ĐOÀN THANH NIÊN CỘNG SẢN HỒ CHÍ MINH

BAN CHẤP HÀNH TP. HỒ CHÍ MINH

CÔNG TRÌNH DỰ THI

GIẢI THƯỞNG SINH VIÊN NGHIÊN CỨU KHOA HỌC EURÉKA

LẦN THỨ 23 NĂM 2021

TÊN CÔNG TRÌNH: ỨNG DỤNG HỌC MÁY TRONG NHẬN DẠNG NHIỄM SẮC THỂ

LĨNH VỰC NGHIÊN CỨU: CÔNG NGHỆ THÔNG TIN

CHUYÊN NGÀNH: TRÍ TUỆ NHÂN TẠO

Nhóm tác giả: Ngô Quốc Hoàng, Lê Thanh Phong, Phan Nguyễn Mai Phương, Lê Văn Tiến.

MỤC LỤC

1. ĐẶT VẤN ĐỀ 5
1.1 Tính cấp thiết của đề tài: 5
1.2 Ý nghĩa khoa học và thực tiễn của đề tài.....6
1.3 Mục tiêu nghiên cứu của đề tài.....6
1.4 Đối tượng nghiên cứu:.....6
1.5 Nội dung thực hiện, phương pháp nghiên cứu: 6
1.5.1 Nghiên cứu các kỹ thuật xử lý ảnh và máy học để phân đoạn đối tượng.6
1.5.2 Phân đoạn nhiễm sắc thể. 6
1.5.3 Cắt ảnh các NST từ kết quả của mục 7
1.5.4 Trích chọn các đặc trưng NST.....7
1.5.5 Phân loại NST.....7
1.5.6 Viết báo cáo tổng kết.....7
2. TỔNG QUAN ĐỀ TÀI.....8
2.1 Tổng quan về nhiễm sắc thể:.....8
2.2 Tổng quan của ứng dụng học máy trong nhận dạng nhiễm sắc thể 8
2.2.1 Giới thiệu về học máy: 8
2.2.2 Học máy trong nhận dạng nhiễm sắc thể.....8
3. PHƯƠNG PHÁP TIẾP CẬN 9
3.1. Inception (GoogleLeNet).....9
3.2. Kiến trúc Resnet 10
3.3. IAI: Giao diện thích ứng hình ảnh.....11
3.4. Thuật toán CDA 12
3.5. CIR-Net 13
4. KẾT QUẢ VÀ THẢO LUẬN.....15
4.1 Dữ liệu.....15
4.2 Tiến độ thực hiện.....15

4.3 Kết quả thực nghiệm 15
5. KẾT LUẬN VÀ ĐỀ NGHỊ 18
5.1 Kết quả khoa học đạt được 18
5.2 Ý nghĩa của dự án.....18
5.3 Hướng phát triển.....18
6. Tài liệu tham khảo.....19

TÓM TẮT

Đề tài: "Ứng dụng học máy trong nhận dạng nhiễm sắc thể" được thực hiện tại trường Đại học Công Nghiệp TPHCM

- Nghiên cứu về hình dạng của nhiễm sắc thể.
- Nghiên cứu về sự biến đổi và bất thường của các nhiễm sắc thể.
- Nghiên cứu phương pháp xử lý hình ảnh nhiễm sắc thể bằng thư viện CV, Tensorflow, Keras,.. trên Google Colab.
- Nghiên cứu các mô hình CIR-Net, kiến trúc ResNet và thuật toán CDA để cải thiện hiệu suất của bộ dữ liệu nhiễm sắc thể

Kết quả thu được:

- Cải thiện hiệu suất cũng như tốc độ nhận dạng nhiễm sắc thể bằng ngôn ngữ máy học.
- Độ nhận diện và chính xác cao.

1. ĐẶT VẤN ĐỀ

1.1 Tính cấp thiết của đề tài:

Đột biến nhiễm sắc thể là hiện tượng có ý nghĩa quan trọng trong sự tiến hoá của loài người. Đây được xem là một trong những nguồn quý giá cho quá trình nghiên cứu các bệnh về di truyền học. Biến đổi nhiễm sắc thể có thể làm thay đổi cấu trúc và số lượng nhiễm sắc thể, sự thay đổi này có ảnh hưởng nhất định đến cơ thể người nhưng hầu hết đều là có hại, chỉ có một số ít là có lợi hoặc không biểu hiện ra bên ngoài.

Ngày nay, vấn đề bất thường ở nhiễm sắc thể vẫn đang là một trong những vấn đề rất được quan tâm. Sự bất thường ấy có thể gây ra những hậu quả không mong muốn, và những đặc điểm này có thể di truyền lại cho các thế hệ con nếu không được phát hiện sớm cũng như nhận dạng được chúng một cách chính xác. Từ đó việc nhận dạng, nghiên cứu cấu trúc nghiễm sắc thể của một người để xác định xem họ có mắc bệnh di truyền nào có liên quan đến nhiễm sắc thể hay không đã trở thành vấn đề nan giải của đội ngũ kỹ sư, bác sĩ.

Theo WHO, lệch bội nhiễm sắc thể giới tính là lệch bội thường gặp chiếm khoảng một nửa của tất cả các dị thường nhiễm sắc thể ở người với tổng tần số 1:400* và hiếm khi gây tử vong nhưng lại để lại ảnh hưởng lớn tới sự phát triển của trẻ sau sinh về giới tính và hình thái cơ thể. Việc quan sát, xét nghiệm cũng như nghiên cứu về sự biến đổi bất thường của các nhiễm sắc thể thường rất mất thời gian do nhiều yếu tố khách quan khác như: cơ sở vật chất, y thiết bị nghiên cứu, kỹ năng chuyên môn,…)

Sự tiến bộ của công nghệ - kỹ thuật và các nhu cầu về Machine Learning (học máy) trong những năm gần đây, nghiên cứu về nhận dạng nhiễm sắc thể cũng đã có được sự phát triển mạnh mẽ. Nhận thấy tính cấp bách của những bệnh di truyền liên quan tới nhiễm sắc thể và tầm quan trọng của Machine Learning trong tương lai. Vì vậy trong đề tài nghiên cứu này, chúng tôi nghiên cứu tiếp cận phân tích nhiễm sắc thể người dựa trên học máy nhằm rút ngắn thời gian cho việc xét nghiệm Karyotype một cách thủ công, cũng như đưa ra những kết quả nhanh hơn, giúp việc nghiên cứu của đội ngũ y bác sĩ sẽ diễn ra thuận lợi hơn để phát hiện và điều trị các vấn đề về di truyền một cách hiệu quả.

1.2 Ý nghĩa khoa học và thực tiễn của đề tài

Về khoa học:

- Tạo nền tảng phát triển các nghiên cứu về sau.
- Tạo tài liệu hỗ trợ học tập trong ngành Khoa học máy tính và Khoa học dữ liệu.

Về thực tiễn: mô hình này phục vụ cho việc nhận dạng nhiễm sắc thể, có thể rút ngắn thời gian cũng như một vài công đoạn trong quá trình xét nghiệm nhiễm sắc thể của các bác sĩ.

1.3 Mục tiêu nghiên cứu của đề tài

Mục tiêu tổng quát:

Nhận dạng, nghiên cứu cấu trúc nghiễm sắc thể của một người để xác định xem họ có mắc bệnh di truyền nào có liên quan đến nhiễm sắc thể hay không

Mục tiêu cụ thể:

- Nghiên cứu ứng dụng các mô hình Machine Learning trong nhận dạng NST với hiệu năng cao hơn, có thể áp dụng trong thực tế.
- Tạo nền tảng phát triển các nghiên cứu về sau.
- Tạo tài liệu hỗ trợ học tập trong ngành Khoa học máy tính và Khoa học dữ liệu

1.4 Đối tượng nghiên cứu:

Bộ dữ liệu hình ảnh nhiễm sắc thể được thu thập từ trung tâm y tế di truyền Quảng Đông.

1.5 Nội dung thực hiện, phương pháp nghiên cứu:

1.5.1 Nghiên cứu các kỹ thuật xử lý ảnh và máy học để phân đoạn đối tượng.

- Cách tiếp cận: Tiếp cận từ cơ sở lý luận kết hợp với đánh giá thực nghiệm đặc thù trong ảnh Y khoa.
- Phương pháp nghiên cứu, kỹ thuật sử dụng: Kết hợp lý thuyết và thực nghiệm. Sử dụng nền tảng công nghệ trên Python.
- Kết quả dự kiến: Báo cáo.

1.5.2 Phân đoạn nhiễm sắc thể.

- Cách tiếp cận: Dùng các kỹ thuật xử lý ảnh kết hợp học máy.
- Phương pháp nghiên cứu, kỹ thuật sử dụng: Kết hợp lý thuyết và thực nghiệm. Sử dụng nền tảng công nghệ trên Python.
- Kết quả dự kiến: Chương trình và báo cáo kết quả thực nghiệm.

1.5.3 Cắt ảnh các NST từ kết quả của mục

- Cách tiếp cận: Dùng các kỹ thuật xử lý ảnh thông thường.
- Phương pháp nghiên cứu, kỹ thuật sử dụng: Kết hợp lý thuyết và thực nghiệm. Sử dụng nền tảng công nghệ trên Python.
- Kết quả dự kiến: Chương trình và báo cáo.

1.5.4 Trích chọn các đặc trưng NST

- Cách tiếp cận: Dùng các kỹ thuật trích chọn đặc trưng.
- Phương pháp nghiên cứu, kỹ thuật sử dụng: Kết hợp lý thuyết và thực nghiệm. Sử dụng nền tảng công nghệ trên Python.
- Kết quả dự kiến: Chương trình và báo cáo.

1.5.5 Phân loại NST.

- Cách tiếp cận: Dùng các kỹ thuật xử lý ảnh dùng máy học.
- Phương pháp nghiên cứu, kỹ thuật sử dụng: Kết hợp lý thuyết và thực nghiệm. Sử dụng nền tảng công nghệ trên Python.
- Kết quả dự kiến: Chương trình và báo cáo.

1.5.6 Viết báo cáo tổng kết

2. TỔNG QUAN ĐỀ TÀI

2.1 Tổng quan về nhiễm sắc thể:

Nhiễm sắc thể là bào quan chứa bộ gen chính của một sinh vật, là cấu trúc quy định sự hình thành protein, có vai trò quyết định trong di truyền tồn tại ở cả sinh vật nhân thực và sinh vật nhân sơ. Trong ngôn ngữ của nhiều quốc gia, khái niệm này được gọi là Chromosome. Chúng chứa thông tin di truyền của người, và được sử dụng để phân tích các bệnh di truyền. Ở người có 23 cặp NST, trong đó có 22 cặp NST thường, còn lại là NST giới tính (nhiễm sắc thể X và Y ở tế bào đực và nhân đôi X ở tế bào cái). Một trong những phương pháp hỗ trợ phân tích bộ NST trên tế bào là Karyotyping. Nó cho biết về số lượng và sự thay đổi về cấu trúc của NST, từ đó cho phép xác định rối loạn di truyền gây nên dị tật.

2.2 Tổng quan của ứng dụng học máy trong nhận dạng nhiễm sắc thể

2.2.1 Giới thiệu về học máy:

Là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kỹ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

2.2.2 Học máy trong nhận dạng nhiễm sắc thể

Trong di truyền tế bào học, Karyotyping là công việc vô cùng khó khăn. Đã có rất nhiều nhà nghiên cứu dành riêng cho việc tạo mẫu tự động bằng cách sử dụng các kỹ thuật tính toán trong nhiều năm qua [3-8].

Mặc dù những nghiên cứu [1], [8], [9], [10] đã đóng góp một phần nào đó cho nhiệm vụ phân loại nhiễm sắc thể. Tuy nhiên, việc phân loại nhiễm sắc thể một cách chính xác và hiệu quả trong ứng dụng lâm sàng (Clinical Application) với điều kiện tập dữ liệu được dán nhãn không đầy đủ vẫn là một nhiệm vụ khó khăn vì những lý do sau:

- *Các biến dạng phong phú về hình dạng nhiễm sắc thể:* Nhiễm sắc thể cùng loại có hình dạng và hướng hoàn toàn khác nhau.
- *Khó khăn để thu thập một lượng lớn dữ liệu được dán nhãn:* Vì hình ảnh nhiễm sắc thể có liên quan nhiều đến quyền riêng tư của bệnh nhân, nên các nhà nghiên cứu rất khó sửa đủ dữ liệu từ các cơ sở y tế để có thể phân loại.
- *Trang thiết bị, cơ sở vật chất:* Đề tài hiện chỉ được training trên những laptop cá nhân, cấu hình không được mạnh mẽ như những thiết bị chuyên dụng để nghiên cứu nên đã mất rất nhiều thời gian để có thể kiểm tra được bộ dữ liệu một cách hoàn chỉnh.

3. PHƯƠNG PHÁP TIẾP CẬN

Để giải quyết vấn đề trên, chúng tôi đã đề ra phương pháp tiếp cận phân loại có tên là **CIR – Net** để tự động phân loại nhiễm sắc thể trong một đường ống từ đầu đến cuối . Một mạng nơ-ron phức hợp tên là **Inception v3** để hỗ trợ phân tích hình ảnh và phát hiện đối tượng. Bên cạnh đó, chúng tôi còn áp dụng kiến thức về **ResNet** (Residual neural network) - mạng lưới thần kinh nhân tạo thuộc loại xây dựng trên các cấu trúc được biết đến từ các tế bào hình chóp trong vỏ não. Nhờ khả năng biểu diễn mạnh mẽ của ResNet, hiệu suất của nhiều ứng dụng thị giác máy, không chỉ các ứng dụng phân loại hình ảnh được tăng cường. Một số ví dụ có thể kể đến là các ứng dụng phát hiện đồ vật và nhận dạng khuôn mặt. Và một số giao diện cũng như các thuật toán khác như **IAI, CDA**,..

3.1. Inception (GoogleLeNet)

Vào năm 2014, các nhà nghiên cứu của google đã đưa ra mạng Inception tham dự cuộc thi ImageNet 2014. Mô hình này khá đặc biệt, không hoàn toàn là các tầng layer nối tiếp gối đầu lên nhau như các mạng trên. Mạng gồm các đơn vị gọi là "inception cell". Thực hiện convolution 1 input với nhiều filter khác nhau rồi tổng hợp lại, theo nhiều nhánh (branch).

Inception có một đặc điểm khá hay là có thêm 2 output phụ. Người ta tin rằng hai output phụ này không quá ảnh hưởng tới chất lượng của mạng trong khi train những epoch đầu. Nó giúp cho việc train diễn ra nhanh hơn khi tối ưu những layer đầu dựa vào các output phụ (trong những epoch đầu).

Bốn lớp max-pooling: dùng để giảm kích thước của dữ liệu đầu vào nhằm giảm sự phức tạp của mô hình và chi phí tính toán.

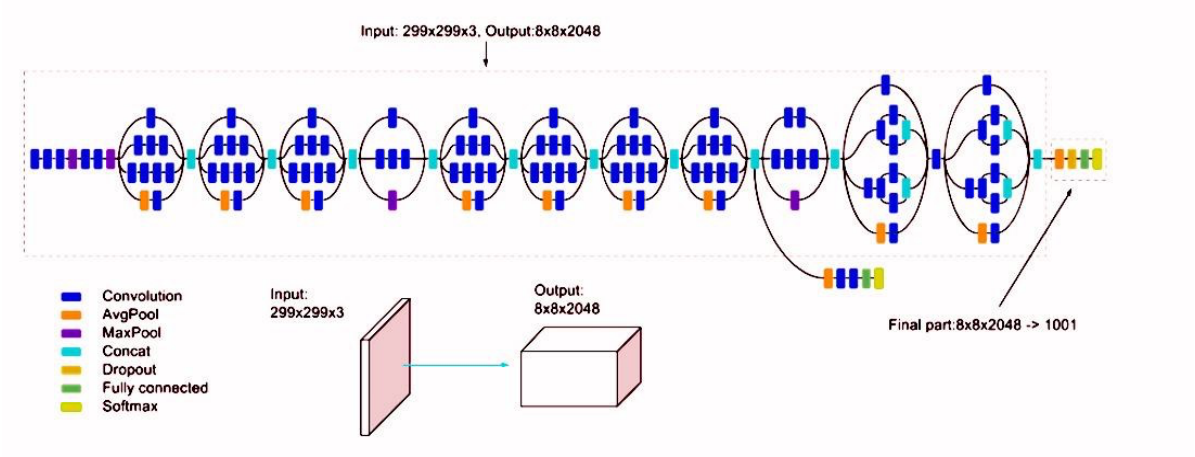
Lớp average-pooling: dùng để cải thiện hiệu suất của mô hình và giảm overfitting.

Lớp dropout (với 40% tỉ lệ đầu ra bị giảm): là 1 phương thức regularization được sử dụng trước những hàm kích hoạt nhằm giảm overfitting.

Hàm softmax: dùng cho lớp đầu ra với 24 đầu ra tương ứng với 24 lớp cần phân loại.

Ngoài ra còn có mạng lưới bổ sung bao gồm những bộ phân loại phụ trợ sau: 1 lớp average-pooling với kernel 5x5 và stride 3, 1 lớp 1x1 convolutional với bộ lọc 127, 1 lớp full-connected, 1 lớp dropout (với 70% tỉ lệ đầu ra bị giảm), cuối cùng là 1 hàm kích hoạt softmax với đầu ra tương ứng 24 loại NST

Dưới đây là kiến trúc mạng Inception:

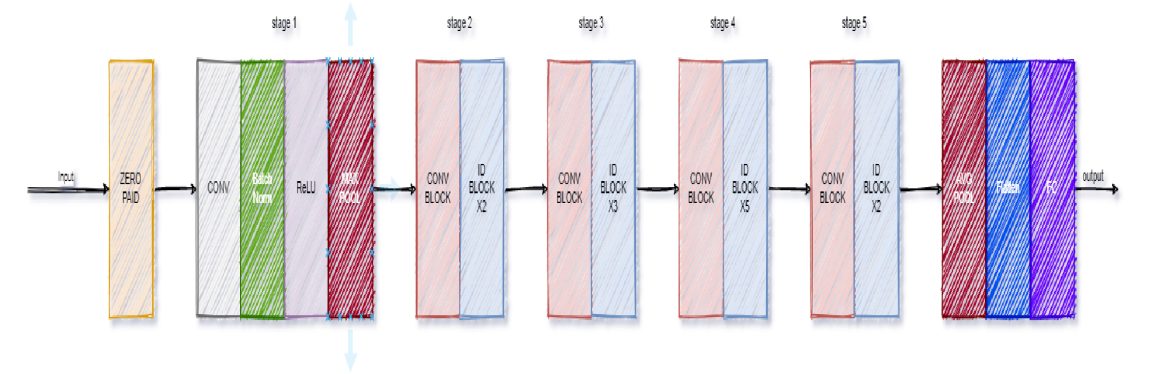


Hình 2. Kiến trúc mạng inception.

3.2. Kiến trúc Resnet

ResNet dùng 152 layer, độ lớn được xem là hơn hẳn so với các kiến trúc trước đó[11]. Kiến trúc này chiến thắng cuộc thi ILSVRC năm 2015, đồng thời đạt được top-5 error rate ở mức 3.6%. Độ chính xác này đạt được bởi sự vận hành đồng bộ của một mạng lưới ResNet , chỉ nguyên một mô hình trong số này đã cho độ chính xác là 4.5%. Việc huấn luyện một kiến trúc gồm 152 layer về cơ bản là không mấy khả thi, trừ khi người ta tìm được cách kết hợp với các cải tiến quan trọng.

ResNet sử dụng phương thức kết nối tắt giữa các layer nhằm mục đích cho phép việc sao chép giữa các layer, đồng thời đề xuất kiến trúc lặp lại đặc tính (so với kiến trúc kế thừa). Các mạng sâu sử dụng bộ nhớ ngắn hạn và các đơn vị hồi quy có tác động đến các nguyên lý tương đương của chuỗi dữ liệu bằng việc cho phép các phần trạng thái được sao chép từ layer này tới layer tiếp theo thông qua các cổng điều chỉnh.



Hình 3. Kiến trúc mạng Resnet

3.3. IAI: Giao diện thích ứng hình ảnh

Trong việc phân loại hình ảnh, việc phân loại một hình ảnh ở dạng cố định chẳng hạn như (224 , 224 , 3) trong nhiệm vụ phân loại ImageNet [12]. Chúng có nghĩa là mỗi hình ảnh của tập dữ liệu nên có 3 kênh màu, mỗi kênh có chiều cao và chiều rộng là 224 pixel. Những hình ảnh không đáp ứng được các yêu cầu của hình dạng, nên chúng ta cần phải cắt hoặc độn những hình ảnh thành hình dạng nhất định bằng thao tác tiền xử lý trước khi đưa vào bộ phân loại. Tuy nhiên, trong ứng dụng lâm sàng, nhiễm sắc thể có hình ảnh ở thang độ xám nên chỉ có một kênh màu. Ngoài ra, các nhiễm sắc thể từ các tập dữ liệu khác nhau có thể có sự khác biệt nhỏ trong hình dạng của chúng, điều này sẽ tạo ra mặt hạn chế khả năng ứng dụng của bộ phân loại này.

Được thúc đẩy bởi những khó khăn trong ứng dụng lâm sàng, chúng tôi đã thiết kế ra môđun giao diện thích ứng hình ảnh (IAI) cho việc phân loại các hình dạng khác nhau của hình ảnh một cách dễ dàng.

$$T = (ht, wt, ct) \quad (1)$$

$$O = (ho, wo, co) \quad (2)$$

Giả sử rằng hình dạng cuối cùng của quá trình phân loại được mô tả như Công thức 1, trong đó ht, wt và ct tương ứng biểu thị chiều cao, chiều rộng và kênh màu của đầu vào. Theo đó chúng tôi sử dụng Công thức 2 biểu thị hình dạng thực tế của hình

ảnh đã cho, trong đó ho, wo và co tương ứng biểu thị chiều cao, chiều rộng và kênh màu của hình ảnh đã cho.

Do đó, chúng tôi thiết kế mô-đun IAI bằng cách sử dụng lớp thần kinh phức hợp con2D() có các tham số bao gồm filter, kernel_size, padding, strides và input_shape. Trong các tham số này, filter là độ sâu đầu ra, kernel_size là hạt nhân của nơ-ron tích chập, padding đề cập đến các pixel và strides biểu thị có bao nhiêu pixel bỏ qua ở thao tác tích hợp tiếp theo. Các tham số này bị hạn chế được biểu diễn như Công thức 3.

$$wt = [wo + 2 * p - ks + 1] \quad (3)$$

$$ht = [ho + 2 * p - ks + 1]$$

Trong đó, khi wt = wo với p = 0, k = 1 và s = 1, vì vậy kiến trúc IAI được xem là một NiN tiêu chuẩn [13]. Khi wt < wo với p = 0 để kiến trúc IAI là một lớp cắt. Mặt khác, với s = 1 để kiến trúc IAI là một lớp đệm *(padding layer)*.

3.4. Thuật toán CDA

Thuật toán CDA là một thuật toán tăng cường hình ảnh. Thuật toán này có lợi ích: không chỉ có thể mở rộng, nâng cao độ chính xác của tập dữ liệu mà quá trình phân loại cũng có thể loại bỏ các tính năng định hướng của nhiễm sắc thể. Điều đó giúp cho việc phân loại không cần trải qua các tiền xử lý đặc biệt nào (ví dụ: xoay,làm thẳng, ...) trong giai đoạn thử nghiệm hoặc ứng dụng lâm sàng

Sau đây là thuật toán CDA: Tăng cường dữ liệu của nhiễm sắc thể

Require:

1 : X, Bộ dữ liệu hình ảnh nhiễm sắc thể

2 : Y, Tiêu đề tương ứng với X

3 : rtest, Tỷ lệ dữ liệu thử nghiệm 10%

4 : rvalid, Tỷ lệ xác thực 90%

5 : function CHROMOSOME_AUGMENTATION(X, Y, rtest, rvalid)

6 : test_set ← {}

```
7 :   train_set ← {}, val_set ← {}
8 :   tmp_set ← {}
9 :   Θ ← {θ0, θ1, . . . }
10:   //Tách tập dữ liệu thành tập training và tập test bởi rtest
11:   for (x, y) in (X, Y ) do
12:       if rand(0, 1) < rtest then
13:           test_set ← test_set ∪ {(x, y)}
14:       else
15:           tmp_set ← tmp_set ∪ {(x, y)}
16:       end if
17:   end for
18:   //Bổ sung tmp_set và tách thành tập training và tập
      testing
19:   for (x, y) in tmp_set do
20:       for θ in Θ do
21:           b ← V ector.random()
22:           x(θ) ← A(θ)x + b
23:           if rand(0, 1) < rvalid then
24:               val_set ← val_set ∪ {(x(θ) , y)}
25:           else
26:               train_set ← train_set ∪ {(x(θ) , y)}
27:           end if
28:       end for
29:   end for
30:   return train_set, val_set,
      test_set
31: end function
```

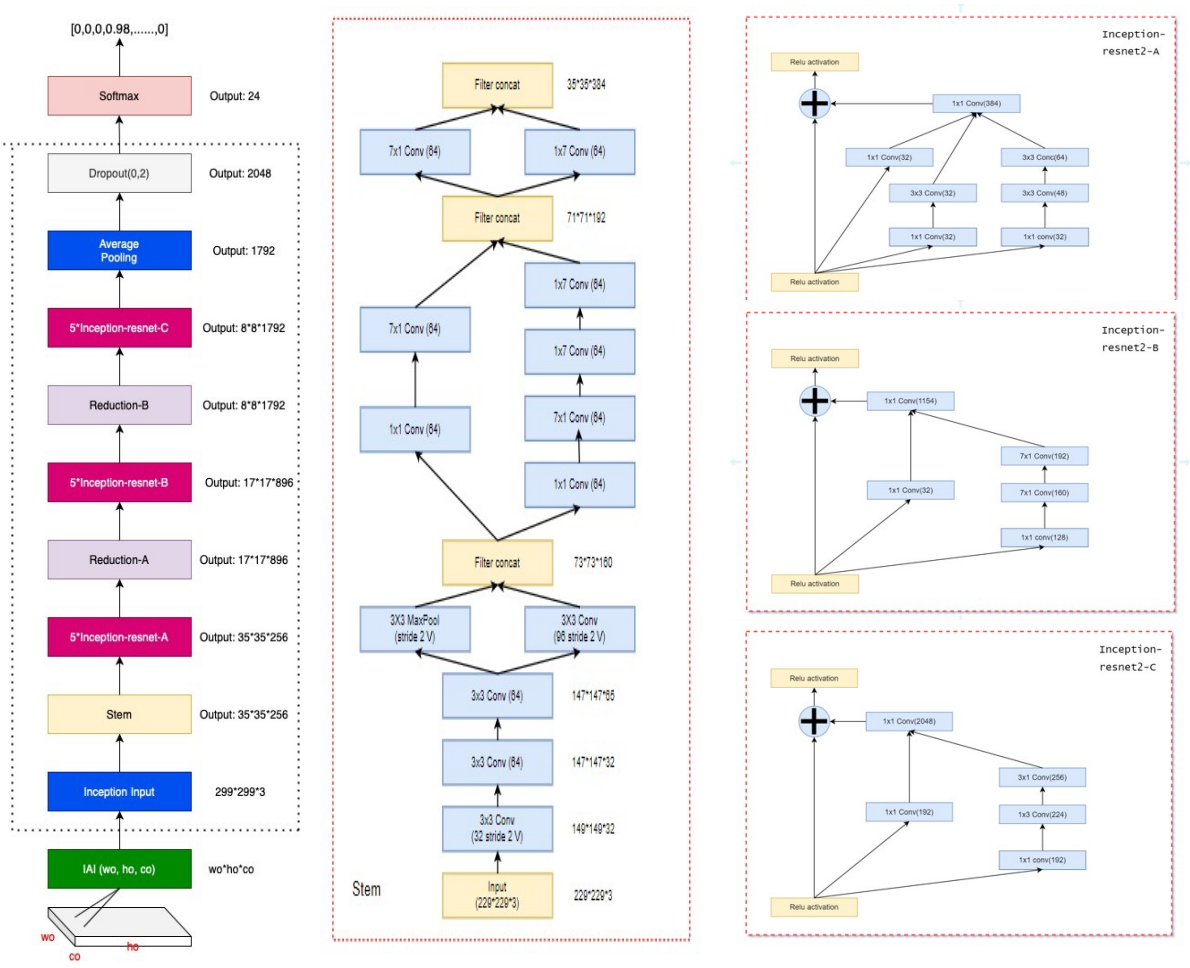
3.5. CIR-Net

CIR-Net được dựa trên kiến trúc **ResNet** và thuật toán **CDA** để cải thiện hiệu suất của bộ dữ liệu. Mô-đun này giúp cải thiện việc sử dụng các tài nguyên của máy tính và được viết bởi Szegedy và cộng sự trong cuộc thi về mạng ILSVRC 2014 [14]. Và sau đó được nâng cấp lên phiên bản mới có tên là Inception-V2(V3) [15]. Nhằm

mở rộng quy mô và các phép tính toán hiệu quả nhất có thể bằng các phép toán tích chập.

Ý tưởng của CIR-Net là để mã hóa các tính năng thưa thớt nhiễm sắc thể một cách tự nhiên. Và sau khi đánh giá các tính năng đó, chúng tôi cho rằng kiến trúc mạng ResNet đủ điều kiện cho nhiệm vụ phân loại nhiễm sắc thể bằng các sửa đổi thích hợp.

Sau đây là mô hình CIR-Net đã được cải tiến dựa trên ResNet và CDA:



Hình 4. Lược đồ tổng thể CIR-Net

Lược đồ tổng thể CIR-Net được sửa đổi từ khuôn khổ Inception-Resnet-v2 [11] bằng cách thêm một mô-đun IAI và sửa đổi các cách của Softmax từ 1000 thành 24. Ở phần giữa, nó là chi tiết của mô-đun Stem trong lược đồ tổng thể. Phần bên phải là lược đồ cho các mô-đun bên trong. Inception-ResNet2-A, Inception-ResNet2-B và Inception-ResNet2-C được mô tả từ trên xuống dưới.

4. KẾT QUẢ VÀ THẢO LUẬN

4.1 Dữ liệu

Tập dữ liệu được lấy từ trung tâm y tế di truyền ở Quảng Đông. Tập dữ liệu khi chúng tôi thu thập đã được xóa các dữ liệu cá nhân.

Dữ liệu gồm 65 karyotypes trong đó có 32 karyotypes đực, 33 karyotypes cái được tách ra thành 2990 nhiễm sắc thể đơn lẻ. Được đánh dấu từ 1 đến 22 cho các nhiễm sắc thể thường và 23,24 cho cặp nhiễm sắc thể giới tính X,Y. Kích thước đơn lẻ của từng nhiễm sắc thể sau khi xử lý 224x224, 8 bit/pixel trên thang độ xám.

Với số lượng dữ liệu ích cũng là một vấn đề khó khăn để đạt được kết quả mà chúng tôi mong muốn.

4.2 Tiến độ thực hiện

Về cơ bản, dự án đã hoàn thành xây dựng thuật toán, xây dựng các tập dữ liệu sơ bộ và đang tiến hành hoàn thành đầy đủ các phần còn thiếu

4.3 Kết quả thực nghiệm

Bằng cách sử dụng bộ công cụ **Keras** [16] và **TensorFlow** [17] chúng tôi đã đào tạo mạng với số lượng ảnh mỗi lần học là 32, mạng đã được đào tạo nhiều lần trên Colab pro. Và đây cũng là khó khăn lớn nhất của chúng tôi, bởi thời gian sử dụng Colab là có hạn.

Tuy nhiên chúng tôi đã cố gắng với nhiều cách như sử dụng **KFold** với 10 khoảng chia, lưu mô hình sau mỗi khoảng. Chúng tôi thực hiện huấn luyện 50 epoch cho mỗi khoảng, mỗi lần huấn luyện cách nhau 1 giờ. Với sự nỗ lực trên chúng tôi đã đưa ra kết quả chính xác lên đến 95,98%.

So sánh kết quả với các thuật toán khác như **Vanilla-CNN**[18] 86,44%, **Siamese Net** [19] 87,63%. Sử dụng thuật toán **CIR-Net**[20] thực sự thành công trong dự án này. Bằng cách chạy thử nghiệm trên các mô hình được công bố [18][19] tuy nhiên do ảnh đầu vào của các mô hình [18][19] là ảnh đã được duỗi thẳng, phương pháp này có thể gây biến dạng thông tin trên nhiễm sắc thể. Vì vậy chúng tôi đã áp dụng thuật toán tăng cường CDA để làm phong phú tập dữ liệu, điều đó có thể gây một số sự khác biệt so với mô hình gốc trong bài báo được công bố. Tất cả đều được thử nghiệm trên một bộ dữ liệu giống nhau và đưa ra kết quả hiệu xuất như trong bảng

Bảng 1

Phương pháp	**Phương pháp tăng cường**	**Precision**	**Recall**	F_1	**Acc**
Vanilla-CNN	Không	0.26	0.22	0.21	22.41%
	Duỗi thẳng	0.82	0.85	0.82	83.78%
	CDA	0.88	0.86	0.87	86.44%
Siamese Net	Không	0.22	0.22	0.21	21.91%
	Duỗi thẳng	0.86	0.87	0.86	86.79%
	CDA	0.88	0.87	0.87	87.63%
CIR-Net	Không	0.43	0.31	0.28	28.42%
	Duỗi thẳng	0.89	0.88	0.88	87.46%
	CDA	0.96	0.96	0.96	95.98%

Hiệu suất của bộ phân loại nhiễm sắc thể được đồng minh chúng tôi đánh giá bằng độ chính xác (acc), giá trị thu hồi (recall) và F1-score (F1) CIR-Net được đánh giá định lượng bằng các chỉ số đó. Để tính toán các chỉ số này, chúng ta cần xác định những điều sau bốn tiêu chí để phù hợp với ngữ cảnh của trình phân loại nhiều lớp như các tiêu chí nghiên cứu khác.

- True Positives (TP_j) : Nhiễm sắc thể được phân loại là loại j nhưng thực tế thuộc loại j
- False Positives (FP_j): Nhiễm sắc thể được phân loại là loại j nhưng thực tế không thuộc loại j
- False Negatives (FN_j): Nhiễm sắc thể được phân loại là loại k($\forall k \neq j$) nhưng thực tế thuộc loại j
- True Negatives (TN_j): Nhiễm sắc thể được phân loại là loại k($\forall k \neq j$) nhưng thực tế không thuộc loại j

Với sự giúp đỡ của TP_j, FP_j, FN_j và TN_j, precision được định nghĩa ở công thức 5, recall là công thức 7, F1 là công thức 9 và acc là công thức 10

$$precision_j = \frac{TP_j}{TP_j + FP_j} \quad (4)$$

$$\text{Precision} = \frac{1}{N_{types}} \sum_{j=1}^{N_{types}} precision_j \quad (5)$$

$$recall_j = \frac{TP_j}{TP_j + FN_j} \quad (6)$$

$$\text{recall} = \frac{1}{N_{types}} \sum_{j=1}^{N_{types}} recall_j \quad (7)$$

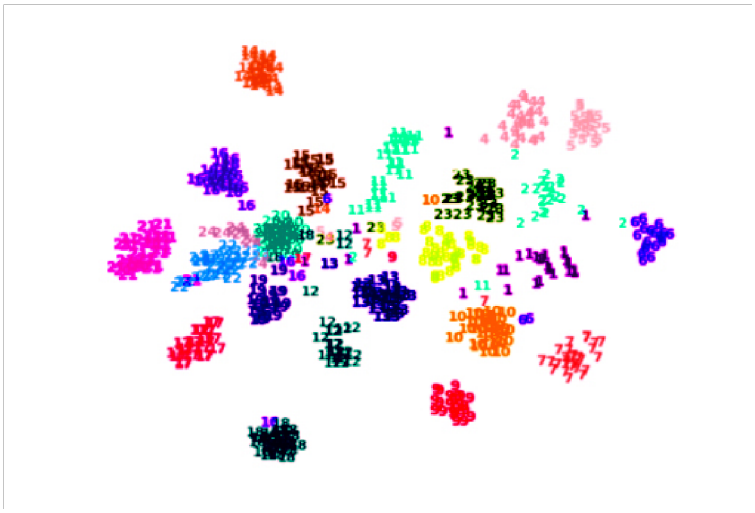
$$F_{1j} = \frac{2 \,.\, precision_j \,.\, recall_j}{precision_j + recall_j} \quad (8)$$

$$F_1 = \frac{1}{N_{types}} \sum_{j=1}^{N_{types}} F_{1j} \quad (9)$$

$$\text{acc} = \frac{1}{N} \sum_{j=1}^{N_{types}} TP_j \quad (10)$$

Trong các phương trình trên, N_{types} bằng 24 biểu thị các loại của nhiễm sắc thể trong khi N biểu thị tổng số sắc độ của một số hình ảnh trong thử nghiệm

Theo như kết quả thống kê ở bảng 1, phương pháp CIR-Net kết hợp với thuật toán tăng cường CDA cho ra độ chính xác 95.98% cũng tức là CIR tốt hơn những phương pháp khác. Nhưng độ chính xác của chúng tôi cao không chỉ do CDA mà còn do Inception-ResNet. Kiến trúc bao gồm ba mô-đun meta khác nhau Inception-resber2-A, Inception-resber2-B, và Inception-resber2-C. Các mô-đun meta tập hợp các nơ-ron phức hợp khác nhau, tạo khả năng trích xuất phong phú hơn nhằm nâng cao tính đặc trưng tạo thuận lợi cho quá trình phân cụm.



Hình 5. Mô tả khả năng phân loại NST.

5. KẾT LUẬN VÀ ĐỀ NGHỊ

5.1 Kết quả khoa học đạt được

Trong nghiên cứu này, chúng tôi đã sử dung mô hình học sâu CIR-Net dựa trên Phương pháp kiến trúc Inception-ResNet để phân tích bộ dữ liệu NST của trung tâm y tế di truyền Quảng Đông. Hiệu suất phân loại của phương pháp CIR-Net do chúng tôi đề xuất tốt hơn so với các phương pháp khác trong tập huấn luyện với bộ dữ liệu không đầy đủ. Phương pháp hoc tăng cường CDA có thể cải thiện cho bộ dữ liệu ít. Và nghiên cứu của chúng tôi lần này sẽ là tiền đề cho các mô hình về sau.

5.2 Ý nghĩa của dự án

Dự án ứng dụng học máy trong nhận dạng nhiễm sắc thể giúp cho các y bác sĩ giảm thiểu thời gian cũng như một vài công đoạn trong quá trình xét nghiệm nhiễm sắc thể của các bác sĩ. Tạo nền tảng phát triển cho các nghiên cứu về sau. Tạo tài liệu hỗ trợ học tập trong ngành Khoa học máy tính và Khoa học dữ liệu.

5.3 Hướng phát triển

Mặc dù kết quả khả quan, tuy nhiên vẫn còn rất nhiều hạn chế. Trong thời gian tới, chúng tôi sẽ tiếp tục thu thập thêm nhiều dữ liệu trong và ngoài nước để nghiên cứu và đem lại kết quả có độ chính xác cao hơn.

6. TÀI LIỆU THAM KHẢO

Trích dẫn tài liệu:

[1] S. Jindal, G. Gupta, M. Yadav, M. Sharma, and L. Vig, "Siamese networks for chromosome classification," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 72–81.

[2] T. Arora and R. Dhir, "A review of metaphase chromosome image selection techniques for automatic karyotype generation," Medical & biological engineering & computing, vol. 54, no. 8, pp. 1147–1157, 2016.

[3] J. Piper and E. Granum, "On fully automatic feature measurement for banded chromosome classification," Cytometry: The Journal of the International Society for Analytical Cytology, vol. 10, no. 3, pp. 242–255, 1989.

[4] G. Agam and I. Dinstein, "Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification," IEEE Transactions on pattern analysis and machine Intelligence, vol. 19, no. 11, pp. 1212–1222, 1997.

[5] P. A. Errington and J. Graham, "Application of artificial neural networks to chromosome classification," Cytometry: The Journal of the International Society for Analytical Cytology, vol. 14, no. 6, pp. 627–639, 1993.

[6] G. Ritter and M. T. Gallegos, "Outliers in statistical pattern recog - nition and an application to automatic chromosome classifica - tion," Pattern Recognition Letters, vol. 18, no. 6, pp. 525–539, 1997.

[7] X. Hu, W. Yi, L. Jiang, S. Wu, Y. Zhang, J. Du, T. Ma, T. Wang and X. Wu, "Classification of metaphase chromosomes using deep convolutional neural network," Journal of Computational Biology, 2019.

[8] M. Sharma, L. Vig et al., "Automatic chromosome classification using deep attention based sequence learning of chromosome bands," in 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1–8.

[9] W. Zhang, S. Song, T. Bai, Y. Zhao, F. Ma, J. Su, and L. Yu, "Chromosome classification with convolutional neural network based deep learning," in 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2018, pp. 1–5.

[10] Y. Qin, J. Wen, H. Zheng, X. Huang, J. Yang, L. Wu, N. Song, Y.-M.Zhu, and G.-Z. Yang, "Varifocal-net: A chromosome classification approach using deep convolutional networks," IEEE transactionson medical imaging, 2

[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inceptionv4, inception-resnet and the impact of residual connections on learning," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009
IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255

[13] M. Lin, Q. Chen, and S. Yan, “Network in network,” arXiv preprint arXiv:1312.4400, 2013.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[16] F. Chollet et al., “Keras: The python deep learning library,” Astro-physics Source Code Library, 2018.

[17] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” arXiv preprint arXiv:1603.04467, 2016.

[18] W. Zhang, S. Song, T. Bai, Y. Zhao, F. Ma, J. Su, and L. Yu, “Chromosome classification with convolutional neural network based deep learning,” in 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2018, pp. 1–5.

[19] S. Jindal, G. Gupta, M. Yadav, M. Sharma, and L. Vig, “Siamese networks for chromosome classification,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 72–81.

[20] Chengchuang Lin, Gansen Zhao, Zhirong Yang, Aihua Yin, Xinming Wang, Li Guo, Hanbiao Chen, Zhaohui Ma, Lei Zhao, Haoyu Luo, Tianxing Wang, Bichao Ding, Xiongwen Pang, Qiren Chen, “CIR-Net: Automatic Classification of Human Chromosome based on Inception-ResNet Architecture”.