UIT Trường Đại học Công nghệ Thông tin
Khoa Khoa học và Kỹ thuật Thông tin

DS108.O21

PHÂN LOẠI TỔN THƯƠNG DA QUA HÌNH ẢNH SOI DA

MÔN HỌC: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU - DS108.O21

GVHD:

TS. Nguyễn Gia Tuấn Anh

CN. Trần Quốc Khánh

Nhóm 08:

Trần Đại Hiển - 22520426

Hồ Ngọc Mai - 22520839

Lê Ngọc Thiên Phúc - 22521117

01 GIỚI THIỆU

Động lực

Phân loại tổn thương da là một lĩnh vực quan trọng trong chẩn đoán da liễu. Việc phân loại thủ công dựa trên kinh nghiệm của bác sĩ có thể gặp nhiều hạn chế như sai sót do con người, tốn thời gian và phụ thuộc vào chuyên môn của bác sĩ.

Mục tiêu

Phát triển hệ thống phân loại tổn thương da mạnh mẽ dựa trên hình ảnh soi da.

Đóng góp

Thu thập và xử lý dữ liệu
Phát triển hệ thống phân loại tự động

Tổng quan đề tài

Bài toán: Sử dụng mô hình máy học để phân loại tổn thương da đa lớp thông qua hình ảnh soi da

Input: Hình ảnh soi da của bệnh nhân, có kích thước WxHx3 pixel, với W và H là số nguyên dương, số 3 thể hiện cho ảnh ở kênh màu RGB.
Output: Nhãn tổn thương da mà bệnh nhân mắc phải tương ứng cùng điểm số tin cậy.

02 NỀN TẢNG NGHIÊN CỨU

Công trình liên quan

1. Phân loại tổn thương da bằng mô hình học sâu: Esteva và cộng sự đã sử dụng kiến trúc Inception v3 CNN và đạt độ chính xác ấn tượng với 86.6%, chứng minh tiềm năng của công nghệ học sâu trong phân loại tổn thương da là vô cùng to lớn.
2. Kỹ thuật tăng cường dữ liệu để nâng cao hiệu suất: Fabio Perez và cộng sự đề xuất kỹ thuật tăng cường dữ liệu sử dụng các phép biến đổi hình học ngẫu nhiên để tăng kích thước và tính đa dạng dữ liệu.
3. So sánh tính chính xác của các mô hình học máy truyền thống: Grignaffini và cộng sự đã đánh giá các kỹ thuật học máy khác nhau để phát hiện và phân loại ung thư da từ nhiều bộ dữ liệu.
4. Phân tích tập trung các loại tổn thương cụ thể: Zafar và cộng sự đã phân loại các tổn thương và cung cấp cái nhìn tổng quan về các kỹ thuật phân tích và chẩn đoán ung thư da từ nhiều bộ dữ liệu khác nhau.

03 BỘ DỮ LIỆU

Nguồn thu thập

Ảnh soi da được thu thập từ hai trang web

- dermnetnz.org thu thập được **107 ảnh**
- www.isic-archive.com thu thập được **8600 ảnh**

DermNet NZ gồm **3** loại tổn thương da

ISIC gồm **7** loại tổn thương da

→ Bộ dữ liệu hoàn chỉnh chứa tổng cộng **8707 ảnh** với **7 loại tổn thương da**

03 BỘ DỮ LIỆU

So sánh bộ dữ liệu

	akiec	bcc	bkl	df	mel	nv	vasc	Tổng
HAM10000	327	514	1099	115	1113	6705	142	10015
ISIC2017	-	-	254	-	374	1372	-	2000
PH2	-	-	-	-	40	160	-	200
Dataset nhóm	1320	1542	1045	300	1200	3000	300	8707

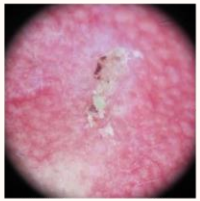
03 BỘ DỮ LIỆU

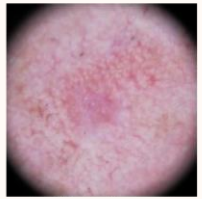
Mô tả dữ liệu

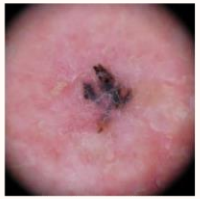
Bộ dữ liệu có tổng cộng **8707** hình ảnh. Bộ dữ liệu bao gồm **bảy loại tổn thương da** như sau: **mel, nv, akiec, bcc, bkl, df, vasc.**

Metadata bao gồm **8707 dòng** và có **7 cột thuộc tính.**

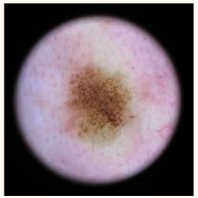
Thuộc tính	Ý nghĩa
image_id	Mã định danh duy nhất cho mỗi hình ảnh
attribution	Nguồn cung cấp hình ảnh
age_approx	Tuổi xấp xỉ của bệnh nhân
anatom_site_general	Vị trí giải phẫu của tổn thương da
dx	Chẩn đoán y tế cho tổn thương da
sex	Giới tính của bệnh nhân
lesion_id	Mã định danh duy nhất cho mỗi tổn thương da trong hình ảnh


actinic keratosis (akiec)

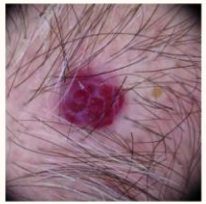

basal cell carcinoma (bcc)


benign keratosis (bkl)


dermatofibroma (df)

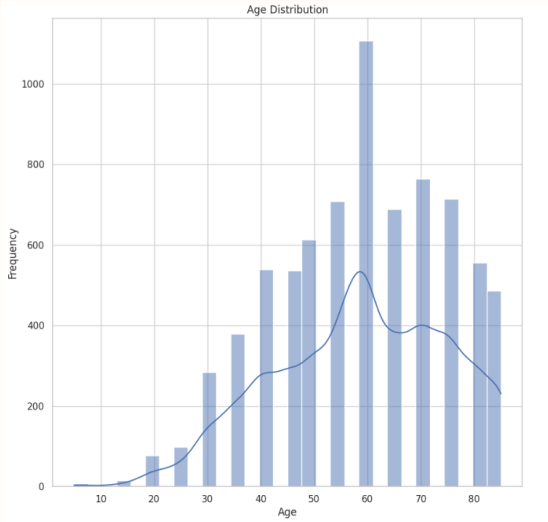

melanoma (mel)


nevus (nv)

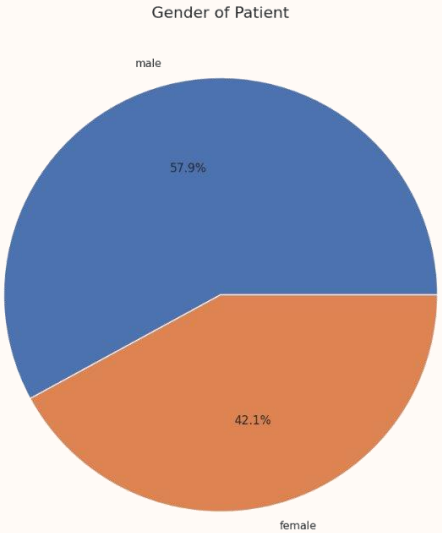

vascular lesion (vasc)

03 BỘ DỮ LIỆU

Khai phá dữ liệu



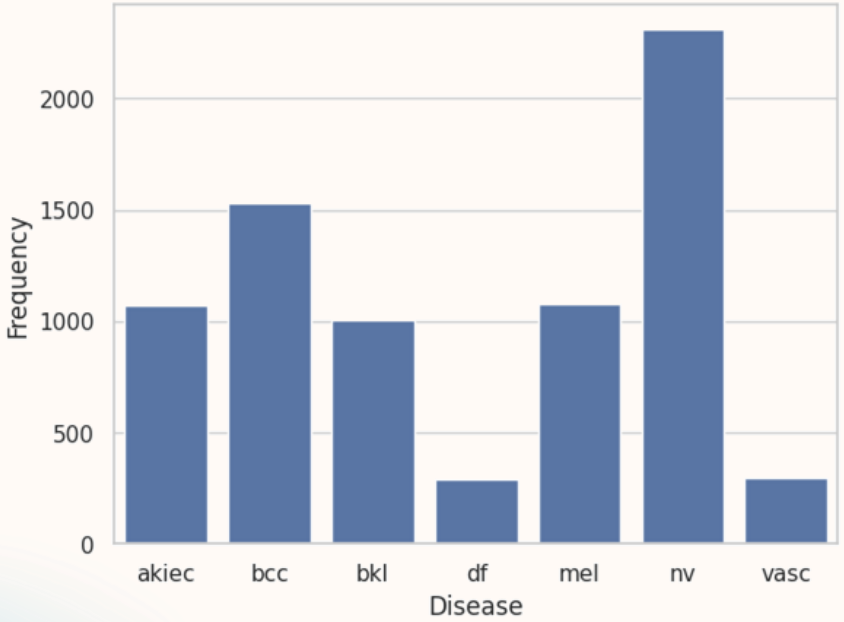
Phần lớn bệnh nhân trong nghiên cứu là nam, chiếm khoảng **57.9%**. Thiểu số bệnh nhân trong nghiên cứu là nữ giới, chiếm khoảng **42,1%**.



Tổn thương da phân bố ở mọi lứa tuổi. Tuy nhiên, số lượng người sau 50 tuổi tăng đáng kể. Nhóm tuổi có tần suất mắc bệnh cao nhất là **50-60 tuổi**.

03 BỘ DỮ LIỆU

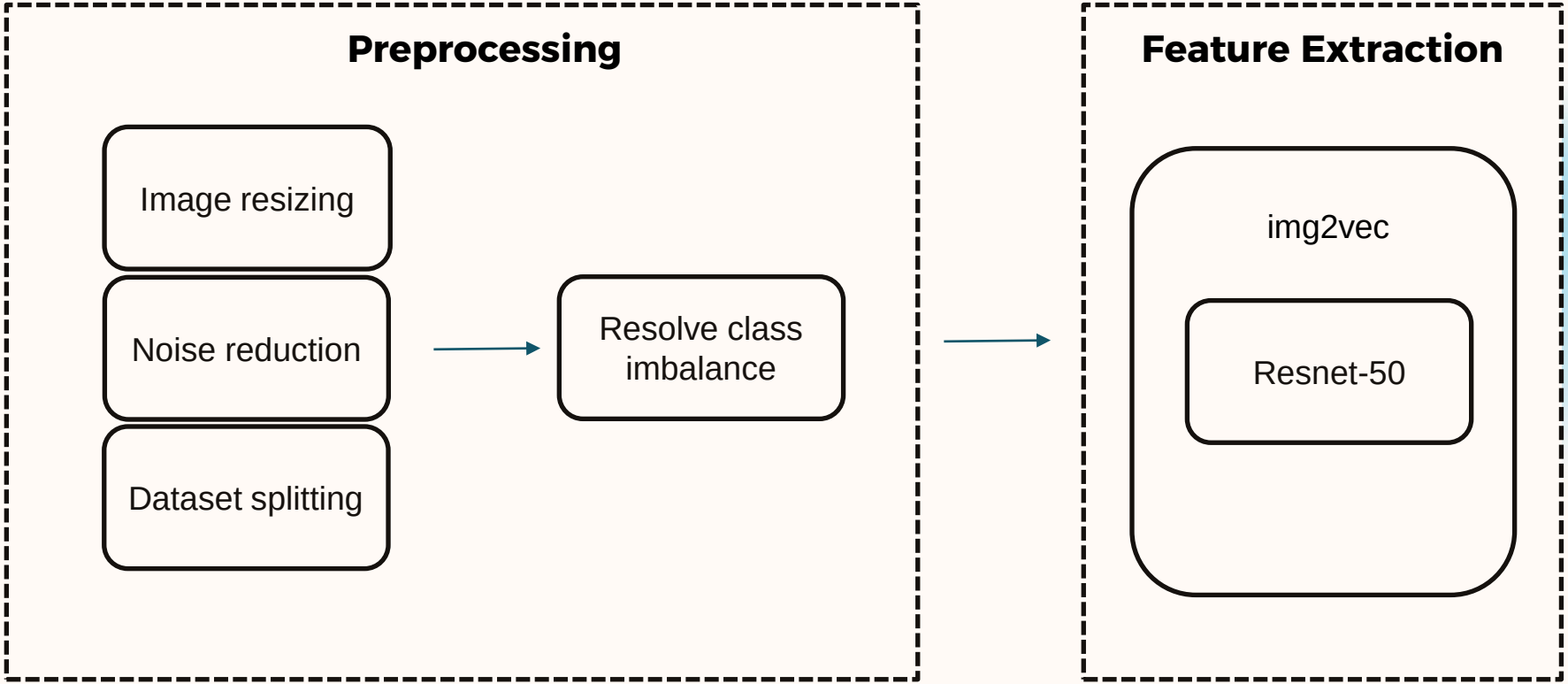
Khai phá dữ liệu



Loại tổn thương da thường gặp nhất là **nv (nevus)** gấp **10** lần so với hai loại tổn thương da có tỉ lệ nhỏ nhất, bao gồm **df (dermatofibroma)** và **vasc (Vascular lesion).**

Tập dữ liệu bị mất cân bằng.

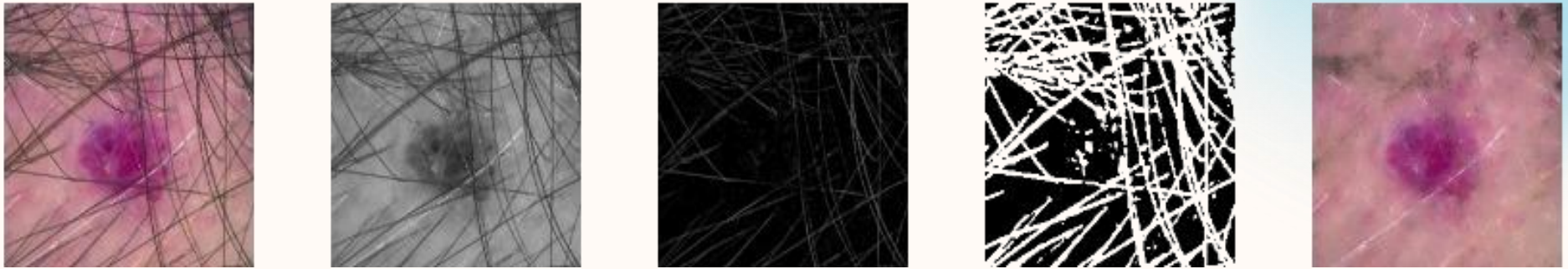
04 TIỀN XỬ LÝ DỮ LIỆU



04 TIỀN XỬ LÝ DỮ LIỆU

Crop image and Noise Reduction: Hair removal

Cropped Image → Gray Scale → Black Hat transform → Binary Theshold → impaint to output image

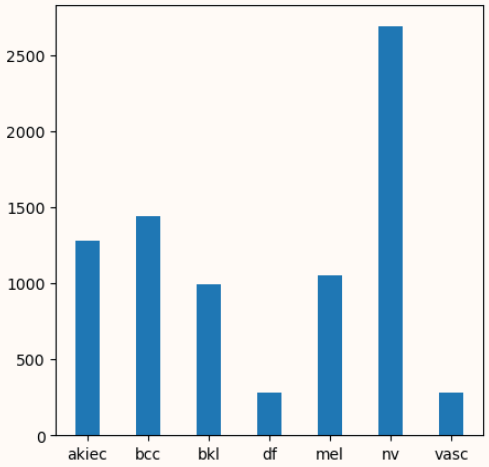


04 TIỀN XỬ LÝ DỮ LIỆU

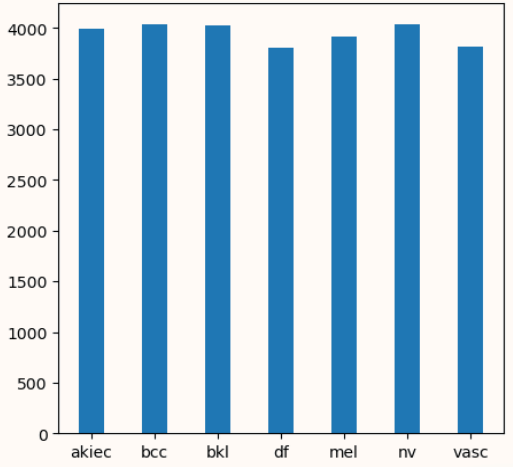
Tách thành train/test và xử lý mất cân bằng dữ liệu trên tập train

Geometric data augmentation

- Xoay ảnh ngẫu nhiên trong khoảng 360°.
- Dịch chuyển ảnh theo chiều ngang và chiều dọc với tỷ lệ tối đa 10%.
- Phóng to hoặc thu nhỏ ảnh trong khoảng 10%.
- Lật ngang, lật dọc ảnh.
- Điền các pixel bị mất bằng các pixel gần nhất



Before

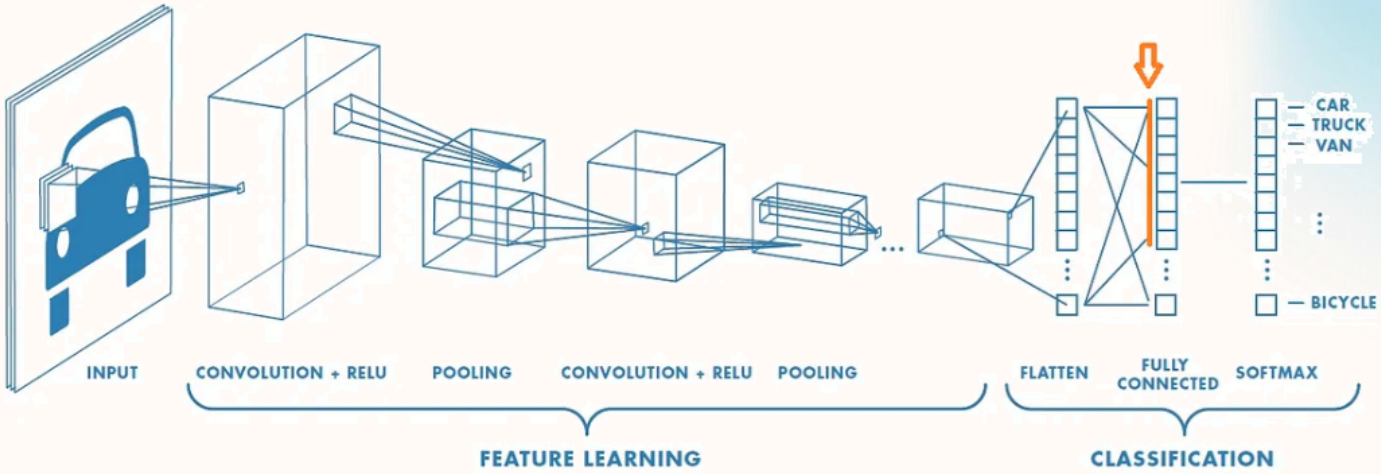


After

04 TIỀN XỬ LÝ DỮ LIỆU

Feature Extraction: Chuyển từ hình ảnh thành các vectors

Sử dụng pre-trained model: Resnet-50 -> 512 features



05 MÔ HÌNH HỌC MÁY

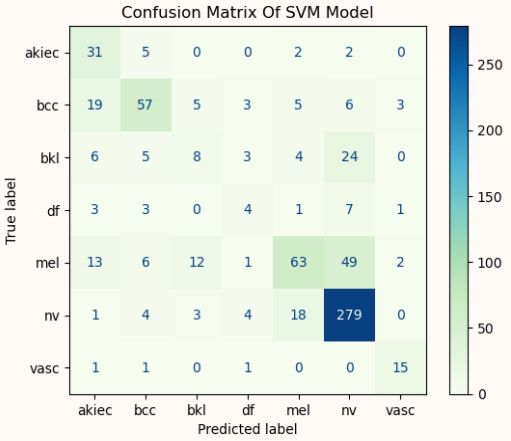
- **Các mô hình học máy mà nhóm sử dụng:**
  - Support Vector Machine
  - Random Forest Classifier
  - Light - GBM Classifier
- **Các độ đo hiệu suất mà nhóm sử dụng:**
  - Recall
  - Precision
  - F1-score

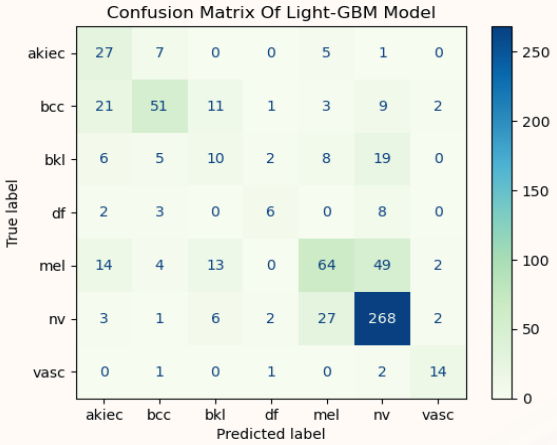
  (các độ đo đều dùng phương pháp macro để tính toán)
- **Confusion Matrix**
  - Trực quan hóa sự hỗn loạn giữa nhãn thực tế và nhãn dự đoán

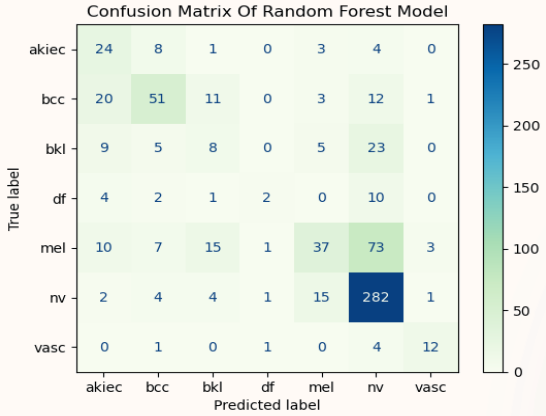
06 KẾT QUẢ THỰC NGHIỆM

Kết quả:

	Recall	Precision	F1-score
SVM	56%	54%	53%
Random Forest	46%	51%	46%
Ligth-GBM	54%	55%	53%







Kết quả hiệu suất của các mô hình đều không được cao, nhưng đa số đều trên 50.

Hiệu suất nhật dạng các nhãn của các mô hình được thể hiện qua ma trận nhầm lẫn.

06 KẾT QUẢ THỰC NGHIỆM

Hạn chế và hướng phát triển của nghiên cứu:

Hạn chế

- Kỹ thuật data augmentation cơ bản
- Dữ liệu thu thập từ khu vực nhất định
- Công cụ xóa tóc còn hạn chế

Hướng phát triển

- Nâng cao hiệu quả của mô hình
- Mở rộng khả năng phân loại
- Phát triển ứng dụng hỗ trợ