

TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN - TRUYỀN THÔNG



NIÊN LUẬN NGÀNH
NGÀNH HỆ THỐNG THÔNG TIN

Đề tài

CHUẨN ĐOÁN CÁC BỆNH Y KHOA DỰA
TRÊN CÁC KỸ THUẬT MÁY HỌC

Sinh viên: Nguyễn Ngọc Trân

Mã số: B2012048

Khóa: K46

Cần Thơ, 04/2024

TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN - TRUYỀN THÔNG

NIÊN LUẬN NGÀNH
NGÀNH HỆ THỐNG THÔNG TIN

Đề tài
**CHUẨN ĐOÁN CÁC BỆNH Y KHOA DỰA
TRÊN CÁC KỸ THUẬT MÁY HỌC**

Người hướng dẫn
TS Nguyễn Thanh Hải

Sinh viên: Nguyễn Ngọc Trân
Mã số: B2012048
Khóa: K46

Cần Thơ, 04/2024

LỜI CẢM ƠN

Trong gần bốn tháng thực hiện đề tài, bản thân em đã nhận được sự hỗ trợ, quan tâm tận tình từ quý thầy cô, người thân và bạn bè, tạo điều kiện cho em có cơ hội học tập và trưởng thành như hôm nay.

Xin gửi lời cảm ơn chân thành đến quý thầy cô Trường Đại học Cần Thơ, đặc biệt là quý thầy cô Trường Công nghệ Thông tin và Truyền thông đã tận tình chỉ bảo và truyền đạt kiến thức cho em, giúp em có được nền tảng kiến thức quý báu, là hành trang giúp em vững bước trên con đường sắp tới.

Em xin gửi lời cảm ơn chân thành nhất đến thầy Nguyễn Thanh Hải, cảm ơn thầy đã tận tình dìu dắt, hướng dẫn, giúp đỡ em rất nhiều trong suốt quá trình nghiên cứu đề tài, cảm ơn cô đã dành nhiều thời gian và công sức hỗ trợ em khoảng thời gian qua.

Bên cạnh những kết quả đã đạt được, đề tài vẫn có nhiều thiếu sót. Rất mong quý thầy cô thông cảm, mong quý thầy cô chỉ bảo, góp ý cho em, vì mỗi ý kiến đóng góp của quý thầy cô đều rất đáng trân trọng và là những kinh nghiệm, kiến thức có thể giúp em hoàn thiện bản thân mình hơn.

Bằng tất cả sự chân thành, một lần nữa cảm ơn mọi người, xin gửi lời chúc sức khỏe và cầu mong cho mọi điều tốt đẹp sẽ đến với mọi người trong tương lai.

Trân trọng!

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU VÀ MÔ TẢ BÀI TOÁN	1
1.1. Mục tiêu đề tài	1
1.2. Mô tả chi tiết đề tài	1
1.3. Các nghiên cứu liên quan	1
1.4. Hướng tiếp cận giải quyết đề tài	2
CHƯƠNG 2. THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP	5
2.1. Kiến trúc tổng quát hệ thống	5
2.2. Xây dựng các mô hình	5
2.3. Giải pháp cài đặt	7
2.4. Dataset	9
CHƯƠNG 3. KIỂM THỬ VÀ ĐÁNH GIÁ	10
3.1. Kịch bản kiểm thử	10
3.2. Kết quả kiểm thử	10
3.3. Thảo luận	16
CHƯƠNG 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	19
4.1. Kết luận	19
4.2. Hướng phát triển	19
TÀI LIỆU THAM KHẢO	20

DANH MỤC HÌNH

Hình 1.1: Các nhãn của tập dữ liệu heaercsv [3]	3
Hình 1.2: Sơ đồ tiền xử lý các tập dữ liệu Polyp	4
Hình 2.1: Sơ đồ về YOLO[12]	5
Hình 2.2: Các bước phân biệt trường hợp mắc bệnh hoặc không bệnh	6
Hình 2.3: YOLO chia hình ảnh thành một grid $S \times S$ để dự đoán [14]	7
Hình 3.1: Kết quả với thuật toán K-NN	11
Hình 3.2: Kết quả với thuật toán LR	11
Hình 3.3: Biểu đồ đánh giá qua các thực nghiệm	12
Hình 3.4: Biểu đồ đánh giá qua các thực nghiệm	13
Hình 3.5: Biểu đồ đánh giá qua các thực nghiệm	14
Hình 3.6: Biểu đồ đánh giá qua các thực nghiệm	14
Hình 3.7: Biểu đồ label đánh giá qua các thực nghiệm	15
Hình 3.8: Biểu đồ labels_correlogram đánh giá qua các thực nghiệm	15
Hình 3.9: Biểu đồ confusion_matrix đánh giá qua các thực nghiệm	16
Hình 3.10: Biểu đồ đánh giá qua các thực nghiệm	16

DANH MỤC BẢNG

Bảng 1.1: Mô tả tiền xử lý dữ liệu trước và sau khi xử lý	4
Bảng 3.1: Bảng kết quả sau train model với 100 epoch	10
Bảng 3.2: Bảng so sánh chỉ số của hai thuật toán K-NN và LR	12
Bảng 3.3: Bảng kết quả kiểm tra model YOLOv9 với các tập dữ liệu	12
Bảng 3.4: Bảng so sánh các chỉ số trên cùng tập dữ liệu dung để xác định polyp so với các nghiên cứu khác	18

DANH MỤC TỪ CHUYÊN NGÀNH

Viết tắt	Giải thích
CRC	Colorectal Cancer
K-NN	K-Nearest Neighbors
LR	Logistic Regression
mAP	mean Average Precision
ML	Machine Learning
DL	Deep Learning
TP	True Positives
FN	False Negatives
FP	False Positives

CHƯƠNG 1. GIỚI THIỆU VÀ MÔ TẢ BÀI TOÁN

1.1. Mục tiêu đề tài

Mục tiêu chính của đề tài là “Nghiên cứu và ứng dụng các mô hình máy học để xác định bệnh tim và nhận biết tế bào ung thư đại trực tràng” sử dụng mô hình cơ bản có thể thực hiện nhiều thuật toán và xử lý dữ liệu đã nhận diện và phát hiện trường hợp mắc bệnh tim. Bên cạnh đó sử dụng mô hình của YOLOv9 c để phát hiện và ngăn ngừa ung thư đại trực tràng.

1.2. Mô tả chi tiết đề tài

Bệnh tim mạch là một trong những nguyên nhân gây tử vong phổ biến nhất trên toàn thế giới hàng năm. Phát hiện sớm các bệnh lý bất thường ở tim có thể tránh được tình trạng tim đột ngột tử vong và các bệnh nguy hiểm khác do bệnh tim gây ra. Ngoài ra, các chẩn đoán bệnh tim sớm giúp bác sĩ biết được tình trạng của bệnh nhân và đưa ra có phương án điều trị hợp lý và sớm.

Ung thư đại trực tràng (CRC) là một loại ung thư ảnh hưởng đến ruột già và là một trong những dạng ung thư nghiêm trọng và phổ biến nhất. Tỷ lệ sống sót sau 5 năm bị ảnh hưởng bởi nhiều yếu tố khác nhau và có thể thay đổi đáng kể tùy thuộc vào giai đoạn ung thư và vị trí của nó ở đại tràng hoặc thực tràng. Trung bình, tỷ lệ sống sót sau 5 năm đối với CRC được ước tính dao động từ 48,6% đến 59,4%[1]. Dự án thống kê đến năm 2020, gần 150.000 người sẽ được chẩn đoán mắc CRC và hơn 50.000 người sẽ không chống chọi được với căn bệnh này. [2]

1.3. Các nghiên cứu liên quan

Gần đây đã có nhiều nghiên cứu dự đoán bệnh tim dựa trên tập dữ liệu [3], sử dụng các phương pháp thống kê và học máy khác nhau. Đáng chú ý, công trình [4] đã áp dụng thuật toán K-NN và Random Forest để dự đoán các trường hợp bị bệnh hoặc không có bệnh. Md. Imam Hossain cùng cộng sự [5] đã làm Các kỹ thuật trí tuệ nhân tạo khác nhau (logistic regression, Naïve Bayes, K-Nearest Neighbor (K-NN), máy vector hỗ trợ (SVM), decision tree, random forest và multilayer perceptron (MLP)) được áp dụng và so sánh cho hai loại tập dữ liệu bệnh tim (tất cả các đặc trưng và các đặc trưng được chọn). Trong các nghiên cứu khác, Jyoti Soni và đồng đội [6] đã thí nghiệm và thực hiện để so sánh hiệu suất của các kỹ thuật khai thác dữ liệu dự đoán trên cùng một tập dữ liệu của Decision Tree và Bayesian Classification được cải thiện sau khi áp dụng phương pháp di truyền để giảm kích thước dữ liệu và tạo ra một tập con tối ưu các thuộc tính cần thiết để dự đoán bệnh tim.

Bên cạnh đó, Nima Tajbakhsh và cộng sự đã đề xuất một phương pháp phát hiện polyp mới dựa trên hình ảnh 3 chiều độc đáo trình bày và mạng lưới thần kinh tích chập.

Phương pháp này học một loạt các đặc điểm của polyp như màu sắc, cấu trúc, hình dạng và thông tin về thời gian ở nhiều tỉ lệ khác nhau, từ đó giúp việc định vị polyp chính xác hơn. Đưa vào một polyp, một tập hợp các mạng nơ-ron tích chập - mỗi mạng chuyên biệt trong một loại đặc điểm - được áp dụng trong vùng lân cận của polyp và sau đó kết quả của họ được tổng hợp để chấp nhận hoặc loại bỏ polyp đó. Trong một nghiên cứu khác của Mehrshad Lalinia và cộng sự [7] họ đã đề xuất một hệ thống phát hiện polyp dựa trên trí tuệ nhân tạo sử dụng mạng YOLO-V8. và xây dựng một bộ dữ liệu đa dạng từ nhiều nguồn công khai khác nhau và tiến hành đánh giá một cách chi tiết. Được biết YOLO-V8 vượt trội hơn so với các mô hình tiên tiến khác về trung bình độ chính xác trung bình. YOLO-V8 cung cấp một sự cân bằng giữa độ chính xác và hiệu suất tính toán.

1.4. Hướng tiếp cận giải quyết đề tài

Nhiệm vụ chính trong việc xác định ung thư đại trực tràng (CRC) trên hình ảnh và dự đoán trường hợp bệnh nhân mắc bệnh tim là tập trung vào việc tìm kiếm và phân loại xử lý các dữ liệu đầu vào của hình ảnh và các dữ liệu và chỉ số của các trường hợp mắc và không mắc các bệnh về tim mạch. Hiện nay có khá nhiều mô hình được sử dụng để giải quyết nhiệm vụ này và đã đạt được kết quả đáng kinh ngạc. Trong nghiên cứu này thực hiện xây dựng các phương pháp chính: sử dụng phương pháp học máy (Machine Learning) và học sâu (Deep Learning) để giải quyết bài toán này:

- Về học máy trong nghiên cứu dự đoán trường hợp bệnh tim này đề xuất sử dụng thuật toán K-Nearest Neighbors (K-NN) thuật toán học máy này sử dụng trong bài toán phân loại và dự đoán và Logistic Regression (LR) được sử dụng để dự đoán xác suất của một biến phụ thuộc nhị phân dựa trên một hoặc nhiều biến độc lập.
- Về học sâu trong nghiên cứu này, YOLOv9 được đề xuất sử dụng đây là một phiên bản cải tiến của YOLO (You Only Look Once), được gọi là YOLOv9, để tự động phát hiện ung thư đại trực tràng trên hình ảnh y tế. YOLOv9 là một mạng nơ-ron tích chập sâu (CNN) được thiết kế để xử lý hình ảnh y tế với độ chính xác cao. Phương pháp này xử lý và nhận diện hình ảnh polyp với độ chính xác và cho ra được kết quả cao và chính xác.

Dưới đây là các phương pháp tiền xử lý dữ liệu để giải quyết nhiệm vụ này:

❖ Thu thập dữ liệu :

Trong nghiên cứu này, chúng tôi sử dụng tập dữ liệu từ heartcsv [3] cho việc huấn luyện và kiểm thử để dự đoán về các trường hợp mắc bệnh tim. Tập dữ liệu heartcsv là tập dữ liệu dùng thử nghiệm train model dự đoán trường hợp mắc bệnh tim.

- Về phần nghiên cứu về ung thư đại trực tràng ở nghiên cứu này tôi sử dụng các tập dữ liệu gồm có : Kvasir-SEG [8]CVC-ClinicDB [9] CVC-ColonDB

[10], ETIS [11] bốn tập dữ liệu này sử dụng để huấn luyện model và đánh giá, để xác định vùng chứa polyp trên hình ảnh được đưa vào.

❖ Tiền xử lý dữ liệu :


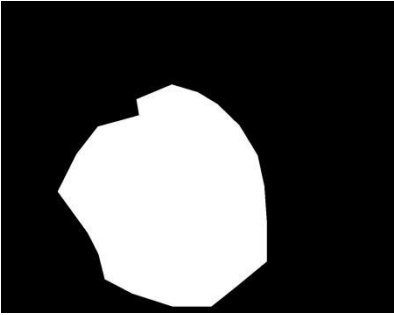
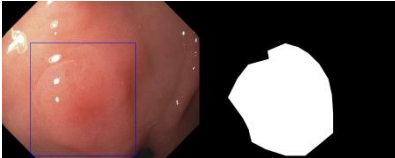
- Về tập dữ liệu heartcsv [3] đầu tiên tôi kiểm tra giá trị thiếu (missing values) và nhận thấy không có giá trị thiếu nào trong tập dữ liệu, tiếp đến là xử lý ngoại lai(outliers) sử dụng phương pháp IQR để tìm và chọn cách áp dụng biến đổi Box-Cox để giảm thiểu ảnh hưởng của ngoại lai thay vì loại bỏ chúng hoàn toàn. Cuối cùng là biến đổi thuộc tính lệch (skewed features transformation) sử dụng để chuyển đổi các thuộc tính liên tục bị lệch về một phía, giúp phân bố dữ liệu giống với phân bố chuẩn hơn.

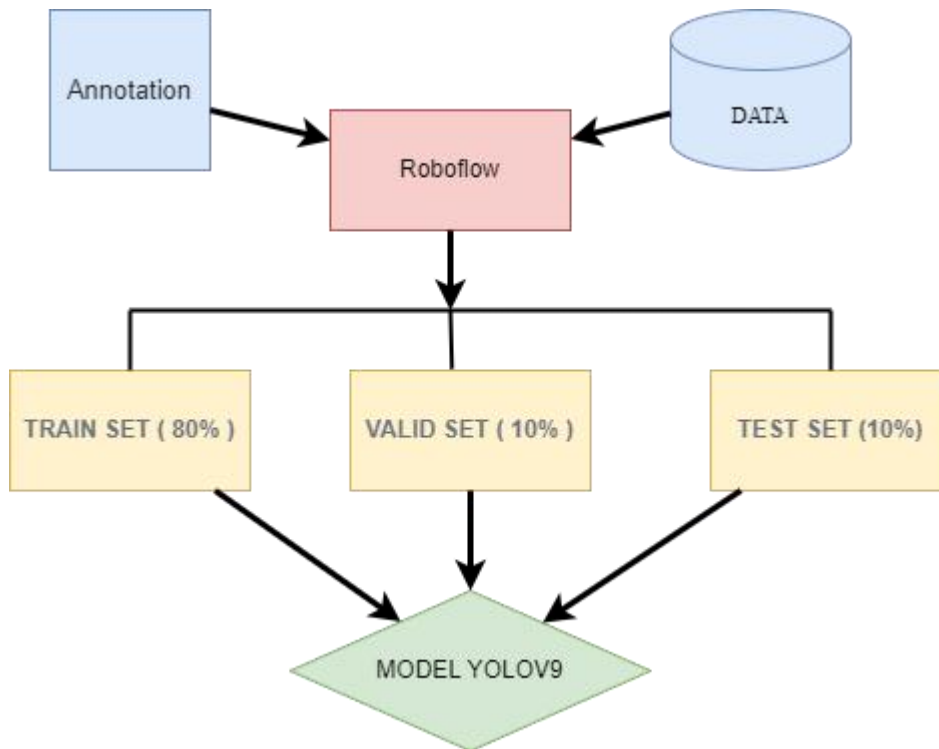
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age         303 non-null    int64
1   sex         303 non-null    int64
2   cp          303 non-null    int64
3   trestbps    303 non-null    int64
4   chol        303 non-null    int64
5   fbs         303 non-null    int64
6   restecg     303 non-null    int64
7   thalach     303 non-null    int64
8   exang       303 non-null    int64
9   oldpeak     303 non-null    float64
10  slope       303 non-null    int64
11  ca          303 non-null    int64
12  thal        303 non-null    int64
13  target      303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Hình 1.1: Các nhãn của tập dữ liệu heartcsv [3]

- Về phần tập dữ liệu hình ảnh để sử dụng cho việc huấn luyện và đánh giá ung thư đại trực tràng tôi gom các tập dữ liệu gồm Kvasir-SEG [8], CVC-ClinicDB [9], CVC-ColonDB [10], ETIS [11] lại làm một bộ dữ liệu chung gồm 2151 hình. Đầu tiên tôi xử lý lại hình ảnh và tạo Annotation từ image và mask của tập dữ liệu tiếp theo là phân loại và chuẩn hóa hình ảnh và sử dụng Roboflow để gắn nhãn hình ảnh cũng như là tăng cường bộ dữ liệu. Tiếp theo điều chỉnh kích thước hình ảnh, chuẩn hóa các giá trị pixel, loại bỏ nhiễu, hoặc cắt phần quan trọng của hình ảnh.

Bảng 1.1: Mô tả tiền xử lý dữ liệu trước và sau khi xử lý

IMAGE	MASK	RESULTS
		 <p>Annotation (polyp 0.14391691394658754 0.26542056074766357 0.6772997032640949 0.9803738317757009)</p>



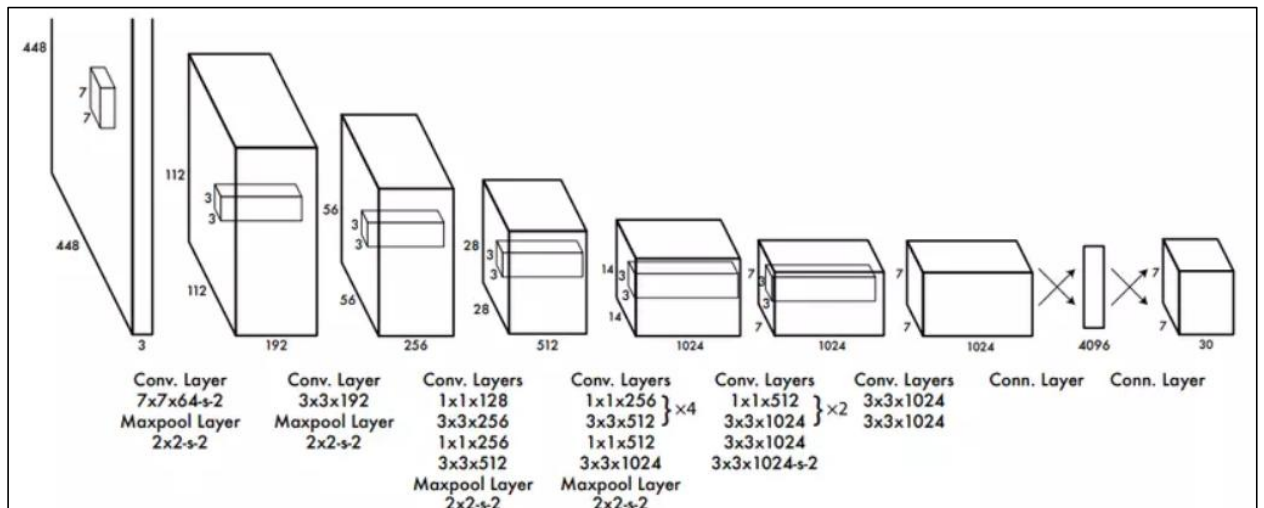
Hình 1.2: Sơ đồ tiền xử lý các tập dữ liệu Polyp

CHƯƠNG 2. THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP

2.1. Kiến trúc tổng quát hệ thống

Trong phần này trình bày các thuật toán KNN (K-Nearest Neighbors) và LR (Linear Regression) để dự đoán bệnh tim với quy trình thực hiện của mô hình chẩn đoán được trình bày. Các bước này bao gồm thu thập dữ liệu, tiền xử lý, đào tạo, tinh chỉnh và thử nghiệm mô hình. Đầu tiên, dữ liệu đầu vào của các mô hình được đọc từ tập dữ liệu heartcsv[3], đồng thời đầu ra là độ chính xác được lưu trữ lại với mục đích so sánh với kết quả dự đoán, đánh giá mô hình.

Về phần kỹ thuật để xác định nhận biết polyp tôi sử dụng kiến trúc của YOLO một mô hình hồi quy để dự đoán vị trí và độ tin cậy của các bounding box chứa vật thể trong ảnh. Mô hình chia ảnh thành lưới ô lưới có kích thước $S \times S$. Mỗi ô lưới trong lưới chịu trách nhiệm phát hiện vật thể nếu tâm của vật thể nằm trong ô đó.

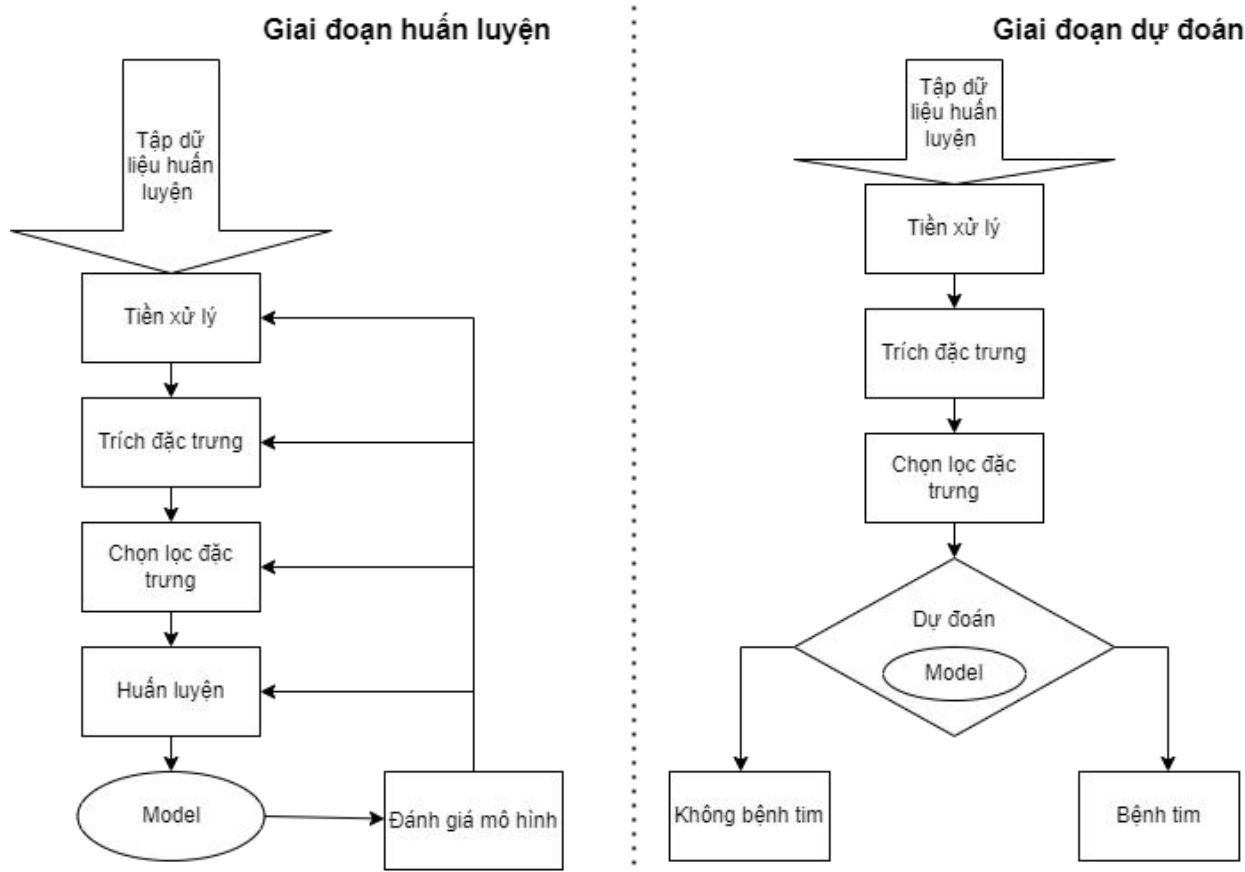


Hình 2.1: Sơ đồ về YOLO [12]

2.2. Xây dựng các mô hình

KNN (K-Nearest Neighbors) là thuật toán đơn giản dùng để phân loại và dự đoán dựa trên việc xác định nhãn của một điểm dữ liệu bằng cách so sánh với các điểm láng giềng gần nhất trong không gian đặc trưng.

Logistic Regression (LR) là một thuật toán học máy giám sát phổ biến được sử dụng cho bài toán phân loại. Mặc dù có tên là "regression," nhưng LR thực chất là một thuật toán phân loại và không phải là một phương pháp hồi quy.

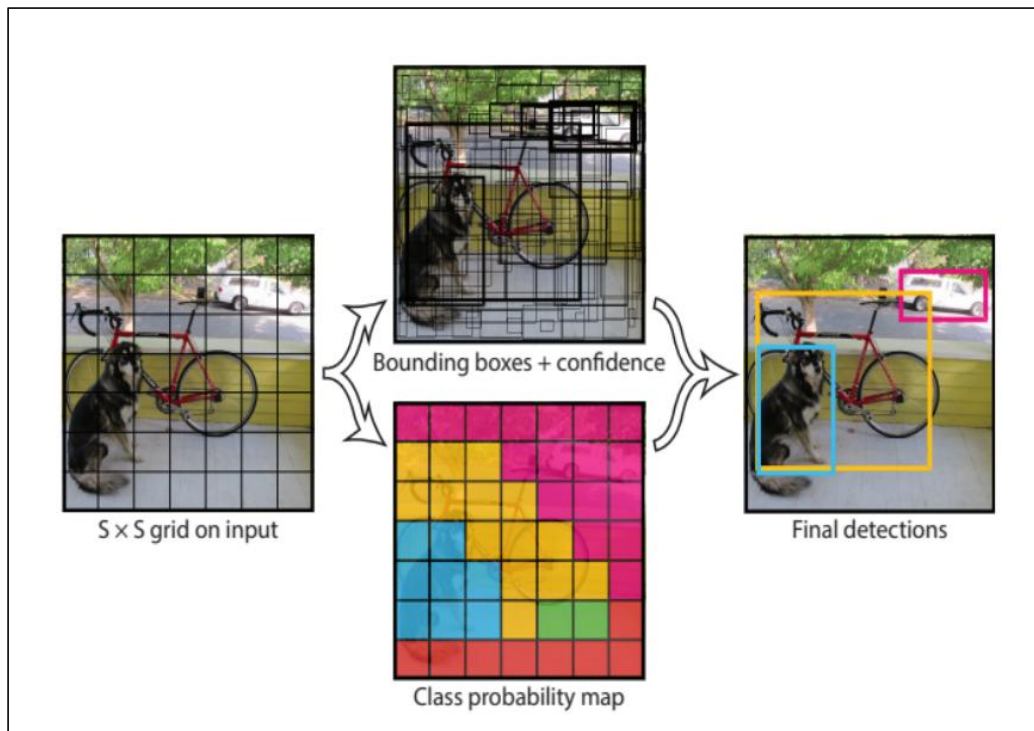


Hình 2.2: Các bước phân biệt trường hợp mắc bệnh hoặc không bệnh

Đầu tiên tôi chuẩn bị tập dữ liệu và tiền xử lý dữ liệu xem xét bộ dữ liệu có thiếu sót không. Sau khi kiểm tra nhận thấy không có giá trị nào bị thiếu và tiếp tục tiến hành đem đi xử lý ngoại lệ trong dữ liệu qua phương pháp IQR thì xác định được bất thường và thay vì loại bỏ tôi quyết định áp dụng Box-Cox để giảm thiểu ảnh hưởng của chúng. Sau đó mã hóa các đặc điểm phân loại thành dạng số để chuẩn bị cho quá trình huấn luyện. Những bước tiền xử lý này sẽ giúp cho bộ dữ liệu sạch và phù hợp từ đó tăng cường hiệu suất và độ chính xác của mô hình.

Chia tập dữ liệu thành hai tập : huấn luyện (training), kiểm tra (testing). Huấn luyện bằng hai thuật toán KNN và LR tôi sử dụng cùng một tập dữ liệu và cùng tiền xử lý dữ liệu giống nhau để training và testing và Đánh giá hiệu suất của thuật toán bằng cách sử dụng các chỉ số như độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity), v.v.

Bên cạnh đó ở bài toán nhận biết polyp trong ung thư đại trực tràng tôi sử dụng kiến trúc của YOLOv9 được giới thiệu vào tháng 2 năm 2024 bởi Chien-Yao Wang và cộng sự. Nó được xây dựng dựa trên các phiên bản trước, kết hợp những tiến bộ trong kỹ thuật học sâu và thiết kế kiến trúc để đạt được hiệu suất vượt trội trong các nhiệm vụ phát hiện đối tượng [13].



Hình 2.3: YOLO chia hình ảnh thành một grid $S \times S$ để dự đoán [14]

2.3. Giải pháp cài đặt

Tất cả các quy trình trên đều được code và áp dụng trên môi trường Jupyter Notebook. Những sổ ghi chép này cung cấp cho người dùng một môi trường tương tác để phân tích dữ liệu và huấn luyện các mô hình máy học. Nó được chạy trên PC với CPU Intel(R) Core(TM) i7-10750H CPU 2.60GHz 2.59 GHz và RAM 8 GB. Để đánh giá hiệu suất của mô hình được đề xuất, nghiên cứu này sử dụng một số ma trận đánh giá như accuracy, precision, recall và F1-score.

Về phần ứng thử đại trực tràng nhận biết khối polyp nghiên cứu được chạy trong môi trường Google Colab với các thông số như sau epochs: 100, batch: 16, imgsz: 640, workers: 8, dropout: 0.0. Nó được chạy trên PC với CPU Intel(R) Core(TM) i7-10750H CPU 2.60GHz 2.59 GHz và RAM 8 GB. Để đánh giá sử dụng các chỉ số đánh giá sau mAP50, Recall, F1-score và Precision.

Công thức tính độ đo mAP:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i$$

Độ đo mAP : mAP là viết tắt của "mean Average Precision", là một độ đo phổ biến được sử dụng trong lĩnh vực nhận dạng và phát hiện đối tượng trong hình ảnh. Đây là một phương pháp đo lường chất lượng của một hệ thống phân loại hoặc phát hiện dựa trên các giá trị Precision và Recall.

Công thức tính độ đo Recall :

$$Recall = \frac{TP}{TP + FN}$$

Recall: True Positives (TP): Số lượng các trường hợp dự đoán đúng và thực sự thuộc về nhóm quan tâm. False Negatives (FN): Số lượng các trường hợp không được dự đoán đúng nhưng thực sự thuộc về nhóm quan tâm.

Công thức tính Accuracy :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP (True Positives - Số lượng dự đoán đúng tích cực): Đây là số lượng các mẫu dữ liệu mà mô hình đã phân loại đúng vào lớp tích cực (positive class), tức là những mẫu thực tế thuộc vào lớp này và mô hình dự đoán đúng. FP (False Positives - Số lượng dự đoán sai tích cực): Đây là số lượng các mẫu dữ liệu mà mô hình đã phân loại sai vào lớp tích cực (positive class), tức là những mẫu thực tế thuộc vào lớp khác nhưng mô hình lại dự đoán chúng là thuộc vào lớp tích cực. FN (False Negatives - Số lượng dự đoán sai tiêu cực): Đây là số lượng các mẫu dữ liệu thực tế thuộc vào lớp tích cực (positive class), nhưng mô hình đã phân loại sai vào lớp tiêu cực (negative class). TN (True Negatives - Số lượng dự đoán đúng tiêu cực): Đây là số lượng các mẫu dữ liệu mà mô hình đã phân loại đúng vào lớp tiêu cực (negative class), tức là những mẫu thực tế thuộc vào lớp này và mô hình dự đoán đúng.

Công thức tính Precision:

$$Precision = \frac{TP}{TP + FP}$$

Precision đo lường tỷ lệ các điểm dự đoán là positive mà thực sự là positive trong tất cả các dự đoán positive. TP (True Positives - Số lượng dự đoán đúng tích cực): Đây là số lượng các mẫu dữ liệu mà mô hình đã phân loại đúng vào lớp tích cực (positive class), tức là những mẫu thực tế thuộc vào lớp này và mô hình dự đoán đúng. FP (False Positives - Số lượng dự đoán sai tích cực): Đây là số lượng các mẫu dữ liệu mà mô hình đã phân loại sai vào lớp tích cực (positive class), tức là những mẫu thực tế thuộc vào lớp khác nhưng mô hình lại dự đoán chúng là thuộc vào lớp tích cực.

Công thức tính F1-score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-score là trung bình điều hòa giữa Precision và Recall. Nó cung cấp một phép đo tổng thể về hiệu suất của mô hình phân loại.

Precision: đo lường tỷ lệ các điểm dự đoán là positive mà thực sự là positive trong tất cả các dự đoán positive.

Recall :Tỉ lệ nhận dạng đúng

2.4. Dataset

Nghiên cứu này sử dụng tập dữ liệu heart gồm có 303 mẫu từ index 0 đến index 302. Trong nghiên cứu này tôi chia tập dữ liệu thành tập huấn luyện và tập kiểm tra. Tập kiểm tra được chiếm tỷ lệ là 20% của tập dữ liệu ban đầu, còn tập huấn luyện chiếm 80% còn lại.

Trong nghiên cứu này sử dụng bộ dữ liệu với 2151 hình ảnh của hình ảnh siêu âm CRC và 2151 Segmentation của hình ảnh siêu âm CRC tương ứng. Bộ dữ liệu này bao gồm 4 bộ dữ liệu được tập hợp lại gồm có Kvasir-SEG [8], CVC-ClinicDB [9], CVC-ColonDB [10], ETIS [11]. Về phần đánh giá sau khi đã huấn luyện tôi kiểm tra theo từng tập dữ liệu tách biệt trong bốn tập dữ liệu trên.

CHƯƠNG 3. KIỂM THỬ VÀ ĐÁNH GIÁ

3.1. Kịch bản kiểm thử

Về phân dự đoán bệnh tim trên kỹ thuật học máy có hai trường hợp :

Trường hợp 1: Thực hiện bằng thuật toán K-Nearest Neighbors(K-NN)

Mô hình K-Nearest Neighbors sau khi được tinh chỉnh siêu tham số với weight là distance và p là 1 đã đạt được độ chính xác là 68.85% trên tập dữ liệu kiểm tra. Đây là một kết quả khá, nhưng có thể cần phải điều chỉnh thêm các siêu tham số khác để cải thiện hiệu suất của mô hình.

Trường hợp 2: Thực hiện bằng thuật toán Logistic Regression (LR)

Mô hình Logistic Regression sau khi được tinh chỉnh siêu tham số với $C = 0.09$ và solver là "liblinear" đã đạt được độ chính xác là 88.52% trên tập dữ liệu kiểm tra. Điều này cho thấy mô hình có khả năng phân loại dữ liệu một cách hiệu quả sau quá trình tinh chỉnh siêu tham số.

Về nghiên cứu xác định polyp :

Tôi sử dụng tập dữ liệu ảnh siêu âm ung thư đại trực tràng sau đó áp dụng mô hình YOLOv9 c để train dữ liệu mà nhóm thu thập được. Từ kết quả thu được, sử dụng để nhận dạng tập dữ liệu test để kiểm tra xem tỉ lệ đạt được của mô hình. Nghiên cứu này sử dụng ngôn ngữ lập trình python, các thực nghiệm chạy trên và máy cục bộ Kết quả mô hình dự đoán trên tập xác thực. Kích thước ảnh đầu vào được thay đổi thành 640 x 640, ngưỡng tin cậy 0,25. Nghiên cứu này sử dụng ngôn ngữ lập trình python, các thực nghiệm chạy trên nền tảng Google colab và máy local Kết quả mô hình dự đoán trên tập xác thực trên YOLOv9 c. Kết quả từ các lần train 100 Epoch có kết quả như sau:

Bảng 3.1: Bảng kết quả sau train model với 100 epoch

YOLO Type	Precision	Recall	F1-score	mAP50	mAP50-95
YOLOv9 c	94,95%	81,9%	88%	85,8%	70,4%

3.2. Kết quả kiểm thử

Dưới đây là kết quả kiểm thử của hai thuật toán K-NN và LR :

KNN Model Scores:					
Test Accuracy: 0.6885					
Confusion Matrix:					
[[21 8]					
[9 23]]					
Classification Report:					
	precision	recall	f1-score	support	
0	0.70	0.72	0.71	29	
1	0.74	0.72	0.73	32	
accuracy			0.72	61	
macro avg	0.72	0.72	0.72	61	
weighted avg	0.72	0.72	0.72	61	
Cross-validated Metrics:					
Accuracy: 0.6734					
Precision: 0.6845					
Recall: 0.7455					
F1: 0.7124					
AUC-ROC: 0.7201					

Hình 3.1: Kết quả với thuật toán K-NN

Logistic Regression Best Parameters:					
C: 0.09					
Test Accuracy: 0.8852					
Confusion Matrix:					
[[25 4]					
[3 29]]					
Classification Report:					
	precision	recall	f1-score	support	
0	0.89	0.86	0.88	29	
1	0.88	0.91	0.89	32	
accuracy			0.89	61	
macro avg	0.89	0.88	0.88	61	
weighted avg	0.89	0.89	0.89	61	
Cross-validated Metrics:					
Accuracy: 0.8348					
Precision: 0.8183					
Recall: 0.9030					
F1: 0.8573					
AUC-ROC: 0.8969					

Hình 3.2: Kết quả với thuật toán LR

Bảng 3.2: Bảng so sánh chỉ số của hai thuật toán K-NN và LR

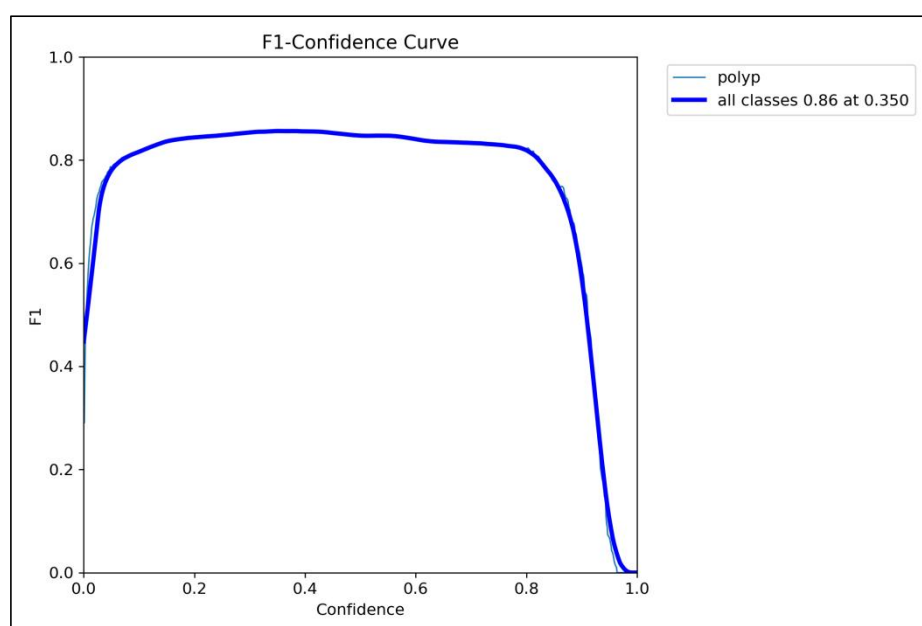
Algorithm	HeartDisease	Accuracy	Precision	Recall	f1-score
K-NN	0	72%	70%	72%	71%
	1		74%	72%	73%
LR	0	89%	89%	86%	88%
	1		88%	91%	89%

Từ Bảng 3.2 nhận thấy được thuật toán LR tối ưu và đạt được chỉ số vượt trội trong công việc dự đoán trường hợp mắc bệnh tim trên cùng tập dữ liệu so với giải thuật K-NN.

Dưới đây là kết quả kiểm tra các tập data của model được train bằng YOLOv9 phiên bản c:

Bảng 3.3: Bảng kết quả kiểm tra model YOLOv9 với các tập dữ liệu

Data	Recall	Precision	F1-score	mAP-50
CVC-ClinicDB	98,7%	97,7%	98,1%	98,9%
ETIS	98%	99,4%	98,6%	99,3%
CVC-ColonDB	96,8%	98,2%	97,4%	98,2%
Kvasir-SEG	80,5%	96%	87,5%	85,8%

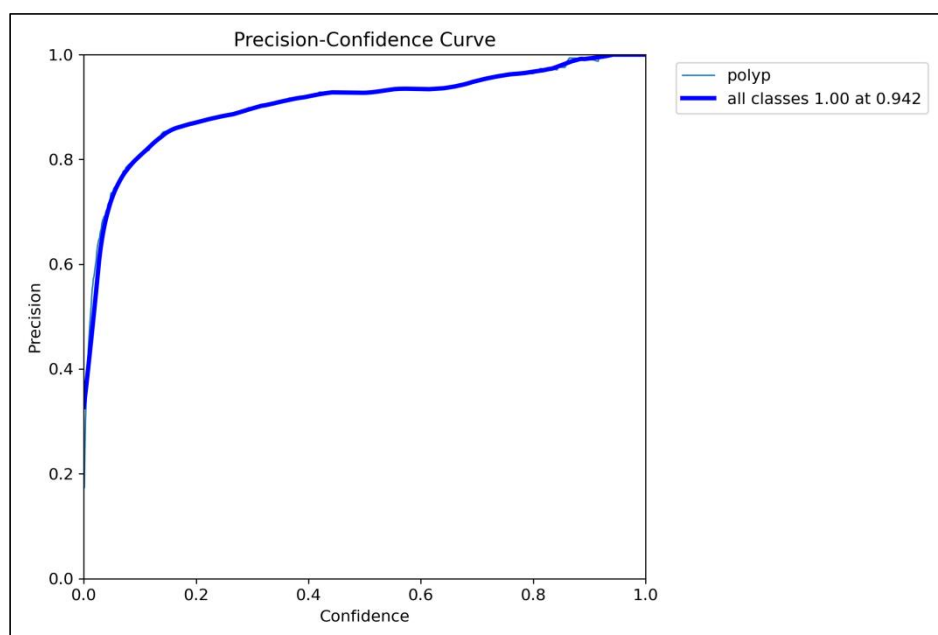


Hình 3.3: Biểu đồ đánh giá qua các thực nghiệm

Biểu đồ Hình 3.3 thể hiện hình ảnh đường cong F1-confidence cho hai trường hợp: phân biệt lớp "polyp" và phân biệt tất cả các lớp.

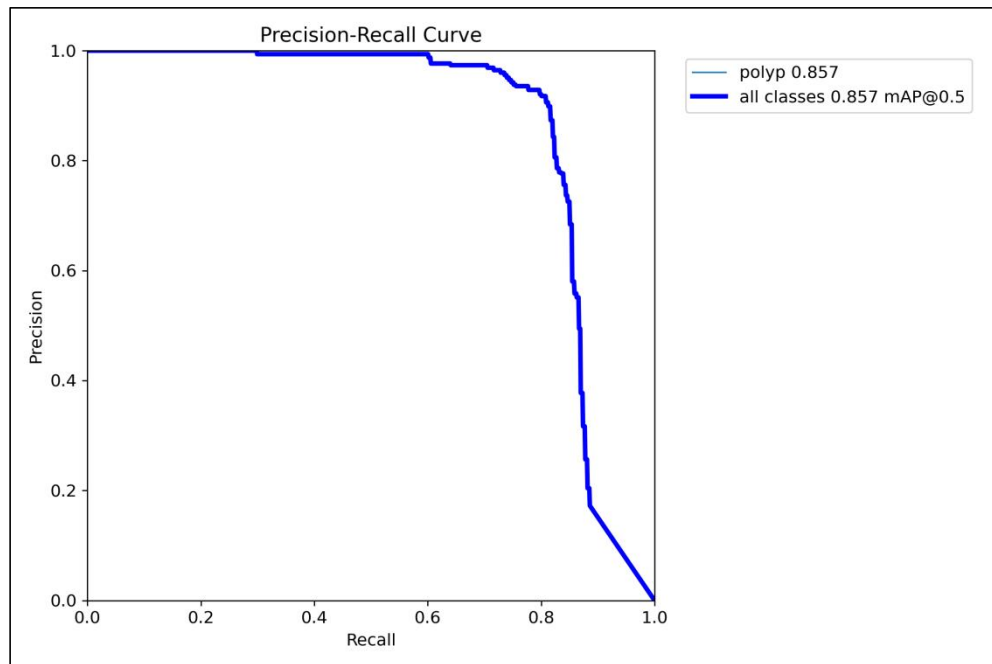
Polyp: Điểm F1 cao nhất đạt 0,86 tại mức độ tự tin 0,350, cho thấy mô hình hoạt động tốt nhất khi phân biệt "polyp" với các lớp khác ở mức độ tự tin này.

Tất cả các lớp: Điểm F1 cao nhất đạt 0,86 tại mức độ tự tin 0,350, cho thấy mô hình hoạt động tốt nhất khi phân biệt tất cả các lớp với nhau ở mức độ tự tin này.



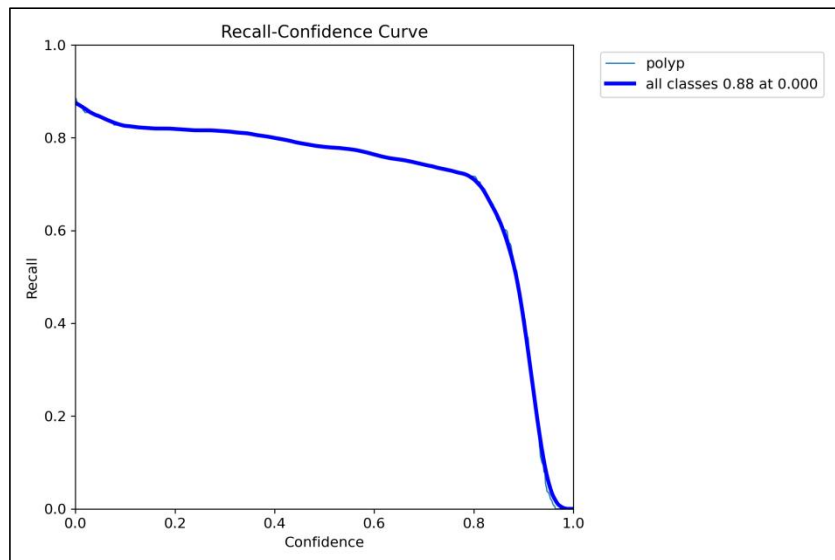
Hình 3.4: Biểu đồ đánh giá qua các thực nghiệm

Hình 3.4 trên là biểu đồ ROC trực quan tổng thể hiệu suất của mô hình theo ngưỡng tự tin Confidence. Mô hình đạt hiệu suất tích cực tối đa tại ngưỡng tự tin là 0.942. Từ 0.2 tới 1 thì mô hình đạt hiệu suất tích cực trên 0.8. Điều này cho thấy, mô hình cho ra hiệu suất cao với những ngưỡng tự tin thấp.



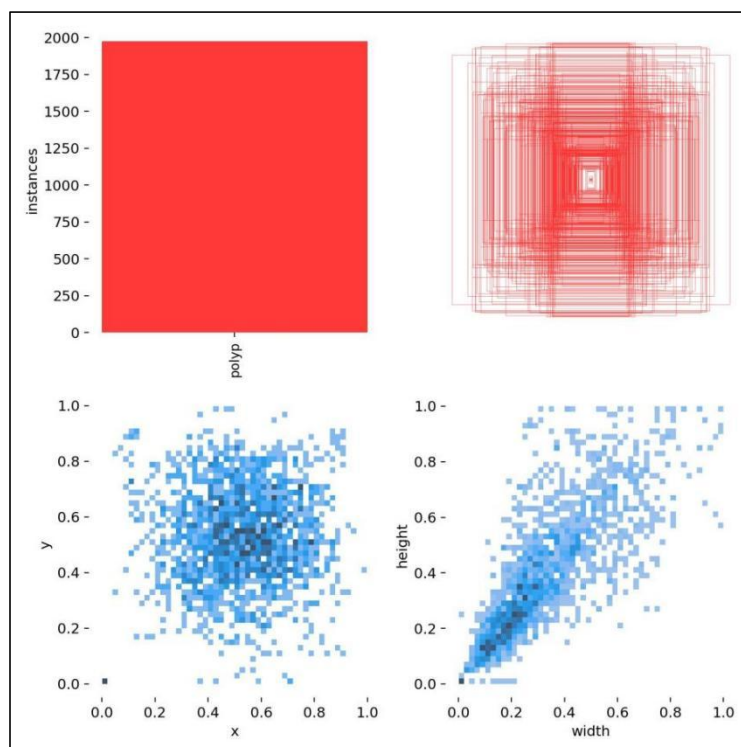
Hình 3.5: Biểu đồ đánh giá qua các thực nghiệm

Hình 3.5 đường cong độ chính xác - thu hồi (precision-recall curve), hay còn gọi là PRC, là một biểu đồ trực quan thể hiện hiệu suất của mô hình phân loại ở các ngưỡng phân loại khác nhau. Nó mô tả mối quan hệ giữa độ chính xác (precision) và độ thu hồi (recall) khi ngưỡng phân loại được điều chỉnh. Đường cong độ chính xác - thu hồi cao và mượt mà, thể hiện độ ổn định cao của thuật toán. Thuật toán phân loại có hiệu suất tốt với độ chính xác cao nhất 0,857, độ thu hồi cao nhất 0,857 và AUC là 0,902. Đường cong độ chính xác - thu hồi cao và mượt mà, thể hiện độ ổn định cao của thuật toán.

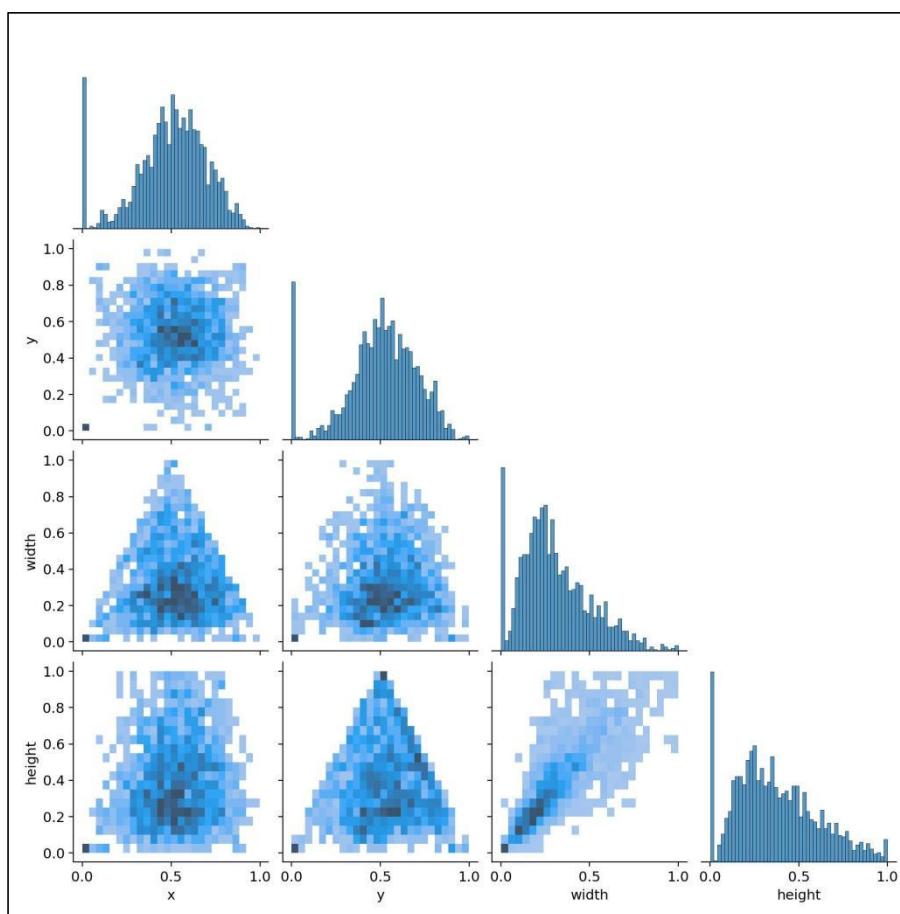


Hình 3.6: Biểu đồ đánh giá qua các thực nghiệm

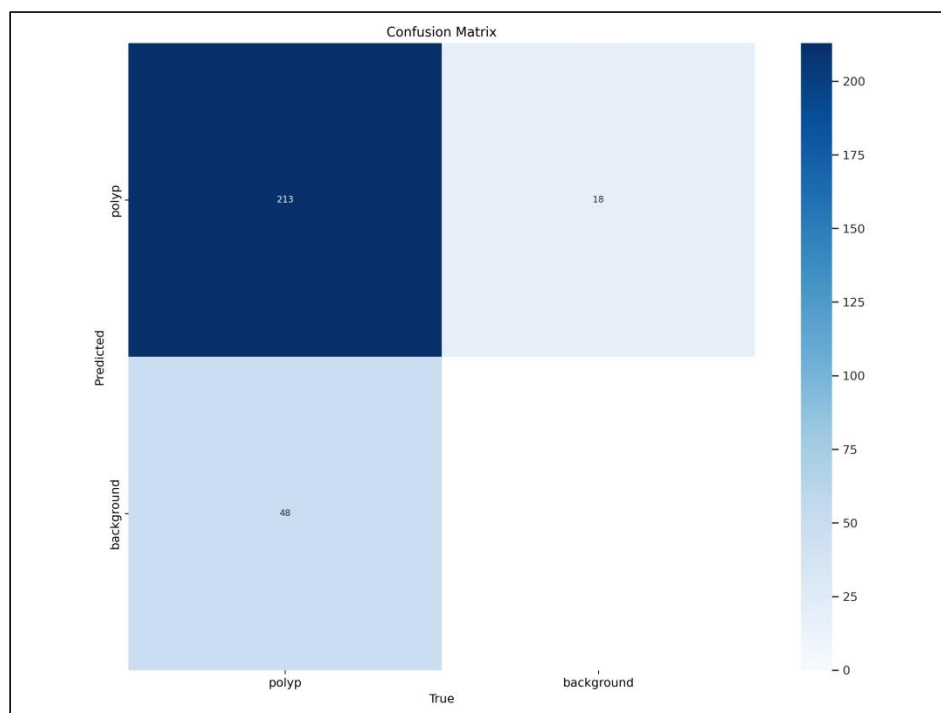
Hình 3.6 là biểu đồ thể hiện Recall-Confidence Curve cho một lớp trong mô hình học máy. Biểu đồ này cho thấy mối quan hệ giữa số lượng lớp và mức độ tự tin. Biểu đồ cho biết mức độ tự tin cao nhất mà mô hình đạt được (0.88) và số lượng lớp tương ứng (0.000).



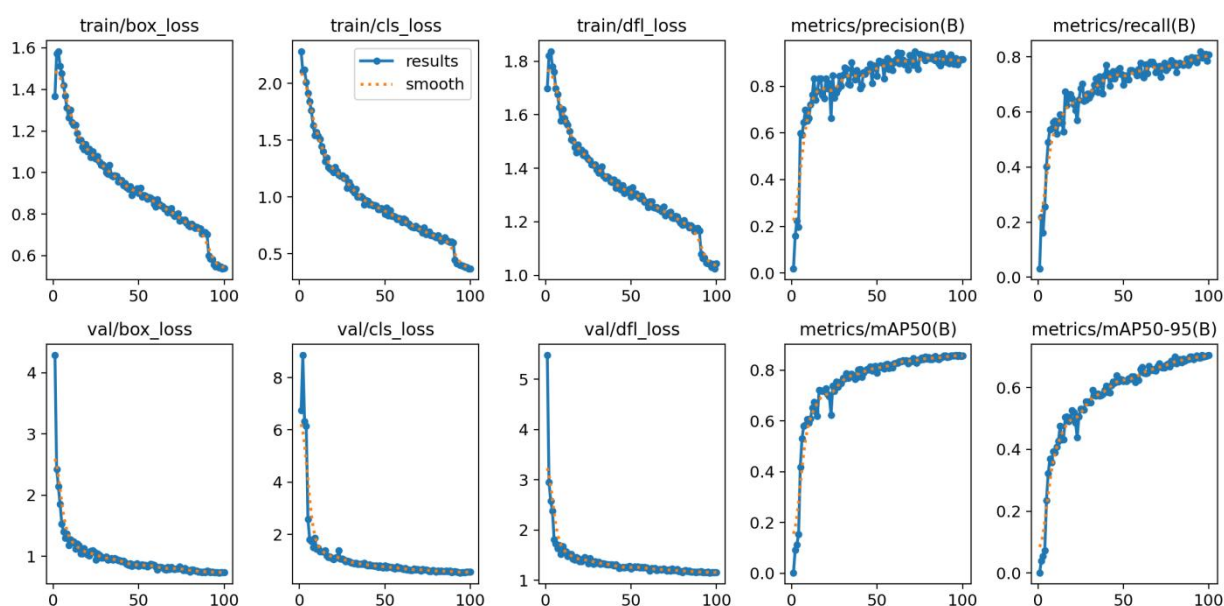
Hình 3.7: Biểu đồ label đánh giá qua các thực nghiệm



Hình 3.8: Biểu đồ labels_correlogram đánh giá qua các thực nghiệm



Hình 3.9: Biểu đồ confusion_matrix đánh giá qua các thực nghiệm



Hình 3.10: Biểu đồ đánh giá qua các thực nghiệm

3.3. Thảo luận

Nhận thấy ở bài toán bệnh tim thuật toán LR có kết quả vượt trội hơn kết quả chạy với thuật toán K-NN và trong ngữ cảnh quan trọng của việc chẩn đoán bệnh tim, mục tiêu chính của tôi là đảm bảo một tỷ lệ nhớ cao cho lớp tích cực. Việc xác định chính xác mọi trường hợp nghi ngờ mắc bệnh tim là rất quan trọng, vì một trường hợp bị bỏ sót có thể có hậu quả nghiêm trọng. Tuy nhiên, trong khi cố gắng đạt được tỷ lệ nhớ cao này, việc duy trì một hiệu suất cân bằng là rất quan trọng để tránh các biện pháp can thiệp y tế

không cần thiết đối với những người khỏe mạnh. Bây giờ, chúng tôi sẽ đánh giá các mô hình của mình dựa trên các tiêu chí y tế quan trọng này.

Về nghiên cứu để xác định vị trí polyp trên hình ảnh bằng bounding boxes dựa vào kết quả sau khi thực nghiệm nhận thấy sự tối ưu của YOLOv9 c đã vượt trội trên cả bốn tập dữ liệu so với các nghiên cứu trước kia dưới đây là bảng so sánh nghiên cứu của tôi với các nghiên cứu trước kia.

Bảng 3.4: Bảng so sánh các chỉ số trên cùng tập dữ liệu dung để xác định polyp so với các nghiên cứu khác

Study	Architecture	Recall	Precision	F1-score
Tajbakhsh et al. [15]	Custom architecture	70%	63%	66%
Zheng et al. [16]	YOLO-V1	74% on ETIS	77.4% on ETIS	75.7% on ETIS
Zhang et al. [17]	Custom architecture based on Single-Shot Multibox Detector (SSD)	76.37%	93.92%	84%
Wang et al. [18]	Custom architecture	80.77% on ETIS	88.89% on ETIS	84.6% on ETIS
Lee et al. [19]	YOLO-V2	96.7% 90.2% on CVC-ClinicDB	97.4% 98.2% on CVC-ClinicDB	97% 94% on CVC-ClinicDB
Qadir et al. [20]	Resnet34 and MDeNet	86.54% on ETIS 91% on CVC-ColonDB	86.12% on ETIS 88.35% on CVC-ColonDB	86.3% on ETIS 89.6% on CVC-ColonDB
Xu et al. [21]	YOLO-V3	75.70% 71.63% on ETIS	85.54% 83.24% on ETIS	7.99% 77% on ETIS
Liu et al. [22]	Custom architecture	87.5% on ETIS	77.8% on ETIS	82.4% on ETIS
Pacal and Karaboga [23]	YOLO-V3, YOLO-V4	76.10% on ETIS for YOLO-V3 79.25% on ETIS for YOLO-V4 82.55% on ETIS for YOLO-V4 modified	79.44% on ETIS for YOLO-V3 81.78% on ETIS for YOLO-V4 91.62% on ETIS for YOLO-V4 modified	77.7% on ETIS for YOLO-V3 80.5% on ETIS for YOLO-V4 86.8% on ETIS for YOLO-V4 modified
Nogueira-Rodríguez et al. [24]	YOLO-V3	87%	89%	88.1%
Wan et al. [25]	YOLO-V5	92.1% on Kvasir-SEG	91.3% on Kvasir-SEG	91.7% on Kvasir-SEG
Pacal et al. [26]	YOLO-V3, YOLO-V4	91.04% on ETIS for YOLO-V3 modified	90.61% on ETIS for YOLO-V3 modified	90.8% on ETIS for YOLO-V3 modified
Li et al. [27]	FCN+U-Net	77% on ClinicDB	90% on ClinicDB	83% on ClinicDB
Tashk et al. [28]	U-Net with post-processed	82.7% on CVC-ClinicDB 90.9% on ETIS 82.4% on CVC-ColonDB	70.2% on CVC-ClinicDB 70.2% on ETIS 62% on CVC-ColonDB	76% CVC-ClinicDB 79.2% on ETIS 70.7% on CVC-ColonDB
Qadir [29]	F-CNN	86.3% on CVC-ClinicDB	73.6% on CVC-ClinicDB	79.4% on CVC-ClinicDB
Akbari et al. [30]	FCN-8s	74.8% on CVC-ColonDB	88.3% on CVC-ColonDB	81% on CVC-ColonDB
Lalinia[31]	YOLO-V8 m	91.2% on CVC-ClinicDB 90.7% on ETIS 91.4% on CVC-ColonDB	95.1% on CVC-ClinicDB 94.6% on ETIS 94.4% on CVC-ColonDB	91.4% on CVC-ClinicDB 91.5% on ETIS 92.1% on CVC-ColonDB
YOLOv9 (our)	YOLOv9 c	CVC-ClinicDB 98,7% ETIS 98% CVC-ColonDB 96,8% Kvasir-SEG 80,5%	CVC-ClinicDB 97,7% ETIS 99,4% CVC-ColonDB 98,2 % Kvasir-SEG 96%	CVC-ClinicDB 98,1% ETIS 99,3% CVC-ColonDB 98,2% Kvasir-SEG 85,8 %

CHƯƠNG 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1. Kết luận

Học máy (ML) đã trở thành một công cụ quan trọng trong lĩnh vực y học và chẩn đoán tự động nhờ khả năng đạt được độ chính xác cao và giảm thời gian, chi phí và công sức. Với số lượng người mắc bệnh tim tăng đáng kể trên toàn cầu, việc chẩn đoán bệnh tim đang trở nên khó khăn đối với các bác sĩ, kể cả những chuyên gia giàu kinh nghiệm.

Trong bài viết này, chúng tôi trình bày mô hình Logistic Regression (LR) được tối ưu hóa để phân loại rối loạn nhịp tim từ các bộ dữ liệu điện tâm đồ một cách chính xác. Chúng tôi đã xem xét mô hình được đề xuất và xác nhận các chỉ số đánh giá như Accuracy, Precision, Recall và F1-score.

Hơn nữa, chúng tôi đã so sánh kết quả của mô hình đề xuất với kết quả của các nghiên cứu gần đây. Với Accuracy đạt 72%, Precision đạt 74%, Recall đạt 72% và F1-score đạt 73%. Trong tương lai, mô hình được đề xuất có thể trở thành một công cụ quan trọng trong việc chẩn đoán chính xác nhiều loại bệnh y tế. Học máy có thể đóng vai trò quan trọng trong việc hỗ trợ các bác sĩ và chuyên gia trong việc đưa ra các quyết định y tế phù hợp.

Học sâu (Deep Learning) đã trở thành một công cụ quan trọng trong lĩnh vực y học và chẩn đoán tự động nhờ khả năng đạt được độ chính xác cao và giảm thời gian, chi phí và công sức. Trong lĩnh vực này, một ứng dụng quan trọng của học sâu là trong việc nhận dạng polyp trong hình ảnh ung thư đại trực tràng.

Về phần xác định vị trí polyp tôi đã thực hiện việc huấn luyện và kiểm tra trên tập dữ liệu và nhận thấy được sự vượt bậc của mô hình YOLOv9 phiên bản c . Cụ thể, các kết quả đạt được cho thấy cả bốn bộ dữ liệu đều đạt được độ chính xác cao trong việc nhận biết khối polyp trong ung thư đại trực tràng.

4.2. Hướng phát triển

Tăng cường thu thập thêm dữ liệu từ nhiều nguồn.

Chạy thực nghiệm các mô hình hiện có trên nhiều bộ tham số khác nhau, đồng thời nghiên cứu và xây dựng thêm các mô hình mới nhất cho việc chuẩn đoán trường hợp mắc bệnh tim và nhận dạng polyp.

TÀI LIỆU THAM KHẢO

- [1] J. Lewis, Y.-J. Cha, and J. Kim, “Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images,” *Sci Rep*, vol. 13, no. 1, p. 1183, Jan. 2023, doi: 10.1038/s41598-023-28530-2.
- [2] “Colorectal cancer statistics, 2020 - Siegel - 2020 - CA: A Cancer Journal for Clinicians - Wiley Online Library.” Accessed: Apr. 26, 2024. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21601>
- [3] “heart.csv.” Accessed: Apr. 26, 2024. [Online]. Available: <https://www.kaggle.com/datasets/arezaei81/heartcsv>
- [4] A. Garg, B. Sharma, and R. Khan, “Heart disease prediction using machine learning techniques,” *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012046, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012046.
- [5] Md. I. Hossain *et al.*, “Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison,” *Iran Journal of Computer Science*, pp. 1–21, Jun. 2023, doi: 10.1007/s42044-023-00148-7.
- [6] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” *International Journal of Computer Applications*, vol. 17, pp. 43–48, Mar. 2011, doi: 10.5120/2237-2860.
- [7] “Colorectal polyp detection in colonoscopy images using YOLO-V8 network | Signal, Image and Video Processing.” Accessed: Apr. 26, 2024. [Online]. Available: <https://link.springer.com/article/10.1007/s11760-023-02835-1>
- [8] D. Jha *et al.*, “Kvasir-SEG: A Segmented Polyp Dataset,” in *MultiMedia Modeling*, vol. 11962, Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, Eds., in Lecture Notes in Computer Science, vol. 11962. , Cham: Springer International Publishing, 2020, pp. 451–462. doi: 10.1007/978-3-030-37734-2_37.
- [9] “Polyp - Grand Challenge,” grand-challenge.org. Accessed: Apr. 26, 2024. [Online]. Available: <https://polyp.grand-challenge.org/CVCClinicDB/>
- [10] “Papers with Code - CVC-ColonDB Benchmark (Medical Image Segmentation).” Accessed: Apr. 26, 2024. [Online]. Available: <https://paperswithcode.com/sota/medical-image-segmentation-on-cvc-colondb>
- [11] “Results from testing on ETIS-Larib polyp data.” ResearchGate. Accessed: Apr. 26, 2024. [Online]. Available: https://www.researchgate.net/figure/Results-from-testing-on-ETIS-Larib-polyp-data_tbl4_333759671

- [12] “Tìm hiểu về YOLO trong bài toán real-time object detection.” Accessed: Apr. 30, 2024. [Online]. Available: <https://viblo.asia/p/tim-hieu-ve-yolo-trong-bai-toan-real-time-object-detection-yMnKMdvr57P>
- [13] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information.” arXiv, Feb. 28, 2024. doi: 10.48550/arXiv.2402.13616.
- [14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection.” arXiv, May 09, 2016. doi: 10.48550/arXiv.1506.02640.
- [15] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks,” 2015, pp. 79–83.
- [16] Y. Zheng *et al.*, “Localisation of Colorectal Polyps by Convolutional Neural Network Features Learnt from White Light and Narrow Band Endoscopic Images of Multiple Databases,” 2018, pp. 4142–4145.
- [17] X. Zhang *et al.*, “Real-time gastric polyp detection using convolutional neural networks,” 2019, pp. 636–643.
- [18] D. Wang *et al.*, “AFP-Net: Realtime Anchor-Free Polyp Detection in Colonoscopy,” . In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI, 2019, pp. 636–643.
- [19] J. Y. Lee *et al.*, “Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets,” 2020.
- [20] H. A. Qadir, Y. Shin, J. Sollhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, “Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction,” 2021.
- [21] J. Xu *et al.*, “Real-time automatic polyp detection in colonoscopy using feature enhancement module and spatiotemporal similarity correlation unit,” 2021.
- [22] X. Liu, X. Guo, Y. Liu, and Y. Yuan, “Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images,” 2021.
- [23] Ishak Pacal and Dervis Karaboga, “A robust real-time deep learning based automatic polyp detection system,” 2022.
- [24] Alba Nogueira-Rodríguez, R. Domínguez, Fernando Campos-Tato, and Jesus Miguel Herrero Rivas, “Real-time polyp detection model using convolutional neural networks,” 2022.
- [25] Jingjing Wan, Bolun Chen, and Yongtao Yu, “Polyp Detection from Colorectum Images by Using Attentive YOLOv5,” 2021.

- [26] Ishak Pacal *et al.*, “An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets,” 2022.
- [27] Qiaoliang Li *et al.*, “Colorectal polyp segmentation using a fully convolutional neural network,” 2017, pp. 1–5.
- [28] Ashkan Tashk, Jürgen Herp, and Esmail Nadimi, “Automatic Segmentation of Colorectal Polyps based on a Novel and Innovative Convolutional Neural Network ApproachAutomatic Segmentation of Colorectal Polyps based on a Novel and Innovative Convolutional Neural Network Approach,” 2019, pp. 384–391.
- [29] Qadi and Hemin Ali Qadir, “Development of Image Processing Algorithms for the AutomaticScreening of Colon Cancer,” 2020.
- [30] Mojtaba Akbari *et al.*, “Polyp Segmentation in Colonoscopy Images Using Fully Convolutional Network,” In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC, 2018, pp. 69–72.
- [31] Mehrshad Lalinia, “Colorectal polyp detection in colonoscopy images using YOLO-V8 network,” 2024, pp. 2047–2058.