

TÓM TẮT BÁO CÁO ĐỒ ÁN CUỐI KÌ MÔN MACHINE LEARNING

* Giảng viên: PGS.TS. Lê Đình Duy,
ThS. Phạm Nguyễn Trường An

ĐỀ TÀI

NHẬN DIỆN KÍ TỰ VIẾT TAY TIẾNG VIỆT

Nhóm sinh viên:

1. Trịnh Tuấn Nam - 19521874
2. Nguyễn Dương Hải - 19521464
3. Phạm Nguyễn Công Danh - 19521324

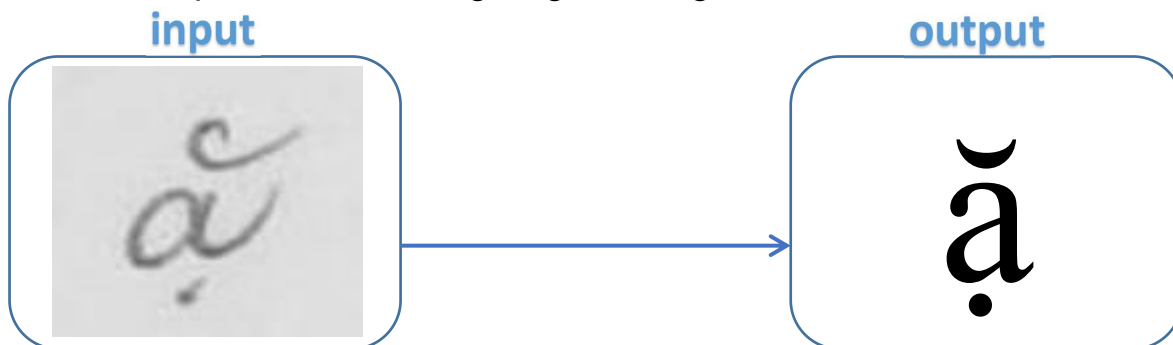
<> [Link Github:](#)

<https://github.com/namt9/CS114.L22.KHCL/tree/main/MACHINE%20LEARNING%20PROJECT>

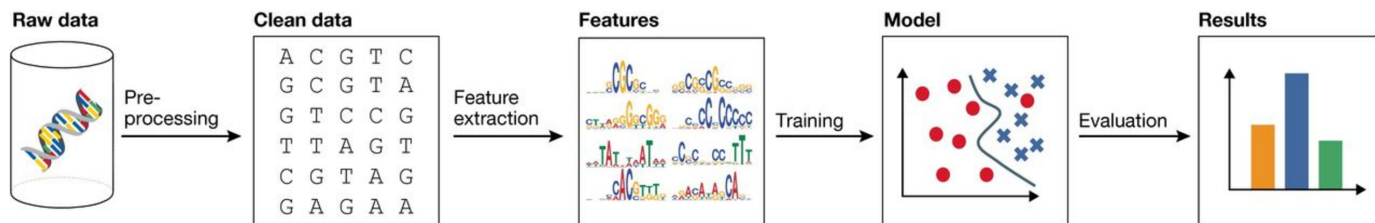
A/ MÔ HÌNH BÀI TOÁN:

* Đề tài “Nhận dạng kí tự viết tay Tiếng Việt”:

- Bài toán thuộc dạng bài toán Classification
- INPUT: ảnh bất kì một chữ cái Tiếng Việt viết tay đúng chuẩn.
- OUTPUT: kết quả chữ cái tương ứng có trong tấm ảnh đó.



B/ QUY TRÌNH TỔNG QUAN:









C/ MÔ TẢ BỘ DỮ LIỆU:

- Nội dung bộ dataset gồm 89 chữ cái viết tay tiếng việt gồm 22 chữ cái la-tin và các chữ cái biến thể bằng cách thêm “dấu” và “thanh”.

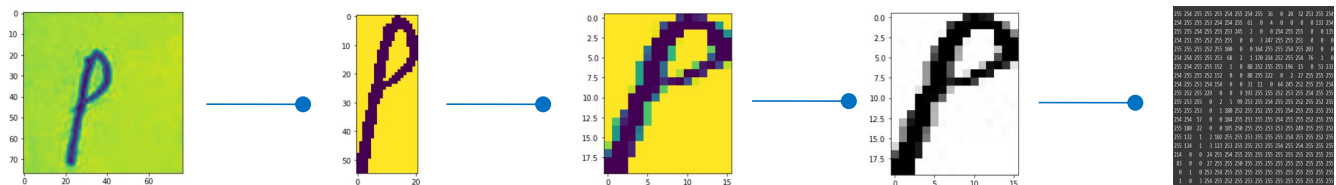
[illegible]

** Datatrain sẽ được được thu thập chung với nhóm bạn Trần Vĩ Hà.*

- Sau khi thu thập chữ theo mẫu trên, nhóm dùng code để cắt viền và cắt thành từng ảnh đơn (chỉ chứa một chữ) và đưa thủ công về từng thư mục tương ứng.

 226.jpg	tôi
 225.jpg	tôi
 224.jpg	tôi
 223.jpg	tôi
 222.jpg	tôi
 221.jpg	tôi

- Sau đó, nhóm crop sát chữ -> resize về 16x20 -> chuyển ảnh sang dạng chữ đen nền trắng -> lọc ảnh “không đạt yêu cầu” và chuyển tất cả về ma trận (320,) sau đó lưu thành 2 file .csv (gồm train và test).



- Kết quả thu được: **+ DataTrain.csv (19229 arrays) trung bình 216 ảnh / chữ**
+ DataTest.csv (5000 arrays) trung bình 55 ảnh / chữ

- Vì data còn ít, nên nhóm không sử dụng chia file valid từ đầu mà sử dụng phương pháp “K-Fold cross validation” để vừa học vừa đánh giá model, giúp model tận dụng được hết nguồn data sẵn có mà không “vô tình” bỏ qua các data quan trọng.

D/ MODEL ÁP DỤNG:

* Đối với bộ data ban đầu: các model nhóm áp dụng đều cho kết quả không cao, cao nhất là **54.53%** ở model MLP Classification.

* Đối với bộ data HOG: kết quả của các model nhìn chung đã cải thiện, cao nhất vẫn là MLP Classification với **57.32%**.

- Sau đó, nhóm có tham khảo và sử dụng **model CNN** và kết quả tăng nhiều hơn so với trước, đạt **68.44%**.

E/ CẬP NHẬP SAU BUỔI BÁO CÁO ĐỒ ÁN ONLINE:

- Cập nhập đầy đủ các nguồn mà nhóm tham khảo để thực hiện đề tài.
(slide 6 + slide 23)
 - Bổ sung hướng phát triển sau này bài toán “Nhận dạng kí tự viết tay Tiếng Việt”.
(slide 23)
 - Bổ sung và mô tả cụ thể quá trình *lọc dữ liệu* sau khi xử lí để có được bộ data cuối cùng.
(slide 9)
-

CẢM ƠN THẦY ĐÃ NHẬN XÉT