ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

Đề tài:

NHẬN BIẾT UNG THƯ DA

Lớp: CS114.N11.KHCL

GIẢNG VIÊN HƯỚNG DẪN: PGS.TS LÊ ĐÌNH DUY
ThS PHẠM NGUYỄN TRƯỜNG AN

SINH VIÊN THỰC HIỆN: ĐỖ THỊ THU TRANG - 20520816
BÙI QUANG PHÚ - 20520273
NGUYỄN THỊ NGỌC NGA - 20521641

TP Hồ Chí Minh, tháng 02 năm 2023

MỤC LỤC

I. TỔNG QUAN:

1. Giới thiệu đề tài:

Ở con người thường tồn tại một số hiện tượng tổn thương da rất phổ biến và lành tính. Những tổn thương này có thể biểu hiện ra bên ngoài như nốt ruồi, tàn nhang, các mảng da thừa, thay đổi sắc tố lành tính, dày sừng tiết bã và các u nhú. Tuy nhiên, trong một số trường hợp, các biểu hiện trên da, đặc biệt là những nốt ruồi biến đổi bất thường, có thể là dấu hiệu cảnh báo ung thư.

Nốt ruồi thường có màu nâu sẫm hoặc đen, có thể xuất hiện ở bất cứ đâu trên cơ thể và có thể xuất hiện đơn lẻ hoặc thành cụm nhỏ, mịn hoặc sần. Đặc biệt, một số nốt ruồi nằm sâu trong da hoặc có thể hơi nổi lên, đôi khi có lông. Với người trưởng thành, số lượng nốt ruồi có thể lên tới 30 - 40. Vì vậy, khi thấy những nốt lạ trên da, người bệnh thường lầm tưởng đó là nốt ruồi bình thường nên không đi khám sớm dẫn đến nhiều biến chứng nặng nề.

Ung thư từ nốt ruồi là loại ung thư tế bào biểu bì nguy hiểm nhất. Chúng phát triển nhanh, và thông thường, một khối u ác tính di căn nhanh chóng, gây ra ung thư thứ phát ở gan, xương, phổi, não và hệ bạch huyết... Ung thư từ nốt ruồi thường bắt đầu từ u kích thước nhỏ, không ngứa, không đau nên ít được để ý. Tuy nhiên, trên thực tế, trong quá trình phát triển, loại u này có thể gây viêm nhiễm, loét, hoại tử, phá hủy tổ chức tại chỗ. Trường hợp biến chứng nặng, khối u này có thể ăn mòn ngón tay, chân, miệng, mắt, mũi... tùy vào vị trí khối u.

Tuy nhiên, nếu bệnh nhân được phát hiện sớm và điều trị kịp thời ngay từ khi nốt ruồi mới chuyển thành ác tính thì tỷ lệ chữa khỏi tương đối cao. Với những yếu tố trên, nhóm mong muốn xây dựng hệ thống *"Nhận biết ung thư da"*, để có thể dự báo dễ dàng và người bệnh có thể được chữa trị sớm hơn.

Cụ thể, người dùng sẽ *nhập* vào hệ thống một bức ảnh có chứa nốt ruồi, hoặc các tổn thương da rõ nét, *đầu ra* của hệ thống là dòng chữ dự báo ví dụ như có ung thư hoặc không ung thư.

2. Mục tiêu dự án:

Trong dự án, nhóm sẽ sử dụng nhiều mô hình khác nhau để so sánh và đánh giá kết quả của những loại mô hình đã chọn. Ba mô hình được sử dụng là Logistic Regression, SVM (Support Vector Machine) và KNN (K-Nearest Neighbors), là ba mô hình phân loại khá phổ biến trong Machine Learning và hoạt động tốt trên pixels.

Đặc biệt, nhóm sử dụng kỹ thuật *Trích chọn đặc trưng (Feature Extraction hoặc Feature Engineering)* để có thêm kết quả ở nhiều khía cạnh khác nhau. Đây là một quá trình quan trọng trong nhiều bài toán thực tế, giúp loại ra những dữ liệu nhiễu và đưa dữ liệu thô với số chiều khác nhau về cùng một chuẩn. Ở phần này, nhóm sẽ sử dụng 2 mô hình là Gaussian Naive Bayes và SVM.

Do đó, trong báo cáo, nhóm sẽ tập trung so sánh phương pháp dựa trên pixel để trích xuất các đặc điểm (features) của hình ảnh soi da về tổn thương da, và phương pháp mới sử dụng kỹ thuật đặc trưng và kỹ thuật khử nhiễu hình ảnh cùng với các phương pháp *Trích chọn đặc trưng* khác để phân loại hình ảnh đầu vào là ác tính hay lành tính. Điều này sẽ giúp chứng minh sự cải thiện trong phân loại hình ảnh và giúp hướng đến quyết định đưa các phương pháp này vào các hệ thống phát hiện bệnh lâm sàng.

II. BỘ DỮ LIỆU:

1. Thu thập dữ liệu:

Vì các tế bào ung thư biểu bì thường xuất hiện sớm dưới hình thức nốt ruồi nhiều nhất, và cũng là loại ung thư tế bào biểu bì nguy hiểm nhất, nên nhóm sẽ tập trung thu thập dữ liệu là hình ảnh của nốt ruồi.

Nhóm sẽ sử dụng và thực hiện chủ yếu dựa trên bộ dữ liệu được thu thập từ ISIC (International Skin Imaging Collaboration), một tổ chức hợp tác giữa ngành công nghiệp và viện nghiên cứu, với mong muốn tạo điều kiện thuận lợi cho việc áp dụng hình ảnh kỹ thuật số của da vào các nghiên cứu tương tự như "Nhận biết ung thư da". Tổ chức được xây dựng nhằm phát hiện và điều trị ở giai đoạn sớm nhất, giúp giảm tỷ lệ tử vong. ISIC được điều hành dưới sự bảo trợ của Trung tâm Ung thư Memorial Sloan Kettering và được hỗ trợ bởi các khoản đóng góp từ thiện (Nhà tài trợ và Đối tác) và hỗ trợ bằng hiện vật từ các thành viên.

- Thông tin về ISIC Challenge:

"ISIC Challenge" là một cuộc thi được tổ chức bởi Học viện Nghiên cứu Ung thư da Quốc tế (International Skin Imaging Collaboration - ISIC) về lập trình nhận diện ung thư da, bao gồm cả việc phân loại và phát hiện ung thư da.

Cuộc thi ISIC Challenge thường được tổ chức hàng năm và mục tiêu của cuộc thi là khuyến khích các nhà nghiên cứu và chuyên gia trong lĩnh vực lập trình, học máy và trí tuệ nhân tạo (AI) phát triển các giải pháp nhận diện ung thư da thông qua các bức ảnh chụp da.

Các đội tham gia sẽ được cung cấp một tập dữ liệu ảnh chụp da để phát triển các giải pháp nhận diện ung thư da. Cuộc thi sẽ đánh giá các giải pháp dựa trên khả năng phát hiện ung thư da và phân loại chúng theo các loại khác nhau. Cuộc thi này đã thu hút sự tham gia của các nhà nghiên cứu, sinh viên và chuyên gia từ khắp nơi trên thế giới.

Thông tin chi tiết về cuộc thi ISIC Challenge và các quy định tham gia có thể được tìm thấy trên trang web của ISIC: https://challenge.isic-archive.com/ .

- Cách ISIC thu thập dataset:

Để tổ chức cuộc thi ISIC Challenge về nhận diện ung thư da, ISIC sử dụng cơ sở dữ liệu hình ảnh lớn về các bệnh lý da được gọi là ISIC Archive. Tập dữ liệu này chứa hàng nghìn hình ảnh chụp da từ các bệnh nhân với các loại bệnh da khác nhau, bao gồm cả ung thư da.

Để thu thập bộ dữ liệu cho cuộc thi ISIC Challenge, ISIC đã sử dụng nhiều nguồn khác nhau. Các nguồn bao gồm các cơ sở y tế, phòng khám da liễu, các tổ chức tài trợ và cộng đồng người dùng ISIC.

Một số ảnh trong bộ dữ liệu được chụp từ các cơ sở y tế và phòng khám da liễu trên toàn thế giới, bao gồm cả ảnh từ các nước có thu nhập thấp và trung bình. Việc thu thập ảnh từ các cơ sở y tế được thực hiện với sự hỗ trợ của các bác sĩ chuyên khoa da liễu và các chuyên gia y tế khác. Các ảnh này bao gồm các loại ung thư da khác nhau, từ các bệnh nhân khác nhau và được chụp bằng nhiều phương pháp khác nhau.

Ngoài ra, ISIC cũng thu thập các ảnh từ cộng đồng người dùng của mình, cho phép bất kỳ ai trên thế giới có thể đóng góp hình ảnh của mình vào bộ dữ liệu. Đây là một cách để thu thập các ảnh hiếm gặp hoặc các dạng ung thư da không phổ biến.

Tất cả các ảnh thu thập được đã được kiểm tra và xác minh bởi các chuyên gia để đảm bảo tính chính xác và đầy đủ. Sau đó, các ảnh đã được gán nhãn bởi các chuyên gia da liễu để phân loại các dạng ung thư da khác nhau. Bộ dữ liệu này được cung cấp miễn phí cho các đội tham gia cuộc thi ISIC Challenge để phát triển và đánh giá các giải pháp nhận diện ung thư da.

Dữ liệu được thu thập từ các nhà cung cấp hình ảnh chụp da, bao gồm các bệnh viện, phòng khám và các trung tâm y tế khác trên toàn thế giới. Các nhà cung cấp hình ảnh này đã cho phép ISIC sử dụng dữ liệu của họ để xây dựng cơ sở dữ liệu ISIC Archive.

Ngoài ra, nhóm nhận được sự hỗ trợ dữ liệu từ kho lưu trữ hình ảnh của Đại học Y - Dược Huế. Tuy không phong phú như bộ dữ liệu từ ISIC nhưng gần gũi với người dân Việt Nam hơn do các tính chất về cơ địa, màu da,…

2. Xây dựng dữ liệu:

Đầu tiên, nhóm thực hiện kiểm định dữ liệu và sàn lọc cơ bản, để chắc chắn rằng tất cả dữ liệu thô tương tự nhau về chất lượng hình ảnh. Các yếu tố như góc chụp, ánh sáng, kích thước,… sẽ được cân nhắc.

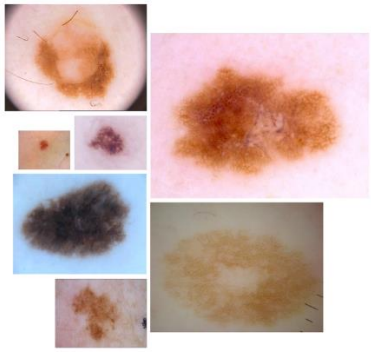
Sau đó, dữ liệu thô đã được làm sạch. Tùy vào việc mô hình có hoặc không sử dụng Feature Engineering sẽ được xử lý dữ liệu khác nhau. Dữ liệu thu thập được sẽ chia thành 3 tập khác nhau: tập huấn luyện (training data), tập kiểm định (validation data) và tập kiểm tra (testing data).

Trong đó, việc huấn luyện mô hình sẽ được thực hiện trên khoảng 700 ảnh (70% của bộ dữ liệu từ ISIC). Tập kiểm tra chứa khoảng 300 ảnh (30% còn lại của bộ dữ liệu từ ISIC). Để tránh tình trạng overfit, nhóm sẽ sử dụng dữ liệu từ Đại học Y - Dược Huế làm tập kiểm định. Phương pháp này có thể đạt được độ chính xác trung bình hơn 50%.

III. MÔ HÌNH KHÔNG SỬ DỤNG FEATURE ENGINEERING

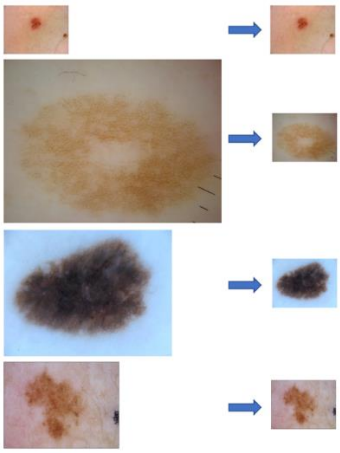
1. Xử lý dữ liệu

Sau khi có bộ dữ liệu khám phá được thu thập trên hình ảnh nốt ruồi lành tính và ác tính thì ta thấy bộ dữ liệu có kích thước chưa được đồng bộ với kích thước tối đa và tối thiểu của hình ảnh lần lượt là (4459,6688) và (450,600) cho cả hai loại hình ảnh.



Hình 3.1.1: Hình ảnh trích xuất từ bộ dataset

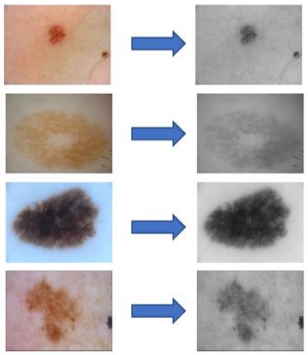
Vì vậy ta sẽ tiến hành thu nhỏ tất cả hình ảnh trong bộ dữ liệu về cùng 1 kích thước 450x600.



Hình 3.1.2: Thu nhỏ ảnh về cùng kích thước

Sau khi đưa các ảnh về cùng một kích thước thì tiếp đến ta sẽ tiến hành chuyển đổi ảnh sang ảnh xám vì các ảnh dữ liệu trong dataset sẽ là ảnh màu RGB gồm 3 kênh màu (red, green,blue) có tổng cộng 3 chiều sẽ gây tốn chi phí tính toán và thời gian đào tạo

model nên ta sẽ đổi ảnh sang kênh màu xám [0..255] để giảm chi phí và thời gian đào tạo mô hình đồng thời các đặc trưng của hình ảnh cũng sẽ được thu nhỏ lại.



Hình 3.1.3: Chuyển ảnh sang ảnh xám

Các giá trị pixel từ các hình ảnh trên sẽ được chuyển đổi sau đó sẽ được trích xuất thành một mảng và sau đó được trích xuất thành một mảng. Tiếp đến mảng đó sẽ được chuyển đổi thành một khung dữ liệu với lớp của mỗi được đính kèm làm cột cuối cùng. Lớp ác tính sẽ được gán nhãn là 0, ngược lại lớp lành tính sẽ được gán nhãn là 1.

	0	1	2	3	4	5	6	7	8	9	...	269991	269992	269993	269994	269995	269996	269997	269998	269999	270000
0	152	84	88	80	89	85	85	82	86	85	...	87	87	85	84	85	86	93	88	148	1
1	95	97	99	100	101	102	103	104	107	108	...	154	155	154	154	155	154	152	152	153	1
2	52	55	57	59	60	60	59	59	59	61	...	13	16	15	14	16	17	16	15	16	1
3	151	85	90	81	90	87	88	85	86	86	...	85	87	87	88	88	84	87	86	152	1
4	40	48	42	39	49	50	43	44	49	47	...	14	16	17	17	17	16	15	16	16	1

	0	1	2	3	4	5	6	7	8	9	...	269991	269992	269993	269994	269995	269996	269997	269998	269999	270000
0	111	112	114	115	115	114	113	112	114	113	...	144	142	140	140	141	141	139	138	138	0
1	3	3	2	1	1	2	3	4	4	6	...	193	192	192	191	189	186	185	186	187	0
2	247	206	211	209	216	213	208	217	213	213	...	215	218	214	214	215	207	222	195	245	0
3	155	156	159	162	164	166	173	181	182	183	...	151	151	147	147	147	143	141	141	140	0
4	154	153	152	151	152	153	155	156	154	155	...	149	150	149	149	149	149	149	150	150	0

Hình 3.1.4: Trích xuất các đặc trưng pixel của ảnh

Quá trình này được thực hiện cho từng lớp và được nối lại với khung dữ liệu chứa 27,000 đặc trưng.

Vì các đặc trưng số trong khung dữ liệu rất lớn, nên việc giảm kích thước sẽ phải được thực hiện và để chuẩn bị cho việc này. Do đó, khung dữ liệu được chia tỷ lệ bình thường và được phân loại.

Sau đó, khung dữ liệu này được chia ngẫu nhiên thành tập huấn luyện và kiểm tra theo tỷ lệ tương ứng là 70/30 với PCA được áp dụng cho tập huấn luyện tính năng để xác định các đặc trưng chiếm 90% phương sai. Hóa ra có 24 đặc trưng chịu trách nhiệm cho việc này trong số 27000 tính năng. Điều này sau đó đã được sử dụng để chuyển đổi đào tạo đặc trưng và thử nghiệm.

2. Các mô hình xử lý & đánh giá

Sau khi xử lý các đặc trưng của ảnh, bước tiếp theo là xác định các thuật toán để phân loại chính xác các đặc tính của hình ảnh thử nghiệm sau khi đã được huấn luyện. Đối với đề tài này thì nhóm chúng em lựa chọn 3 thuật toán sau để huấn luyện: Logistic Regression, Support vector machine, KNN.

a. Logistic Regression

Thuật toán logistic regresion là một trong những thuật toán phân loại phổ biến trong machine learning. Nên nhóm chúng em sẽ áp dụng thuật toán này để tiến hành phân loại các dữ liệu đã được xử lý ở trên.

Logistic regression là một thuật toán dựa trên thống kê. Trong bài toán này thì input đầu vào (1 vài đặc trưng của hình ảnh nốt ruồi) và trả về kết quả đầu ra (Y) đại diện cho kết quả dự đoán rằng có ung thư hay không.

Về mặt toán học Logistic Regression là một loại thuật toán supervised learning tính toán mối quan hệ giữa các feature trong input và output dựa trên hàm logistic/sigmoid. Mặc dù gọi là Logistic Regression nhưng thuật toán này không dự đoán ra giá trị thực như các thuật toán Regression khác, Logistic Regression được dùng để dự đoán ra một kết quả nhị phân (với giá trị 0/1 hay -1/1 hay True/False) dựa vào input của nó. Nhưng Logistic Regression cũng có một chút giống với Linear Regression trong quá trình xây dựng model.

Trong Linear Regression output y được tính toán bằng tổng tích giữa các biến input và hệ số w của model

$$y = h_0x_0 + h_1x_1 + h_2x_2 + \cdots + h_nx_n$$

Mục đích của Linear Regression là ước tính giá trị cho các hệ số mô hình c, w1, w2, w3… .wn và fit nó với dữ liệu huấn luyện với hàm loss tối thiểu và dự đoán đầu ra y.

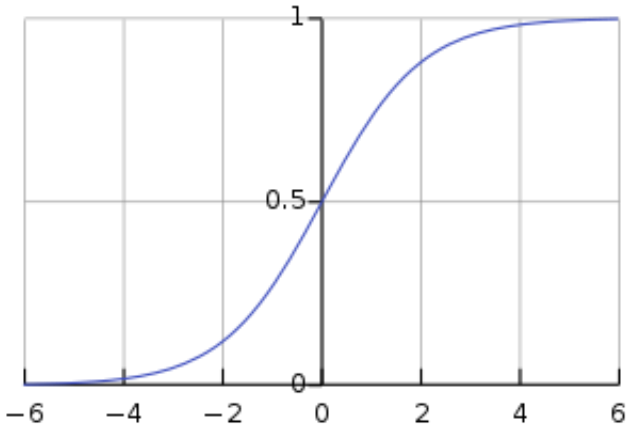
Logistic Regression thực hiện điều tương tự, nhưng với một bổ sung. Nó chạy kết quả thông qua một hàm non-linear (phi tuyến tính) đặc biệt được gọi là hàm logistic hoặc hàm sigmoid để tạo ra đầu ra là một xác suất p.

Công thức hồi quy của model Logistic Regression:

Với $log(p_1 - p)$ được gọi là logit(p) hay còn gọi là log-odds ta sẽ tính được xác suất p như sau:

$$p = \frac{1}{1 + e^{-(h_0 + h_1x_1 + \cdots + h_nx_n)}}$$

Logistic được ký hiệu là S_0 là hàm sigmoid với đầu ra là một số có giá trị từ 0 đến 1 được định nghĩa với công thức sau: $S_0(t) = \frac{1}{1+exp(-t)}$ với đồ thị được biểu thị bên dưới.



Hình 3.2.1: Biểu diễn đồ thị của hàm S_0

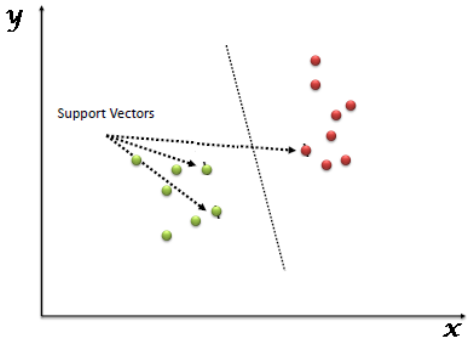
Chú ý rằng $S_0(t) < 0.5$ khi $t < 0$ và $S_0(t) \geq 0.5$ khi $t \geq 0$ mỗi khi Logistic Regression tính xác xuất P model sẽ đưa ra dự đoán với công thức sau:

$$\begin{cases} 1 \; Nếu \; p < 0.5 \\ 0 \; Nếu \; p \geq 0.5 \end{cases}$$

Với y^ là kết quả dự đoán.

b. Support Vector Machine (SVM)

Support Vector Machine hay còn được gọi là SVM là một thuật toán học có giám sát, cũng là một trong những thuật toán phổ biến thường được sử dụng cho việc phân loại trong machine learning.

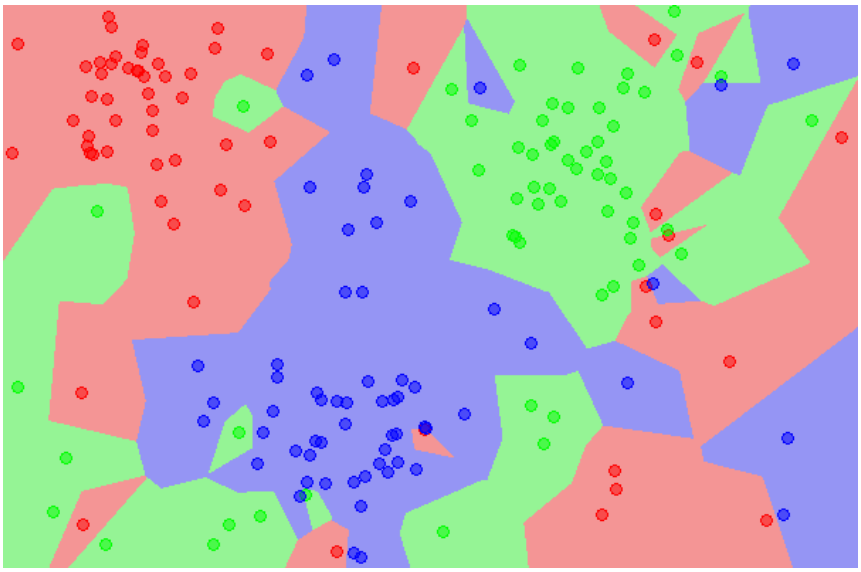


Hình 3.2.2: Hình minh họa SVM

Cũng giống như Logistic Regression ở trên thì ở thuật toán này ta cũng sẽ biểu diễn các đặc trưng của hình ảnh nốt ruồi trên một mặt phẳng không gian n chiều và sau đó ta sẽ đi tìm đường thẳng phân chia các lớp với nhau được gọi là hyper-plane.

c. K-Nearest Neighbors (KNN)

K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression.



Hình 3.2.3: Hình minh họa KNN

Trong bài toán phân loại nốt ruồi này, khi label của một điểm dữ liệu mới khi được thêm vào được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label.

3. Đánh giá

Để đánh giá các mô hình trên thì nhóm đã đưa ra các hướng đánh giá sau. Nhóm sẽ tiến hành chia bộ dữ liệu ra 70% training và 30% dùng cho testing.

Sau khi hiện đo các chỉ số như Accuracy, Precision, Recall, f1-score, Cross-Validation score thì thu được kết quả như sau:

Classifier	Accuracy	Precision	Recall	F1-score	Cross-Validation score
Logistic regression	0.522	0.48	0.609	0.537	0.6476
SVM	0.533	0.490	0.609	0.543	0.623
KNN	0.533	0.491	0.731	0.588	0.657

Nhìn vào kết quả trên ta có thể thấy được thuật toán tốt nhất trong 3 thuật toán trên là SVM và KNN với Accuracy = 0.533 tuy nhiên, KNN cho ra Precision và Recall cao nhất. Thấp nhất là Logistic Regression với Accuracy = 0.522.

Bên cạnh các thông số trên thì trong bài toán phân loại người ta cũng thường đánh giá theo một phương pháp khác đó là sử dụng confision matrix.

Classifier	True Positive	Flase Positive	True Negative	Flase Negative
Logistic regression	27.78%	30 %	24.44%	17.78%
SVM	27.78%	28.89 %	25.56%	17.78%
KNN	33.33%	34.44 %	20%	12.22%

Nhưng xét trong confusion matrix thì KNN lại là thuật toán dự đoán nốt ruồi lành tính tốt nhất với 33.33%.

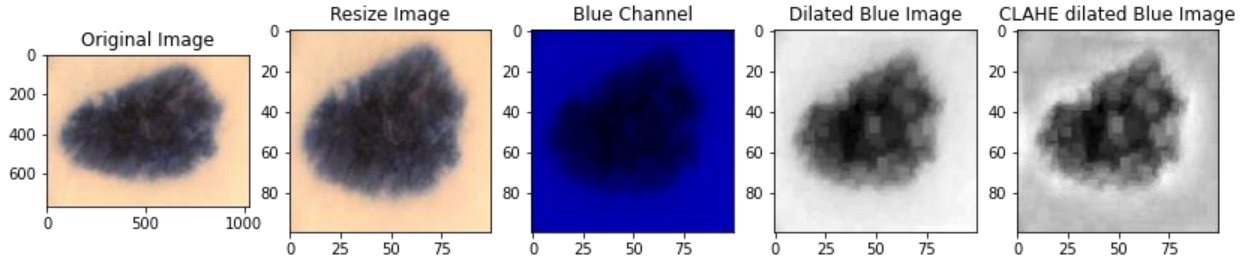
IV. MÔ HÌNH SỬ DỤNG FEATURE ENGINEERING

Đầu vào của bài toán đóng vai trò quan trọng trong việc tăng hiệu suất của bài toán. Do đó, chúng tôi sử dụng thêm các phương pháp xử lý và khử nhiễu, cụ thể như sau: Chuyển đổi không gian màu, tăng cường độ tương phản, loại bỏ hình ảnh dư thừa và phân đoạn. Với bộ dữ liệu được xử lý này, sẽ sử dụng mô hình Gaussian Navie Bayes và SVM để phân loại.

1. Xử lý dữ liệu

Hình ảnh đầu vào sẽ được thay đổi kích thước thành 100x100 pixel. Điều này sẽ giúp giảm kích thước tính năng của hình ảnh, cải thiện hiệu quả tính toán. Đồng thời nén các hình ảnh để dễ dàng hơn trong việc lưu trữ.

Sau đó, các hình ảnh sẽ bị loại bỏ kênh màu xanh lá và đỏ, chỉ giữ lại kênh màu xanh lam (Blue Channel). Thông qua cách này, hình ảnh của nốt ruồi sẽ trở nên rõ ràng hơn.



Các hình ảnh được giãn ra (Dilated) bằng cách sử dụng phân đoạn hình thái trong kênh màu xanh lam. Nó mở rộng ranh giới của các phần tử tiền cảnh và giảm số lượng các phần tử nền trong một hình ảnh. Pixel đầu ra sẽ là giá trị lớn nhất trong số tất cả các pixel nằm trong kernel. Điều này sẽ làm tăng vùng trắng trong ảnh. Quá trình này cũng chuyển hình ảnh sang thang độ xám.

Sau khi làm giãn hình ảnh, cân bằng biểu đồ thích ứng hạn chế độ tương phản (CLAHE) được áp dụng để cải thiện độ tương phản của hình ảnh cục bộ. Phương pháp này áp dụng hostogram, mỗi hình ảnh tương ứng với phần riêng biệt của hình ảnh và sử dụng để phân phối lại các giá trị pixel của hình ảnh, từ đó cân bằng sự giãn nở được thực hiện trước đó.

Quá trình cuối cùng để trích xuất các tính năng từ các hình ảnh đã xử lý là bằng cách sử dụng Biểu đồ độ dốc (HOG). Quá trình này trích xuất gradient và hướng của các cạnh hình ảnh. Gradient của hình ảnh cũngthay đổi mạnh với các giá trị pixel của ảnh và điều này được giảm thiểu bằng cách chuẩn hóa dữ liệu.

Sau đó 10000 features được giảm xuống còn khoảng 4300 features. Các features sẽ được lưu dưới dạng dataframe và được dùng trong mô hình phân loại.

2. Các mô hình xử lý và đánh giá

a. Gaussian Navie Bayes

Định lý Bases

Công thức chỉ ra xác suất của A xảy ra nếu B cũng xảy ra, ta viết là **P(A|B)**. Và nếu ta biết xác suất của B xảy ra khi biết A, ta viết là **P(B|A)** cũng như xác suất độc lập của A và B.

- P(A|B) là "xác suất của A khi biết B"
- P(A) là xác suất xảy ra của A
- P(B|A) là "xác suất của B khi biết A"
- P(B) là xác suất xảy ra của B

Ví dụ, P(lửa) là xác suất có lửa, P(khói) là xác suất ta nhìn thấy khói. Ta sẽ có những trường hợp sau:

P (Lửa | Khói) có nghĩa là tần suất có lửa khi chúng ta nhìn thấy khói. P (Khói | Lửa) có nghĩa là chúng ta thường thấy khói khi có lửa.

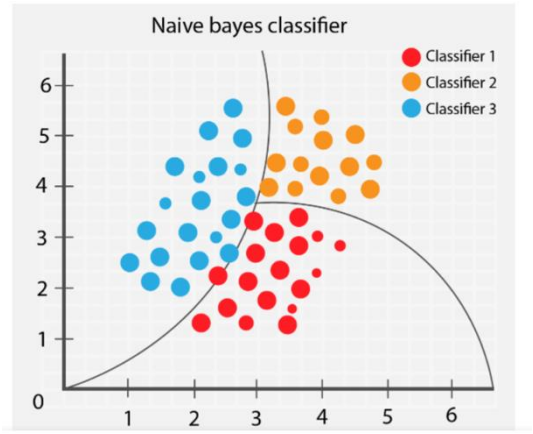
Công thức sẽ cho chúng ta biết được điều gì xảy ra tiếp theo nếu ta đã biết một điều.

Ví dụ: Một đám cháy nguy hiểm là có xác suất là 1% nhưng khói lại khá phổ biến là 10% (từ các nhà máy) và 90% đám cháy nguy hiểm tạo ra khói. Vậy ta có:

P (Lửa | Khói) = P (Lửa) P (Khói | Lửa) = 1% x 90% = 9% P (Khói) 10%

Trong trường hợp này, 9% khả năng thấy khói có nghĩa là có một đám cháy nguy hiểm.

Phân loại Naive Bayes:



Naive Bayes là một thuật toán phân loại cho các vấn đề phân loại nhị phân (hai lớp) và đa lớp. Kỹ thuật này dễ hiểu nhất khi được mô tả bằng các giá trị đầu vào nhị phân hoặc phân loại.

Thuật toán Naive Bayes tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất.

Tuy nhiên, ta cần lưu ý giả định của thuật toán Naive Bayes là các yếu tố đầu vào được cho là độc lập với nhau.

Gaussian Navie Bayes

Mô hình này được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.

Với mỗi chiều dữ liệu i và một class c, x_i tuân theo một phân phối chuẩn có kỳ vọng μ_{ci} và phương sai σ_{ci}^2:

$$p(x_i|c)=p(x_i|\mu_{ci},\sigma_{ci}^2)=\frac{1}{\sqrt{2\pi\sigma_{ci}^2}}exp(-\frac{(x_i-\mu_{ci})^2}{2\sigma_{ci}^2})$$

Trong đó, bộ tham số $\theta=\{\mu_{ci},\sigma_{ci}^2\}$ được xác định bằng Maximum Likelihood:

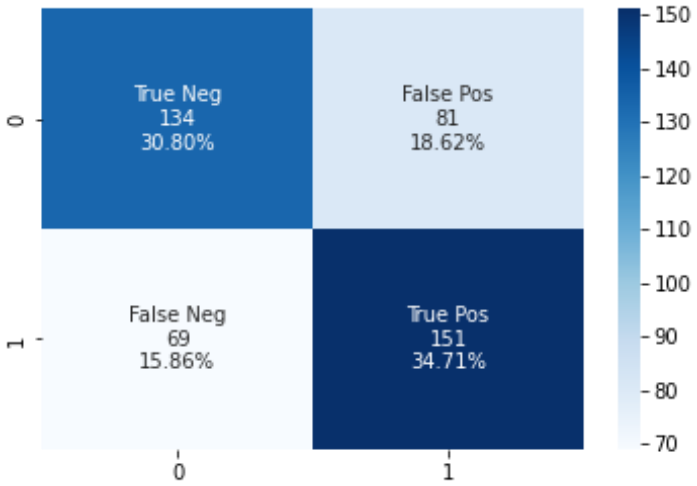
$$(\mu_{ci},\sigma_{ci}^2)=argmax\ \prod_{n=1}^{N}p(x_i^n|\mu_{ci},\sigma_{ci}^2)$$

Đây là cách tính của thư viện sklearn. Chúng ta cũng có thể đánh giá các tham số bằng MAP nếu biết trước priors của μ_{ci} và σ_{ci}^2.

b. Support Vector Machine (SVM)

SVM được sử dụng và được đánh dựa trên các độ đo accuracy, precision, recall and f-1 score, giống như phần trước. Kết quả cho ra khá tốt, Accuracy đạt 0.65%, ngoài ra precision cũng đạt 0.68, recall đạt 0.68 và F1-score đạt 0.66. về chi tiết các thông tin về độ tương quan giữa TF, NF, TP, NP cũng được thể hiện dưới biểu đồ dưới đây.

3. Đánh giá



Classifier	True Positive	Flase Positive	True Negative	Flase Negative
Gaussian Navie Bayes	27.78%	26,67%	27,78%	17,78%
SVM	26,67%	23,33%	31,11%	18,89%

V. THỬ NGHIỆM

Dựa vào các thông số từ các mô hình, kết quả đến từ KNN là kết quả tốt nhất. Do đó, chúng tôi sẽ sử dụng mô hình KNN đã được đào tạo để thử nghiệm trên bộ dữ liệu tự thu thập được.

index	Accuracy	Precission	Recall	F1 Score	Validation
Logistic Regression	0.6551724137931034	0.6711111111111111	0.6651982378854625	0.6681415929203539	0.6640652300524169
KNN	0.6735632183908046	0.6888888888888889	0.6828193832599119	0.6858407079646017	0.694535041739468
SVM	0.6574712643678161	0.6875	0.6299559471365639	0.6574712643678161	0.6591343241858081
Gaussian naive bayes (FE)	0.6689655172413793	0.7065217391304348	0.5909090909090909	0.6435643564356436	0.7034459328285769
SVM (FE)	0.6551724137931034	0.6508620689655172	0.6863636363636364	0.6681415929203539	0.6365851291011454

Với bộ dữ liệu tự thu thập, kết quả cho ra như sau:

```
Accuracy: 0.6551724137931034
Precision: 0.6508620689655172
Recall: 0.6863636363636364
F1-score: 0.6681415929203539
```

VI. TỔNG KẾT

Ngay cả với một bộ dữ liệu nhỏ gồm hình ảnh (300) và không có mạng nơ-ron, việc thực hiện các phương pháp xử lý hình ảnh mới và khử nhiễu trên tập dữ liệu hình ảnh và sử dụng bộ phân loại máy học có giám sát đã cung cấp độ chính xác xác thực chéo là 0,757. Điều này cho thấy tiềm năng thực sự trong việc phát hiện và dự đoán bệnh.

Kết quả cũng chỉ ra rằng các đặc điểm quan trọng nhất để xác định ung thư là cường độ và hình dạng của các tổn thương trên da. Độ chính xác cũng được cải thiện với một tập dữ liệu lớn hơn. Điều này cũng được thể hiện thông qua kết quả ghi lại với bộ dữ liệu 300 hình ảnh và 1400 hình ảnh.