tài**

Cải thiện vấn đề quá nhiệt khi sử dụng đèn halogen:

- Thay thế đèn halogen bằng nguồn sáng khác có bước sóng phù hợp với cảm biến.
- Sử dụng nhiều giải pháp tản nhiệt giúp hạ nhiệt cho đèn halogen khi đèn được sử dụng.

Cải thiện nguồn dữ liệu quang phổ đầu vào của thiết bị:

- Thay đổi vị trí, khoảng cách của đèn halogen và cảm biến giúp cảm biến gần vật mẫu hơn.
- Sử dụng giải pháp khác hoặc thiết bị khác để đo đặc giữ liệu.

Tối ưu kích thước của thiết bị:

- Sử dụng máy tính nhúng có kích thước nhỏ hơn.
- Thay đổi thiết bị đo đặc có kích thước nhỏ gọn hơn.

Tích hợp thêm những mô hình của nhiều loại trứng gia cầm khác: trứng gà ta, trứng vịt, trứng cút hoặc trứng đà điểu.

Tìm hiểu thêm về sự thay đổi chất lượng trứng gia cầm ở nhiệt độ khác nhau: nghiên cứu sự khác nhau của trứng được bảo quản tại những môi trường khác nhau. Mục tiêu, sản phẩm được sử dụng tại nhiều vùng khí hậu khác nhau.

## TÀI LIỆU THAM KHẢO

- [1] Y. Akter, A. Kasim, H. Omar, and A. Q. Sazili, "Effect of storage time and temperature on the quality characteristics of chicken eggs," *Journal of Food, Agriculture & Environment*, vol. 12, no. 3-4, pp. 87-92, 2014.
- [2] V. G. Narushin, M. N. Romanov, and D. K. Griffin, "A novel egg quality index as an alternative to Haugh unit score," *Journal of Food Engineering*, vol. 289, p. 110176, 2021.
- [3] W. Yongwei, J. Wang, B. Zhou, and Q. Lu, "Monitoring storage time and quality attribute of egg based on electronic nose," *Analytica Chimica Acta*, vol. 650, no. 2, pp. 183-188, 2009.
- [4] M. Aboonajmi, S. Setarehdan, A. Akram, T. Nishizu, and N. Kondo, "Prediction of poultry egg freshness using ultrasound," *International Journal of food properties*, vol. 17, no. 9, pp. 1889-1899, 2014.
- [5] A. Giunchi, A. Berardinelli, L. Ragni, A. Fabbri, and F. A. Silaghi, "Non-destructive freshness assessment of shell eggs using FT-NIR spectroscopy," *Journal of food engineering*, vol. 89, no. 2, pp. 142-148, 2008.
- [6] J. Zhao, H. Lin, Q. Chen, X. Huang, Z. Sun, and F. Zhou, "Identification of egg's freshness using NIR and support vector data description," *Journal of food Engineering*, vol. 98, no. 4, pp. 408-414, 2010.
- [7] J. Coronel-Reyes, I. Ramirez-Morales, E. Fernandez-Blanco, D. Rivero, and A. Pazos, "Determination of egg storage time at room temperature using a low-cost NIR spectrometer and machine learning techniques," *Computers and Electronics in Agriculture*, vol. 145, pp. 1-10, 2018.
- [8] S. Suktanarak and S. Teerachaichayut, "Non-destructive quality assessment of hens' eggs using hyperspectral images," *Journal of food engineering*, vol. 215, pp. 97-103, 2017.
- [9] J. Cruz-Tirado, M. L. da Silva Medeiros, and D. F. Barbin, "On-line monitoring of egg freshness using a portable NIR spectrometer in tandem with machine learning," *Journal of Food Engineering*, vol. 306, p. 110643, 2021.
- [10] R. Haugh, "The Haugh Unit for Measuring Egg Quality. US Egg Poultry. Mag., 43: 522-555, 572-573," ed, 1937.
- [11] W. J. Foley, A. McIlwee, I. Lawler, L. Aragones, A. P. Woolnough, and N. Berding, "Ecological applications of near infrared reflectance spectroscopy—a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance," *Oecologia*, vol. 116, no. 3, pp. 293-305, 1998.
- [12] J. Lammertyn, A. Peirs, J. De Baerdemaeker, and B. Nicolai, "Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment," *Postharvest Biology and Technology*, vol. 18, no. 2, pp. 121-132, 2000.
- [13] R. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied spectroscopy*, vol. 43, no. 5, pp. 772-777, 1989.
- [14] Å. Rinnan, F. Van Den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201-1222, 2009.
- [15] H. Martens, S. Jensen, and P. Geladi, "Multivariate linearity transformation for near-infrared reflectance spectrometry," in *Proceedings of the Nordic symposium on applied statistics*, 1983: Stokkand Forlag Publishers Stavanger, Norway, pp. 205-234.