Tạp chí Khoa học Trường Đại học Cần Thơ
Số chuyên đề: Thủy sản
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jvn.2021.078

KIỂM THỬ GIẢI THUẬT AI TRONG NHẬN DIỆN BỆNH TÔM QUA HÌNH ẢNH

Quách Luyl Đa[1], Phan Trọng Nghĩa[2], Trần Thanh Hùng[2] và Nguyễn Chí Ngôn[2*]

[1]*Trường Đại học FPT Cần Thơ*

[2]*Khoa Công nghệ, Trường Đại học Cần Thơ*

[*]*Người chịu trách nhiệm về bài viết: Nguyễn Chí Ngôn (email: ncngon@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 22/02/2021
Ngày nhận bài sửa: 06/04/2021
Ngày duyệt đăng: 01/06/2021

Title:
Testing AI algorithms in images-based identification of shrimp diseases

Từ khóa:
K láng giềng gần nhất, hồi qui tuyến tính đa thức, Naïve Bayes, rừng ngẫu nhiên, bệnh tôm, SURF

Keywords:
K nearest neighbors, multinomial logistic regression, Naïve Bayes, random forest, shrimp diseases, SURF

ABSTRACT

Artificial intelligence (AI) is often used in the classification of images. In this study, AI algorithms have been used in combining with SURF features, K-mean clustering on a 6-class shrimp disease dataset. In order to find the most appropriate model for image classification of shrimp diseases, the study has been tested on four AI models including Multinomial Logistic Regression, Naïve Bayes, K Nearest Neighbors, and Random Forest. Criteria for evaluating the accuracy of these models include Precision, Recall and F_1. Testing results when applying with initial feature dataset show a low accuracy that the best model is Random Forest algorithm, with Recall evaluation criterion of 47.7%. The study has been continued to conduct random combinations of 4 clusters classified by K-mean algorithm, the results indicate that the Random Forest model can get highest accuracy of 85.9% by Recall criteria.

TÓM TẮT

Trí tuệ nhân tạo thường được dùng trong việc phân loại hình ảnh. Trong nghiên cứu này, các giải thuật trí tuệ nhân tạo được sử dụng kết hợp với các đặc trưng SURF, phân cụm dữ liệu với K-mean trên bộ dữ liệu bệnh tôm 6 lớp. Nhằm tìm kiếm giải thuật thích hợp nhất trong việc phân loại bệnh tôm qua hình ảnh, nghiên cứu đã tiến hành kiểm thử trên 4 giải thuật trí tuệ nhân tạo, gồm: giải thuật hồi qui logic, Naïve Bayes, K láng giềng gần nhất và rừng ngẫu nhiên. Tiêu chí đánh giá độ chính xác của các giải thuật này gồm precision, recall và F_1. Kết quả thử nghiệm khi áp dụng trên các tập đặc trưng cho thấy đạt tỷ lệ thấp, độ chính xác cao nhất là giải thuật rừng ngẫu nhiên với tiêu chí đánh giá recall là 47,7%. Nghiên cứu tiếp tục tiến hành kết hợp ngẫu nhiên của 4 cụm được phân loại bởi giải thuật K-mean, kết quả thu được với độ chính xác cao nhất theo tiêu chí recall cho giải thuật rừng ngẫu nhiên là 85,9%.

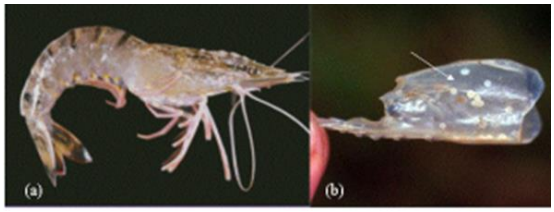
1. GIỚI THIỆU

Trí tuệ nhân tạo (artificial intelligence - AI), học máy (machine learning – ML) hay học sâu (deep learning - DL) là những thuật ngữ thường được sử dụng ngày nay. Trong đó, ML là một hướng nghiên cứu của khoa học máy tính và là một phần trong hệ thống của trí tuệ nhân tạo, dễ dàng tích hợp các loại dữ liệu khác nhau (Roell et al., 2020). Trong khi đó, DL là một nhánh cụ thể của ML với việc sử dụng các giá trị dữ liệu phân cấp, trong đó có việc chuyển đổi thông tin giữa các bước khác nhau thành các biểu diễn phức tạp hơn của dữ liệu (Goodfellow et al., 2016). Cuối cùng, AI là một nhánh của khoa học máy tính, được dùng để nghiên cứu và xây dựng phần mềm và máy móc thông minh (Zahraee, 2016).

Việc ứng dụng AI vào phân lớp hình ảnh được ứng dụng mạnh mẽ trong khoảng thời gian gần đây.

Phân loại hình ảnh là kỹ thuật được sử dụng để trích xuất thông tin từ hình ảnh, nhãn và pixel từ hình ảnh. Để thực hiện phân loại, các hình ảnh cùng đối tượng sẽ được cung cấp kết hợp với một sơ đồ phân loại thích hợp và khi đủ số lượng mẫu huấn luyện thì hiệu quả phân loại sẽ càng cao. Do đó, hệ thống phân loại phụ thuộc vào yêu cầu của người dùng thông qua việc bố trí sơ đồ phân loại thích hợp (Lu et al., 2007). Phân loại ảnh có nhiều cách tiếp cận khác nhau bằng cách sử dụng các giải thuật của máy học, mà phổ biến là mạng nơ-ron nhân tạo, hệ chuyên gia và logic mờ,... Quá trình tiền xử lý ảnh bao gồm các thao tác: lựa chọn mẫu, tiền xử lý hình ảnh, trích xuất đặc trưng, lựa chọn giải thuật, xử lý sau phân loại và đánh giá độ chính xác của giải thuật. Trong đó, quá trình lựa chọn mẫu và tiền xử lý có vai trò quan trọng, ảnh hưởng đến độ chính xác của giải thuật phân loại.

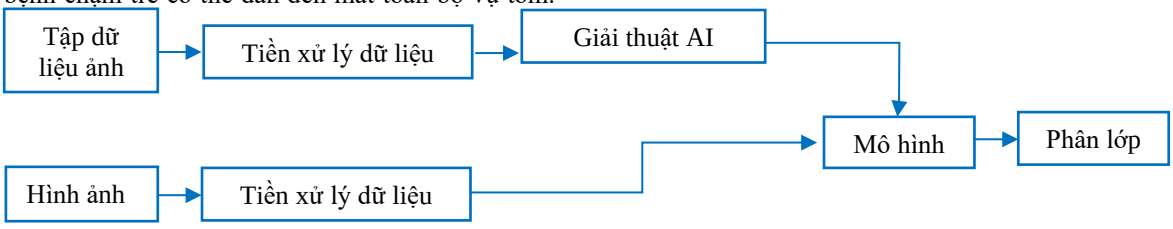
Ở Việt Nam, ngành nuôi tôm đóng góp một vị thế quan trọng nhưng kèm theo nhiều thách thức. Năm 2018, tổng diện tích nuôi của khu vực đồng bằng sông Cửu Long là 720.000 ha với tổng sản lượng 745.000 tấn, chiếm 2/3 tổng số tôm nuôi toàn quốc. Trong 7 tháng đầu năm 2019, Việt Nam xuất khẩu tôm đạt 1,8 tỷ USD (Cát Tường, 2019). Tuy nhiên, vấn đề dịch bệnh là điều không thể tránh khỏi. Trong năm 2012, bệnh tử vong sớm (EMS - early mortality syndrome) đã gây thiệt hại 1/6 diện tích nuôi tôm (Nguyen, 2015). Trong khoảng thời gian từ năm 2013 đến 2016 (Pongthanapanich et al, 2019), theo báo cáo thống kê của FAO cho biết các bệnh thường gặp là hoại tử gan cấp tính (AHPND - acute hepatopancreatic necrosis disease), EMS, virus đốm trắng (WSSV - white spot syndrome virus, được minh họa trên Hình 1), phân trắng (WFS - white feces syndrome) và virus HPV (Hepatopancreatic Parvovirus). Tuy nhiên, thời gian ủ bệnh và tái phát bệnh lại khác nhau trong quá trình nuôi, như trường hợp WSSV thường nhận thấy sự xuất hiện nhiễm trùng trong khoảng thời gian rộng (25 đến 60 ngày). Sự phát hiện và can thiệp điều trị bệnh chậm trễ có thể dẫn đến mất toàn bộ vụ tôm. Trường hợp bệnh AHPND trong năm 2015 được báo cáo là 5.875 ha, gây thiệt hại ước tính hơn 25,98 triệu đô la Mỹ. Do vậy, việc tìm kiếm các kỹ thuật thích hợp để xác định sớm bệnh tôm là một chủ đề nghiên cứu hấp dẫn.



Hình 1. Bệnh đốm trắng (Durand et al., 1997)

Áp dụng kỹ thuật phân loại hình ảnh dùng AI đã được nhiều nghiên cứu quan tâm như: ứng dụng giải thuật ImageNet cho việc phân loại cây thuốc nam (Duong-Trung et al., 2019) ; Bao et al. (2019) sử dụng giải thuật Niblack để phát hiện, xác định và loại bỏ tôm bị bệnh vàng đầu YHV. Ghasemi-Varnamkhasti et al. (2016) phát hiện tôm bệnh WSSV sử dụng kỹ thuật phân cụm K-Means. Một số ứng dụng khác tập trung phát hiện đánh giá độ tươi của tôm (Okpala, 2014), xác định vỏ mềm và âm thanh tôm (Liu et al., 2016).

Hiện tại, chưa có nghiên cứu chuyên sâu trong việc ứng dụng công nghệ thông tin để phân loại bệnh tôm dựa trên hình ảnh tổng hợp của nhiều loại bệnh. Vì vậy, nghiên cứu được thực hiện nhằm đánh giá các giải thuật AI trong việc phân loại hình ảnh bệnh tôm như Hình 2. Việc phân loại hình ảnh sử dụng một hàm $y = f(x)$ phân biệt để ánh xạ từ dữ liệu đầu vào thành lớp đích. Với việc sử dụng giải thuật trích xuất đặc trưng cục bộ SURF (Bay et al., 2006) sinh ra vector đầu vào có dạng $<x_1, x_2, \ldots, x_n>$ và y là tập hữu hạn các nhãn dữ liệu $<y_1, y_2, \ldots y_c>$ để tạo ra được giải thuật phân loại gần đúng f' (Hastie et al., 2009). Trong giai đoạn tiền xử lý dữ liệu, giải thuật K-Means được áp dụng (Likas et al., 2003) để sửa chữa, biến đổi hoặc tập hợp con để lựa chọn các đặc trưng phù hợp với dự định phân loại. Các giải thuật AI được sử dụng là hồi quy tuyến tính đa thức, K láng giềng gần nhất, rừng ngẫu nhiên và Naïve Bayes



Hình 2. Sơ đồ huấn luyện hệ thống phân loại bệnh tôm dựa trên hình ảnh

Các giải thuật AI được kiểm thử trong nghiên cứu này được đánh giá bằng các chỉ tiêu Precision, Recall và F1. Tương tự, giải thuật K-mean được sử dụng để thực hiện việc chia bộ dữ liệu đã lấy đặc trưng SURF ra làm 4 cụm và kết hợp ngẫu nhiên trong 6 tập bệnh tôm, tạo ra 4.096 lần kiểm thử để đánh giá được các phần dữ liệu quan trọng trong việc nhận diện đã được phát hiện hay chưa.

2. TIỀN XỬ LÝ DỮ LIỆU

2.1. Lựa chọn dữ liệu ảnh tôm

Dữ liệu tôm bệnh được nghiên cứu sưu tầm thông qua website của Nguyễn Chí Ngôn và ctv. (2019). Dữ liệu được thu thập từ người nông dân nuôi tôm, bị ảnh hưởng bởi nhiều yếu tố như:

– Chất lượng hình ảnh: Nông dân sử dụng nhiều loại điện thoại khác nhau nên camera được sử dụng cũng khác nhau; nhiều điện thoại có chức năng làm đẹp ảnh chụp bằng phần mềm nên cũng gây nhiều khó khăn cho quá trình nhận diện và phân loại ảnh bệnh.

– Môi trường chụp ảnh: Người nông dân chụp ảnh trong nhiều môi trường có ánh sáng khác nhau; môi trường nước ao nuôi khác nhau cũng ảnh hưởng đến chất lượng hình ảnh.

– Loại tôm được chụp: Hiện nay, khu vực Đồng bằng sông Cửu Long nuôi tôm sú và tôm thẻ là chủ yếu, nên việc chụp ảnh bệnh trên 2 loại tôm này cũng cho hình ảnh và sự thể hiện bệnh qua ảnh khác nhau.

Dữ liệu ảnh bệnh tôm thu về gồm 5 tập ảnh tôm bệnh ứng với 5 loại bệnh và 1 tập ảnh tôm khỏe mạnh. Sau khi loại bỏ nền ảnh, tập dữ liệu hình ảnh được được mô tả như Bảng 1.

Bảng 1. Thống kê số lượng mẫu bệnh tôm thu được

STT	Ảnh bệnh	Nhãn bệnh	Số lượng ảnh
1		Tôm bị đen mang	125
2		Tôm bị đốm đen	102
3		Tôm bị đốm trắng	166
4		Tôm bị hoại tử cơ	115
5		Tôm bị hoại tử gan	61
6		Tôm bình thường	75
Tổng cộng			644

2.2. Đặc trưng cục bộ SURF

Đặc trưng cục bộ SURF (Bay et al., 2006) sử dụng bộ mô tả bất biến, nhanh chóng và hiệu quả với việc áp dụng bộ lọc hộp Haar. Bộ lọc hộp này sử dụng một phép chập được tính toán nhanh chóng bằng cách sử dụng hình ảnh phân tích. Kết quả quả thu được bằng cách sử dụng các phản hồi Wavelet theo hướng ngang – dọc. Đầu tiên, vector đặc trưng V_j được tạo ra bởi những vùng phụ j trong việc sử dụng các phản hồi Wavelet trong khu vực tiểu vùng 4 × 4, được mô tả như (1).

$$V_j=\left[\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right] \quad (1)$$

Trong đó, d_x và d_y là phản hồi của Wavelet Haar theo các hướng ngang – dọc. Bộ mô tả SURF với 64 vector đặc trưng được tạo ra từ mỗi vùng của ảnh. Trong nghiên cứu này, bộ mô tả SURF được sử dụng cho việc trích xuất các đặc trưng từ các ảnh màu thay vì biểu diễn theo thang độ xám. Kết quả của quá trình lấy đặc trưng được minh họa như Hình 3 và số lượng đặc trưng SURF cho từng tập bệnh được trình bày trong Bảng 2.

Hình 3. Ảnh mẫu sau khi lấy đặc trưng SURF

Bảng 2. Số đặc trưng ảnh thu được trên mỗi bệnh và số liệu sau khi chia cụm

STT	Nhãn bệnh	Số đặc trưng SURF				
		Tổng cộng	Cụm 1	Cụm 2	Cụm 3	Cụm 4
1	Tôm bị đen mang	2.779	793	702	581	703
2	Tôm bị đốm đen	2.548	1.240	372	401	535
3	Tôm bị đốm trắng	3.549	1.763	503	769	514
4	Tôm bị hoại tử cơ	2.342	1.074	347	412	509
5	Tôm bị hoại tử gan	1.653	164	764	482	243
6	Tôm bình thường	1.659	170	722	517	250
Tổng cộng		14.530	5.204	3.410	3.162	2.754

3. CÁC GIẢI THUẬT AI ĐƯỢC SỬ DỤNG ĐỂ PHÂN LOẠI

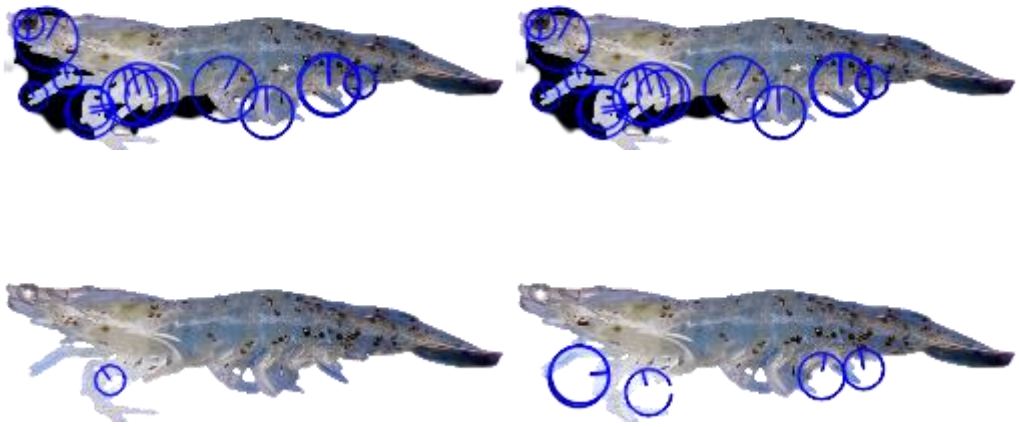
3.1. Giải thuật phân cụm K-Means

MacQueen (1967) đã đề xuất giải thuật phân cụm K-means. Đây là một giải thuật học không giám sát được sử dụng để phân loại tập dữ liệu thành K nhóm. Giải thuật này tiến hành bằng cách chọn K trung tâm cụm ban đầu và sau đó tinh chỉnh lặp đi lặp lại sao cho:

– Mỗi d_i sẽ được gán cho cụm trung tâm gần nó nhất.

– Mỗi trung tâm cụm C_j được cập nhật để trở thành giá trị trung bình của các thể hiện cấu thành nó.

Giải thuật sẽ dừng khi không có sự thay đổi nào trong việc gán các thể hiện cho các cụm. Trong nghiên cứu này, K-means được sử dụng để chọn 4 cụm ngẫu nhiên từ từng tập dữ liệu bệnh trên tôm để tiến hành đánh giá từng cụm với nhau, nhằm đánh giá giải thuật cũng như độ khả thi của nghiên cứu với việc đề xuất giải thuật phù hợp với dữ liệu hiện tại. K-means được áp dụng vào việc phân chia đặc trưng SURF làm 4 cụm khác nhau, với số lượng được thể hiện ở Bảng 2, để tạo ra bộ dữ liệu kết hợp. Minh họa ảnh chia cụm trên bệnh đốm đen được trình bày trên Hình 5.



Hình 4. Ảnh mẫu của bệnh đốm đen sau khi phân chia làm 4 cụm bằng K-means

3.2. Giải thuật hồi quy tuyến tính đa thức

Đối với bài toán nhận dạng mẫu nhiều lớp thì giải thuật được dùng là hồi quy đa thức (multinomial logistic regression - MLR) được giới thiệu bởi McCullagh et al. (1989). Một ước lượng hậu kỳ về xác suất một mẫu thuộc về mỗi lớp trong *c* lớp rời rạc là kết quả đầu ra của giải thuật MLR. Giải thuật MLR sử dụng xác suất mang lại nhiều lợi thế thực tế như đặt ra ngưỡng loại bỏ, điều chỉnh các tần số tương đối không bằng nhau trong tập huấn luyện và trong hoạt động, hoặc áp dụng để dự đoán nhằm giảm thiểu rủi ro mong đợi (Cawley et al., 2007).

Trong MLR, mục tiêu *y* là một biến có phạm vi trên 2 lớp, để xác định xác suất của y trong mỗi lớp tiềm năng $c \epsilon C, p(x)$. Khi đó, để tính xác suất $p(x)$ ta sử dụng hàm softmax. Trong nghiên cứu này, hàm softmwax nhận vector $z = [z_1, z_2, \ldots, z_k]$ với *k* giá trị, khi đó softmax được định nghĩa như sau:

$$soft\max(z) = \left[\frac{e^{z_1}}{\sum_{i=1}^{k} e^{z_i}}, \frac{e^{z_2}}{\sum_{i=1}^{k} e^{z_i}}, \ldots, \frac{e^{z_k}}{\sum_{i=1}^{k} e^{z_i}}\right] \quad (2)$$

3.3. Giải thuật Naïve Bayes

Trong AI, giải thuật Naïve Bayes được xem là một giải thuật phân loại sử dụng giải thuật xác suất Bayes trong công thức (3) hoạt động dựa trên các giả định độc lập, điều này có nghĩa là xác suất của một thuộc tính không ảnh hưởng đến xác xuất của thuộc tính kia (Al-Sharafat, 2009). Tuy nhiên, kết quả của phân loại Naïve Bayes thường cho độ chính xác cao.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (3)$$

Với *X* là các vector đặc trưng SURF và *y* là các nhãn bệnh tôm thì công thức (3) có:

– $P\ y|X$: xác suất của mục tiêu y với điều kiện có đặc trưng X.

– $P\ X|y$: xác suất của đặc trưng X khi đã biết mục tiêu y

– $P\ y$: xác suất tiền nhiệm của mục tiêu y

– $P\ X$: xác suất tiền nhiệm của đặc trưng X

Việc sử dụng giải thuật nhằm đánh giá 2 giả thiết đặt ra:

– Các đặc trưng SURF đưa vào giải thuật là độc lập với nhau. Điều này có nghĩa là sự thay đổi của một đặc trưng SURF không ảnh hưởng đến các đặc trưng còn lại.

– Các đặc trưng đưa vào giải thuật dự đoán bệnh tôm có ảnh hưởng ngang nhau đối với đầu ra của mục tiêu.

Khi đó, hàm mục tiêu *y* để $P\ X|y$ đạt cực đại trở thành:

$$y = \arg\max_y P(y) = \prod_{i=1}^{n} P(x_i|y) \quad (4)$$

3.4. Giải thuật K láng giềng

Giải thuật K láng giềng (K Nearest Neighbors – KNN) không có quá trình học, khi dự đoán nhãn của phần tử dữ liệu mới đến. Giải thuật KNN đi tìm k láng giềng của nó từ tập dữ liệu học, rồi sau đó thực hiện phân lớp phần tử mới đến. Kết quả của giải thuật còn phụ thuộc vào việc chọn độ đo khoảng cách (Goldberger, 2004). Trong nghiên cứu này, dữ liệu được sinh ra là các vector đặc trưng SURF. Do đó, ma trận chuyển đổi tuyến tính tối ưu có kích thước m × n, với n là thành phần và m là tính chất, tối đa hóa tổng trên tất cả các mẫu *i* được tính xác xuất p_i mà *i* được lựa chọn phân loại theo (5).

$$\arg\max \sum_{i=0}^{N-1} p_i \quad (5)$$

Với $N = n$ mẫu và p_i là xác suất của mẫu *i* được phân loại chính xác theo quy tắc ngẫu nhiên láng giềng gần nhất trong không gian như sau:

$$p_i = \sum_{j \in C_i} p_{\text{ij}} \quad (6)$$

Với C_i là tập hợp các điểm trong cùng lớp mẫu i và p_{ij} là softmax trên khoảng cách Euclid trong không gian theo (7):

$$p_{\text{ij}} = \frac{\exp\left(-\left\|L_{x_i} - L_{x_j}\right\|^2\right)}{\sum_{k \neq i} \exp-\left(\left\|L_{x_i} - L_{x_k}\right\|^2\right)}, p_{\text{ij}} = 0 \quad (7)$$

Với việc sử dụng độ đo khoảng cách Mahalanobis như (8)

$$\left\|L_{x_i} - L_{x_j}\right\|^2 = x_i - x_j^{\ T} \ x_i - x_j \quad (8)$$

Trong đó, $M = L^T L$ là một ma trận bán xác định dương đối xứng với ma trận đặc trưng n × n.

3.5. Giải thuật Random Forest

Giải thuật rừng ngẫu nhiên (Random forest - RF) tạo ra một tập hợp các cây quyết định không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap, tại mỗi nút phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính. Việc không cắt nhánh của giải thuật RF nhằm giữ cho thành phần lỗi bias thấp và dùng tính ngẫu nhiên để điều khiển tính tương quan giữa các cây. Giải thuật RF học nhanh, giảm được lỗi tốt và có độ chính xác cao, đáp ứng được yêu cầu thực tiễn trong vấn đề phân loại, hồi qui và phát hiện những phần tử đặc biệt (Breiman, 2001). Trong giải thuật RF (Hình 5) có các tham số sau:

– Tập dữ liệu học *S* có *m* phần tử và *n* thuộc tính, giải thuật RF xây dựng *T* cây quyết định một cách độc lập nhau.

– Giải thuật cây quyết định thứ *t* được xây dựng trên tập ngẫu nhiên Bootstrap thứ *t* (lấy mẫu m phần tử có hoàn lại từ tập học *S*).

– Tại nút trong, chọn ngẫu nhiên *n'* thuộc tính và tính toán phân hoạch tốt nhất dựa trên thuộc tính này. Trong giải thuật này, thuộc tính phân hoạch tốt nhất được lựa chọn theo công thức entropy và độ lợi thông tin. Giả sử p_i là xác suất mà phần tử trong tập dữ liệu *S* thuộc lớp C_i(i=1,k) thì độ đo hỗn loạn thông tin trước khi phân hoạch được tính theo (9). Với việc sử dụng thuộc tính *A* phân hoạch dữ liệu *D* thành *v* thành phần, thì độ đo hỗn loạn sau khi phân hoạch được tính theo (10). Sau khi tính toán độ đô hỗn loại của thuộc tính và tập dữ liệu *S*, độ lợi thông tin khi lựa chọn thuộc tính *A* phân hoạch dữ liệu *D* thành *v* thành phần theo (11).

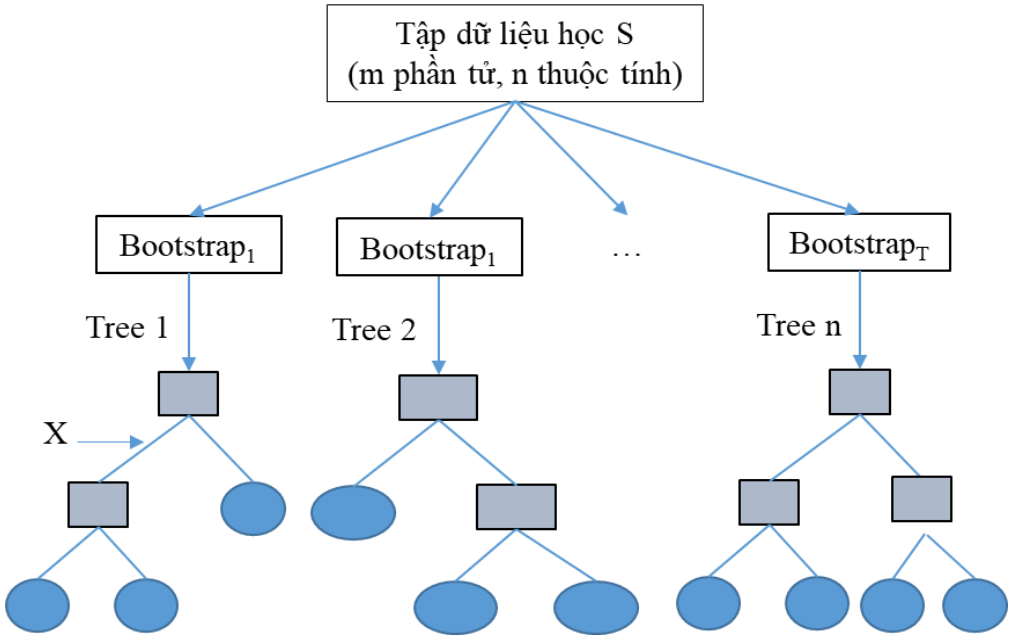
– Kết thúc quá trình xây dựng *T* giải thuật cơ sở, chiến lược bình chọn số đông trong $\{\hat{y}_1(x), \hat{y}_2(x), \ldots, \hat{y}_T(x)\}$ để phân lớp một phần tử mới đến hoặc giá trị trung bình cho bài toán hồi quy được tính như (12).

$$Info\ \ S\ \ = -\sum_{i=1}^{k} p_i\ log_2\ \ p_i \quad (9)$$

$$Info_A\ \ S\ \ = \sum_{j=1}^{v} \frac{\left|S_J\right|}{\left|S\right|} \times Info\ \ S_j \quad (10)$$

$$Gain\ \ A\ \ = Info\ \ S\ \ - Info_A\ \ S \quad (11)$$

$$\hat{y}_1\ \ x\ \ + \hat{y}_2\ \ x\ \ + \ldots + \hat{y}_T\ \ x\ \ \ / T \quad (12)$$



Hình 5. Giải thuật rừng ngẫu nhiên RF

4. ĐÁNH GIÁ ĐỘ CHÍNH XÁC PHÂN LOẠI

Để đánh giá độ chính xác phân loại, giải thuật phát hiện bất thường (Anomaly detection – AD) (Powers, 2011) được sử dụng. AD hoạt động như giải thuật nhận dạng mẫu và phân loại nhị phân. Nó nhận ra một số mẫu nhất định để phân loại nó là bình thường hay bất thường. Đối với giải thuật này, các tiêu chí Recall, Precision và F_1 thường được sử dụng để đánh giá hiệu quả của giải thuật học mà phân lớp dữ liệu nhị phân không cân bằng, chúng được định nghĩa trong (13), (14) và (15).

$$precision = \frac{TP}{TP + FP} \quad (13)$$

$$recall = \frac{TP}{TP + FN} \quad (14)$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \quad (15)$$

Trong đó:

– TP: Tổng số phần tử lớp dương được giải thuật dự đoán là lớp dương.

– FN: Tổng số phần tử lớp dương được giải thuật dự đoán là lớp âm (Biểu thị âm tính giả).

– TN: Tổng số phần tử lớp âm được giải thuật dự đoán là lớp âm.

– FP: Tổng số phần tử lớp âm được giải thuật dự đón là lớp dương (Biểu thị dương tính giả).

5. KẾT QUẢ

Dữ liệu được mô tả tại mục 2.1 với 644 ảnh được chia làm 6 lớp, sau khi thực hiện xử lý với đặc trưng SURF, tiến hành xử lý làm 2 bộ dữ liệu riêng biệt được mô tả trong Bảng 2, với mô tả như sau:

– Bộ dữ liệu 1: Bộ dữ liệu với 14.530 vector đặc trưng SURF.

– Bộ dữ liệu 2: Bộ dữ liệu được sử dụng K-mean để phân ra làm 4 cụm với tổng số lượng dữ liệu lần lượt là 5.204, 3.410, 3.162 và 2.754. Các bộ dữ liệu này được kết hợp ngẫu nhiên 4 cụm của từng loại bệnh với nhau. Việc kết hợp này sinh ra 4.096 mẫu huấn luyện và kiểm thử khác nhau.

Nghiên cứu tiến hành kiểm thử với các giải thuật hồi qui tuyến tính, Naïve Bayes, K láng giềng gần nhất và RF trên 2 bộ dữ liệu với 70% dùng để huấn luyện và 30% dùng để kiểm thử. Kết quả kiểm thử ở Bảng 3 cho thấy trong 3 giải thuật, RF có độ chính xác cao nhất và thấp nhất là giải thuật hồi qui tuyến tính. Bảng 4 mô tả kết quả chính xác lớn nhất sau khi kiểm thử trên 4.096 mẫu huấn luyện, điều này cho kết quả với độ chính xác cao hơn kết quả ở Bảng 3 rất nhiều, độ chính xác cao nhất là RF với độ chính xác Precision, Recall và F1 lần lượt là 85,2 – 85,9 – 85,4.

Bảng 3. Kết quả kiểm thử trên bộ dữ liệu 1 không có K-mean

Giải thuật	Precision (%)	Recall (%)	F_1 (%)
Hồi qui tuyến tính	28,1	32,2	26,2
Naïve Bayes	31,1	26,5	27,0
K láng giềng gần nhất	40,7	41,0	39,6
RF	47,6	47,7	46,5

Bảng 4. Kết quả chính xác lớn nhất trên bộ dữ liệu 2 (Có sử dụng K-Mean)

Giải thuật	Precision (%)			Recall (%)			F_1 (%)		
	Max	Min	AVG	Max	Min	AVG	Max	Min	AVG
Hồi qui tuyến tính	82,9	28,0	60,7	84,9	30,5	63,2	81,0	22,9	59,8
Naïve Bayes	81,4	28,5	62,1	80,1	27,7	60,7	80,1	26,8	60,6
K láng giềng gần nhất	85,2	37,2	66,8	85,6	39,1	67,2	84,0	36,2	66,1
RF	85,2	43,4	70,7	85,9	43,9	70,9	85,4	41,6	69,9

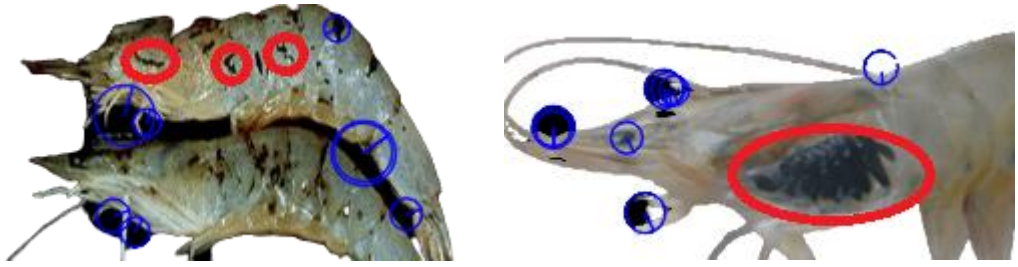
6. THẢO LUẬN

Với kết quả được thể hiện ở Bảng 3 và 4, kết quả nghiên cứu được đánh giá như sau:

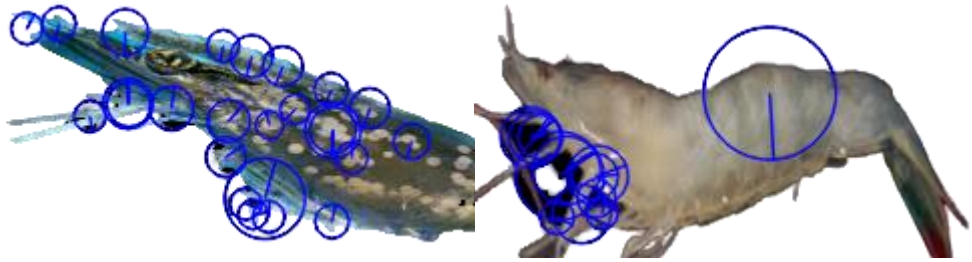
– Bộ dữ liệu được sử dụng có nhiễu, dẫn tới kết quả dự đoán với độ chính xác không được cao như mong đợi. Điều này đến từ các yếu tố như đã trình bày ở mục 2.1 về việc lựa chọn dữ liệu phân loại. Việc sử dụng đặc trưng SURF kết hợp với giải thuật K-means để tách các điểm được xem là bệnh ra khỏi những điểm ảnh nhiễu làm số lượng mẫu tăng lên, giúp cho kết quả dự đoán có độ chính xác cao hơn.

– Số lượng đặc trưng chưa đồng đều giữa các bệnh: Các vector đặc trưng xuất hiện nhiều trên các bộ dữ liệu tôm bệnh đen mang, đốm đen, đốm trắng và hoại tử cơ. Bên cạnh đó, một số bệnh thể hiện thông qua màu sắc như đen mang, đốm đen còn chưa chính xác như Hình 6.

– Giải thuật RF có độ chính xác cao nhất, đạt 85,9% theo tiêu chí đánh giá Recall. Kết quả nhận diện được thể hiện như Hình 7.



Hình 6. Một số ảnh khi lấy đặc trưng SURF bị lỗi



Hình 7. Kết quả sau khi nhận diện được và thể hiện lại trên tôm bệnh

7. KẾT LUẬN

Nghiên cứu này đã thu được 644 hình ảnh gồm: hình ảnh của 5 loại bệnh tôm và hình ảnh tôm khỏe mạnh, từ nhiều nguồn khác nhau, để chia làm 6 lớp dữ liệu. Sau khi tiền xử lý, 2 bộ dữ liệu thu được gồm: 14.530 mẫu dùng đặc trưng SURF và 4.096 mẫu dùng Kmeans. Việc kiểm thử các giải thuật AI trong nhận diện bệnh tôm được tiến hành trên 4 giải thuật, gồm: giải thuật hồi qui tuyến tính, Naïve Bayes, K láng giềng gần nhất và RF. Các giải thuật

này được huấn luyện trên 70% số mẫu của bộ dữ liệu và được kiểm tra trên 30% số mẫu còn lại. Các tiêu chí được dùng để đánh giá độ tin cậy của giải thuật bao gồm: Precision, Recall và F1. Kết quả kiểm thử cho thấy giải thuật RF có độ chính xác cao nhất, đạt 85,9% theo tiêu chí đánh giá Recall.

Trong thời gian tới, để cải thiện độ tin cậy của giải thuật nhận dạng, một số biện pháp sau cần được áp dụng như: (i) tăng số dữ liệu mẫu huấn luyện; (ii) phân chia bộ dữ liệu theo từng nhóm đặc trưng khác nhau như màu sắc, hình dạng,…; (iii) tiếp tục áp dụng giải thuật học sâu trong huấn luyện dữ liệu và nhận diện bằng cơ chế attention. Cơ chế attention là một cơ chế giúp giải thuật có thể tập trung vào các phần quan trọng trên dữ liệu, bằng việc tạo ra một giải thuật liên kết với các điểm căn chỉnh để đánh lại trọng số các trạng thái ẩn của mã hóa.

LỜI CẢM TẠ

Nghiên cứu này được tài trợ một phần từ Dự án nâng cấp trường Đại học Cần Thơ VN14-P6 được hỗ trợ bởi nguồn vốn ODA của Chính phủ Nhật Bản.

TÀI LIỆU THAM KHẢO

Al-Sharafat, W.S. & Reyadh Naoum (2009). Development of Genetic-based Machine Learning for Network Intrusion Detection. *Inter. J. of Computer and Information Engineering*, *3*(7), 1677-1681. DOI: 10.5281/zenodo.10.5281/zenodo.1060305

Bao, T.Q., Cuong, T.C., Tu, N.D. & Hieu, L.T. (2019). Designing the Yellow Head Virus Syndrome Recognition Application for Shrimp on an Embedded System. *Exchanges: The Interdisciplinary Research Journal*, *6*(2), 48-63. DOI: https://doi.org/10.31273/eirj.v6i2.309

Bay H., Tuytelaars T. & Van Gool L. (2006). SURF: Speeded Up Robust Features. In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006, *Lecture Notes in Computer Science*, vol 3951. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11744023_32

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cát Tường (2019). Vietnam shrimp exports started to reverse, *website of the Directorate of Fisheries*, Ministry of Agriculture and Rural Development, issued 22-Aug-2019.

Cawley, G. C., Talbot, N. L. C. & Girolami, M. (2007). Sparse multinomial logistic regression via Bayesian L1 regularisation. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems*, vol. 19 (pp. 209-216). MIT Press.

Duong-Trung, Nghia, Luyl-Da Quach & Chi-Ngon Nguyen (2019). Learning deep transferability for several agricultural classification problems. *Inter. J. of Advanced Computer Science and Applications*, 10(1), 58 – 67. http://dx.doi.org/10.14569/IJACSA.2019.0100107

Durand, S., Lightner, D. V., Redman, R. M. & Bonami, J. R. (1997). Ultrastructure and morphogenesis of white spot syndrome baculovirus (WSSV). *Diseases of Aquatic Organisms*, *29*(3), 205-211.

Ghasemi-Varnamkhasti, M., Goli, R., Forina, M., Mohtasebi, S.S., Shafiee, S. & Naderi-Boldaji, M. (2016). Application of image analysis combined with computational expert approaches for shrimp freshness evaluation. *International Journal of Food Properties*, *19*(10), 2202-2222. DOI: 10.1080/10942912.2015.1118386

Goldberger, J., Hinton, G. E., Roweis, S. T. & Salakhutdinov, R. R. (2004). *Neighbourhood components analysis.* 17th Inter. Conf. on Neural Information Processing Systems, December 2004 (pp. 513-520). DOI: 10.5555/2976040.2976105

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016). *Deep Learning*. Cambridge: MIT press, 800 pages.

Hastie, T., Tibshirani, R. & Friedman, J.H., 2009. The elements of statistical learning: data mining, Inference and Prediction, 2nd edn. Springer, New York, USA, 533 pages.

Likas, A., Vlassis, N. and Verbeek, J.J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461. DOI: 10.1016/S0031-3203(02)00060-2

Liu, Z., Cheng, F. & Zhang, W. (2016). Identification of soft shell shrimp based on deep learning. In *2016 ASABE Annual International Meeting*, 162455470, American Society of Agricultural and Biological Engineers. DOI:10.13031/aim.20162455470

Lu, D. & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *Inter. J. of Remote sensing*, *28*(5), 823-870. DOI: 10.1080/01431160600746456.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Fifth Symposium on Math, Statistics, and Probability.* Berkeley, CA, University of California Press: 281–297.

Nguyen, T. B. T. (2015). Good Aquaculture Practices (VietGAP) and Sustainable Aquaculture Development in Viet Nam. In Romana-Eguia et.al. (2015), *Resource enhancement and sustainable aquaculture practices in Southeast Asia: challenges in responsible production of aquatic species:*

proceedings of the international workshop on resource enhancement and sustainable aquaculture practices in Southeast Asia 2014 (pp. 85-92). Aquaculture Department, Southeast Asian Fisheries Development Center.

Nguyễn Chí Ngôn, Dương Trung Nghĩa & Quách Luyl Đa (2019). *Thu thập dữ liệu tôm bệnh*/ Truy cập 11/08/2020. https://sites.google.com/view/shrimp-image-collection/home

Okpala, C.O.R., Choo, W.S. & Dykes, G.A. (2014). Quality and shelf life assessment of Pacific white shrimp (Litopenaeus vannamei) freshly harvested and stored on ice. *LWT-Food Science and Technology*, *55*(1), 110-116. DOI: 10.1016/j.lwt.2013.07.020

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models. *Monographs on Statistics and Applied Probability*, *37*, Chapman & Hall/CRC, 2nd edition, 532 pages. ISBN: 9780412317606.

Pongthanapanich, T., Nguyen, K. A. T., & Jolly, C. M. (2019). Risk management practices of small intensive shrimp farmers in the Mekong Delta of Viet Nam. *FAO Fisheries and Aquaculture Circular*, (C1194), I-20.

Powers, David Martin (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Inter. J. of Machine Learning Technology*, *2*(1), 37-63.

Roell, Y. E., Beucher, A., Møller, P. G., Greve, M. B., & Greve, M. H. (2020). Comparing a Random-Forest-Based Prediction of Winter Wheat Yield to Historical Yield Potential. *Agronomy*, *10*(3), 395.

Zahraee, S.M., Assadi, M.K. & Saidur, R. (2016). Application of artificial intelligence methods for hybrid energy system optimization. *Renewable and sustainable energy reviews*, 66, 617-630. DOI: 10.1016/j.rser.2016.08.028.