**ĐẠI HỌC ĐÀ NẴNG**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỒ ÁN TỐT NGHIỆP**

**NGÀNH: CÔNG NGHỆ THÔNG TIN**

**CHUYÊN NGÀNH: KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI:**

**ĐÁNH GIÁ CHẤT LƯỢNG CÁ DỰA TRÊN PHỔ CẬN HỒNG NGOẠI**

Người hướng dẫn: **TS. Ninh Khánh Duy**

Sinh viên thực hiện: **Trương Công Văn**

Số thẻ sinh viên: **102190148**

Lớp**: 19TCLC-DT3**

**Đà Nẵng, 06/2023**

**Nhận xét của người hướng dẫn**

........................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

*Đà Nẵng, ngày    tháng    năm 2022*

        Giảng viên hướng dẫn

**TS. Ninh Khánh Duy**

**Nhận xét của người phản biện**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

*Đà Nẵng, ngày    tháng    năm 2023*

Giảng viên phản biện

**TÓM TẮT**

Tên đề tài: Đánh giá chất lượng cá dựa trên phổ cận hồng ngoại

Sinh viên thực hiện: Trương Công Văn

Số thẻ SV: 102190148 Lớp: 19TCLC-DT3

**Nội dung tóm tắt đề tài**

Việc kiểm soát chất lượng của cá là một vấn đề cần được quan tâm vì chất lượng cá ảnh hưởng trực tiếp đến sức khỏe của người tiêu dùng. Việc này càng trở nên cần thiết hơn đặc biệt là trong bối cảnh hiện nay khi cá đã và đang trở thành một loại thực phẩm ngày càng được ưa chuộng và vấn đề vệ sinh an toàn thực phẩm đang được các cơ quan quản lý và người tiêu dùng rất quan tâm. Những chất hóa học phổ biến hay được sử dụng để tẩm ướp cá nhằm duy trì độ tươi của cá như hàn the và urê để lại những tác hại nghiêm trọng cho sức khỏe người sử dụng.

Hiện nay, phổ NIR được xem là một trong những công cụ phân tích nhanh và không phá hủy hữu hiệu vì thời gian đáp ứng nhanh, có thể phân tích tại hiện trường, quá trình xử lý mẫu đơn giản, có khả năng phân tích nhiều cấu tử cùng một lúc, thân thiện với môi trường và chi phí thiết bị tương đối thấp. Do đó, việc sử dụng phổ cận hồng ngoại kết hợp với các phương pháp học máy sẽ là cách tối ưu để đánh giá chất lượng cá.

Mục đích của đề tài đánh giá chất lượng cá dựa trên phổ cận hồng ngoại là tìm hiểu tổng quan quan về bài toán, tập trung và phân tích các phương pháp tiếp cận, áp dụng các phương pháp học máy vào việc phát hiện độc tố và dự đoán độ tươi của cá.

|  |  |
| --- | --- |
| ĐẠI HỌC ĐÀ NẴNG  **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  KHOA CÔNG NGHỆ THÔNG TIN | **CỘNG HÒA XÃ HÔI CHỦ NGHĨA VIỆT NAM**  Độc lập - Tự do - Hạnh phúc |

**NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP**

Họ tên sinh viên: Trương Công Văn Số thẻ sinh viên: 102190148

Lớp: 19TCLC-DT3 Khoa: Công nghệ thông tin Ngành: Khoa học dữ liệu và trí tuệ nhân tạo

1. *Tên đề tài đồ án:*

Đánh giá chất lượng cá dựa trên phổ cận hồng ngoại

1. *Đề tài thuộc diện:* ☐ *Có ký kết thỏa thuận sở hữu trí tuệ đối với kết quả thực hiện*
2. *Các số liệu và dữ liệu ban đầu:*

Dữ liệu được giáo viên hướng dẫn TS. Ninh Khánh Duy cung cấp.

1. *Nội dung các phần thuyết minh và tính toán:*

* Chương 1 : Tổng quan về đề tài : giới thiệu và pháp biểu bài toán.
* Chương 2 : Cơ sở lý thuyết: trình bày các lý thuyết liên quan đến bài toán.
* Chương 3 : Đánh giá chất lượng cá dựa trên phổ cận hồng ngoại : trình bày cách quá trình huấn luyện mô hình kết hợp với các phương pháp xử lí dữ liệu và trích xuất đặc trưng dể tìm được phương pháp tối ưu.

1. *Các bản vẽ, đồ thị ( ghi rõ các loại và kích thước bản vẽ ):*

Không có

1. *Họ tên người hướng dẫn:* TS. Ninh Khánh Duy
2. *Ngày giao nhiệm vụ đồ án: 10/04/2023*
3. *Ngày hoàn thành đồ án: 26/06/2023*

|  |  |
| --- | --- |
|  | *Đà Nẵng, ngày 26 tháng 06 năm 2023* |
| **Trưởng Bộ môn** …………………….. | **Người hướng dẫn** |

# LỜI NÓI ĐẦU

Đầu tiên, tôi xin bày tỏ lòng biết ơn của mình tới thầy giáo TS. Ninh Khánh Duy, người đã từng bước hướng dẫn, giúp đỡ tôi, đồng thời đưa ra những góp ý, gợi ý về hướng giải quyết hợp lý trong quá trình thực hiện đồ án tốt nghiệp của mình.

Tiếp đến, tôi xin chân thành cảm ơn các thầy cô giảng viên Khoa công nghệ thông tin của trường Đại Học Bách khoa - Đại học Đà Nẵng đã dìu dắt, truyền đạt cho tôi cả về kiến thức chuyên môn và tinh thần học tập để tôi có đủ trang bị thực hiện đồ án tốt nghiệp của mình. Ngoài ra, tôi xin cảm ơn gia đình đã luôn hỗ trợ dành thời gian và điều kiện tốt nhất cho tôi để hoàn thành đồ án tốt nghiệp.

Tuy có nhiều cố gắng trong quá trình học tập, cũng như trong quá trình làm đồ án tốt nghiệp nhưng cũng không thể tránh khỏi những thiếu sót, do đó tôi rất mong được sự góp ý kiến quý báu của tất cả các thầy cô cũng như tất cả các bạn để kết quả đồ án của được hoàn thiện hơn.

Một lần nữa, tôi xin chân thành cảm ơn.

# CAM ĐOAN

Tôi xin cam đoan :

* Toàn bộ nội dung trong đề tài này do tôi thực hiện dưới sự hướng dẫn của thầy Ninh Khánh Duy.
* Mọi tham khảo dùng trong báo cáo đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian, địa điểm công bố,
* Nếu có bất cứ sao chép không hợp lệ, vi phạm quy chế đào tạo hay gian trá tôi xin hoàn toàn chịu trách nhiệm.

Sinh viên thực hiện

**Trương Công Văn**

**MỤC LỤC**

[LỜI NÓI ĐẦU i](#_Toc139198703)

[CAM ĐOAN ii](#_Toc139198704)

[DANH SÁCH CÁC BẢNG, HÌNH VẼ v](#_Toc139198705)

[DANH SÁCH CÁC KÝ HIỆU, CHỮ VIẾT TẮT vii](#_Toc139198706)

[MỞ ĐẦU 1](#_Toc139198707)

[Chương 1: TỔNG QUAN VỀ ĐỀ TÀI 2](#_Toc139198713)

[1.1 Giới thiệu 2](#_Toc139198714)

[1.2 Phát biểu bài toán 2](#_Toc139198715)

[Chương 2: CƠ SỞ LÝ THUYẾT 3](#_Toc139198716)

[2.1 Phổ cận hồng ngoại và phương pháp phân tích nhanh 3](#_Toc139198717)

[2.1.1 Phổ cận hồng ngoại và ứng dụng 3](#_Toc139198718)

[2.1.2 Phương pháp phân tích nhanh 3](#_Toc139198719)

[2.2 Máy đo phổ cận hồng ngoại DLP NIRscan Nano EVM 4](#_Toc139198720)

[2.2.1 Giới thiệu 4](#_Toc139198721)

[2.2.2 Một số đặc điểm khi sử dụng 5](#_Toc139198722)

[2.2.3 Cơ chế của hệ thống quang học 6](#_Toc139198723)

[2.3 Các kĩ thuật tiền xử lý 8](#_Toc139198724)

[2.3.1 Hiệu chỉnh tán xạ nhân 9](#_Toc139198725)

[2.3.2 Standard Normal Variate 10](#_Toc139198726)

[2.3.3 Chuẩn hóa min max 10](#_Toc139198727)

[2.3.4 Chuẩn hóa mạnh với ngoại lệ 10](#_Toc139198728)

[2.3.5 Bộ lọc Savitzky-Golay 10](#_Toc139198729)

[2.4 Kĩ thuật xử lý dữ liệu mất cân bằng SMOTE 11](#_Toc139198730)

[2.5 Các phương pháp học máy 12](#_Toc139198731)

[2.5.1 Suport Vector Machine 12](#_Toc139198732)

[2.5.2 K láng giềng gần nhất 15](#_Toc139198733)

[2.5.3 Cây quyết định 16](#_Toc139198734)

[2.5.4 Rừng ngẫu nhiên 17](#_Toc139198735)

[2.5.5 Mạng nơ ron tích chập 18](#_Toc139198736)

[2.5.6 Partial Least Squares Regression (PLSR) 20](#_Toc139198737)

[Chương 3: ĐÁNH GIÁ CHẤT LƯỢNG CÁ DỰA TRÊN PHỔ CẬN HỒNG NGOẠI 21](#_Toc139198738)

[3.1 Cách thu thập dữ liệu phổ cận hồng ngoại của cá 21](#_Toc139198739)

[3.2 Mô hình tổng quan của bài toán 22](#_Toc139198740)

[3.3 Đánh giá nồng độ Ure có trong cá 23](#_Toc139198741)

[3.3.1 Giới thiệu 23](#_Toc139198742)

[3.3.2 Thống kê dữ liệu 23](#_Toc139198743)

[3.3.3 Tiền xử lý dữ liệu 24](#_Toc139198744)

[3.3.4 Trích xuất đặt trưng 26](#_Toc139198745)

[3.3.5 Các thuật toán học máy có giám sát 28](#_Toc139198746)

[3.3.6 Kết quả thực nghiệm 30](#_Toc139198747)

[3.4 Đánh giá nồng độ Histamine có trong cá 35](#_Toc139198748)

[3.4.1 Giới thiệu 35](#_Toc139198749)

[3.4.2 Thống kê dữ liệu 36](#_Toc139198750)

[3.4.3 Tiền xử lý dữ liệu 37](#_Toc139198751)

[3.4.4 Trích xuất đặt trưng 39](#_Toc139198752)

[3.4.5 Các thuật toán học máy có giám sát 40](#_Toc139198753)

[3.4.6 Kết quả thực nghiệm 41](#_Toc139198754)

[KẾT LUẬN 47](#_Toc139198755)

[TÀI LIỆU THAM KHẢO 48](#_Toc139198755)

# DANH SÁCH CÁC BẢNG, HÌNH VẼ

[Hình 2. 1 Nguyên lý, phương pháp và lĩnh vực ứng dụng của phổ NIR. 4](#_Toc138327782)

[Hình 2. 2 Máy đo phổ cận hồng ngoại DLP NIRscan Nano EVM. 5](#_Toc138327783)

[Hình 2. 3 Xử lý ánh sáng số DLP dụng linh kiện vi gương số DMD. 6](#_Toc138327784)

[Hình 2. 4 Hệ thống quang học của DLP NIRscan Nano EVM. 6](#_Toc138327785)

[Hình 2. 5 Góc nhìn từ trên xuống của hệ thống quang học. 8](#_Toc138327786)

[Hình 2. 6 Kĩ thuật OverSampling. 11](#_Toc138327787)

[Hình 2. 7 Siêu mặt phẳng tối ưu và biên. 12](#_Toc138327788)

[Hình 2. 8 Công thức tính khoảng cách 15](#_Toc138327789)

[Hình 2. 9 Ví dụ minh họa K-nearest Neighbor 15](#_Toc138327790)

[Hình 2. 10 Cấu trúc của cây quyết đinh. 16](#_Toc138327791)

[Hình 2. 11 Một trình tự CNN để phân loại các chữ số viết tay 18](#_Toc138327792)

[Hình 2. 12 Ví vụ về phép tích chập trên ma trận 3x3 với kernel 2x2 19](#_Toc138327793)

[Hình 2. 13 Max pooling với cửa sổ gộp có kích thước 3x3 với stride = 1. 19](#_Toc138327794)

[Hình 2. 14 Hai loại gộp với kích thước cửa sổ gộp là 2x3, stride = 2. 20](#_Toc138327795)

[Hình 3.1 Cách thu thập dữ liệu phổ NIR. 21](#_Toc138327796)

[Hình 3.2 Phổ hấp thụ của 12 vị trí đo khác nhau của mẫu cá 650. 22](#_Toc138327797)

[Hình 3.3 Mô hình tổng quan của bài toán 23](#_Toc138327798)

[Hình 3.4 Phân bố giá trị của nồng độ Ure trong cá. 25](#_Toc138327799)

[Hình 3.5 Phân bố mẫu phổ có nồng độ Ure trên và dưới LOD của cả tập dữ liệu. 25](#_Toc138327799)

[Hình 3.6 Phân bố mẫu phổ có nồng độ Ure trên và dưới LOD sau khi cân bằng SMOTE của tập huấn luyện. 25](#_Toc138327800)

[Hình 3.7 Phân bố mẫu phổ có nồng độ Ure trên và dưới LOD của tập kiểm thử. 25](#_Toc138327801)

[Hình 3.8 Minh họa phổ NIR gốc. 26](#_Toc138327802)

[Hình 3.9 Phổ chuẩn hóa Min-Max, Robust, MSC, SNV 26](#_Toc138327803)

[Hình 3.10 Minh họa phổ đã chuẩn hóa SNV sau đó qua bộ lọc SG, bộ lọc SG lấy đạo hàm bậc 1 và bậc 2. 27](#_Toc138327804)

[Hình 3.11 Mô hình tổng quan của thuật toán PLSR. 29](#_Toc138327805)

[Hình 3.12 Mô hình CNN đề xuất. 30](#_Toc138327805)

[Hình 3.13 Biến thiên của hàm mất mát và độ chính xác của mô hình CNN trên tập huấn luyện và xác nhận. 31](#_Toc138327806)

[Hình 3.14 Ma trận nhầm lẫn tổng của 5 fold ở vị trí đo ngoài da, đuôi. 35](#_Toc138327807)

[Hình 3.15 Phân bố giá trị của nồng độ Histamine trong cá. 37](#_Toc138327799)

[Hình 3.16 Phân bố mẫu phổ Histamine trên và dưới LOD của cả tập dữ liệu. 37](#_Toc138327808)

[Hình 3.17 Phân bố mẫu phổ có nồng độ Histamine trên và dưới LOD sau khi cân bằng SMOTE của tập huấn luyện. 38](#_Toc138327809)

[Hình 3.18 Phân bố mẫu phổ có nồng độ Histamine trên và dưới LOD của tập kiểm thử. 38](#_Toc138327810)

[Hình 3.19 Minh họa phổ NIR gốc trong tập dữ liệu Histamine. 39](#_Toc138327811)

[Hình 3.20 Phổ chuẩn hóa Min-Max, Robust, MSC, SNV của các mẫu phổ trong tập dữ liệu Histamine. 39](#_Toc138327812)

[Hình 3.21 Minh họa phổ đã chuẩn hóa robust sau đó qua bộ lọc SG, bộ lọc SG lấy đạo hàm bậc 1 và bậc 2. 40](#_Toc138327813)

[Hình 3.22 Mô hình CNN đề xuất để đánh giá nồng độ Histamine có trong cá. 41](#_Toc138327814)

[Hình 3.23 Biến thiên của hàm mất mát và độ chính xác của mô hình CNN trên tập huấn luyện và xác nhận. 42](#_Toc138327815)

[Hình 3.24 Ma trận nhầm lẫn tổng của 5 fold ở vị trí đo ngoài da, lưng. 46](#_Toc138327816)

[Bảng 3.1 Số lượng mẫu cá và phổ hấp thụ NIR dùng để phân loại nồng độ Ure. 24](#_Toc138327817)

[Bảng 3.2 Mô tả các loại vecto đặc trưng để đánh giá chất lượng cá 28](#_Toc138327818)

[Bảng 3.3 Giá trị tham chiếu của nồng độ Ure được phân tích bằng phương pháp tiêu chuẩn. 29](#_Toc138327819)

[Bảng 3.4 Độ chính xác của mô hình CNN khi sử dụng các loại vector đặc trưng. 31](#_Toc138327819)

[Bảng 3.5 Tập hợp các tham số tối ưu cho từng mô hình học máy truyền thống. 32](#_Toc138327820)

[Bảng 3.6 Kết quả kiểm thử của các thuật toán học máy đề xuất. 33](#_Toc138327821)

[Bảng 3.7 Kết quả R2 và MSE của thuật toán PLSR cho dữ liệu Ure. 33](#_Toc138327821)

[Bảng 3.8 Độ chính xác của dữ liệu kiểm thử Ure ở từng vị trí đo khi dùng phương pháp đánh giá repeated hold-out và cross-validation. 34](#_Toc138327822)

[Bảng 3.9 Số lượng mẫu cá và phổ hấp thụ NIR dùng để đánh giá nồng độ Histamine 36](#_Toc138327823)

[Bảng 3.10 Giá trị tham chiếu của nồng độ Histamine được phân tích bằng phương pháp tiêu chuẩn. 43](#_Toc138327824)

[Bảng 3.11 Độ chính xác của mô hình CNN trên các loại vector đặc trưng. 43](#_Toc138327824)

[Bảng 3.12 Tập hợp các tham số tối ưu cho từng mô hình học máy truyền thống dùng để đánh giá Histamine. 43](#_Toc138327825)

[Bảng 3.13 Kết quả R2 và MSE của thuật toán PLSR cho dữ liệu Histamine. 44](#_Toc138327826)

[Bảng 3.14 Kết quả kiểm thử của các thuật toán học máy đề xuất. 45](#_Toc138327826)

[Bảng 3.15 Độ chính xác của dữ liệu kiểm thử Histamine ở từng vị trí đo khi dùng phương pháp đánh giá repeated hold-out và cross-validation. 45](#_Toc138327827)

# DANH SÁCH CÁC KÝ HIỆU, CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Viết tắt** | **Tiếng Anh** | **Tiếng Việt** |
| NIR | Near Infrared | Cận hồng ngoại |
| NIRS | Near Infrared Spectroscopy | Phổ cận hồng ngoại |
| SVM  KNN  CNN  SNV  MSC  SMOTE  DLP  EVM  DMD  SG | Support Vector Maching  K-Nearest Neighbors  Convolutional Neural Network  Stradard Normal Variate  Multiplicative Scatter Correction  Synthetic Minority Oversampling Technique  Digital Light Processing  Evaluation Module  Digital Micromirror Device  Savitzky-Golay | Máy Vecto hỗ trợ  K láng giềng gần nhất  Mạng nơ ron tích chập  Chuẩn biến  Hiệu chỉnh tán xạ nhân  Xử lí ánh sáng số  Mô đun phát triển  Linh kiện vi gương số  Savitzky-Golay |
|  |  |  |

# MỞ ĐẦU

1. **Mục đích**

Ứng dụng ứng dụng các mô hình học máy và học sâu kết hợp với các kĩ thuật trích xuất đặc trưng từ phổ cận hồng ngoại nhằm xác định độ độc tố trong cá một cách tiện lợi nhanh chóng, có thể thực hiện tại hiện trường và ít tốn kém hơn các phương pháp phân tích hóa-lý (ví dụ như High performance liquid chromatography-HPLC).

1. **Mục tiêu**

Nghiên cứu, ứng dụng các mô hình học máy và học sâu kết hợp với các kĩ thuật trích xuất đặc trưng từ phổ cận hồng ngoại (NIR) nhằm xác định nhanh độ tươi của cá và phát hiện độc tố trong cá.

1. **Phạm vi và đối tượng nghiên cứu**

Phạm vi và đối tượng nghiên cứu sẽ là 03 loại cá: nục, ngừ và thu, đều được thu mua tại cảng cá để đảm bảo chúng có cùng nguồn gốc vì yếu tố này có thể ảnh hưởng một số đến dữ liệu quang phổ.

1. **Phương pháp nghiên cứu**

Tìm hiểu các phương pháp đã có thông qua các bài báo tổng hợp phương pháp, đào sâu vào từng phương pháp con để hiểu cách hoạt động cũng như biết được đầu vào của từng phương pháp. Sau khi đã hiểu được các phương pháp đó thì chọn ra một hướng đi cụ thể và vận dụng những kiến thức đã có tự đề xuất ra mô hình mới. Tiến hành cài đặt mô hình đề xuất đó, tính toán các các kết quả thu được so sánh với các mô hình đã có.

1. **Cấu trúc của đồ án tốt nghiệp**

Gồm các phần sau:

* Tổng quan đề tài
* Cơ sở lý thuyết
* Đánh giá chất lượng cá dựa trên phổ cận hồng ngoại

# Chương 1: TỔNG QUAN VỀ ĐỀ TÀI

### 1.1 Giới thiệu

Việc kiểm soát chất lượng của cá là một vấn đề cần được quan tâm vì chất lượng cá ảnh hưởng trực tiếp đến sức khỏe của người tiêu dùng. Việc này càng trở nên cần thiết hơn đặc biệt là trong bối cảnh hiện nay khi cá đã và đang trở thành một loại thực phẩm ngày càng được ưa chuộng và vấn đề vệ sinh an toàn thực phẩm đang được các cơ quan quản lý và người tiêu dùng rất quan tâm. Những chất hóa học phổ biến hay được sử dụng để tẩm ướp cá nhằm duy trì độ tươi của cá như hàn the và urê để lại những tác hại nghiêm trọng cho sức khỏe người sử dụng. Theo thông tư số 27/2012/TT-BYT của Bộ Y tế về hướng dẫn việc quản lý phụ gia thực phẩm, urê không nằm trong danh mục phụ gia được phép sử dụng trong thực phẩm hay nói cách khác, việc sử dụng urê để bảo quản thực phẩm hoàn toàn bị pháp luật nghiêm cấm. Trong thực tế, người tiêu dùng có thể xác định được cá đã qua tẩm ướp urê bằng việc quan sát bằng mắt thường, tuy nhiên phương pháp này đòi hỏi nhiều kinh nghiệm và có phần không được khách quan. Bên cạnh đó, vẫn có những phương pháp hóa-lý (ví dụ như sắc ký lỏng hiệu năng cao HPLC) giúp xác định chính xác nồng độ các độc tố bên trong cá, nhưng những phương pháp này cần thiết bị phân tích hiện đại, tốn nhiều thời gian cũng như chi phí thí nghiệm và không thể thực hiện tại hiện trường.

Chính vì vậy, tôi đề xuất sử dụng mô hình học máy và học sâu kết hợp với các kĩ thuật trích xuất đặc trưng từ phổ cận hồng ngoại (NIR) nhằm phát hiện độ độc tố trong cá một cách tiện lợi và nhanh chóng.

### 1.2 Phát biểu bài toán

Nhiệm vụ của bài toán đánh giá chất lượng cá dựa trên phổ cận hồng ngoại này sẽ là xác định nhanh nồng độ ure, histamine có trong cá và phân lọai mẫu cá thành hai loại: trên và dưới ngưỡng phát hiện (level of detection hay LOD) được thiết lập ở 1000 mg/kg cho ure và 100 mg/kg cho histamine vì đây là giới hạn của nồng độ ure, histamine được cho phép tiêu thụ.

# Chương 2: CƠ SỞ LÝ THUYẾT

## **2.1 Phổ cận hồng ngoại và phương pháp phân tích nhanh**

### 2.1.1 Phổ cận hồng ngoại và ứng dụng

Phổ cận hồng ngoại (Near Infrared Spectroscopy – NIR) là một loại quang phổ sử dụng năng lượng photon trong dải năng lượng từ 2.65 x 10-19 đến 7.96 x 10-20 J, tương ứng với dải bước sóng 780nm đến 2500nm.

Phổ cận hồng ngoại được sử dụng phổ biến trong các lĩnh vực phân tích thành phần hoá sinh của vật thể, là một phương pháp nhằm tìm ra các quy luật liên hệ giữa tính chất sinh học và hóa học với các phổ phát xạ hay hấp thụ của vật thể. Quang phổ được ứng dụng trong việc tìm lại các tính chất của vật chất từ quang phổ quan sát được.

### 2.1.2 Phương pháp phân tích nhanh

Phương pháp phân tích nhanh dựa trên phổ cận hồng ngoại NIR (Near-Infrared spectroscopy) là phương pháp sử dụng bức xạ điện từ trong khoảng bước sóng từ 780 nm đến 2500 nm để chiếu vào mẫu và thu phổ hấp thụ/phản xạ để từ đó phân tích các thành phần hoặc tính chất mong muốn từ thực phẩm. Phổ hấp thụ/phản xạ thu được là do các liên kết hóa học trong các hợp chất hữu cơ có thể hấp thụ/phản xạ năng lượng khi nó được chiếu vào nguồn bức xạ điện từ. Ứng với sóng hồng ngoại, các liên kết hóa học có thể hấp thụ năng lượng của bức xạ hồng ngoại liên quan đến các liên kết trong các nhóm chức có chứa hydro như O-H, C-H, N-H. Năng lượng hấp thụ sẽ tạo nên các dao động của các liên kết này. Mức độ dao động phụ thuộc vào khối lượng phân tử, hình dạng (cấu trúc không gian), độ “cứng” của liên kết… Từ đó, phổ hấp thụ/phản xạ thu được của các mẫu sẽ có những đặc trưng rất khác nhau.

Máy đo phổ NIR chứa nguồn sáng, hệ thống tách chùm (bộ chọn bước sóng), bộ phát hiện mẫu, bộ phát hiện quang học và hệ thống xử lý/phân tích dữ liệu (tuỳ chọn, có thể có hoặc không). Các bộ phận này có thể có các đặc điểm khác nhau và nên được lựa chọn dựa trên mục đích sử dụng của chúng để tạo ra một thiết bị hiệu quả và nhất quán. Hầu hết các hệ thống phổ NIR hoạt động ở chế độ truyền dẫn (transmission), phản xạ (reflection), phản xạ khuếch tán (diffuse reflectance), hoặc chuyển hướng (transflectance) tùy thuộc vào loại thiết bị đang được sử dụng và loại phân tích đang được thực hiện. Ban đầu, phổ của các mẫu được thu thập bằng máy đo phổ NIR. Sau khi thu thập dữ liệu, phân tích hóa học (chemometric analysis) được thực hiện để tạo ra mô hình hiệu chuẩn (calibration model) để đánh giá (các) thành phần mục tiêu bằng cách sử dụng các dải bước sóng có liên quan trong phổ NIR. Bước này rất quan trọng vì độ chính xác ở giai đoạn này đảm bảo rằng (các) mô hình hiệu chuẩn cuối cùng đảm bảo độ tin cậy cao. Do đó, nhược điểm lớn của phương pháp phân tích nhanh dùng phổ NIR là nó luôn yêu cầu dữ liệu tham chiếu để phân tích định lượng, điều này đòi hỏi phải sử dụng phân tích hóa học thông qua các dụng cụ hoá phân tích thông thường [8].

Hiện nay, phổ NIR được xem là một trong những công cụ phân tích nhanh và không phá hủy hữu hiệu vì thời gian đáp ứng nhanh, có thể phân tích tại hiện trường, quá trình xử lý mẫu đơn giản, có khả năng phân tích nhiều cấu tử cùng một lúc, thân thiện với môi trường và chi phí thiết bị tương đối thấp. Phổ NIR được sử dụng trong phân tích định tính (phân loại thực phẩm, xác thực nguồn gốc về địa lý, giống, phương pháp chế biến) như táo, dầu ô liu, cà phê, gạo… Còn trong phân tích định lượng, phổ NIR cũng được sử dụng để phân tích hàm lượng ẩm, protein, chất béo, carbohydrate, flavonols, phenols, acid hữu cơ, glycerol, melanine, nitrate… trong nhiều loại nông sản và thực phẩm như gạo, ca cao, cà phê, sữa, nước ép trái cây… Một số nghiên cứu về lĩnh vực này đã được mô tả ngắn gọn trong Phần 1 của Chương Mở đầu. Tổng quan về phổ NIR và các ứng dụng trong công nghiệp thực phẩm được trình bày đầy đủ trong bài báo của Osborne [7]. Hình 1.1 minh hoạ tổng quan nguyên lý, các phương pháp và lĩnh vực ứng dụng của phổ NIR [8].

A picture containing text, screenshot, font, diagram

Description automatically generated

Hình 2.1 Nguyên lý, phương pháp và lĩnh vực ứng dụng của phổ NIR [8].

## **2.2 Máy đo phổ cận hồng ngoại DLP NIRscan Nano EVM**

### 2.2.1 Giới thiệu

Mặc dù có nhiều thiết bị/cảm biến đo phổ NIR được thương mại hoá và có sẵn trên thị trường như SCiO của Consumer Physics, chúng tôi chọn sử dụng thiết bị đo phổ NIR (còn gọi là máy quét NIR) cầm tay nhãn hiệu DLP NIRscan Nano EVM của hãng Texas Instrument [9] để thu phổ NIR của trái cây vì sự linh hoạt (có sẵn các công cụ phần mềm và thư viện đi kèm), giá thành phải chăng (khoảng 1000 USD/máy) và hiệu suất khá cao của thiết bị. Mô-đun phát triển (Evaluation Module - EVM) này chứa mọi thứ mà một nhà thiết kế cần để bắt đầu phát triển máy quang phổ dựa trên công nghệ xử lý ánh sáng số (Digital Light Processing hay DLP). Công nghệ DLP cho phép máy phân tích quang phổ cầm tay sử dụng trong thực phẩm, dược phẩm, dầu khí, y tế, an ninh và các ngành công nghiệp mới nổi khác và cung cấp hiệu năng ở mức phòng thí nghiệm trong lĩnh vực này.

Mô-đun EVM này chứa linh kiện vi gương số (digital micromirror device – DMD) DLP2010NIR, bộ điều khiển số DLPC150 và các linh kiện quản lý nguồn tích hợp DLPA2005. DLP2010NIR được tối ưu hoá hoạt động ở các bước sóng từ 700 đến 2500 nm. Tuy nhiên, thiết bị DLP NIRscan Nano EVM chỉ hỗ trợ hoạt động với các bước sóng trong dải 900-1700 nm.

**A close-up of a computer chip

Description automatically generated with low confidence**

Hình 2.2 Máy đo phổ cận hồng ngoại DLP NIRscan Nano EVM.

### 2.2.2 Một số đặc điểm khi sử dụng

* Máy đo quang phổ giúp xác định nhiều loại lượng nguyên tố có trọng vật liệu.
* Phân tích được mẫu chất rắn, bột, lỏng mà không cần xử lý hóa học hay gia công mẫu trước.
* Phân tích được nhiều loại vật liệu.
* Phân tích chính xác có thể từ vài ppm đến 100%.
* Gọn nhẹ, có thể di chuyển dễ dàng.

### 2.2.3 Cơ chế của hệ thống quang học

Máy đo phổ dựa trên công nghệ xử lý ánh sáng số DLP sử dụng một linh kiện vi gương số DMD để lựa chọn bước sóng và một bộ phát hiện đơn điểm (single point detector) như thể hiện trong Hình 1.4. Bằng cách quét tuần tự qua các cột (bật các dãy pixel cụ thể nào đó) của DMD, một bước sóng ánh sáng tương ứng sẽ được dẫn (direct) đến bộ phát hiện và được thu lại.

**A picture containing text, diagram, line, screenshot

Description automatically generated**

Hình 2.3 Xử lý ánh sáng số DLP dụng linh kiện vi gương số DMD.

*Diagram

Description automatically generated*

Hình 2.4 Hệ thống quang học của DLP NIRscan Nano EVM.

Hệ thống quang học được gắn nằm trên hệ thống điện tử của máy đo phổ (Hình 4). Mô-đun phản chiếu gồm hai đèn dây tóc vonfram băng rộng. Mẫu vật được đặt sát cửa sổ phía trước bằng sapphire của đầu phản xạ (reflectance head). Trong quá trình quét phổ, mẫu hấp thụ một lượng ánh sáng NIR nào đó và phản xạ một cách khuếch tán lượng ánh sáng không được hấp thụ vào hệ thống quang học. Lượng ánh sáng bị hấp thụ ở mỗi bước sóng phụ thuộc vào cấu tạo phân tử của vật chất và đặc trưng cho vật chất đó nên được gọi là đặc trưng hóa học (chemical fingerprint).

Ánh sáng phản xạ lại từ mẫu vật được thu thập bởi thấu kính thu thập và tập trung vào mô-đun quang thông qua khe đầu vào (input slit). Kích thước khe được chọn để cân bằng độ phân giải bước sóng với SNR của máy đo phổ. Máy đo phổ này sử dụng khe rộng 25 µm và cao 1,8 mm. Ánh sáng đi qua khe được chuẩn trực bởi bộ thấu kính đầu tiên, đi qua bộ lọc thông sóng dài 885 nm, và sau đó chạm vào một cách tử phản xạ (reflective grating). Cách tử này, kết hợp với thấu kính hội tụ, phân tán ánh sáng thành các bước sóng cấu thành. Các thấu kính hội tụ tạo thành hình ảnh của khe tại DLP2010NIR DMD. Các bước sóng khác nhau của hình ảnh khe này được trải theo chiều ngang trên DLP2010NIR DMD.

Hệ thống quang học hình ảnh các bước sóng 900nm đến một đầu của DMD và 1700nm ở đầu kia, với tất cả các bước sóng khác được phân tán liên tục ở giữa. Khi các cột DMD cụ thể được chọn là bật hoặc nghiêng về vị trí 17 ° dương, năng lượng được phản ánh bởi các cột được chọn sẽ được dẫn qua hệ thống quang học thu thập đến máy dò InGaAs pixel đơn. Tất cả các cột DMD khác được chọn là tắt hoặc nghiêng về vị trí 17 ° âm, chuyển hướng các bước sóng không được chọn xuống dưới cùng của động cơ ánh sáng và tránh xa đường quang của máy dò để không ảnh hưởng đến phép đo bước sóng đã chọn.

Để cho phép dung sai cơ học ở vị trí khe, góc cách tử và vị trí DMD, hình ảnh khe DLP NIRscan Nano trên DMD được lấp đầy trong trục phân tán 10% ở mỗi đầu và được lấp đầy trong trục trực giao. Điều này dẫn đến khoảng (1700 - 900nm) / (854 \* 0,8 pixels) = 1,17nm mỗi pixel trên DMD. Trong quá trình sản xuất, hiệu chuẩn được thực hiện giữa các bước sóng và vị trí cột của chúng trên DMD. Do số lượng 15 cột DMD thường không chia hết cho số nhóm bước sóng mong muốn, nên DLP NIRscan Nano duy trì hằng số chiều rộng cột trong quá trình quét, nhưng bước qua mảng DMD bằng một lượng khác với chiều rộng của cột. Số lượng bước này phụ thuộc vào chiều rộng và số lượng mẫu mong muốn (điểm bước sóng).

Khi chúng ta thực hiện các phép đo với máy quang phổ, máy quang phổ sẽ phát ra đèn hồng ngoại trên vật liệu. Tùy thuộc vào vật liệu, phản ứng của nó ở các bước sóng khác nhau là khác nhau, cảm biến InGaAS thu thập ánh sáng và chuyển đổi cường độ này thành tín hiệu điện, sau đó thành các giá trị số nhờ ADC. Giá trị này được gọi là cường độ.

Hình 5 thể hiện góc nhìn từ trên xuống của hệ thống quang học. Mô-đun phản xạ của thiết bị hoạt động bằng cách chiếu sáng mẫu được thử ở một góc sao cho các phản xạ gương không được thu thập, trong khi thu thập và tập trung các phản xạ khuếch tán vào khe. Các đèn chiếu sáng được chỉ định là đèn kết thúc thấu kính vì đầu trước của bóng đèn thủy tinh được tạo thành một thấu kính hướng nhiều ánh sáng hơn từ dây tóc đến vùng thử nghiệm mẫu. Các hình chữ nhật màu xanh lá cây đại diện cho đèn kết thúc ống kính. Các hình nón màu vàng đậm là ánh sáng được chiếu ra từ đèn. Mỗi đèn tạo ra một chùm ánh sáng ở góc 40 độ giao nhau qua cửa sổ sapphire ở khoảng 0,75 mm. Có khoảng dung sai khoảng 0,25 mm đối với giao điểm của chùm tia do dung sai cơ học của khung máy, các biến thể của đầu thấu kính từ đèn sang đèn, các biến thể của hình dạng đèn và vị trí của đèn. Đèn kết thúc thấu kính tập trung chùm sáng ở khoảng cách 3 mm so với đèn và tạo kích thước điểm bao phủ cửa sổ mẫu sapphire.

**Diagram, engineering drawing

Description automatically generated**

Hình 2.5 Góc nhìn từ trên xuống của hệ thống quang học.

Ống kính thu thập tập hợp ánh sáng từ vùng có đường kính 2,5 mm ở cửa sổ mẫu. Kích thước của vùng thu thập được khớp với kích thước điểm chiếu sáng danh nghĩa được tạo bởi các đèn cuối thấu kính. Điều này đòi hỏi mẫu phải được đặt trực tiếp vào cửa sổ sapphire, nơi hai đường dẫn nguồn sáng góc giao nhau với hình nón tầm nhìn của thấu kính. Nếu mẫu được dịch chuyển ra xa khỏi cửa sổ, mẫu có thể không nhận đủ ánh sáng để hệ thống thực hiện quét phổ một cách chính xác.

## **2.3 Các kĩ thuật tiền xử lý**

Tiền xử lý dữ liệu quang phổ cận hồng ngoại (NIR) đã trở thành một phần không thể thiếu trong mô hình hiệu chuẩn hóa học. Mục tiêu của quá trình tiền xử lý là loại bỏ các hiện tượng vật lý trong quang phổ để cải thiện hồi quy đa biến tiếp theo, mô hình phân loại hoặc phân tích thăm dò. Các kỹ thuật tiền xử lý được sử dụng rộng rãi nhất trong quang phổ cận hồng ngoại có thể được chia thành hai loại: phương pháp hiệu chỉnh tán xạ và đạo hàm phổ.

Các phương pháp tiền xử lý hiệu chỉnh tán xạ đầu tiên bao gồm hiệu chỉnh tán xạ nhân (Multiplicative Scatter Correction – MSC), biến chuẩn (Standard Normal Variate – SNV) [4] và chuẩn hóa.

Đạo hàm phổ được đại diện bởi kĩ thuật: đạo hàm đa thức Savitzky-Golay (SG). Kĩ thuậtnày làm mịn phổ trước khi tính đạo hàm để giảm tác động bất lợi đến tỉ lệ tín hiệu trên nhiễu.

Các kĩ thuật tiền xử lý đều có mục tiêu là giảm độ biến thiên không được mô hình hóa trong dữ liệu để tăng đặc trưng trong phổ. Điều này có thể đạt được bằng cách sử dụng một kỹ thuật tiền xử lý phù hợp, nhưng luôn có nguy cơ áp dụng sai loại hoặc áp dụng quá trình tiền xử lý quá nghiêm ngặt sẽ loại bỏ thông tin có giá trị. Rất khó để đánh giá lựa chọn tiền xử lý phù hợp trước khi xác thực mô hình, nhưng nói chung, không nên thực hiện một số bước tiền xử lý và theo yêu cầu tối thiểu, tiền xử lý phải duy trì hoặc giảm độ phức tạp của mô hình hiệu quả.

### 2.3.1 Hiệu chỉnh tán xạ nhân

Hiệu chỉnh tán xạ nhân (MSC) có thể là kĩ thuật tiền xử lý được sử dụng rộng rãi nhất cho phổ cận hồng ngoại (NIR). MSC lần đầu được giới thiệu bởi Martens và cộng sự vào năm 1983 và sau đó được đóng góp thêm bởi Geladi và cộng sự năm 1985. Ý tưởng đằng sau MSC là loại bỏ các hiệu ứng tán xạ không tốt khỏi ma trận dữ liệu trước khi mô hình hóa dữ liệu. MSC bao gồm hai bước:

1. Ước tính các hệ số hiệu chỉnh

x subscript o r g end subscript space equals space b subscript 0 space end subscript plus space b subscript r e f comma 1 end subscript space cross times space x subscript r e f end subscript space plus space e

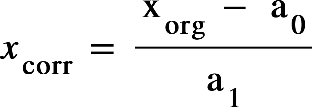
1. Hiệu chỉnh phổ



Trong đó xorg là một mẫu phổ ban đầu được đo bằng thiết bị NIR, xref là phổ tham chiếu được sử dụng để tiền xử lý toàn bộ tập dữ liệu, e là phần không được mô hình hóa của xorg, xcorr là phổ đã được hiệu chỉnh, b0 và bref,1 là các tham số vô hướng, khác nhau đối với mỗi mẫu.

### 2.3.2 Standard Normal Variate

Tiền xử lý SNV có lẽ là phương pháp được áp dụng nhiều thứ hai để hiệu chỉnh tán xạ dữ liệu NIR. Định dạng cơ bản của SNV và chuẩn hóa hiệu chỉnh giống như định dạng cho MSC truyền thống:



Với SNV, a0 là giá trị trung bình của phỗ mẫu cần hiệu chỉnh, a1 là độ lệch chuẩn của phổ mẫu.

### 2.3.3 Chuẩn hóa min max

Chuẩn hóa min-max là phương pháp đơn giản nhất trong việc co giãn phạm vi của đặc trưng bằng việc co giãn chúng về phạm vi [0,1] hoặc [-1,1].

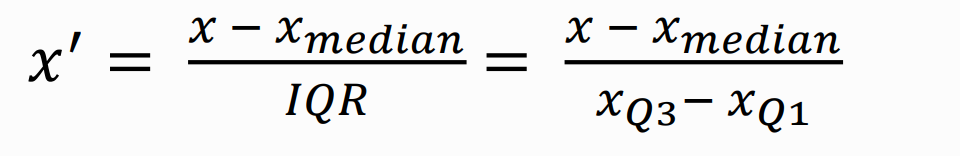
A picture containing text, font, line, screenshot

Description automatically generated

Trong đó x là giá trị ban đầu, x’ là giá trị sau khi chuẩn hóa, min(x) là giá trị nhỏ nhất của đặc trưng và max(x) là giá trị lớn nhất của đặc trưng.Chuẩn hóa mạnh với ngoại lệ.

### 2.3.4 Chuẩn hóa mạnh với ngoại lệ

Phương pháp chuẩn hóa mạnh với ngoại lệ (robust) sẽ dụng các mẫu ít bị ảnh hưởng bởi các giá trị ngoại lệ là trung vị (median) và IQR (interquartile range) để chuẩn hóa.



Trong đó x là giá trị ban đầu, x’ là giá trị sau khi chuẩn hóa xQ1 là 1st quartile của x, xQ3 là 3rd quartile của x.

### 2.3.5 Bộ lọc Savitzky-Golay

Bộ lọc Savitzky-Golay là một bộ lọc kỹ thuật số có thể được áp dụng cho một tập hợp các điểm dữ liệu kỹ thuật số nhằm mục đích làm mượt dữ liệu, tức là tăng độ chính xác của dữ liệu mà không làm méo đi xu hướng tín hiệu. Bộ lọc này loại bỏ nhiễu tần số cao khỏi dữ liệu và có lợi thế giữ nguyên hình dạng và đặc trưng ban đầu của tín hiệu tốt hơn so với các phương pháp lọc khác, chẳng hạn như kỹ thuật trung bình di động.Các bước thực hiện thuật toán Savitzky-Golay:

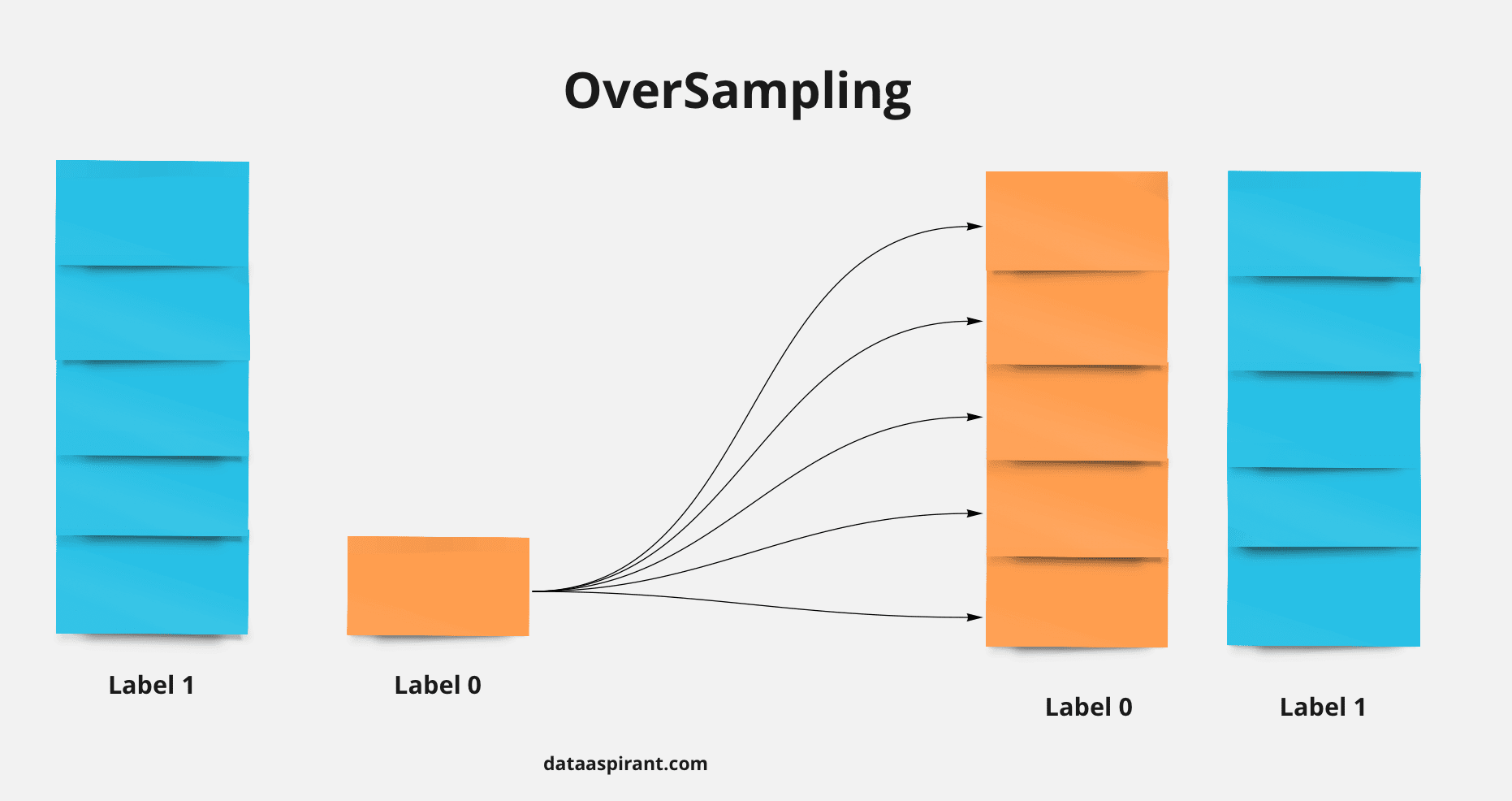
* Xác định kích thước cửa sổ (window size) và bậc của đa thức (polynomial order).
* Xác định trọng số cho từng điểm dữ liệu trong cửa sổ.
* Tính toán giá trị mới cho điểm dữ liệu ở giữa cửa sổ bằng cách sử dụng trọng số đã xác định ở bước 2 và các giá trị của các điểm dữ liệu trong cửa sổ.

## **2.4 Kĩ thuật xử lý dữ liệu mất cân bằng SMOTE**

Phương pháp SMOTE (Synthetic Minority Oversampling Technique) đã được đề xuất trong một bài báo năm 2002 trên Journal of Artificial Intelligence. Phương pháp này dùng để cải tiến để xử lý dữ liệu mất cân bằng trong các bài toán phân loại.

SMOTE là một thuật toán thực hiện tăng cường dữ liệu bằng cách tạo các điểm dữ liệu tổng hợp dựa trên các điểm dữ liệu gốc. SMOTE có thể được coi là một phiên bản nâng cao của quá trình lấy mẫu (oversampling) hoặc là một thuật toán cụ thể để tăng cường dữ liệu. Ưu điểm của SMOTE là bạn không tạo các bản sao, mà là tạo các điểm dữ liệu tổng hợp hơi khác so với các điểm dữ liệu gốc. Thuật toán SMOTE hoạt động như sau:

* Chọn một mẫu ngẫu nhiên từ lớp thiểu số.
* Tìm k láng giềng gần nhất cho mẫu đã được chọn
* Chọn một trong các láng giềng gần nhất và tạo ra một mẫu mới bằng cách lấy sự khác biệt giữa các đặc trưng của láng giềng được chọn và đặc trưng của mẫu ban đầu, sau đó nhân với một số ngẫu nhiên trong khoảng [0, 1].
* Lặp lại các bước 1-3 cho đến khi dữ liệu được cân bằng.



Hình 2.6 Kĩ thuật OverSampling.

## **2.5 Các phương pháp học máy**

### 2.5.1 Suport Vector Machine

Support Vector Machine là một thuật toán trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau nhằm để phân loại và phân tích hồi quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Thuật toán SVM chia hai lớp dữ liệu bằng một siêu mặt phẳng d–1 chiều khi số chiều của dữ liệu huấn luyện là d. Hình xxx là ví dụ phân tách dữ liệu thuộc hai lớp sử dụng SVM. Trong đó, w.x–b=0 là siêu mặt phẳng thể hiện sự phân tách dữ liệu.

A picture containing line, diagram, parallel, circle

Description automatically generated

Hình 2.7 Siêu mặt phẳng tối ưu và biên.

Cho trước tập huấn luyện được biểu diễn trong không gian véc tơ, trong đó mỗi văn bản là một điểm, phương pháp SVM dạng chuẩn tìm ra một siêu mặt phẳng 30 quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng với lớp (+) và lớp (-). Hiệu quả xác định siêu mặt phẳng này được quyết định bởi khoảng cách của điểm gần mặt phẳng nhất của mỗi lớp. Khoảng cách càng lớn thì mặt phẳng quyết định càng tốt đồng nghĩa với việc phân loại càng chính xác và ngược lại. Mục đích cuối cùng của phương pháp là tìm khoảng cách biên lớn nhất.

* **Mục đích:** Dùng để phân lớp dữ liệu mới thuộc lớp nào.
* **Dữ liệu vào:** Cho trước một tập dữ liệu huấn luyện đã gán nhãn thuộc lớp -1 hoặc +1.
* **Dữ liệu ra:** Dùng để tìm ra nhãn cho các dữ liệu mới.
* **Mô tả thuật toán:** Ta có một tập huấn luyện D, một tập gồm n điểm có dạng. 

Với mang giá trị +1 hoặc −1, ngầm định lớp chứa điểm . Mỗi là một véc tơ thực p - chiều. Ta cần tìm siêu mặt phẳng có lề lớn nhất chia tách các điểm thuộc lĩnh vực quan tâm và được gán nhãn = 1 và các điểm thuộc lĩnh vực không quan tâm và được gán nhãn = −1. Mỗi siêu mặt phẳng đều có thể được viết như là một tập các điểm thỏa mãn phương trình:

w . x – b = 0

Với ( . ) kí hiệu cho tích vô hướng. Véc tơ trọng số w, nó vuông góc với siêu mặt phẳng . Tham số xác định độ lệch của siêu mặt phẳng từ nó đến vectơ w.

Phương trình trên tương đương với phương trình sau:



Tương đương với công thức :

A picture containing font, text, white, screenshot

Description automatically generated

Với w = là bộ hệ số siêu mặt phẳng hay là vectơ trọng số, C là độ dịch, khi thay đổi w và C thì hướng và khoảng cách từ gốc tọa độ đến siêu mặt phẳng thay đổi.

Chúng ta cần chọn w và b để cực đại lề, hay khoảng cách giữa hai siêu mặt phẳng song song sao cho chúng càng xa càng tốt trong khi vẫn phân chia tốt dữ liệu. Các siêu mặt phẳng ấy được xác định bằng:

w . x – b = 1

và

w . x – b = -1

Để ý rằng nếu một dữ liệu huấn luyện có thể được chia tách một cách tuyến tính (bằng một đường thẳng), chúng ta có thể chọn hai siêu mặt phẳng của lề sao cho không có điểm nào ở giữa chúng và cố gắng cực đại khoảng cách giữa chúng. Bằng hình học, chúng ta tìm được khoảng cách giữa hai siêu mặt phẳng là w 2 và ta muốn cực tiểu giá trị ||w||. Để tránh các điểm dữ liệu rơi vào bên trong lề, chúng ta thêm vào các điều kiện sau, với mỗi i ta có:

 nếu = 1 (đối với lớp thứ nhất)

hoặc

 nếu = -1 (đối với lớp thứ hai)

Có thể viết gọn lại như sau :



Ta có thể gom chúng lại trong một bài toán tối ưu:

Cực tiểu (theo w, b): với điều kiện (với mọi i = 1,..., n)



Các vấn đề tối ưu hóa được trình bày trong phần trước là khó khăn để giải quyết bởi vì nó phụ thuộc vào ||w||, chỉ tiêu w, trong đó bao gồm một căn bậc hai. May mắn thay, nó có thể làm thay đổi phương trình bằng cách thay thế với

(Yếu tố 1/2 đang được sử dụng để thuận tiện trong toán học) mà không thay đổi các giải pháp (tối thiểu của bản gốc và phương trình sửa đổi có w và b). Đây là vấn đề tối ưu hóa một phương trình bậc hai. Rõ ràng hơn:

Cực tiểu (trong w, b) : (với bất kỳ i = 1,...,n)



Đưa về phương trình Lagrange với số nhân αi [7]:

A picture containing font, text, white, handwriting

Description automatically generated

Máy véc tơ hỗ trợ SVM là quá trình tìm ra các siêu mặt phẳng phụ thuộc vào tham số vectơ trọng số w và độ dịch C. Mục tiêu của phương pháp SVM là ước lượng w và C để cực đại hoá lề giữa các lớp dữ liệu dương và âm. Các giá trị khác nhau của lề cho ta các họ siêu mặt phẳng khác nhau và lề càng lớn thì năng lực của máy học càng giảm. Như vậy, cực đại hoá lề thực chất là việc tìm một máy học có năng lực nhỏ nhất. Quá trình phân loại là tối ưu khi sai số phân loại là cực tiểu. Sau khi đã tìm được phương trình của siêu mặt phẳng bằng thuật toán SVM, áp dụng công thức này để tìm ra nhãn lớp cho các dữ liệu mới. Nếu dữ liệu học không tách rời tuyến tính, thêm biến và thay phương trình trên bằng phương trình:

A picture containing text, font, white, handwriting

Description automatically generated

Từ đó ta có phương trình tổng quát của siêu mặt phẳng tìm ra được bởi thuật toán SVM là:



Với i = 1,…,n. Trong đó n là số dữ liệu huấn luyện.

### 2.5.2 K láng giềng gần nhất

K-Nearest Neighbor (KNN) là phương pháp phân lớp các đối tượng dựa vào khoảng cách gần nhất. Mỗi đối tượng được xác định bằng K láng giềng của nó bằng cách đo khoảng cách giữa chúng và các điểm làng giềng dưới độ đo Euclid, Manhattan hoặc Minkowski.

A picture containing text, font, diagram, handwriting

Description automatically generated

Hình 2.8 Công thức tính khoảng cách

A picture containing text, screenshot, circle, design

Description automatically generated

Hình 2.9 Ví dụ minh họa K-nearest Neighbor

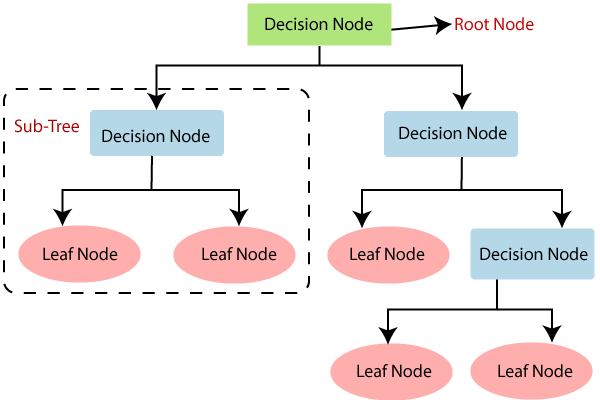
Các bước thực thực hiện thuật toán:

1. Xác định giá trị tham số K (số láng giềng gần nhất)
2. Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng cách Euclid)
3. Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với điểm cần phân lớp.
4. Lấy tất cả các lớp của K láng giềng gần nhất đã xác định
5. Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho điểm cần phân lớp.

### 2.5.3 Cây quyết định

Cây quyết định là một trong những thuật toán học máy được sử dụng rộng rãi nhất do tính dễ hiểu của nó. Thuật toán này nhằm mục tiêu xây dựng các quy tắc hay luật lệ quyết định dựa theo cấu trúc cây với mỗi bộ luật tương ứng với nhánh lá của cây.

Decision Trees gồm 3 phần chính: 1 node gốc (root node), những node lá (leaf nodes) và các nhánh của nó (branches). Node gốc là điểm bắt đầu của cây quyết định và cả hai node gốc và node chứa câu hỏi hoặc tiêu chí để được trả lời. Nhánh biểu diễn các kết quả của kiểm tra trên nút. Ví dụ câu hỏi ở node đầu tiên yêu cầu câu trả lời là “yes” hoặc là “no” thì sẽ có 1 node con chịu trách nhiệm cho phản hồi là “yes”, 1 node là “no”.



Hình 2.10 Cấu trúc của cây quyết đinh.

Các bước thực hiện để tạo nên một cây quyết định:

1. Chọn thuộc tính có độ lợi thông tin (information gain) cao nhất để phân chia tập dữ liệu hiện tại.
2. Tạo một nút cho thuộc tính được chọn.
3. Phân chia tập dữ liệu hiện tại thành các tập con nhỏ hơn dựa trên giá trị của thuộc tính được chọn.
4. Lặp lại quá trình 1-3 cho mỗi tập con nhỏ hơn cho đến khi các tập con chỉ chứa các mẫu thuộc cùng một lớp.

Công thức tính entropy:

A picture containing font, white, text, graphics

Description automatically generated

Trong đó S là tập dữ liệu ban đầu, pi là xác xuất thứ của lớp con i.

Công thức tính độ lợi thông tin:

A black text on a white background

Description automatically generated with medium confidence

Trong đó Entropy(S) là hệ số entropy của tập dữ liệu S. Sv là tập con của S với giá trị thuộc tính A bằng v. S là tập dữ liệu ban đầu.

### 2.5.4 Rừng ngẫu nhiên

Rừng ngẫu nhiên là một thuật toán học có giám sát. Như bạn có thể thấy từ tên của nó, nó tạo ra một khu rừng một cách ngẫu nhiên. “Khu rừng” mà anh ta tạo ra là một tập hợp các cây quyết định, trong hầu hết các trường hợp, được đào tạo với phương pháp đóng bao . Ý tưởng chính của phương pháp đóng gói là sự kết hợp của các mô hình học tập làm tăng kết quả chung.Nói một cách đơn giản: thuật toán rừng ngẫu nhiên tạo ra một số cây quyết định và kết hợp chúng để có được một dự đoán chính xác hơn và ổn định hơn.

Thuật toán rừng ngẫu nhiên bổ sung thêm tính ngẫu nhiên cho mô hình khi xây dựng cây. Thay vì tìm kiếm tính năng tốt nhất khi phân vùng các nút, nó sẽ tìm kiếm tính năng tốt nhất trong một tập hợp con ngẫu nhiên của các tính năng. Quá trình này tạo ra rất nhiều sự đa dạng, thường dẫn đến việc tạo ra các mô hình tốt hơn

Các bước thực hiện để tạo nên một rừng ngẫu nhiên:

1. Chọn ngẫu nhiên một tập con của tập dữ liệu.
2. Xây dựng một cây quyết định trên tập con này.
3. Lặp lại các bước 1-2 cho số lượng cây quyết định được chọn trước.
4. Đưa ra dự đoán cho mỗi cây được tạo ra trong giai đoạn 2 và kết hợp các dự đoán này để có được kết quả cuối cùng1.

### 2.5.5 Mạng nơ ron tích chập

Giống với mạng nơ-ron (Neural Networks), mạng nơ-ron tích chập (Convolutional Neural Networks) cũng bao gồm nhiều lớp được kết nối với nhau. Từng lớp cũng bao gồm nhiều nơ-ron với các trọng số .Quá trình huấn luyện) của mạng nơ-ron tích chập cũng bao gồm 2 phần: foward propagation - lan truyền xuôi, duyệt qua mạng theo chiều thuận (sử dụng dữ liệu đầu vào tính toán đầu ra của từng lớp sử dụng các trọng số tương ứng) và backward propagation - lan truyền ngược, duyệt qua mạng theo chiều ngược (sử dụng các đầu ra đã tính toán và sai số tương ứng để cập nhật các tham số của từng lớp).

A picture containing text, screenshot, diagram, line

Description automatically generated

Hình 2.11 Một trình tự CNN để phân loại các chữ số viết tay

Lớp tích chập (Conv) chính là lớp chiếm phần lớn trong một mạng nơ-ron tích chập. Lớp tích chập sẽ tính toán đầu ra bằng cách trượt một kernel qua từng vùng của đầu vào và thực hiện tính toán với kernel trên vùng đấy. Ví dụ ở hình 2 đầu vào là một ma trận 3×3 và một kernel 2×2, kernel sẽ trượt dần trên từng vùng từ trái sang phải, từ trên xuống dưới rồi thực hiện phép nhân từng phần tử tương ứng và tính tổng để nhận được kết quả đầu ra. Ở vùng đầu tiên (tô xanh ở Hình 2) khi trượt kernel qua sẽ thực hiện phép tính 0x0 + 1x1 + 2x3 + 3x4 = 19, các vùng tiếp theo tính tương tự.

A picture containing screenshot, line, diagram, number

Description automatically generated

Hình 2.12 Ví vụ về phép tích chập trên ma trận 3x3 với kernel 2x2

Lớp gộp (pooling layer) thường được sử dụng giữa các lớp tích chập dùng để giảm kích thước của dữ liệu mà vẫn giữ được các thuộc tính quan trọng. Đồng thời, việc giảm kích thước dữ liệu cũng giúp việc tính toán dễ dàng hơn. Khác với lớp tích chập, lớp gộp là lớp không có tham số khi lớp này không có kernel mang tham số mà chỉ có những cửa sổ gộp. Cửa số gộp cũng trượt trên dữ liệu giống như lớp tích chập, quy tắc sải bước (stride) và đệm (padding). Chỉ khác một điểm cửa sổ gộp không mang tham số, và các phép toán thực hiện ở lớp gộp thường là: gộp cực đại (max pooling) và gộp trung bình (average pooling). Gộp cực đại là giữ lại phần tử có giá trị lớn nhất của vùng mà cửa sổ gộp trượt qua, còn gộp trung bình là tính trung bình tất cả các phần tử trong vùng. Hình 10 minh họa rõ hơn về cách hoạt động của hai loại gộp này.

A picture containing rectangle, square

Description automatically generated

Hình 2.13 Max pooling với cửa sổ gộp có kích thước 3x3 với stride = 1.

A picture containing text, diagram, line, screenshot

Description automatically generated

Hình 2.14 Hai loại gộp với kích thước cửa sổ gộp là 2x3, stride = 2.

Lớp kết nối đầy đủ (fully connected layer) được dùng để học các tổ hợp phi tuyến tính từ các đặc trưng trích xuất từ kết quả ma trận tích chập đầu ra. Lớp kết nối đầy đủ thường nằm cuối của mạng nơ ron tích chập để tính toán đầu ra tương ứng với mô hình.

### 2.5.6 Partial Least Squares Regression (PLSR

Partial Least Squares Regression (PLSR) là một phương pháp hồi quy đa biến được sử dụng để dự đoán một biến phụ thuộc dựa trên một tập các biến độc lập (đặc trưng) liên quan. PLSR kết hợp các phép tách tham số và phân tích thành phần chính (PCA) để tạo ra các thành phần mới từ ma trận đặc trưng ban đầu và ma trận biến phụ thuộc.Dưới đây là cách hoạt động của PLSR:

1. Tạo thành phần chính (PCs): PLSR bắt đầu bằng việc tạo ra các thành phần chính từ ma trận đặc trưng ban đầu. Quá trình này tương tự như phân tích thành phần chính (PCA) và nhằm giảm chiều dữ liệu ban đầu thành các thành phần chính mới. Tuy nhiên, khác với PCA, PLSR đồng thời sử dụng thông tin về biến phụ thuộc để xác định các thành phần chính này.
2. Tạo các thành phần PLSR: Sau khi tạo thành phần chính, PLSR bắt đầu tạo ra các thành phần PLSR từ cả ma trận đặc trưng và biến phụ thuộc. Quá trình này tương tự như việc tìm các trục hợp tuyến tính giữa ma trận đặc trưng và ma trận biến phụ thuộc để tạo ra các thành phần mới. Mỗi thành phần PLSR được tạo ra dựa trên mối quan hệ tuyến tính giữa các thành phần chính tương ứng.
3. Tối ưu hồi quy: Khi đã có các thành phần PLSR, PLSR sử dụng phương pháp hồi quy tối thiểu bình phương để tìm một mô hình hồi quy tốt nhất cho biến phụ thuộc. Quá trình này tối thiểu hóa sai số bình phương giữa giá trị dự đoán và giá trị thực tế của biến phụ thuộc.

# Chương 3: ĐÁNH GIÁ CHẤT LƯỢNG CÁ DỰA TRÊN PHỔ CẬN HỒNG NGOẠI

## **3.1 Cách thu thập dữ liệu phổ cận hồng ngoại của cá**

Đề tài này khảo sát 3 loại cá là cá nục, cá thu và cá ngừ. Dữ liệu phổ NIR của các loại cá được thu thập bằng thiết bị DLP NIRscan Nano EVM (mô tả ở phần 2.2) và phần mềm đi kèm của máy. Máy quét NIR được kết nối với một máy tính thông qua cáp USB. Mẫu cá cần đo phổ được đặt sát mặt kính sa-phia của thiết bị. Cách bố trí này (Hình 3.1) được giữ nguyên trong tất cả các lần thu thập dữ liệu. Các mẫu cá sẽ được thu thập tại cảng cá. Để thay đổi nồng độ Ure trong cá thì cá mới mua về sẽ được tẩm với Ure trộn với đá xay theo tỷ lệ 1% trong thùng xốp, sau mỗi 0 giờ, 3 giờ, 6 giờ, 9 giờ, 12 giờ cá sẽ được lấy ra và đo phổ NIR. Ngoài ra, để khảo sát tính độc lập của các phương pháp đề xuất theo vị trí đo trên mẫu cá thì mỗi mẫu cá sẽ được đo tại mười hai vị trí khác nhau, trong đó tám vị trí không cần phá hủy mẫu gồm: gáy, lưng, đuôi, bụng, mắt 1, mắt 2, mang 1, mang 2 và bốn vị trí phải phá hủy mẫu gồm: gáy, lưng, đuôi, bụng và mỗi vị trí sẽ đo 5 lần. Hình 3.2 thể hiện sự khác biệt rõ ràng của phổ hấp thụ khi được đo ở các vị đo khác nhau của cùng một mẫu cá.

A picture containing person, tool, cutting, indoor

Description automatically generated

Hình 3.1 Cách thu thập dữ liệu phổ NIR.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Hình 3.2 Phổ hấp thụ của 12 vị trí đo khác nhau của mẫu cá 650.

Trong mỗi lần đo, mẫu cá hấp thụ một lượng ánh sáng cận hồng ngoại phát ra từ máy quét, phần ánh sáng không bị hấp thụ được phản xạ vào thấu kính của máy quét và được mô-đun xử lý quang học chuyển thành phổ chứa 228 bước sóng trong dải từ 900 nm đến 1700 nm. Dữ liệu phổ được truyền từ thiết bị đo đến máy tính bằng cáp USB dưới dạng các tệp có định dạng .csv chứa loại phổ hấp thụ (absorption). Về mặt vật lý, phổ hấp thụ cho biết mức độ các phân tử vật chất trong cá hấp thụ sóng điện từ (ở đây là ánh sáng cận hồng ngoại), vì vậy nó có thể phản ánh độ tươi cũng như độ độc tố của cá ở mức độ nào đó.

## **3.2 Mô hình tổng quan của bài toán**

Dữ liệu phổ NIR thô sẽ được xử lý các giá trị không phải số, hay các số vô hạn bằng cách lấy giá trị trung bình của 4 giá trị bên cạnh. Dữ liệu sau đó được chia thành: tập huấn luyện (80%), tập dữ liệu kiểm thử được chia 20% và tất cả dữ liệu phổ của một mẫu cá sẽ nằm trong tập kiểm thử để tránh bị rò rỉ dữ liệu. Sau đó dữ liệu huấn luyện sẽ được cân bằng dữ liệu bằng phương pháp SMOTE (ngoại trừ dữ liệu kiểm thử). Sau đó tổng hợp dữ liệu huấn luyện đã cân bằng và dữ liệu kiểm thử thành một tập dữ liệu lớn để tìm các bước sóng tối ưu cho từng thuật toán. Tập dữ liệu huấn luyện, kiểm thử sẽ được chuẩn hóa bằng các kĩ thuật: chuẩn hóa minmax, robust, MSC, SNV. Sau khi được chuẩn hóa dữ liệu sẽ được đưa qua bộ lọc Savitzky-Golay (SG) độ dài của cửa sổ bộ lọc được đặt bằng 25 và bậc của đa thức được đặt bằng 5. Dữ liệu đã được làm trơn và lọc bớt nhiễu sẽ được dùng để huấn luyện các mô hình học máy và kiểm thử chúng.

A picture containing text, diagram, line, screenshot

Description automatically generated

Hình 3.3 Mô hình tổng quan của bài toán

## **3.3 Đánh giá nồng độ Ure có trong cá**

### 3.3.1 Giới thiệu

Hiện nay trên thị trường đặc biệt là ở những nơi xa biển như vùng, nông thôn, miền núi để giữ cho cá và hải sản được tươi lâu hơn và đẹp mắt hơn người bán đã sử dụng phân đạm ure để ướp cá. Ure là phân bón hóa học dùng trong nông nghiệp, có tác dụng kìm hãm sự phát triển của vi khuẩn. Vì giá thành rất rẻ nên không ít người kinh doanh thủy hải sản tươi sống đã dùng phân ure nhằm giữ cho thực phẩm tươi lâu không bị ươn thối. Việc làm này tiềm ẩn rất nhiều nguy cơ gây hại cho sức khỏe cho người tiêu dùng.

Khi ăn phải cá có dư lượng ure cao thì người ăn có thể bị ngộ độc cấp tính với các biểu hiện đau bụng, buồn nôn, tiêu chảy, chóng mặt… thậm chí có thể dẫn đến tử vong. Dù với hàm lượng ít, ure khi vào cơ thể sẽ tích tụ về lâu dài, gây ngộ độc mạn tính, biểu hiện là thường đau đầu không rõ nguyên nhân, mất ngủ, giảm trí nhớ, ảnh hưởng đến gan, thận, biếng ăn, suy nhược cơ thể. LOD (Level Of Detection) của ure trong đề tài này là **1000** (mg/kg).

### 3.3.2 Thống kê dữ liệu

Trong đề tài này, chúng tôi đã thực hiện phân loại nồng độ Ure của 3 loại cá là cá nục, cá thu và cá ngừ. Mỗi loại cá đều được mua từ cùng một cảng cá để đảm bảo chúng có cùng nguồn gốc vì yếu tố này có thể ảnh hưởng một số đến dữ liệu quang phổ. Mỗi mẫu cá sẽ được sử dụng thiết bị đo DLP NIRscan Nano EVM để quét phổ tại 12 vị trí trên một mẫu cá, mỗi vị trí 5 lần.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Loại cá | Số mẫu cá | Số mẫu phổ dưới ngưỡng Ure | Số mẫu phổ trên ngưỡng Ure | Tổng số mẫu phổ |
| Nục 2 | 57 | 2405 | 1020 | 3425 |
| Ngừ 1 | 52 | 2760 | 360 | 3120 |
| Thu 1 | 26 | 1185 | 360 | 1545 |
| Nục 1 | 6 | 0 | 360 | 360 |
| **Tổng** | **141** | **6350** | **2100** | **8450** |

Bảng 3. 1 Số lượng mẫu cá và phổ hấp thụ NIR dùng để phân loại nồng độ Ure.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Hình 3.4 Phân bố giá trị của nồng độ Ure trong cá

### 3.3.3 Tiền xử lý dữ liệu

Dữ liệu cho bài toán đang bị mất cân bằng, số lượng mẫu phổ dưới ngưỡng LOD của Ure là 6350 (75,1%) và số lượng mẫu phổ vượt ngưỡng LOD là 2100 (24.9%) như Hình 10, vì vậy tôi đề xuất xử dụng phương pháp SMOTE đã được trình bày trong phần 2.4 để khắc phục tình trạng này.

A picture containing text, diagram, screenshot, plot

Description automatically generated

Hình 3.5 Phân bố mẫu phổ có nồng độ Ure trên và dưới LOD của cả tập dữ liệu.

Phương pháp SMOTE chỉ được sử dụng để chia dữ liệu cho tập huấn luyện, còn tập dữ liệu kiểm thử sẽ là các dữ liệu thật và đã được chia cân bằng. Mỗi một con cá bao gồm các mẫu phổ của nó sẽ được chia vào tập huấn luyện hoặc kiểm thử.

A white background with black text

Description automatically generated with low confidence

Hình 3.6 Phân bố mẫu phổ có nồng độ Ure trên và dưới LOD sau khi cân bằng SMOTE của tập huấn luyện.

A picture containing screenshot, diagram, plot, text

Description automatically generated

Hình 3.7 Phân bố mẫu phổ có nồng độ Ure trên và dưới LOD của tập kiểm thử.

Kỹ thuật tiền xử lý chuẩn hoá phổ được áp dụng lên phổ NIR. Việc chuẩn hoá phổ được thực hiện nhằm để loại bỏ các biến thiên trong dữ liệu phổ do kích thước phân tử của mẫu quả và sự tán xạ ánh sáng gây ra. Các kĩ thuật chuẩn hóa phổ trong đề tài này bao gồm: chuẩn hóa Min-Max, chuẩn hóa Robust, MSC, SNV. Một ví dụ về phổ trước và sau khi chuẩn hóa được thể hiện trên Hình 3.8 và 3.9.

A picture containing text, diagram, plot, line

Description automatically generated

Hình 3.8 Minh họa phổ NIR gốc.

A picture containing text, diagram, plot, line

Description automatically generated

Hình 3.9 Phổ chuẩn hóa Min-Max, Robust, MSC, SNV

### 3.3.4 Trích xuất đặt trưng

Các phổ NIR đã tiền xử lý của mỗi loại cá sẽ được đưa qua bộ lọc Savitzky-Golay (SG) để làm trơn phổ chuẩn hóa mà không làm sai lệch xu hướng của nó, sau đó tính các Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ đã làm trơn để thu được vectơ đặc trưng mở rộng. Đối với bộ lọc SG được sử dụng trong đề tài này, độ dài của cửa sổ bộ lọc được đặt bằng 25 và bậc của đa thức được đặt bằng 5. Hình 3.9 cho thấy một ví dụ về phổ đã tiền xử lý, làm trơn SG và các Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ đã tiền xử lý của các mẫu cá.

**A picture containing text, diagram, plot, line

Description automatically generated**

Hình 3.10 Minh họa phổ đã chuẩn hóa SNV sau đó qua bộ lọc SG, bộ lọc SG lấy đạo hàm bậc 1 và bậc 2.

|  |  |  |
| --- | --- | --- |
| **Loại đặc trưng** | **Kích thước vector** | **Mô tả** |
| x\_sg0\_msc | 228 | Phổ được chuẩn hóa MSC và làm mượt SG |
| x\_sg0\_snv | 228 | Phổ được chuẩn hóa SNV và làm mượt SG |
| x\_sg0\_minmax | 228 | Phổ được chuẩn hóa min max và làm mượt SG |
| x\_sg0\_robust | 228 | Phổ được chuẩn hóa robust và làm mượt SG |
| x\_sg1\_msc | 228 | Làm mượt SG và lấy đạo hàm 1 của phổ được chuẩn hóa MSC |
| x\_sg1\_snv | 228 | Làm mượt SG và lấy đạo hàm bậc 1 của phổ được chuẩn hóa SNV |
| x\_sg1\_minmax | 228 | Làm mượt SG và lấy đạo hàm bậc 1 của phổ được chuẩn hóa min max |
| x\_sg1\_robust | 228 | Làm mượt SG và lấy đạo hàm bậc 1 của phổ được chuẩn hóa robust |
| x\_sg2\_msc | 228 | Làm mượt SG và lấy đạo hàm bậc 2 của phổ được chuẩn hóa MSC |
| x\_sg2\_snv | 228 | Làm mượt SG và lấy đạo hàm bậc 2 của phổ được chuẩn hóa SNV |
| x\_sg2\_minmax | 228 | Làm mượt SG và lấy đạo hàm bậc 2 của phổ được chuẩn hóa min max |
| x\_sg2\_robust | 228 | Làm mượt SG và lấy đạo hàm bậc 2 của phổ được chuẩn hóa robust |
| x\_sg1\_sg2\_msc | 456 | Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ được chuẩn hóa MSC |
| x\_ sg1\_sg2\_snv | 456 | Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ được chuẩn hóa SNV |
| x\_sg1\_sg2\_minmax | 456 | Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ được chuẩn hóa min max |
| x\_ sg1\_sg2\_robust | 456 | Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ được chuẩn hóa robust |

Bảng 3.2 Mô tả các loại vecto đặc trưng để đánh giá chất lượng cá

### 3.3.5 Các thuật toán học máy có giám sát

Trong giai đoạn huấn luyện, các đặc trưng phổ của các mẫu cá được sử dụng làm dữ liệu huấn luyện và hai mức độ trên hoặc dưới LOD được sử dụng làm nhãn. Đề tài này sử dụng cả phương pháp tiếp cận học máy truyền thống và học sâu để xây dựng các mô hình phân loại và so sánh hiệu suất của chúng. Đối với phương pháp tiếp cận học máy truyền thống, 04 thuật toán đã được thử nghiệm bao gồm cây quyết định (decision tree), rừng ngẫu nhiên (random forest) máy vector hỗ trợ (Support vector machine), Partial Least Squares Regression (PLSR) và k-láng giềng gần nhất (k-nearest neighbors).

Đối với mô hình hồi quy PLSR tập dữ liệu huấn luyện sẽ không sử dụng phương pháp tăng cường dữ liệu và dữ liệu của nồng độ Ure sẽ được giảm dải biến thiên bằng phương pháp SNV.

**A picture containing text, screenshot, diagram, number

Description automatically generated**

Hình 3.11 Mô hình tổng quan của thuật toán PLSR.

|  |  |  |
| --- | --- | --- |
|  | Số lượng mẫu | Trung bình ± Độ lệch chuẩn |
| Tập huấn luyện | 6770 | 4.198e-17 ± 0.999 |
| Tập kiểm thử | 1680 | 3.383e-17 ± 1.0 |

Bảng 3.3 Giá trị tham chiếu của nồng độ Ure được phân tích bằng phương pháp tiêu chuẩn.

Đối với phương pháp tiếp cận học sâu, tôi đề xuất xử dụng mạng nơ-ron tích chập (CNN) có kiến trúc được minh họa trong Hình 3.12.Vector đầu vào chứa N nơ-ron đại diện cho vector đặc trưng có kích thước Nx1 (N có thể là 228 hoặc 456 tùy thuộc vào loại đặc trưng đã chọn). Mô hình CNN bao gồm ba lớp tích chập , mỗi lớp được theo sau bởi các lớp pooling và dropout. Các lớp tích chập có lần lượt là 16, 32, 64 bộ lọc với nhân kích thước lần lượt là 3x1, 5x1, 11x1 và hàm kích hoạt ReLU. Trong các lớp pooling, max pooling được sử dụng với kích thước pool là 2x1 và độ dài stride là 2. Các lớp dropout, rate được sử dụng là 0.01. Một lớp làm phẳng được sử dụng để chuyển dữ liệu 2 chiều thành 1 chiều. Sau đó được đưa qua 2 lớp Dense có 256 nơ-ron hàm kích hoạt ReLu và Dense 128 nơ-ron hàm kích hoạt ReLu, 1 lớp Dropout với rate 0.01 và cuối cùng là lớp Dense chứa 2 nơ-ron với hàm kích hoạt softmax được sử dụng để dự đoán nhãn đầu ra.

A picture containing text, screenshot, diagram, line

Description automatically generated

Hình 3.12 Mô hình CNN đề xuất.

### 3.3.6 Kết quả thực nghiệm

#### *3.3.6.1 Kết quả của mô hình CNN đề xuất*

Các thử nghiệm được thực hiện bằng cách sử dụng Keras, một API học sâu chạy trên nền tảng máy học TensorFlow. Bộ dữ liệu quang phổ thu thập được được chia ngẫu nhiên thành ba tập con để huấn luyện, xác nhận và kiểm thử các mô hình CNN theo tỷ lệ tương ứng là 60%, 20%, 20%. Trong mỗi tập con, số lượng mẫu giữa các loại trái cây cũng như giữa các mức độ tươi được giữ cân bằng. Trong mô hình CNN đã đề xuất này, tôi đã sử dụng trình tối ưu hóa Adam với tốc độ học là 0,0005. Tốc độ học được thiết lập để giảm đi 20% nếu giá trị của hàm mất mát không giảm trong 20 vòng lặp huấn luyện và mô hình sẽ dừng lại nếu hàm mất mát không giảm sau 50 vòng lặp huấn luyện. Hình 3.13 cho thấy hàm mất mát và độ chính xác của mô hình CNN thay đổi như thế nào trên các tập huấn luyện và tập xác nhận qua các vòng lặp huấn luyện (epoch) khi thử với vectơ đặc trưng x\_sg0\_snv. Trong trường hợp này, tôi đã dừng quá trình huấn luyện sau 200 epoch để ngăn chặn hiện tượng quá khớp vì hàm mất mát trên tập xác nhận đã hội tụ tại thời điểm này. Tôi đã áp dụng cùng một quy trình như trên để huấn luyện mô hình CNN cho các đặc trưng còn lại.

**A picture containing plot, text, diagram, line

Description automatically generated**

Hình 3. 13 Biến thiên của hàm mất mát và độ chính xác của mô hình CNN trên tập huấn luyện và xác nhận.

Bảng 3.4 thể hiện kết quả thực nghiệm đầy đủ của mô hình CNN đề xuất trên các loại vectơ đặc trưng. Mô hình đạt được hiệu suất tốt nhất khi sử dụng vectơ đặc trưng x\_sg0\_msc và x\_sg1\_sg2\_robust (độ chính xác là 65,5% trên tập kiểm thử).

|  |  |  |  |
| --- | --- | --- | --- |
| Vector đặc trưng | Tập huấn luyện | Tập xác nhận | Tập kiểm thử |
| x\_sg0\_msc | 99.2 | 91.3 | **65.5** |
| x\_sg0\_snv | 99.3 | 92.0 | 64.9 |
| x\_sg0\_minmax | 98.7 | 93.6 | 63.8 |
| x\_sg0\_robust | 98.8 | 90.3 | 65.1 |
| x\_sg1\_msc | 99.7 | 92.7 | 64.8 |
| x\_sg1\_snv | 98.9 | 91.3 | 64.9 |
| x\_sg1\_minmax | 99.5 | 92.3 | 64.1 |
| x\_sg1\_robust | 99.3 | 90.6 | 65.3 |
| x\_sg2\_msc | 98.5 | 89.5 | 62.8 |
| x\_sg2\_snv | 98.7 | 89.3 | 62.7 |
| x\_sg2\_minmax | 97.1 | 87.6 | 62.9 |
| x\_sg2\_robust | 99.3 | 90.5 | 61.7 |
| x\_sg1\_sg2\_msc | 99.6 | 92.7 | 63.7 |
| x\_ sg1\_sg2\_snv | 99.4 | 92.3 | 64.7 |
| x\_sg1\_sg2\_minmax | 99.0 | 91.5 | 64.4 |
| x\_ sg1\_sg2\_robust | 99.3 | 92.8 | 65.5 |

Bảng 3.4 Độ chính xác của mô hình CNN khi sử dụng các loại vector đặc trưng.

#### *3.3.6.2 Kết quả của các mô hình học máy truyền thống*

Các thực nghiệm huấn luyện được thực hiện bằng cách sử dụng bộ công cụ học máy scikit-learn. Vì các tham số mô hình có thể ảnh hưởng đáng kể đến hiệu suất của các thuật toán học máy này nên các thủ tục tìm kiếm lưới (grid search) trên tập huấn luyện đã được cân bằng để tìm ra các tham số tối ưu cho mô hình. Bảng 3.5 liệt kê các tham số tối ưu sẽ được sử dụng cho từng mô hình học máy

|  |  |
| --- | --- |
| **Mô hình** | **Bộ tham số** |
| Cây quyết định | - Tiêu chuẩn đo: gini  - Độ sâu tối đa của cây: 5  - Số lượng mẫu tối thiểu tại một nút lá: 7 |
| Rừng ngẫu nhiên | - Số cây trong rừng: 210  - Tiêu chuẩn đo: gini  - Độ sâu tối đa của cây: 5  - Số lượng mẫu tối thiểu tại một nút lá: 7 |
| K láng giềng gần nhất | - Số lượng láng giềng: 29 |
| Máy vector hỗ trợ | - Kiểu nhân: rbf  - Tham số regularization (C): 1000 |
| Partial Least Squares Regression | - n-components: 35 |

Bảng 3.5 Tập hợp các tham số tối ưu cho từng mô hình học máy truyền thống.

Bảng 3.6 tổng kết độ chính xác phân lớp trên tập kiểm thử của mô hình tối ưu cho từng thuật toán học máy truyền thống theo các loại đặc trưng (có so sánh với mô hình CNN đề xuất). Có thể quan sát thấy rằng mô hình SVM giữ được độ chính xác cao nhất khi dùng đặc trưng x\_sg0\_snv (74,4%) và ổn định với các loại vector đặc trưng.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Vector đặt trưng | SVM | KNN | Decision Tree | Random Forest | CNN |
| x\_sg0\_msc | 73.5 | 66 | 64.2 | 68.9 | 65.5 |
| x\_sg0\_snv | **74.4** | 66.9 | 64.5 | 68.9 | 64.9 |
| x\_sg0\_minmax | 73.3 | 65.4 | 62.2 | 63.8 | 63.8 |
| x\_sg0\_robust | 72.3 | 64.8 | 63.6 | 67.5 | 65.1 |
| x\_sg1\_msc | 69.8 | 65.1 | 60.3 | 68.6 | 64.8 |
| x\_sg1\_snv | 68.7 | 64.5 | 58.9 | 68.9 | 64.9 |
| x\_sg1\_minmax | 68.9 | 64.3 | 63.8 | 66.5 | 64.1 |
| x\_sg1\_robust | 68.5 | 63.2 | 59.9 | 65.2 | 65.3 |
| x\_sg2\_msc | 69 | 60.3 | 62 | 66 | 62.8 |
| x\_sg2\_snv | 70 | 61.8 | 62.7 | 65.9 | 62.7 |
| x\_sg2\_minmax | 69.3 | 63.5 | 64.3 | 66.1 | 62.9 |
| x\_sg2\_robust | 69.8 | 59.9 | 58.4 | 65.8 | 61.7 |
| x\_sg1\_sg2\_msc | 69.5 | 64.5 | 58.7 | 68.2 | 63.7 |
| x\_ sg1\_sg2\_snv | 68.4 | 63.8 | 60.5 | 68.3 | 64.7 |
| x\_sg1\_sg2\_minmax | 68.8 | 64.5 | 64.3 | 66.5 | 64.4 |
| x\_ sg1\_sg2\_robust | 68.5 | 63.1 | 58.5 | 64.9 | 65.5 |

Bảng 3.6 Kết quả kiểm thử của các thuật toán học máy đề xuất.

|  |  |  |
| --- | --- | --- |
| Vector đặc trưng | MSE | R2 |
| x\_sg0\_msc | 0.87 | 0.13 |
| x\_sg0\_snv | 0.83 | 0.17 |
| x\_sg0\_minmax | 0.82 | 0.18 |
| x\_sg0\_robust | 0.82 | 0.18 |
| x\_sg1\_msc | 0.87 | 0.13 |
| x\_sg1\_snv | 0.84 | 0.16 |
| x\_sg1\_minmax | 0.82 | 0.18 |
| x\_sg1\_robust | 0.81 | 0.19 |
| x\_sg2\_msc | 0.86 | 0.14 |
| x\_sg2\_snv | 0.84 | 0.16 |
| x\_sg2\_minmax | 0.81 | 0.19 |
| x\_sg2\_robust | 0.82 | 0.18 |
| x\_sg1\_sg2\_msc | 0.87 | 0.13 |
| x\_ sg1\_sg2\_snv | 0.84 | 0.16 |
| x\_sg1\_sg2\_minmax | 0.82 | 0.18 |
| x\_ sg1\_sg2\_robust | 0.82 | 0.18 |

Bảng 3.7 Kết quả R2 và MSE của thuật toán PLSR cho dữ liệu Ure.

Từ bảng 3.7 ta có thể thấy R2 cao nhất là 0.19 khi sử dụng đặc trưng x\_sg2\_minmax để huấn luyện, tuy nhiên với R2 là 0.19 thì mô hình PLSR chỉ phù hợp với tập dữ liệu ở mức 19%.

Vì mô hình SVM có độ chính xác cao nhất nên tôi sẽ dùng mô hình SVM với đặc trưng x\_sg0\_snv để kiểm tra độ chính xác của tập kiểm thử với các vị trí đo khác nhau kết hợp với phương pháp repeated hold out và cross validation để xác định vị trí đo cho kết quả tốt nhất. Từ bảng 3.8 có thể thấy kết quả dự đoán tối nhất nằm ở vị trí ‘Ngoài da, đuôi’ với độ chính xác 78.3%.

|  |  |  |
| --- | --- | --- |
| **Vị trí đo** | **Repeated hold-out** | **Cross validation** |
| Ngoài da, gáy | 72.0 | 71.5 |
| Ngoài da, lưng | 66.2 | 71.7 |
| Ngoài da, đuôi | 71.0 | **78.3** |
| Ngoài da, bụng | 72.0 | 76.1 |
| Mang 1 | 61.4 | 63.7 |
| Mang 2 | 67.5 | 68.0 |
| Mắt 1 | 58.4 | 61.0 |
| Mắt 2 | 65.0 | 68.4 |
| Trong thịt, gáy | 63.6 | 66.1 |
| Trong thịt, lưng | 63.9 | 65.9 |
| Trong thịt, đuôi | 60.0 | 63.6 |
| Trong thịt, bụng | 60.4 | 60.9 |

Bảng 3.8 Độ chính xác trung bình của dữ liệu kiểm thử Ure ở từng vị trí đo khi dùng phương pháp đánh giá repeated hold-out và cross-validation.

**A screenshot of a graph

Description automatically generated with low confidence**

Hình 3.14 Ma trận nhầm lẫn tổng của 5 fold ở vị trí đo ngoài da, đuôi.

Từ ma trận nhận nhầm lẫn tổng ở vị trí ngoài da, đuôi trên tập kiểm thử của mô hình SVM ta thu được độ nhạy, độ đặc hiệu, độ chính xác như sau:

- Độ chính xác =

- Độ nhạy (sensitivity / recall) =

- Độ đặc hiệu (specificity) =

Trong đó:

- TP: Số mẫu phổ được dự đoán là dưới LOD và có nhãn thực là dưới LOD

- FP: Số mẫu phổ được dự đoán là dưới LOD và có nhãn thực là trên LOD

- TN: Số mẫu phổ được dự đoán là trên LOD và có nhãn thực là trên LOD

- FN: Số mẫu phổ được dự đoán là trên LOD và có nhãn thực là dưới LOD

Từ kết quả độ nhạy và độ đặc hiệu có thể thấy rằng mô hình dự đoán các mẫu phổ dưới LOD là rất tốt (82,8%), nhưng còn các mẫu phổ trên LOD rất cần dự đoán chính xác thì vẫn chưa tốt chỉ đạt 64,6%.

## **3.4 Đánh giá nồng độ Histamine có trong cá**

### 3.4.1 Giới thiệu

Histamine là một loại amin sinh học có tính chịu nhiệt cao, không bị phát hủy khi nấu chín, đông lạnh, hun khói, tiệt trùng thực phẩm. Trong tự nhiên, histamine được tạo thành từ kết quả của sự chuyển hóa từ histidine thành histamine bởi các vi khuẩn sản sinh ra men histidine decarboxylase. Khi cá còn sống các vi khuẩn này tồn tại trong mang cá, ruột cá và không gây hại cho cá. Nhưng khi cá chết, vi khuẩn sinh trưởng và lây lan vào thịt cá, sản xuất men chuyển hóa histidine thành histadine trong thịt cá.

Ngộ độc histamine xảy ra khi cơ thể ăn phải cá biển chứa một lượng histamine cao vượt quá khả năng chấp nhận được của cơ thể. Ở người có cơ địa nhạy cảm thì chỉ cần ăn phải một lượng histamine nhỏ đã có thể bị di ứng. Nếu lượng ăn vào từ 1500 mg - 4000 mg, người ăn có biểu hiện như nhức đầu, đau bụng, tiêu chảy, nổi ban. Trong trường hợp nặng có thể xuất hiện co thắt phế quản (khó thở), mạch nhanh, hạ huyết áp gây nguy hiểm tới tính mạng. Ngộ độc xảy ra sau khoảng 20 - 30 phút sau khi ăn cá biển chứ histamine. LOD (Level Of Detection) của histamine trong đề tài này là **100** (mg/kg)

### 3.4.2 Thống kê dữ liệu

Trong đề tài này, chúng tôi đã thực hiện phân loại nồng độ histamine của 3 loại cá là cá nục, cá thu và cá ngừ. Mỗi loại cá đều được mua từ cùng một cảng cá để đảm bảo chúng có cùng nguồn gốc vì yếu tố này có thể ảnh hưởng một số đến dữ liệu quang phổ.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Loại cá | Số mẫu cá | Mẫu phổ dưới ngưỡng Histamine | Mẫu phổ trên ngưỡng Histamine | Tổng số mẫu phổ |
| Ngừ 1 | 70 | 2755 | 1440 | 4195 |
| Nục 1 | 50 | 1200 | 1800 | 3000 |
| Ngừ 2 | 8 | 480 | 0 | 480 |
| Thu 1 | 30 | 1260 | 540 | 1800 |
| **Tổng** | **158** | **5695** | **3780** | **9475** |

Bảng 3.9 Số lượng mẫu cá và phổ hấp thụ NIR dùng để đánh giá nồng độ Histamine

A picture containing text, screenshot, diagram, plot

Description automatically generated

Hình 3.15 Phân bố giá trị của nồng độ Histamine trong cá

### 3.4.3 Tiền xử lý dữ liệu

Dữ liệu cho bài toán đang bị mất cân bằng, số lượng mẫu phổ dưới ngưỡng LOD của Ure là 5695 (60,1%) và số lượng mẫu phổ vượt ngưỡng LOD là 3780 (39.9%) như Hình 10, vì vậy tôi đề xuất xử dụng phương pháp SMOTE đã được trình bày trong phần 2.4 để khắc phục tình trạng này.

A picture containing screenshot, diagram, text, plot

Description automatically generated

Hình 3.16 Phân bố mẫu phổ Histamine trên và dưới LOD của cả tập dữ liệu.

Phương pháp SMOTE chỉ được sử dụng để chia dữ liệu cho tập huấn luyện, còn tập dữ liệu kiểm thử sẽ là các dữ liệu thật và đã được chia cân bằng. Mỗi một con cá bao gồm các mẫu phổ của nó sẽ được chia vào tập huấn luyện hoặc kiểm thử.

A white background with black text

Description automatically generated with low confidence

Hình 3.17 Phân bố mẫu phổ có nồng độ Histamine trên và dưới LOD sau khi cân bằng SMOTE của tập huấn luyện.

A picture containing screenshot, diagram, text, plot

Description automatically generated

Hình 3.18 Phân bố mẫu phổ có nồng độ Histamine trên và dưới LOD của tập kiểm thử.

Kỹ thuật tiền xử lý chuẩn hoá phổ được áp dụng lên phổ NIR. Việc chuẩn hoá phổ được thực hiện nhằm để loại bỏ các biến thiên trong dữ liệu phổ do kích thước phân tử của mẫu quả và sự tán xạ ánh sáng gây ra. Các kĩ thuật chuẩn hóa phổ trong đề tài này bao gồm: chuẩn hóa Min-Max, chuẩn hóa Robust, MSC, SNV. Một ví dụ về phổ trước và sau khi chuẩn hóa được thể hiện trên Hình 3.19 và 3.20.

A picture containing text, plot, diagram, line

Description automatically generated

Hình 3.19 Minh họa phổ NIR gốc trong tập dữ liệu Histamine.

A picture containing text, diagram, plot, line

Description automatically generated

Hình 3.20 Phổ chuẩn hóa Min-Max, Robust, MSC, SNV của các mẫu phổ trong tập dữ liệu Histamine.

### 3.4.4 Trích xuất đặt trưng

Các phổ NIR đã tiền xử lý của mỗi loại cá sẽ được đưa qua bộ lọc Savitzky-Golay (SG) để làm trơn phổ chuẩn hóa mà không làm sai lệch xu hướng của nó, sau đó tính các Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ đã làm trơn để thu được vectơ đặc trưng mở rộng. Đối với bộ lọc SG được sử dụng trong đề tài này, độ dài của cửa sổ bộ lọc được đặt bằng 25 và bậc của đa thức được đặt bằng 5. Hình 3.21 cho thấy một ví dụ về phổ đã tiền xử lý, làm trơn SG và các Làm mượt SG và lấy đạo hàm bậc 1 và bậc 2 của phổ đã tiền xử lý của các mẫu cá.

A picture containing text, diagram, plot, line

Description automatically generated

Hình 3.21 Minh họa phổ đã chuẩn hóa robust sau đó qua bộ lọc SG, bộ lọc SG lấy đạo hàm bậc 1 và bậc 2.

### 3.4.5 Các thuật toán học máy có giám sát

Trong giai đoạn huấn luyện, các đặc trưng phổ của các mẫu cá được sử dụng làm dữ liệu huấn luyện và hai mức độ trên hoặc dưới LOD được sử dụng làm nhãn. Đề tài này sử dụng cả phương pháp tiếp cận học máy truyền thống và học sâu để xây dựng các mô hình phân loại và so sánh hiệu suất của chúng. Đối với phương pháp tiếp cận học máy truyền thống, 04 thuật toán đã được thử nghiệm bao gồm cây quyết định (decision tree), rừng ngẫu nhiên (random forest) máy vector hỗ trợ (Support vector machine) , Partial Least Squares Regression (PLSR) và k-láng giềng gần nhất (k-nearest neighbors).

Đối với mô hình hồi quy PLSR tập dữ liệu huấn luyện sẽ không sử dụng phương pháp tăng cường dữ liệu và dữ liệu của nồng độ Histamine sẽ được giảm dải biến thiên bằng phương pháp SNV.

|  |  |  |
| --- | --- | --- |
|  | Số lượng mẫu | Trung bình ± Độ lệch chuẩn |
| Tập huấn luyện | 7680 | -6.333e-17 ± 1.0 |
| Tập kiểm thử | 1795 | -7.401e-17 ± 1.0 |

Bảng 3.10 Giá trị tham chiếu của nồng độ Histamine được phân tích bằng phương pháp tiêu chuẩn.

Đối với phương pháp tiếp cận học sâu, tôi đề xuất xử dụng mạng nơ-ron tích chập (CNN) có kiến trúc được minh họa trong Hình 3.22. Vector đầu vào chứa N nơ-ron đại diện cho vector đặc trưng có kích thước Nx1 (N có thể là 228 hoặc 456 tùy thuộc vào loại đặc trưng đã chọn). Mô hình CNN bao gồm bốn lớp tích chập, mỗi lớp được theo sau bởi các lớp pooling và dropout. Các lớp tích chập có lần lượt là 16, 32, 64 bộ lọc với nhân kích thước lần lượt là 3x1, 5x1, 11x1 và hàm kích hoạt ReLU. Trong các lớp pooling, max pooling được sử dụng với kích thước pool là 2x1 và độ dài stride là 2. Các lớp dropout, rate được sử dụng là 0.01. Một lớp làm phẳng được sử dụng để chuyển dữ liệu 2 chiều thành 1 chiều. Sau đó được đưa qua 2 lớp Dense có 256 nơ-ron hàm kích hoạt ReLu và Dense 128 nơ-ron hàm kích hoạt ReLu, 1 lớp Dropout với rate 0.01 và cuối cùng là lớp Dense chứa 2 nơ-ron với hàm kích hoạt softmax được sử dụng để dự đoán nhãn đầu ra.

A picture containing text, screenshot, diagram, line

Description automatically generated

Hình 3.22 Mô hình CNN đề xuất để đánh giá nồng độ Histamine có trong cá.

### 3.4.6 Kết quả thực nghiệm

#### *3.4.6.1 Kết quả của mô hình CNN đề xuất*

Các thử nghiệm được thực hiện bằng cách sử dụng Keras, một API học sâu chạy trên nền tảng máy học TensorFlow. Bộ dữ liệu quang phổ thu thập được được chia ngẫu nhiên thành ba tập con để huấn luyện, xác nhận và kiểm thử các mô hình CNN theo tỷ lệ tương ứng là 60%, 20%, 20%. Trong mỗi tập con, số lượng mẫu giữa các loại trái cây cũng như giữa các mức độ tươi được giữ cân bằng. Trong mô hình CNN đã đề xuất này, tôi đã sử dụng trình tối ưu hóa Adam với tốc độ học là 0,0005. Tốc độ học được thiết lập để giảm đi 20% nếu giá trị của hàm mất mát không giảm trong 20 vòng lặp huấn luyện và mô hình sẽ dừng lại nếu hàm mất mát không giảm sau 50 vòng lặp huấn luyện. Hình 3.23 cho thấy hàm mất mát và độ chính xác của mô hình CNN thay đổi như thế nào trên các tập huấn luyện và tập xác nhận qua các vòng lặp huấn luyện (epoch) khi thử với vectơ đặc trưng x\_ sg0\_snv. Trong trường hợp này, tôi đã dừng quá trình huấn luyện sau hơn 200 epoch để ngăn chặn hiện tượng quá khớp vì hàm mất mát trên tập xác nhận đã hội tụ tại thời điểm này. Tôi đã áp dụng cùng một quy trình như trên để huấn luyện mô hình CNN cho các đặc trưng còn lại.

A close-up of a graph

Description automatically generated with low confidence

Hình 3.23 Biến thiên của hàm mất mát và độ chính xác của mô hình CNN trên tập huấn luyện và xác nhận.

Bảng 3.11 thể hiện kết quả thực nghiệm đầy đủ của mô hình CNN đề xuất trên các loại vectơ đặc trưng. Mô hình đạt được hiệu suất tốt nhất khi sử dụng vectơ đặc trưng x\_sg0\_snv (độ chính xác là 73.611% trên tập kiểm thử).

|  |  |  |  |
| --- | --- | --- | --- |
| Vector đặc trưng | Tập huấn luyện | Tập xác nhận | Tập kiểm thử |
| x\_sg0\_msc | 99.4 | 94.7 | 73.1 |
| x\_sg0\_snv | 99.8 | 96.2 | **73.6** |
| x\_sg0\_minmax | 98.9 | 94.6 | 72.4 |
| x\_sg0\_robust | 99.7 | 94.7 | 72.5 |
| x\_sg1\_msc | 97.0 | 88.5 | 70.6 |
| x\_sg1\_snv | 99.8 | 91.0 | 71.8 |
| x\_sg1\_minmax | 98.0 | 89.6 | 71.7 |
| x\_sg1\_robust | 98.4 | 90.1 | 70.9 |
| x\_sg2\_msc | 97.2 | 89.3 | 71.8 |
| x\_sg2\_snv | 98.9 | 89.2 | 72.4 |
| x\_sg2\_minmax | 95.7 | 87.9 | 70.6 |
| x\_sg2\_robust | 95.4 | 89.2 | 70.7 |
| x\_sg1\_sg2\_msc | 98.7 | 90.5 | 72.7 |
| x\_ sg1\_sg2\_snv | 99.6 | 90.3 | 71.2 |
| x\_sg1\_sg2\_minmax | 99.6 | 91.8 | 72.8 |
| x\_ sg1\_sg2\_robust | 99.2 | 91.5 | 72.8 |

Bảng 3.11 Độ chính xác của mô hình CNN trên các loại vector đặc trưng.

#### 3.4.6.2 Kết quả của các mô hình học máy truyền thống

Các thực nghiệm huấn luyện được thực hiện bằng cách sử dụng bộ công cụ học máy scikit-learn. Vì các tham số mô hình có thể ảnh hưởng đáng kể đến hiệu suất của các thuật toán học máy này nên các thủ tục tìm kiếm lưới (grid search) trên tập huấn luyện đã được cân bằng để tìm ra các tham số tối ưu cho các mô hình. Bảng 3.12 liệt kê các tham số tối ưu sẽ được sử dụng cho từng mô hình học máy.

|  |  |
| --- | --- |
| **Mô hình** | **Bộ tham số** |
| Cây quyết định | - Tiêu chuẩn đo: gini  - Độ sâu tối đa của cây: 17  - Số lượng mẫu tối thiểu tại một nút lá: 18 |
| Rừng ngẫu nhiên | - Số cây trong rừng: 220  - Tiêu chuẩn đo: gini  - Độ sâu tối đa của cây: 17  - Số lượng mẫu tối thiểu tại một nút lá: 18 |
| K láng giềng gần nhất | - Số lượng láng giềng: 11 |
| Máy vector hỗ trợ | - Kiểu nhân: rbf  - Tham số regularization (C): 1000 |
| Partial Least Squares Regression | - n-components: 29 |

Bảng 3.12 Tập hợp các tham số tối ưu cho từng mô hình học máy truyền thống dùng để đánh giá Histamine.

|  |  |  |
| --- | --- | --- |
| Vector đặc trưng | MSE | R2 |
| x\_sg0\_msc | 1.23 | -0.23 |
| x\_sg0\_snv | 0.73 | 0.27 |
| x\_sg0\_minmax | 0.69 | **0.31** |
| x\_sg0\_robust | 0.76 | 0.24 |
| x\_sg1\_msc | 0.78 | 0.22 |
| x\_sg1\_snv | 0.73 | 0.27 |
| x\_sg1\_minmax | 0.69 | 0.31 |
| x\_sg1\_robust | 0.75 | 0.25 |
| x\_sg2\_msc | 0.79 | 0.21 |
| x\_sg2\_snv | 0.73 | 0.27 |
| x\_sg2\_minmax | 0.7 | 0.3 |
| x\_sg2\_robust | 0.75 | 0.25 |
| x\_sg1\_sg2\_msc | 0.79 | 0.21 |
| x\_ sg1\_sg2\_snv | 0.73 | 0.27 |
| x\_sg1\_sg2\_minmax | 0.7 | 0.3 |
| x\_ sg1\_sg2\_robust | 0.75 | 0.25 |

Bảng 3.13 Kết quả R2 và MSE của thuật toán PLSR cho dữ liệu Histamine.

Từ bảng 3.13 ta có thể thấy R2 cao nhất là 0.31 khi sử dụng đặc trưng x\_sg0\_minmax để huấn luyện, tuy nhiên với R2 là 0.31 thì mô hình PLSR chỉ phù hợp với tập dữ liệu ở mức 31%.

Bảng 3.14 tổng kết độ chính xác phân lớp trên tập kiểm thử của mô hình tối ưu cho từng thuật toán học máy truyền thống theo các loại đặc trưng (có so sánh với mô hình CNN đề xuất). Có thể quan sát thấy rằng mô hình SVM giữ được độ chính xác cao nhất khi dùng đặc trưng x\_sg1\_snv (74,1%) và ổn định với các loại vector đặc trưng. Trong khi đó mô hình CNN đề xuất vẫn có độ chính xác khá cao, đạt độ chính xác 73,6% khi được huấn luyện trên đặc trưng x\_sg0\_snv.

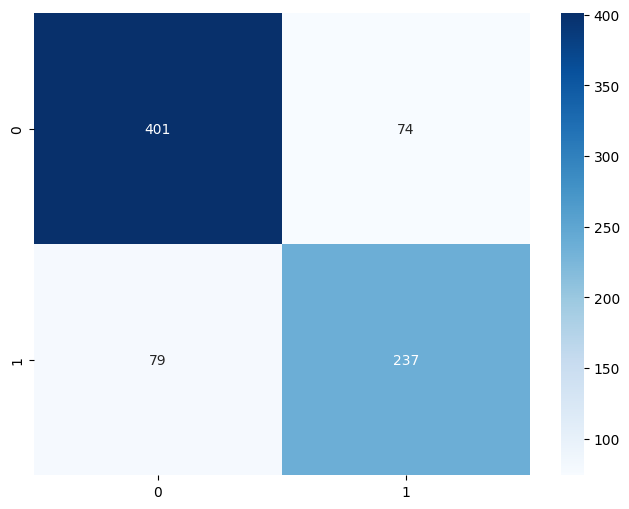
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Vector đặt trưng | SVM | KNN | Decision Tree | Random Forest | CNN |
| x\_sg0\_msc | 69.5 | 65.9 | 67.7 | 72 | 73.1 |
| x\_sg0\_snv | 71.9 | 65.6 | 64.3 | 71.7 | 73.6 |
| x\_sg0\_minmax | 71.4 | 64.8 | 63.4 | 68.2 | 72.4 |
| x\_sg0\_robust | 70.8 | 65.7 | 67.3 | 71.6 | 72.5 |
| x\_sg1\_msc | 71.7 | 66.9 | 65.4 | 70.5 | 70.6 |
| x\_sg1\_snv | **74.1** | 67.1 | 67.2 | 70.9 | 71.8 |
| x\_sg1\_minmax | 71.9 | 67.1 | 68.1 | 70.7 | 71.7 |
| x\_sg1\_robust | 73.1 | 67.1 | 68.4 | 71.2 | 70.9 |
| x\_sg2\_msc | 69.4 | 64.7 | 66.2 | 69.9 | 71.8 |
| x\_sg2\_snv | 71.5 | 65.2 | 66.1 | 69.6 | 72.4 |
| x\_sg2\_minmax | 70.8 | 64.2 | 66.5 | 68 | 70.6 |
| x\_sg2\_robust | 71.7 | 65.7 | 67.8 | 69.9 | 70.7 |
| x\_sg1\_sg2\_msc | 71 | 67.2 | 66 | 71.7 | 72.7 |
| x\_ sg1\_sg2\_snv | 73.9 | 67.7 | 66.2 | 71.2 | 71.2 |
| x\_sg1\_sg2\_minmax | 72 | 66.9 | 67.5 | 70.4 | 72.8 |
| x\_ sg1\_sg2\_robust | 73.1 | 67.2 | 66.9 | 71.2 | 72.8 |

Bảng 3.14 Kết quả kiểm thử của các thuật toán học máy đề xuất.

Vì mô hình SVM có độ chính xác cao nhất nên tôi sẽ dùng mô hình SVM với đặc trưng x\_sg1\_snv để kiểm tra độ chính xác của tập kiểm thử với các vị trí đo khác nhau kết hợp với phương pháp repeated hold-out và cross-validation để xác định vị trí đo cho kết quả tốt nhất. Từ bảng 3.15 có thể thấy kết quả dự đoán tối nhất nằm ở vị trí ‘Ngoài da, đuôi’ với độ chính xác 80,7%.

|  |  |  |
| --- | --- | --- |
| **Vị trí đo** | **Repeated hold-out** | **Cross validation** |
| Ngoài da, gáy | 82.7 | 80.2 |
| Ngoài da, lưng | 80.0 | **80.7** |
| Ngoài da, đuôi | 77.9 | 78.2 |
| Ngoài da, bụng | 78.2 | 78.6 |
| Mang 1 | 80.4 | 77.6 |
| Mang 2 | 78.1 | 77.2 |
| Mắt 1 | 72.7 | 73.9 |
| Mắt 2 | 75.4 | 77.4 |
| Trong thịt, gáy | 76.1 | 72.3 |
| Trong thịt, lưng | 75.5 | 73.2 |
| Trong thịt, đuôi | 79.7 | 74.4 |
| Trong thịt, bụng | 60.0 | 60.9 |

Bảng 3.15 Độ chính xác trung bình của dữ liệu kiểm thử Histamine ở từng vị trí đo khi dùng phương pháp đánh giá repeated hold-out và cross-validation.

****

Hình 3.24 Ma trận nhầm lẫn tổng của 5 fold ở vị trí đo ngoài da, lưng.

Từ ma trận nhận nhầm lẫn tổng ở vị trí ngoài da, đuôi trên tập kiểm thử của mô hình SVM ta thu được độ nhạy, độ đặc hiệu, độ chính xác như sau:

- Độ chính xác =

- Độ nhạy (sensitivity / recall) =

- Độ đặc hiệu (specificity) =

Trong đó:

- TP: Số mẫu phổ được dự đoán là dưới LOD và có nhãn thực là dưới LOD

- FP: Số mẫu phổ được dự đoán là dưới LOD và có nhãn thực là trên LOD

- TN: Số mẫu phổ được dự đoán là trên LOD và có nhãn thực là trên LOD

- FN: Số mẫu phổ được dự đoán là trên LOD và có nhãn thực là dưới LOD

Từ kết quả độ nhạy và độ đặc hiệu có thể thấy rằng mô hình dự đoán các mẫu phổ dưới LOD là rất tốt (84,4%), nhưng còn các mẫu phổ trên LOD rất cần dự đoán chính xác thì vẫn chưa tốt chỉ đạt 75.0%.

# KẾT LUẬN

Tôi đã trình bày nghiên cứu việc đánh giá nồng độ Histamine, Ure có trong cá dựa trên phổ hấp thụ NIR của cá. Trong thời gian bảo quản 12 giờ kể từ lúc mua về, nhiều loại cá có ít hoặc không thay đổi về hình thức bên ngoài. Điều này làm cho các kĩ thuật về thị giác máy tính khó xử lý. Do đó, công nghệ quang phổ NIR cho thấy lợi thế của nó trong trường hợp này nhờ khả năng đặc trưng hóa các đặc tính bên trong của cá. Tôi đã tiến hành thử nghiệm nhiều sự kết hợp giữa các loại vector đặc trưng và mô hình học máy để tìm kiếm phương án tối ưu. Kết quả cho thấy mô hình SVM cho kết quả tốt hơn các mô hình khác bất kể loại vector đặc trưng nào được sử dụng. Đặc biệt, mô hình SVM dự đoán chính xác 78,3% ở vị trí ‘ngoài da, đuôi’ cho việc phát hiện nồng độ Ure trên hay dưới ngưỡng LOD và 80.7% ở vị trí ‘Ngoài da, lưng’ cho việc phát hiện nồng độ Histamine trên hay dưới ngưỡng LOD. Hiệu suất phân lớp thu được đã được thử nghiệm trên nhiều loại cá, điều này cho thấy tiềm năng của việc sử dụng NIR là rất cao.

Với kết quả thử nghiệm như trên, tôi đề xuất chỉ nên sử dụng vị trí đo ‘Ngoài da, đuôi’ để phân loại nồng độ Ure có trong cá và vị trí đo ‘Ngoài da, lưng’ để phân loại Histamine có trong cá thay vì dùng hết tất cả các vị trí đo.

Trong tương lai, tôi dự định sẽ thử nghiệm các mô hình này trên tập dữ liệu lớn hơn và với những loại độc tố và chất bảo quản bị cấm khác. Ngoài ra, tôi vẫn sẽ tiếp tục nghiên cứu và cải thiện hiệu quả của mô hình dự đoán đã triển khai.

**TÀI LIỆU THAM KHẢO**

[1] Mishra, P., Passos, D.: A synergistic use of chemometrics and deep learning

improved the predictive performance of near-infrared spectroscopy models for

dry matter prediction in mango fruit. Chemometrics and Intelligent Laboratory

Systems 212, 104287, 8 pages, 2021

[2] Jun-Hu Cheng, Da-Wen Sun, Hongbin Pu and Zhiwei Zhu, “Development of hyperspectral imaging coupled with chemometric analysis to monitor K value for evaluation of chemical spoilage in fish fillets”, Food Chemistry, Vol. 185, pp. 245–253, 2015.

[3] Zhe Xu, Xiaomin Zhao, Xi Guo, and Jiaxin Guo, "Deep learning application for predicting soil organic matter content by VIS-NIR spectroscopy", Computational Intelligence and Neuroscience, Article ID 3563761, 11 pages, 2019.

[4] Yiping Jiao, Zhichao Li, Xisong Chen, Shumin Fei, "Preprocessing methods for near-infrared spectrum calibration" 2020

[5] Ninh Khánh Duy, Đoàn Thị Ngọc Cảnh, Nguyễn Thị Thanh Xuân, "Determination of Fruit Freshness Using Near-Infrared Spectroscopy and Machine Learning Techniques", 2022

[6] Ông Nguyễn Uyên Nhi, "A STUDY ON MULTIVARIATE DATA PROCESSING AND MACHINE LEARNING METHODS FOR FISH QUALITY CONTROL", 2023

[7] Brian G. Osborne, "Near-infrared Spectroscopy in Food Analysis", BRI Australia Ltd, North Ryde, Australia, 2006.

[8] Sohn, Soo-In, Subramani Pandian, Young-Ju Oh, John-Lewis Z. Zaukuu, Hyeon-Jung Kang, Tae-Hun Ryu, Woo-Suk Cho, Youn-Sung Cho, EunKyoung Shin, and Byoung-Kwan Cho, "An Overview of Near Infrared Spectroscopy and Its Applications in the Detection of Genetically Modified Organisms", International Journal of Molecular Sciences, Vol. 22, No. 18: 9940, 2021.

[9] DLP NIRscan Nano EVM User’s Guide. Retrieved June 21, 2020 from https://www.ti.com/lit/ug/dlpu030g/dlpu030g.pdf?ts=1592754432447&ref\_u rl= https%253A%252F%252Fwww.google.com%252F.