ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

ĐỀ CƯƠNG LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

PHÂN LỚP NHANH HÌNH ẢNH TẾ BÀO BẠCH CẦU DỰA VÀO ĐẶC TRƯNG SÂU VÀ TỐI ƯU HÓA BẦY ĐÀN

HỘI ĐỒNG HỆ THỐNG THÔNG TIN

GVHD: TS. LÊ HỒNG TRANG
GVPB: THS. ĐỖ THANH THÁI

SVTH: THÁI TIỂU PHƯƠNG 1710250
NGUYỄN BÌNH YÊN 1714074

TP. Hồ Chí Minh, 12/2020

LỜI CAM ĐOAN

Chúng tôi xin cam đoan rằng, ngoại trừ các kết quả tham khảo từ các công trình khác có liên quan đã ghi rõ trong đề cương luận văn, các nội dung trình bày trong đề cương này là do chính chúng tôi thực hiện và chưa có phần nội dung nào được nộp để lấy bằng cấp ở một trường khác.

TP. HCM, NGÀY 28 THÁNG 12 NĂM 2020.

LỜI CẢM ƠN

Đầu tiên, chúng tôi xin gửi lời cảm ơn chân thành nhất đến TS. Lê Hồng Trang, người thầy đã giúp đỡ chúng tôi trong quá trình chuẩn bị đề cương luận văn suốt học kỳ vừa qua. Xin cảm ơn sự hướng dẫn tận tình của thầy, chính nhờ những kiến thức và kinh nghiệm mà thầy truyền đạt đã giúp chúng tôi có được cái nhìn tổng quan về đề tài và định hướng đúng đắn cũng như xác định kế hoạch thực hiện trong giai đoạn làm luận văn tiếp theo.

Xin gửi lời cảm ơn đến các thầy cô trường Đại học Bách Khoa, đặc biệt là các thầy cô bộ môn trong khoa Khoa học và Kỹ thuật Máy tính đã truyền đạt những kiến thức quý báu trong bốn năm học qua.

Cuối cùng, chúng tôi xin cảm ơn gia đình, bạn bè, những người đã giúp đỡ, hỗ trợ chúng tôi hết mình trong suốt thời gian hoàn thành chương trình bậc Đại học.

TÁC GIẢ.

TÓM TẮT

Phân loại tế bào thông qua xử lý hình ảnh gần đây đã thu hút được sự quan tâm dựa trên quan điểm xây dựng các công cụ chẩn đoán với sự hỗ trợ của máy tính đối với các bệnh rối loạn máu như bệnh bạch cầu. Để đi đến kết luận chẩn đoán bệnh và mức độ tiến triển, việc xác định tế bào ác tính với độ chính xác cao là rất quan trọng. Thách thức được đặt ra là phải chẩn đoán bệnh sớm để tăng khả năng chữa khỏi bệnh của các đối tượng mắc bệnh ung thư. Mặc dù các phương pháp tiên tiến như kỹ thuật đếm tế bào dòng chảy là sẵn có, chúng rất đắt tiền và không được cung cấp rộng rãi trong các phòng thí nghiệm giải phẫu hoặc bệnh viện, đặc biệt là ở những vùng nông thôn. Mặt khác, một giải pháp dựa trên máy tính có thể được triển khai dễ dàng với chi phí thấp hơn nhiều. Do đó, đề tài này là một nỗ lực nhằm xây dựng một bộ phân loại tự động khắc phục các vấn đề liên quan đến việc triển khai máy móc phức tạp, hỗ trợ các nhà nghiên cứu bệnh lý học và ung thư học đưa ra các suy luận nhanh hơn dựa trên dữ liệu.

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

MỤC LỤC

1

TỔNG QUAN

Trong chương này, chúng tôi xin giới thiệu sơ lược về nội dung đề tài và cấu trúc đề cương.

Mục lục

1.1 Giới thiệu đề tài

Tế bào bạch cầu (White Blood Cell - WBC) Leukaemia được sản sinh quá mức trong tủy xương, việc phát hiện các tế bào ác tính dựa trên hình ảnh hiển vi là rất quan trọng. Các công cụ hỗ trợ của máy tính có thể rất hữu ích trong việc tự động hóa quá trình phân đoạn và nhận dạng tế bào. Việc xác định tế bào ác tính so với tế bào bình thường từ hình ảnh hiển vi là khó khăn vì về mặt hình thái, cả hai loại tế bào đều có vẻ giống nhau.

Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) là một trong những phương pháp tiên tiến nhất hiện nay dùng để phân loại hình ảnh, nhưng chi phí tính toán cho việc huấn luyện và hiện thực là khá cao. Do đó, đề cương này sẽ đề suất một phương pháp lai cải tiến để phân lớp hình ảnh bạch cầu một cách hiệu quả. Trích xuất đặc trưng sâu bằng một kiến trúc CNN hiện đại được huấn luyện trước trên tập dữ liệu ImageNet và cải tiến một giải thuật tối ưu hóa bầy đàn để lựa chọn đặc trưng là cách tiếp cận mà đề cương đề xuất để huấn luyện một mô hình phân lớp đạt được độ chính xác cao và giảm độ phức tạp tính toán.

1.2 Mục tiêu và phạm vi đề tài

Mục tiêu của đề tài này bao gồm:

- Trích xuất đặc trưng sâu từ những bộ dữ liệu hình ảnh hiển vi của tế bào bạch cầu.
- Áp dụng một thuật toán tối ưu hóa bầy đàn để lựa chọn đặc trưng.
- Đề xuất một phương pháp cải thiện thuật toán tối ưu hóa để cải thiện hiệu năng cho bài toán phân lớp, tập trung cải thiện về thời gian tính toán.
- Hiện thực mô hình, kết hợp với một giải thuật để phân lớp hình ảnh tế bào.
- So sánh, đánh giá kết quả đạt được với các phương pháp khác.

1.3 Cấu trúc đề cương luận văn

Chương 1. Tổng quan. Trong chương này, chúng tôi xin giới thiệu sơ lược về nội dung đề tài và cấu trúc đề cương

Chương 2. Phân lớp hình ảnh tế bào bạch cầu. Trong chương này, chúng tôi sẽ trình bày bài toán phân lớp hình ảnh tế bào bạch cầu, một số phương pháp tiếp cận đã có cho bài toán này hiện nay và những thách thức phổ biến của chúng.

Chương 3. Tối ưu hóa bầy đàn. Trong chương này, chúng tôi sẽ trình bày tổng quan về tối ưu hóa bầy đàn, một số giải thuật nổi bật sử dụng trí thông minh bầy đàn và những kết quả ứng dụng từ các công trình có liên quan, cụ thể là tối ưu hóa trong bài toán phân lớp.

Chương 4. Phân loại hình ảnh tế bào bạch cầu dùng tối ưu hóa bầy đàn với các đặc trưng sâu. Chương này sẽ đưa ra khái niệm của đặc trưng sâu, từ đó làm rõ phương pháp xây dựng mô hình phân loại hình ảnh tế bào bạch cầu được chúng tôi đề xuất trong luận văn, bao gồm hai bước: trích xuất đặc trưng sâu bằng một CNN hiện đại và lựa chọn đặc trưng sử dụng một cải tiến của giải thuật Salp Swarm Algorithm (SSA).

Chương 5. Hiện thực và những kết quả ban đầu. Trong chương này, chúng tôi xin trình bày mô tả các tập dữ liệu sẽ sử dụng, các tiêu chuẩn đánh giá và một số kết quả từ mô hình ban đầu.

Chương 6. Tổng kết. Trong chương này, chúng tôi xin trình bày các công việc đã thực hiện trong quá trình làm đề cương luận văn và các công việc sẽ thực hiện khi làm luận văn.

2

PHÂN LỚP HÌNH ẢNH TẾ BÀO BẠCH CẦU

Trong chương này, chúng tôi sẽ trình bày bài toán phân lớp hình ảnh tế bào bạch cầu, một số phương pháp tiếp cận đã có cho bài toán này hiện nay và những thách thức phổ biến của chúng.

Mục lục

2.1 Phân lớp hình ảnh tế bào bạch cầu

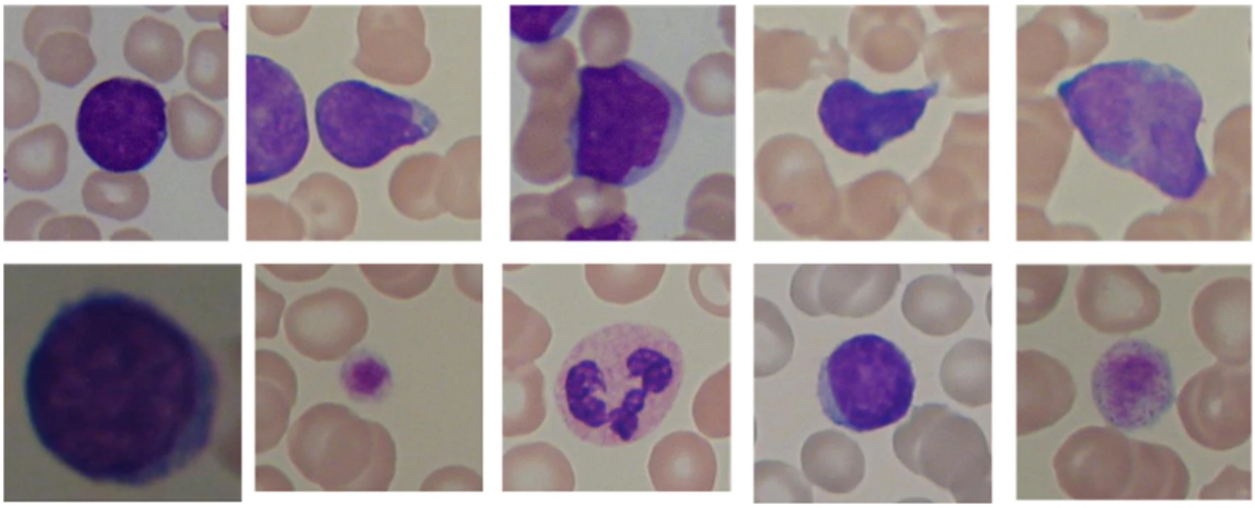
Máu chủ yếu chứa ba loại tế bào: hồng cầu, tiểu cầu và bạch cầu. Tế bào hồng cầu đóng vai trò quan trọng trong việc vận chuyển oxy từ tim đến các mô và mang đi carbon dioxide. Chúng chiếm tới 50% thể tích máu. Tế bào bạch cầu (White Blood Cells) có chức năng quan trọng đối với hệ miễn dịch, vì chúng bảo vệ cơ thể chống lại nhiễm trùng và bệnh tật. Do đó việc phân loại tế bào bạch cầu là rất quan trọng. Hàng triệu người bị ảnh hưởng bởi bệnh bạch cầu (Leukemia), được xem như là một khối u ác tính. Nó hình thành từ tủy xương và sau đó phân phối đến tế bào máu trong toàn bộ cơ thể. Thông thường, bạch cầu phát triển dựa trên nhu cầu cơ thể, nhưng trong trường hợp bệnh bạch cầu, chúng được sản sinh một cách bất thường và kém hiệu quả. Mặc dù có thể được phát hiện bởi vẻ ngoài giống như màu tím sẫm, việc phân tích và xử lý thêm rất phức tạp do những thay đổi về hình dạng và kết cấu. Hơn nữa, tế bào bạch cầu được bao quanh bởi các thành phần máu khác như hồng cầu và tiểu cầu nên việc phân biệt bằng hình dạng và kích thước trở nên không có hiệu quả.

Phân loại tế bào thông qua xử lý hình ảnh gần đây đã thu hút được sự quan tâm dựa trên quan điểm xây dựng các công cụ chẩn đoán với sự hỗ trợ của máy tính đối với các bệnh rối loạn máu như bệnh bạch cầu. Để đi đến kết luận chẩn đoán bệnh và mức độ tiến triển, việc xác định tế bào ác tính với độ chính xác cao là rất quan trọng. Các công cụ hỗ trợ của máy tính có thể rất hữu ích trong việc tự động hóa quá trình phân đoạn và nhận dạng tế bào. Việc xác định tế bào ác tính so với tế bào bình thường từ hình ảnh hiển vi là khó khăn vì về mặt hình thái, cả hai loại tế bào đều có vẻ giống nhau.

Điều quan trọng là phải chẩn đoán bệnh sớm để tăng khả năng chữa khỏi bệnh của các đối tượng mắc bệnh ung thư. Mặc dù các phương pháp tiên tiến như kỹ thuật đếm tế bào dòng chảy là sẵn có, chúng rất đắt tiền và không được cung cấp rộng rãi trong các phòng thí nghiệm giải phẫu hoặc bệnh viện, đặc biệt là ở những vùng nông thôn. Mặt khác, một giải pháp dựa trên máy tính có thể được triển khai dễ dàng với chi phí thấp hơn nhiều.

Do đó, đây là một nỗ lực nhằm xây dựng một bộ phân loại tự động khắc phục các vấn đề liên quan đến việc triển khai máy móc phức tạp, hỗ trợ các nhà nghiên cứu bệnh lý học và ung thư học đưa ra các suy luận nhanh hơn dựa trên dữ liệu.

2.2 Một số tiếp cận phân lớp hình ảnh tế bào bạch cầu



Hình 2.1: Một số mẫu hình ảnh tế bào bạch cầu lành tính (trên) và ác tính (dưới) trích từ tập dữ liệu ALL-IDB2

Tế bào bạch cầu lympho (Lymphocytes) có hình dạng bình thường, hạt nhân của chúng có các cạnh trơn nhẵn, ngược lại tế bào lympho của bệnh nhân tăng lympho bào cấp tính (Acute Lymphoblastic Leukaemia - ALL) có ít cạnh bao bình thường hơn và xuất hiện các khoang nhỏ trong tế bào chất, được gọi là không bào, và các hạt tròn trong nhân của chúng, được gọi là nucleoli. Phân lớp ALL là một bài toán phân lớp mẫu dựa trên các đặc trưng trích xuất từ hình ảnh hiển vi của phết máu. Phân lớp là một kỹ thuật học có giám sát sử dụng dữ liệu huấn luyện để huấn luyện mô hình và kiểm tra trên tập dữ liệu test để tính toán hiệu suất của mô hình. Một số bộ phân lớp được sử dụng để chẩn đoán ALL:

2.2.1 SVM (Máy vectơ hỗ trợ)

Máy vector hỗ trợ là một trong những thuật toán được sử dụng rộng rãi để phát hiện bệnh bạch cầu. Thuật toán này được sử dụng để tối ưu hóa siêu phẳng phân loại dữ liệu đã cho dựa trên các đặc trưng của chúng. Lý do chính đằng sau việc lựa chọn SVM để phát hiện bệnh bạch cầu là bởi vì SVM là một bộ phân loại nhị phân có thể phân loại hiệu quả giữa tế bào bình thường và tế bào blast. [1]

2.2.2 K-Nearest Neighbor (k lân cận gần nhất)

KNN là một kỹ thuật phân loại và hồi quy được sử dụng rộng rãi, sử dụng phương pháp học tập lười và không tham số để phân loại dữ liệu khác nhau. Thuật toán láng giềng gần nhất, phân loại được thực hiện bằng cách bỏ phiếu từ các láng giềng gần nhất. Dựa trên biểu quyết này, các đối tượng sẽ được gán cho lớp có liên quan của chúng. Với phân loại tế bào bệnh bạch cầu nguyên bào lympho cấp tính, bộ phân loại k-NN được sử dụng có kết quả phân loại tốt hơn cho tế bào bình thường và tế bào blast. [2]

2.2.3 Random Forest (Rừng ngẫu nhiên)

Rừng ngẫu nhiên là một kỹ thuật phân loại hiệu quả sử dụng phương pháp học tập hợp để phân loại đối tượng từ một vectơ đầu vào. RF bao gồm các tập hợp cây khác nhau thực hiện việc bỏ phiếu để lựa chọn lớp và do đó việc lựa chọn sẽ dựa trên lớp có số phiếu tối đa. Mặc dù rừng ngẫu nhiên phụ thuộc vào các tham số ngẫu nhiên do đó nó có thể cho các kết quả khác nhau, nhưng bộ phân loại này vẫn được ưa thích vì nó khắc phục được vấn đề overfitting. Để phát hiện bệnh bạch cầu nguyên bào lympho cấp tính, rừng ngẫu nhiên được coi là bộ phân loại mạnh hơn để phát hiện các tế bào máu bình thường và bị nhiễm bệnh. [3]

2.2.4 ANFIS (Adaptive Neuro Fuzzy Inference System)

Hệ thống suy luận mờ thần kinh thích ứng là một hệ thống phân loại tốt khác được sử dụng rộng rãi để phân loại hình ảnh y tế. Đây là một phương pháp rất mạnh vì nó là sự kết hợp của mạng nơron nhân tạo và logic mờ. Mặc dù độ phức tạp cao hơn các bộ phân loại khác như rừng ngẫu nhiên, ANFIS là bộ phân loại end-to-end so với rừng ngẫu nhiên thực hiện phân loại bằng cách sử dụng các đặc trưng nhất định, vì vậy nếu các đặc trưng không đủ mạnh thì rừng ngẫu nhiên không thể phân loại chính xác. Do đó ANFIS giúp phân loại tốt hơn và chẩn đoán sớm nhiều bệnh. Nghiên cứu gần đây cho thấy ANFIS đã được sử dụng để phân loại tế bào blast ở bệnh nhân ung thư máu bằng phân tích hình ảnh máu hiển vi và cho các kết quả thích hợp. [4]

2.2.5 Naive Bayes

Naive Bayes là một bộ phân loại hiệu quả được sử dụng để phân loại dữ liệu bằng cách áp dụng định lý Bayes. Trong bộ phân loại này, giá trị của các đặc trưng được giả định là độc lập với nhau. Để phát hiện bệnh bạch cầu nguyên bào lympho cấp tính, Naive Bayes đã được sử dụng trong những nghiên cứu gần đây, và cho kết quả phân loại hiệu quả tế bào bình thường và tế bào blast từ hình ảnh hiển vi tế bào máu. [5]

2.2.6 Multilayer Perceptron (MLP)

Multilayer perceptron là một mô hình mạng nơ-ron nhân tạo chứa nhiều lớp nút và mỗi lớp được kết nối với nút tiếp theo của nó. Trong MLP, mỗi nút đầu vào đại diện cho dữ liệu đầu vào và tất cả các nút khác là nơron cung cấp đầu ra bằng cách sử dụng hàm kích hoạt (activation function). MLP được sử dụng rộng rãi trong các bài toán quy mô lớn vì thiết kế đơn giản và khả năng tính toán mạnh mẽ. Nghiên cứu gần đây cho thấy MLP đã được sử dụng để phát hiện bệnh bạch cầu nguyên bào lympho cấp tính, trong đó nó sử dụng scaled conjugate gradient (SCG) để huấn luyện hình ảnh tế bào máu và trên cơ sở mô hình được huấn luyện này thực hiện phân loại các tế bào bình thường và ung thư. [6]

2.2.7 Mạng nơron xác suất (PNN)

Trong bộ phân loại PNN, cửa sổ Parzen và một hàm không tham số được sử dụng để ước lượng hàm phân phối xác suất của mỗi lớp, qua đó xác suất dữ liệu đầu vào có thể được tính bằng cách áp dụng quy tắc Bayes, lớp có xác suất cao nhất sẽ được gán cho dữ liệu đầu vào. Để phát hiện bệnh bạch cầu nguyên bào lympho cấp tính, mạng nơron xác suất đã được sử dụng để phân loại hiệu quả các tế bào bình thường và tế bào blast.

2.3 Những thách thức phổ biến của các phương pháp hiện nay

Học sâu sử dụng Convolution Neural Networks (CNN) hiện là lựa chọn tốt nhất trong các ứng dụng hình ảnh y tế như phát hiện và phân loại. Trong khi CNN đạt được kết quả tốt nhất trên các tập dữ liệu lớn, chúng đòi hỏi rất nhiều dữ liệu và tài nguyên tính toán để huấn luyện. Trong nhiều trường hợp, tập dữ liệu bị hạn chế và có thể không đủ để huấn luyện CNN từ đầu. Để tận dụng sức mạnh của CNN và đồng thời giảm chi phí tính toán, có thể sử dụng phương pháp học chuyển tiếp (transfer learning). Một số mạng thần kinh huấn luyện trước đã giành chiến thắng trong các cuộc thi quốc tế như VGGNet, Resnet, Nasnet, Mobilenet, Inception và Xception. Trong một đánh giá về các kiến trúc CNN khác nhau được thực hiện, học chuyển tiếp đã đạt được hiệu suất cao nhất về phân loại hạch bạch huyết ngực-bụng (LN) cũng như phân loại bệnh phổi kẽ (ILD). Các tác giả [7] đã sử dụng phân loại gộp trung bình để phân biệt tế bào ác tính với tế bào lành tính sau khi trích xuất các đặc trưng từ hình ảnh ung thư vú bằng cách sử dụng kiến trúc CNN huấn luyện trước đưa vào một lớp phân loại được kết nối đầy đủ. Kết quả thử nghiệm cho thấy độ chính xác của mô hình vượt trội hơn tất cả các phương pháp tiếp cận khác của CNN trong việc phát hiện và phân loại các khối u vú dựa trên hình ảnh tế bào học. Học chuyển tiếp đã khắc phục những hạn chế của các mô hình được công bố trước đây để phát hiện ung thư vú trong hình ảnh tế bào học.

Những cách tiếp cận này có điểm chung là chúng sử dụng một số lượng lớn các đặc trưng (lên đến 100.000) từ các mô hình CNN được huấn luyện. Điều này không hiệu quả về thời gian và tài nguyên tính toán vì nhiều đặc trưng trong số này là dư thừa và không có giá trị. Hơn nữa, độ chính xác của trình phân loại có thể được cải thiện từ việc giới hạn số lượng các đặc trưng.

3

TỐI ƯU HÓA BẦY ĐÀN

Trong chương này, chúng tôi sẽ trình bày tổng quan về tối ưu hóa bầy đàn, một số giải thuật nổi bật sử dụng trí thông minh bầy đàn và những kết quả ứng dụng từ các công trình có liên quan, cụ thể là tối ưu hóa trong bài toán phân lớp.

Mục lục

3.1 Tổng quan về tối ưu hóa bầy đàn

Tập tính thông minh bầy đàn của những nhóm côn trùng và động vật trong tự nhiên như đàn chim, kiến, cá, bầy ong,... đã thu hút sự chú ý của các nhà nghiên cứu. Hành vi của những loài côn trùng và động vật này được gọi là hành vi bầy đàn. Các nhà côn trùng học đã nghiên cứu hiện tượng tập thể này để mô hình hóa bầy đàn sinh học trong khi các kỹ sư ứng dụng những mô hình này như một framework để giải quyết những bài toàn phức tạp trong thế giới thực. Nhánh trí tuệ nhân tạo này xử lý hành vi tập thể của bầy đàn thông qua sự tương tác phức tạp giữa các cá thể mà không có sự giám sát, thường được gọi là trí tuệ bầy đàn. Bonabeau đã định nghĩa trí tuệ bầy đàn là "bất kỳ nỗ lực nào để thiết kế thuật toán hay phân phối các phương pháp giải quyết vấn đề đều lấy cảm hứng từ hành vi tập thể của cộng đồng côn trùng đất và những loài động vật khác. Trí thông minh bầy đàn có một số ưu điểm như khả năng mở rộng, bỏ qua lỗi, thích ứng, tốc độ, tính mô đun, tính tự trị và tính song song.

Các thành phần quan trọng của trí thông minh bầy đàn là khả năng tự tổ chức và phân công lao động. Trong một hệ thống tự tổ chức, mỗi đơn vị phản ứng với các kích thích cục bộ một cách độc lập và có thể hành động cùng nhau để hoàn thành một nhiệm vụ toàn cục, thông qua sự phân công lao động để tránh sự giám sát tập trung. Do đó, toàn bộ hệ thống có thể thích ứng hiệu quả với những thay đổi trong và ngoài quần thể.

Một số thuật toán bầy đàn đã được phát triển bằng việc kết hợp các quy tắc xác định và tính ngẫu nhiên, bắt chước theo hành vi của nhóm côn trùng và động vậy trong tự nhiên. Đặc biệt, những đàn côn trùng và các nhóm động vật cung cấp một tập hợp ẩn dụ phong phú để thiết kế các giải thuật tối ưu hóa bầy đàn. Chính những thực thể hợp tác này là những hệ thống phức tập được cấu tạo từ các cá thể với những nhiệm vụ hợp tác khác nhau mà trong đó, mỗi thành viên có xu hướng tái tạo những hành vi chuyên biệt tùy thuộc vào giới tính. Những phương pháp dựa trên hành vi xã hội của đàn chim, cá như giải thuật Particle Swarm Optimization (PSO), dựa trên hành vi của đàn kiến trong quá trình tìm kiếm thức ăn như Ant Colony Optimization (ACO) hay dựa trên những chuyển động quần thể để tìm nguồn thức ăn của loài động vật biển Salp Swarm Algorithm (SSA) đã chứng minh sức mạnh trong các vấn đề liên quan đến sinh học. Những thuật toán này hiện được chứng minh là đã cung cấp các kết quả phù hợp trong rất nhiều ứng dụng thực tế.

Mỗi cá thể đều nắm giữ một kết quả ứng viên. Quá trình tìm kiếm bắt đầu bằng việc khởi tạo ngẫu nhiên quần thể phù hợp với bài toán. Tiếp theo là đánh giá chất lượng kết quả ứng viên được đề xuất bởi từng cá thể. Bước thứ ba là tạo ra tập con mới từ một giải thuật đến bất kỳ giải thuật nào khác. Việc tạo ra những tập ứng viên mới dựa trên nguồn cảm hứng của giải thuật. Sự phát triển của các thuật toán bắt đầu từ PSO, sau đó là ACO, SSA và những thuật toán gần đây. PSO, ACO và SSA được lựa chọn rộng rãi cho các thuật toán tìm kiếm meta-heuristic được ứng dụng cho bài toán phân lớp với số chiều lớn.

3.2 Một số phương pháp nổi bật trong tối ưu hóa bầy đàn

3.2.1 Particle Swarm Optimization (PSO)

PSO là một kỹ thuật nổi bật trong số các kỹ thuật trí thông minh bầy đàn, mạnh mẽ và thống trị nhất, được Kennedy và Eberhart giới thiệu vào năm 1995 [8], được lấy cảm hứng từ các hành vi xã hội được tìm thấy trong một đàn chim di cư hoặc bầy cá để giải quyết các vấn đề tối ưu hóa. Trong PSO, một bầy chứa nhiều thực thể được gọi là các hạt liên kết với nhau giữa các nhóm để tìm ra giải pháp tốt nhất trong khi di chuyển trong không gian tìm kiếm rộng lớn. Mỗi hạt chứa một vị trí, một lời giải thích hợp của bài toán và một vận tốc, thường là một vectơ thứ n của các giá trị số. Ngoài ra, vận tốc có cấu trúc không thể phân biệt được với vị trí, biểu thị tốc độ và hướng mà hạt chuyển động trong lần lặp tiếp theo. Trong mỗi lần lặp lại, vận tốc của một hạt chủ yếu được cập nhật dựa trên kinh nghiệm của chính chúng (tốt nhất địa phương, pbest) và kinh nghiệm của những hạt khác xung quanh chúng (tốt nhất toàn cầu, gbest). Phương trình (1) và (2) được sử dụng để cập nhật vận tốc và vị trí của mọi hạt:

$$(1) : v_{id}(t+1) = \omega \times v_{id}(t) + r_1 c_1 (p_{id}(t) - x_{id}(t)) + r_2 c_2 (p_{gd}(t) - x_{id}(t))$$

$$(2) : x_{id}(t+1) = x_{id}(t) + v_{id}(t+1)$$

trong đó $v_{id}(t)$ và $x_{id}(t)$ lần lượt là vận tốc và vị trí của hạt i theo chiều d tại thời điểm t. p_{id} và p_{gd} là các vị trí tốt nhất cục bộ và tốt nhất toàn cục trong chiều thứ d. c_1 và c_2 là hai hệ số dương được đặt tên là hệ số học, và r_1 và r_2 là hai hàm tùy ý phân bố đều trong [0, 1]. Toán tử ω là quán tính được sử dụng như một cải tiến để kiểm soát tác động của lịch sử vận tốc trước đây lên vận tốc hiện tại và cũng đóng vai trò cân bằng giữa tìm kiếm toàn cục và tìm kiếm cục bộ. Vận tốc tối đa được xác định trước, v_{max} trong khoảng [- v_{max}, v_{max}] thường giới hạn các giá trị vận tốc.

Một đặc điểm thú vị của PSO là nó không sử dụng gradient của hàm, do đó, các hàm mục tiêu không cần phải độc lập. Hơn nữa, PSO cơ bản đơn giản một cách đáng kinh ngạc. Thêm các biến thể vào hiện thực gốc ban đầu có thể giúp nó thích ứng với các vấn đề phức tạp hơn. PSO đã được ứng dụng thành công trong nhiều lĩnh vực khác ngoài thông tin sinh học bao gồm các ứng dụng công nghiệp và hệ thống điện.

3.2.2 Ant Colony Optimization (ACO)

ACO được giới thiệu bởi Dorigo và các đồng nghiệp của ông để tìm đường đi ngắn nhất giữa tổ kiến và nguồn thức ăn. Điều này được thực hiện bằng cách sử dụng các đường mòn pheromone, mà kiến lưu trữ tại bất kỳ điểm nào chúng di chuyển, như một hình thức giao tiếp gián tiếp. ACO cũng được tạo ra để giải quyết các vấn đề tối ưu hóa tổ hợp rời rạc như vấn đề người thương gia du hành - travelling salesman problem (TSP). Việc xem xét quá nhiều sự phức tạp của thông tin sinh học như sắp xếp trình tự, lập bản đồ gen và lựa chọn đặc trưng của dữ liệu biểu hiện gen rất giống với TSP. ACO là giải pháp hợp lý cho các vấn đề tối ưu hóa thông tin sinh học có quy mô lớn.

Ý tưởng cơ bản là giải quyết bài toán tìm kiếm đường đi lý tưởng trong một đồ thị có trọng số sử dụng các đặc tính của loài kiến. Điểm mấu chốt trong việc cải tiến ACO là quyết định đặc

trưng phù hợp hoàn toàn dựa trên những thành phần đồ thị của bài toán sẽ được sử dụng với đường dẫn pheromone trạng thái bất thường và quyết định cách mà loài kiến sẽ lạm dụng các thành phần tiềm năng này khi xây dựng các giải pháp mới.

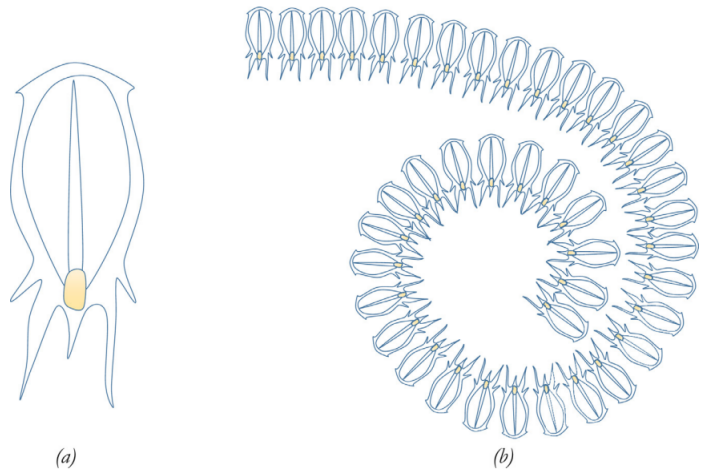
Một số điểm yếu của các tiếp cận ACO hiện tại:

1) Sự phức tạp về không gian của việc tập hợp pheromone trong bộ nhớ lõi là rất bất thường và giải pháp nổi bật nhất hiện nay để giải quyết vấn đề này là sử dụng các giá trị pheromone ứng viên không thể phục hồi quá mức trong nhiều điều kiện.

2) Các biến thể gần đây của ACO sử dụng các thủ tục tìm kiếm cục bộ trong thuật toán tiêu chuẩn của họ; tuy nhiên họ chưa sẵn sàng sử dụng thông tin pheromone để hoạt động hiệu quả hơn.

3) Sự phức tạp về thời gian của việc chọn nước đi tiếp theo là rất lớn.

3.2.3 Salp Swarm Algorithm (SSA)



Hình 3.1: (a) Cá thể Salp; (b) Chuỗi Salp

SSA là một phương pháp tối ưu hóa mô phỏng hành vi kiếm ăn của Salpidae, một loài động vật biển không xương sống phù du. Salp di chuyển và kiếm ăn theo một hành vi được gọi là chuỗi salp, một ví dụ về hành vi bầy đàn. SSA bắt đầu bằng cách chia dân số thành hai phần: những salp đi trước, gọi là những cá thể dẫn đầu (leader) và phần còn lại, những cá thể theo sau (follower). Những cá thể này thay đổi vị trí của chúng để tìm kiếm nguồn thức ăn. Để biểu diễn chuyển động này, phương trình 1 được sử dụng để cập nhật vị trí của những leader:

$$(1): x_j^1 = \begin{cases} F_j + c_1((ub_j - lb_j) \times c_2 + lb_j) & c_3 \leq 0 \\ F_j - c_1((ub_j - lb_j) \times c_2 + lb_j) & c_3 > 0 \end{cases}$$

trong đó x_j^1 biểu thị vị trí của leader theo chiều j. F_j là mục tiêu (nguồn thức ăn) trong chiều thứ j.ub_j và lb_j lần lượt là cận trên và cận dưới.c_2 và c_3 là các số ngẫu nhiên trong khoảng $[0,1]$. Thông số c_1 được sử dụng để cân bằng giữa các bước exploration và exploitation, xuất phát từ công thức sau:

$(2) : c_1 = 2e^{-(\frac{4t}{t_{max}})^2}$

trong đó lần lặp hiện tại là t của t_{max}. Đầu tiên, những leader được cập nhật vị trí, sau đó vị trí của các follower được cập nhật theo phương trình sau:

$(3) : x_j^i = \frac{1}{2}(x_j^i + x_j^{i-1})$

trong đó x_j^i biểu thị vị trí của follower thứ i với i > 1. Mã giả của thuật toán SSA:

1: Initialize a population X.
2: **repeat**
3: Compute the objective function for each solution x_i.
4: Update the best salp (solution) ($F = X^b$).
5: Update c_1 using Eq. (2).
6: **for** $i = 1 : N$ **do**
7: **if** $i == 1$ **then**
8: Update the position of salp using Eq. (1)
9: **else**
10: Update the position of salp using Eq. (3)
11: **end if**
12: **end for**
13: **until** ($t < t_{max}$)
14: Return the best solution F.

Hình 3.2: Mã giả thuật toán Salp Swarm Algorithm

3.3 Một số kết quả ứng dụng

Ta sẽ nghiên cứu các công trình có liên quan trong lĩnh vực lựa chọn đặc trưng sử dụng trí thông minh bầy đàn để phân lớp trên những dữ liệu có số chiều lớn. Những công trình dưới đây áp dụng các thuật toán tìm kiếm bầy đàn đã được trình bày ở phần trước.

3.3.1 PSO trong bài toán phân lớp

Những nghiên cứu hiện tại đã chỉ ra rằng PSO là một phương pháp có khả năng lựa chọn đặc trưng. Tuy nhiên, chúng thường là tối ưu cục bộ, nhất là đối với các bài toán lựa chọn gen với không gian tìm kiếm lớn. Vì vậy, nhiều cách tiếp cận đã được đề xuất để giải quyết vấn đề này.

Trong [9], khoảng cách hamming được áp dụng để so sách với phép đo khoảng cách Euclid thông thường. Hai đối tượng khác nhau có thể trông giống nhau trong một không gian đặc trưng rộng lớn; do đó, công trình này khẳng định rằng khoảng cách Euclid có thể không phù hợp với dữ liệu nhiều chiều. Tập con những đặc trưng quan trọng được chọn lọc bằng cập nhật vận tốc trong PSO nhị phân (HDBPSO). Vận tốc được cập nhật bởi khoảng cách Hamming đã được điều chỉnh như là một thước đo độ gần. Kết quả được đánh giá dựa trên các chỉ số hợp lệ và độ chính xác của phân lớp. Dữ liệu nhị phân được sử dụng trong phương pháp đề xuất để so sánh với GA và GA đa mục tiêu. Kết quả cho thấy HDBPSO thể hiện xuất sắc so với những giải thuật khác.

Yiyuan Chen và các đồng nghiệp của ông [10] đã đề xuất một kỹ thuật lựa chọn đặc trưng dựa trên độ tin cậy và hiệu quả về chi phí (CCFS) sử dụng PSO nhị phân. Trong CCFS, hai điểm mới được đề cập: độ tin cậy đặc trưng và chi phí đặc trưng. Trong hai ý tưởng này, độ tin cậy của đặc trưng được sử dụng để cập nhật vị trí của hạt và nâng cao hiêu suất của lựa chọn đặc trưng. Chi phí đặc trưng được tích hợp vào công thức của hàm mục tiêu. Kết quả thí nghiệm trên tập dữ liệu ung thư phổi được so sánh với ba phương pháp lọc khác nhau như PCFS, InfoGain, CFS, và các phương pháp đóng gói như Best First, GSFS và GSBS. Các kết quả thí nghiệm chứng minh phương pháp CCFS đề xuất đã giúp cải thiện độ chính xác của việc học và giảm số đặc trưng được chọn cũng như chi phí của chúng.

Binh et. al [11] phát triển một giải thuật lựa chọn đặc trưng lai trong một quá trình tiến hóa đơn để đạt được những tập con đặc trưng nhỏ hơn với độ chính xác tốt trong thời gian ngắn hơn. Trong cách tiếp cận này, PSO dựa trên tìm kiếm heuristic cục bộ cho rằng độ đo sự không đảm bảo đối xứng được đề xuất sẽ cải thiện giải pháp và một hàm mục tiêu lai mới được áp dụng để tính toán độ tốt của những đặc trưng được chọn. Những kết quả thí nghiệm trên tám tập dữ liệu có số chiều lớn đã chỉ ra rằng các tập con có số đặc trưng ít hơn đạt được độ chính xác phân loại tốt hơn đáng kể so với tập đặc trưng ban đầu. Thuật toán lựa chọn đặc trưng lai này đã cho ra nhiều đặc trưng liên quan hơn trong thời gian ngắn hơn với chi phí tính toán giảm khi so sánh với các phương pháp lựa chọn đặc trưng dựa trên PSO khác.

Threshold Controlled Binary Particle Swarm Optimization (TC-BPSO) cùng với Multi-Class Support Vector Machine (MC-SVM) được phát triển trong [12] để tận dụng tối đa độ chính xác phân loại với chi phí thấp nhất các feature được chọn có thể có. TC-BPSO được sử dụng để lựa chọn đặc trưng trong khi MC-SVM được dùng để tính độ chính xác phân lớp. Điểm mạnh của

thuật toán này là loại bỏ những đặc trưng không liên quan hoặc bị lặp lại để giảm kích thước của đặc trưng. Kết quả từ các thí nghiệm trên mười tập dữ liệu khác nhau cho thấy phương pháp này đã vượt qua tất cả về lựa chọn đặc trưng, độ chính xác và độ phức tạp tính toán.

3.3.2 ACO trong bài toán phân lớp

Các thuật toán con kiến rất dễ hiện thực và có rất nhiều ứng dụng, nhưng hiệu suất của chúng giảm đáng kể khi giải quyết các bài toán có quy mô lớn. Thuật toán được đề xuất trong [13] là một phương pháp kết hợp tìm kiếm cục bộ dựa trên ACO với độ đo không đảm bảo đối xứng để tìm ra tập hợp con của các tính năng tối ưu. Điểm mạnh của thuật toán lựa chọn đặc trưng lai này là nó đánh giá các đặc trưng tốt nhất. Kết quả thử nghiệm trên bộ dữ liệu biểu hiện gen lớn cho thấy phương pháp ACO-LS hoạt động tốt hơn khi so sánh với các phương pháp hiện có; những kết quả này được đo lường về mặt độ chính xác dự đoán, kích thước đặc trưng nhỏ hơn và tính hiệu quả. Ngoài ra, cách tiếp cận này cũng xem xét sự cân bằng giữa tìm kiếm địa phương và di truyền theo hướng cải thiện chất lượng và hiệu quả tìm kiếm của ACO-LS.

Thuật toán bầy ong lai (ABA) kết hợp hai giải thuật Ant Colony Optimization (ACO) và Artificial Bee Colony (ABC) trong hệ thống chuyên gia mờ để mã hóa các biến giải pháp bằng cách sử dụng dạng sửa đổi của biểu diễn đã được giới thiệu trong [14]. Bộ quy tắc tối ưu của tối ưu hóa tổ hợp được hình thành bởi hiện thực ACO trong công trình này. Sự biểu diễn của hàm thành viên như là một số liên tục được thực hiện bởi ABC. Phương pháp lai này hiện thực trên một số bộ dữ liệu biểu hiện gen bao gồm Receiver Operating Characteristic (ROC). Để so sánh hiệu suất của cách tiếp cận được đề xuất với các thuật toán khác, giá trị của diện tích dưới đường cong ROC được sử dụng. Các kết quả thu được các thí nghiệm được báo cáo là có giá trị tốt nhất khi so sánh với BCGA, RCGA, PSO và GA.

Amirreza Rouhi và các cộng sự của ông [15] đã phát triển một phương pháp lai dựa trên thuật toán đàn kiến nhị phân (BACO) để giảm kích thước đặc trưng bằng cách kết hợp một số phương pháp lọc và sau đó Advanced BACO meta-heuristic sẽ được áp dụng cho các tập hợp đặc trưng đã rút gọn để chọn tập hợp con tính năng tốt nhất. Năm tập dữ liệu microarray có số chiều lớn nổi tiếng đã được triển khai để đo lường hiệu suất của phương pháp. Tỷ lệ lỗi phân loại và số lượng các tính năng được chọn làm thước đo đánh giá và so sánh hiệu suất của các thuật toán đã thử nghiệm. Kết quả thu được bởi bộ phân loại Naïve Bayes đã chứng minh tính hiệu quả của phương pháp đề xuất đối với dữ liệu microarray có số chiều lớn.

4

PHÂN LOẠI HÌNH ẢNH TẾ BÀO BẠCH CẦU DÙNG TỐI ƯU HÓA BẦY ĐÀN VỚI CÁC ĐẶC TRƯNG SÂU

Chương này sẽ đưa ra khái niệm của đặc trưng sâu, từ đó làm rõ phương pháp xây dựng mô hình phân loại hình ảnh tế bào bạch cầu được chúng tôi đề xuất trong luận văn, bao gồm hai bước: trích xuất đặc trưng sâu bằng một CNN hiện đại và lựa chọn đặc trưng sử dụng một cải tiến của giải thuật Salp Swarm Algorithm (SSA).

Mục lục

4.1 Học sâu và đặc trưng sâu

Đặc trưng sâu là phản hồi nhất quán của một nút hoặc lớp trong mô hình phân cấp với đầu vào cung cấp phản hồi có liên quan đến đầu ra cuối cùng của mô hình. Một đặc trưng được coi là "sâu hơn" so với một đặc trưng khác tùy thuộc vào mức độ kích hoạt sớm trong cây quyết định hoặc các bộ phân loại khác.

Trong mạng nơ-ron được thiết kế để phân loại hình ảnh, nó được đào tạo trên một tập hợp các hình ảnh tự nhiên và học các bộ lọc (đặc trưng), chẳng hạn như bộ phát hiện đường viền và cạnh hình ảnh từ các lớp trước đó. Các lớp "sâu hơn" có thể phản hồi và tạo các bộ lọc đặc trưng của riêng mình cho các mẫu phức tạp hơn ở đầu vào, chẳng hạn như kết cấu, hình dạng hoặc biến thể của các đặc trưng được xử lý trước đó.

Vì vậy, mặc dù một mạng được đào tạo thông thường có các nút lọc có thể xác định một đặc điểm cụ thể chẳng hạn như khuôn mặt, chúng sẽ không thể phân biệt được sự khác biệt giữa một khuôn mặt và bất kỳ vật thể tròn nào tương tự. Tuy nhiên, phản hồi từ một lớp sâu hơn trong hệ thống phân cấp của thuật toán đóng vai trò như một bộ lọc đặc trưng mà mô hình có thể sử dụng để không chỉ phân biệt khuôn mặt với các hình ảnh không phải khuôn mặt mà còn tạo bộ phân loại mới trong quá trình phân loại.

4.2 Trích xuất đặc trưng sâu sử dụng CNN

4.2.1 Mạng nơron tích chập - Convolutional Neural Network (CNN)

Mạng nơron tích chập (CNN) là một thuật toán học sâu có khả năng nhận ra và phân loại các đặc trưng của hình ảnh trong thị giác máy tính. Nó là một mạng nơron nhiều lớp được thiết để để phân tích những đầu vào thị giác và thực hiện các tác vụ như phân lớp hình ảnh, phân loại và phát hiện vật thể. CNN được sử dụng cho các ứng dụng học sâu trong lĩnh vực chăm sóc sức khỏe, chẳng hạn như phân tích hình ảnh y tế.

Có hai thành phần chính đối với một CNN:

- Một công cụ tích chập để phân tách những đặc trưng của hình ảnh phân tích.
- Một lớp được kết nối đầy đủ sử dụng đầu ra của tầng tích chập để dự đoán những mô tả chính xác nhất của hình ảnh

Kiến trúc mạng nơ-ron cơ bản: Kiến trúc mạng nơ-ron được lấy cảm ứng từ phương thức tổ chức và chức năng của vỏ não thị giác và được thiết kế để bắt chước mô hình kết nối của các tế bào thần kinh trong não người. Các tế bào thần kinh trong CNN được chia thành một cấu trúc ba chiều, với mỗi tập hợp nơ-ron phân tích một vùng nhỏ hoặc đặc trưng của hình ảnh. Nói cách khác, mỗi nhóm tế bào thần kinh chuyên nhận dạng một phần của bức ảnh. CNN sử dụng những dự đoán từ các tầng và cho ra kết quả cuối cùng biểu diễn một vectơ của những điểm xác suất để thể hiện khả năng mà một đặc trưng cụ thể thuộc về một lớp nhất định.

Một CNN bao gồm các tầng chính:

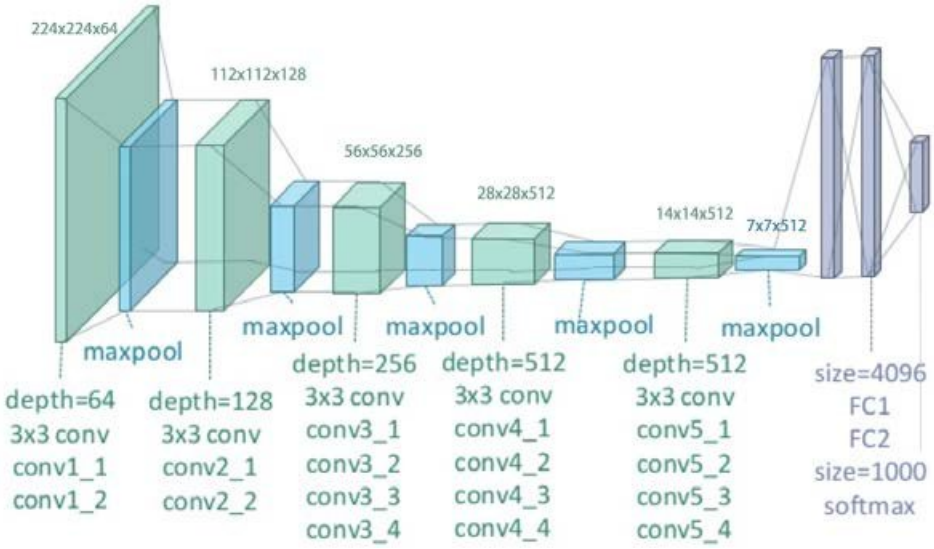
- Tầng tích chập: tạo ra một bản đồ đặc trưng để dự đoán xác suất thuộc lớp cho từng đặc trưng bằng cách áp dụng một bộ lọc quét toàn bộ hình ảnh, một vài pixel cùng lúc.
- Tầng gộp: điều chỉnh lượng thông tin mà tầng tích chập tạo ra cho từng đặc trưng và duy trì những thông tin quan trọng nhất (quá trình của tầng tích chập và tầng gộp thường lặp lại nhiều lần)
- Tầng input được kết nối đầy đủ: làm phẳng các đầu ra được tạo ra ở những tầng trước để chuyển chúng thành một vectơ đơn để có thể sử dụng làm đầu vào cho tầng kế tiếp.
- Tầng được kết nối đầy đủ: áo dụng các trọng số lên đầu vào được tạo ra bởi phân tích đặc trưng để dự đoán nhãn chính xác.
- Tầng output được kết nối đầy đủ: tạo ra những giá trị xác suất cuối cùng để xác định nhãn cho hình ảnh.

Kiến trúc của CNN là yếu tố quyết định để xác định hiệu suất và tính hiệu quả của nó. Cách cấu trúc các tầng, những phần tử nào được sử dụng trong từng tầng và cách thiết kế chúng sẽ thường ảnh hưởng đến tốc độ và độ chính xác của mạng khi thực hiện các tác vụ. Một số kiến trúc CNN phổ biến: LeNet-5 (1988), AlexNet (2012), GoogleNet (2014), VGGNet (2014),...

4.2.2 Mô hình VGG19

Simonyan và Zisserman của Đại học Oxford đã tạo ra CNN 19 tầng (16 tầng tích chập và 3 tầng được kết nối đầy đủ), sử dụng bộ lọc 3x3 với sải bước và đệm là 1, cùng với các tầng gộp tối đa với sải bước 2, được gọi là mô hình VGG19 (đặt theo tên nhóm Visual Geometry Group của Đại học Oxford). So với AlexNet, VGG19 là một mạng tích chập sâu hơn với nhiều tầng hơn. Để giảm số lượng thông số trong các mạng sâu, nó sử dụng những bộ lọc 3x3 trong tất cả các tầng tích chập và được sử dụng tốt nhất với tỷ lệ lỗi 7.3%. Mô hình VGG19 là một trong những nghiên cứu có ảnh hưởng nhất vì nó củng cố quan điểm rằng CNN phải có một mạng các lớp sâu để thể hiện phân cấp của dữ liệu trực quan có thể hoạt động. Sâu và đơn giản.

Mô hình VGG19 với tổng cộng 138 triệu tham số được huấn luyện trên một tập con của cơ sở dữ liệu ImageNet. VGG19 được huấn luyện trên hơn một triệu hình ảnh với 1000 vật thể khác nhau, ví dụ bàn phím, chuột, bút chì và nhiều loài động vật. Kết quả là, mô hình đã học được một lượng biểu diễn đặc trưng phong phú cho nhiều loại hình ảnh.



Hình 4.1: Biểu diễn kiến trúc mạng của VGG19 (conv = tích chập, FC = kết nối đầy đủ)

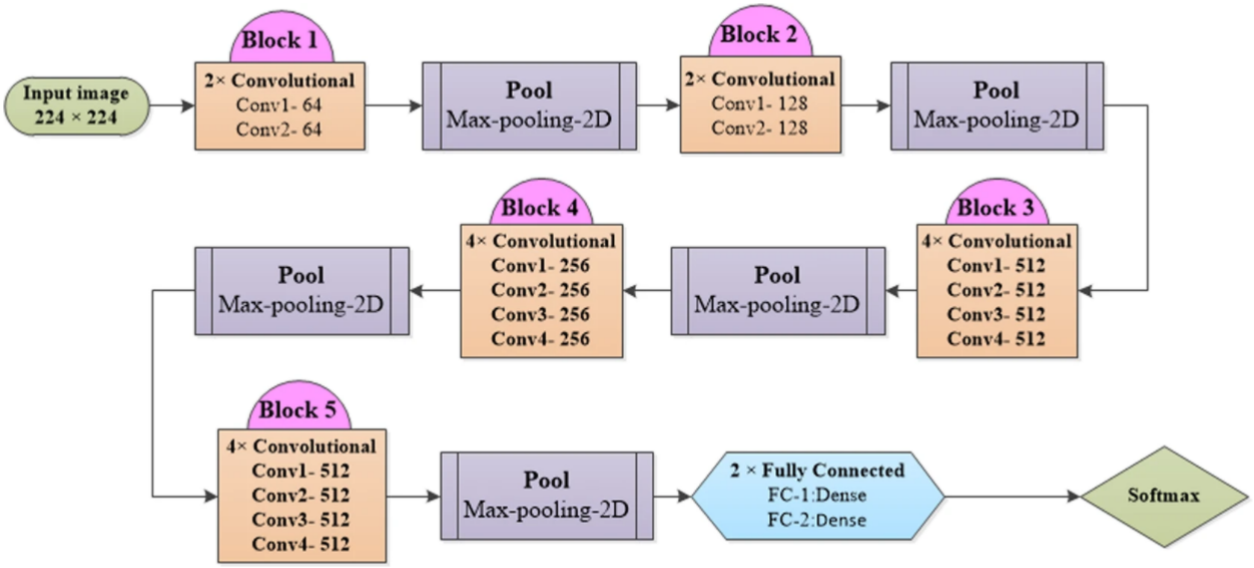
Hàm ReLU được sử dụng làm hàm kích hoạt. Với x là biến độc lập, hàm có công thức như sau:

$$ReLU(x) = \begin{cases} 0x & \text{với } x < 0 \\ xx & \text{với } x \leq 0 \end{cases}$$
$$= max(0, x)$$

So với các hàm kích hoạt khác (tanh và sigmoid), ReLU có những ưu điểm sau: Đối với các hàm tuyến tính, ReLU thể hiện rõ hơn, đặc biệt là trong các mạng học sâu; đối với các hàm phi tuyến, vì gradient của ReLU không thay đổi trong khoảng không âm nên không xảy ra vấn đề gradient biến mất, điều này giữ cho tốc độ hội tụ của mô hình ở trạng thái ổn định.

4.2.3 Đề xuất phương pháp trích xuất đặc trưng của hình ảnh tế bào bạch cầu

Ý tưởng chính của học chuyển tiếp kết hợp với mạng CNN rất sâu mà nhóm chúng tôi đề xuất là sử dụng một mạng sâu được huấn luyện fit trên tập dữ liệu lớn như ImageNet (khoảng 1,2 triệu hình ảnh với 50.000 hình ảnh khác để xác thực và 100.000 hình ảnh để thử nghiệm, trên 1000 vật thể khác nhau), và áp dụng nó để giải quyết một bài toán phân loại ảnh khác. Vì mạng đã học được các đặc trưng có liên quan từ tập dữ liệu huấn luyện chung, nên mạng có các đặc trưng cơ sở để tập trung giải quyết bài toán phân lớp của một loại hình ảnh cụ thể. Trong đề cương này, chúng tôi đề xuất sử dụng một kiến trúc CNN phổ biến và đáng tin cậy, VGG19. Số lượng kênh (chiều rộng của các tầng tích chập) tương đối nhỏ, từ 64 ở tầng đầu tiên tăng lên 512, tăng theo hệ số 2 sau mỗi tầng gộp cực đại. Tầng đầu vào có kích thước cố định là 224 × 224 pixel. Khi mỗi hình ảnh được chuyển qua một chồng các tầng tích chập, một sải bước (stride) sẽ được thêm vào để duy trì độ phân giải không gian. Công đoạn gộp được thực hiện bởi 5 tầng gộp cực đại trên một cửa sổ cụ thể với sải bước theo một vài nhưng không phải tất cả các tầng tích chập. Một chồng các tầng tích chập với độ sâu thay đổi ở các kiến trúc khác nhau được theo dõi bởi ba tầng kết nối đầy đủ với 4096 kênh ở hai tầng đầu, trong khi tầng cuối sẽ thực hiện việc phân lớp. Trong trường hợp của chúng ta, tầng này chỉ chứa hai kênh (một cho mỗi lớp). Tầng cuối cùng là tầng softmax. Tất cả những tầng ẩn đều có một phi tuyến chỉnh lưu. Với mỗi ảnh X thuộc loại T của dữ liệu huấn luyện, tham số tối ưu là hàm mất mát Weighted Cross Entropy (WCE) nhị phân.



Hình 4.2: Cấu trúc tầng VGGNet và các thông số tương ứng

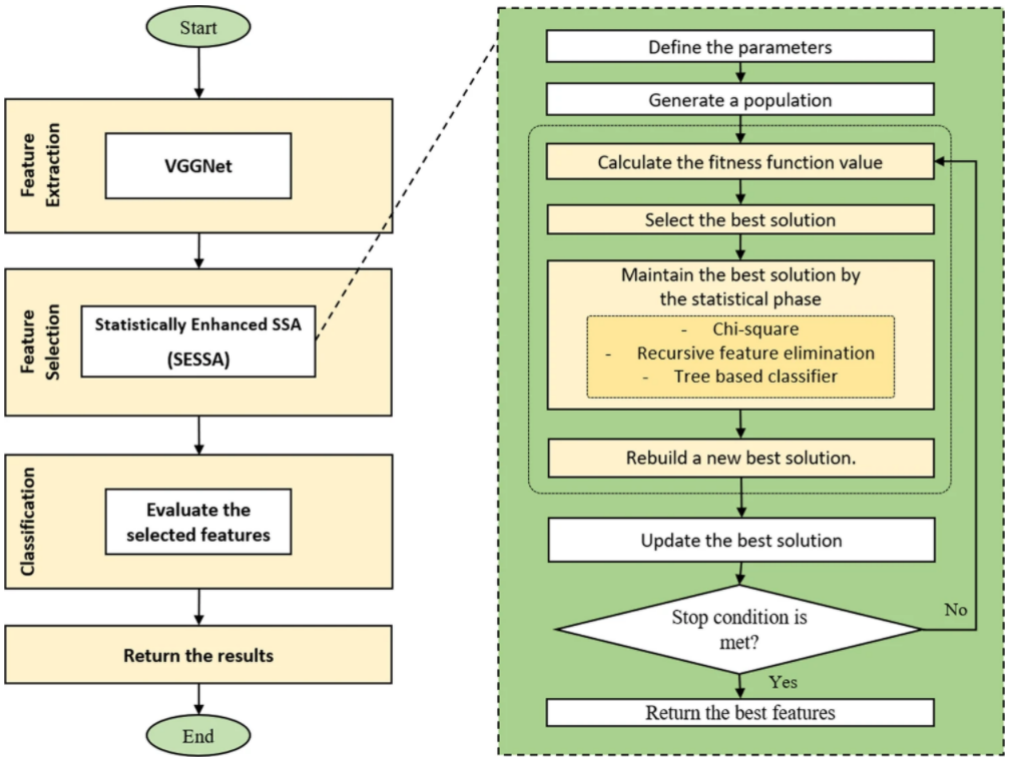
Vì kích thước của hình ảnh đầu vào là (224, 224, 3) nên tầng cuối cùng được tạo ra từ VGGNet có kích thước (7, 7, 512), nghĩa là VGGNet trả về một vector đặc trưng có kích thước 7×7×512 = 25088 đặc trưng. Để học chuyển tiếp với VGGNet, trước tiên chúng tôi lưu các tính năng được trích xuất từ mô hình huấn luyện, sau đó huấn luyện một mô hình (top model) để phân loại dữ liệu bằng các đặc trưng này, cuối cùng kết hợp dữ liệu huấn luyện và mô hình VGGNet với top model để đưa ra dự đoán.

4.3 Lựa chọn đặc trưng sử dụng Salp Swarm Algorithm (SSA)

Sau khi trích xuất đặc trưng bằng CNN, chúng tôi sẽ tiến hành lựa chọn đặc trưng để chỉ sử dụng những đặc trưng có đóng góp nhiều nhất cho mô hình. Có ba lợi ích của việc thực hiện lựa chọn đặc trưng:

- Giảm thời gian huấn luyện (ít đặc trưng hơn đồng nghĩa với việc thuật toán sẽ được huấn luyện nhanh hơn)
- Cải thiện độ chính xác (ít dữ liệu gây hiểu lầm giúp tăng độ hiệu quả của mô hình phân loại)
- Giảm over-fitting (xác suất phân loại thành công cao hơn)

Phương pháp lựa chọn đặc trưng này sẽ cải thiện thuật toán SSA cơ bản bằng cách áp dụng những phép toán thống kê để loại bỏ những đặc trưng nhiễu và không có ích cho việc phân loại, đồng thời làm tăng hiệu quả tính toán và độ ổn định của mô hình.
Mô hình tổng thể của phương pháp lựa chọn đặc trưng:



Hình 4.3: Flowchart của đề xuất cải thiện giải thuật SSA

Những phép toán thống kê được sử dụng để cải thiện SSA:

4.3.1 Loại bỏ đặc trưng quan hệ

Chi-square được dùng để loại bỏ những đặc trưng có quan hệ với nhau bằng cách đo sự phụ thuộc giữa các đặc trưng, tính toán giữa từng đặc trưng cho tất cả các lớp dựa trên công thức:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

trong đó, O_k là giá trị đặc trưng output và E_k là giá trị đặc trưng mong muốn. k đặc trưng có điểm số cao nhất sẽ được giữ lại là những đặc trưng phù hợp nhất.

4.3.2 Loại bỏ đặc trưng đệ quy - RFE

Đây là một cách tiếp cận tối ưu hóa tham lam để tìm ra tập con đặc trưng hiệu quả bằng cách sử dụng một mô hình hồi quy. Thuật toán chọn đặc trưng tốt nhất dựa vào các hệ số gán cho đặc trưng, sau đó lặp lại quá trình này với tập các đặc trưng khác cho đến khi tất cả các đặc trưng của tập dữ liệu bị loại bỏ. Các đặc trưng được xếp hạng dựa trên thời điểm chúng bị loại bỏ. Thuật toán bắt đầu với mô hình hồi quy đầy đủ chứa tất cả P đặc trưng và loại bỏ đặc trưng có hữu ích thấp nhất sau mỗi vòng lặp, theo các bước sau: ($\hat{f}^P$ biểu thị một mô hình có P đặc trưng)

- Với mỗi $k = P, P-1, \ldots, 1$ loại bỏ đặc trưng có hệ số hồi quy chuẩn hóa thấp nhất.
- Fit một mô hình mới $\hat{f}^{P-1}$ và tính độ chính xác chéo (cross validation) cho bài toán phân lớp. Đối với bài toán hồi quy, ta sử dụng các chỉ số AIC, BIC và độ thẩm định chéo R^2.
- Chọn mô hình tốt nhất từ $\hat{f}^P, \hat{f}^{P-1}, \ldots, \hat{f}^0$ dựa trên các chỉ số đã tính toán. Giải thuật RFE sử dụng hồi quy logistic với K feature được chọn thủ công.

4.3.3 Bộ phân loại độ quan trọng của đặc trưng dựa trên cây quyết định

Các phương pháp phân lớp dựa trên cây quyết định rất phổ biến nhờ vào độ chính xác cao, dễ sử dụng và tính mạnh mẽ của nó. Như đã biết, mỗi nút trên cây quyết định là một điều kiện của một thuộc tính để chia tập dữ liệu thành hai phần, những kết quả giống nhau sẽ thuộc cùng một tập hợp. Độ đo để chọn điều kiện tối ưu cục bộ được gọi là độ vẩn đục (impurity). Quá trình huấn luyện cây sẽ tính toán độ vẩn đục trọng số của cây mà từng thuộc tính có thể làm giảm. Từ đó tính được độ giảm độ vẩn đục trung bình trên mỗi đặc trưng, và dùng chúng để xếp hạng các đặc trưng. Tuy nhiên, khi tập dữ liệu chứa hai hay nhiều hơn đặc trưng có quan hệ với nhau và không có sự ưu tiên đặc trưng này hơn đặc trưng kia, dẫn đến khả năng chúng được chọn là như nhau. Khi một trong chúng được chọn, tầm quan trọng của những đặc trưng còn lại sẽ bị giảm ngay lập tức, độ vẩn đục mà chúng có thể loại bỏ đã bị loại bỏ bởi đặc trưng đầu tiên được chọn. Vấn đề này đã được giải quyết bằng việc loại bỏ những đặc trưng quan hệ từ bước 1.

Hàm mục tiêu được sử dụng trong đề tài này là root mean square error (RMSE). RMSE được áp dụng để tính toán sự khác biệt (sai số bình phương) giữa kết quả đầu ra và mục tiêu cho từng tập hợp con các đặc trưng. Do đó, giá trị nhỏ hơn của RMSE cho thấy kết quả đầu ra tốt hơn.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$$

trong đó, n là tổng số phần tử của tập, y và x biểu thị mục tiêu và kết quả đầu ra, $\bar{y}$ và là giá trị trung bình của y.

5

HIỆN THỰC VÀ NHỮNG KẾT QUẢ BAN ĐẦU

Trong chương này, chúng tôi xin trình bày mô tả các tập dữ liệu sẽ sử dụng, các tiêu chuẩn đánh giá và một số kết quả từ mô hình ban đầu.

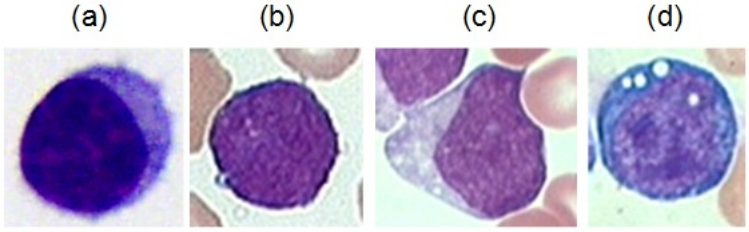
Mục lục

5.1 Mô tả dữ liệu

5.1.1 Đặc điểm hình thái của phôi bào ALL

Việc phân lớp tế bào lympho trong hình ảnh hiển vi khá phức tạp vì ngay cả một chuyên gia cũng gặp khó khăn khi phân loại một số loại tế bào lympho. Trên thực tế, những khía cạnh đặc biệt về hình thái của phôi bào ALL và tế bào lympho bình thường không khác nhau nhiều.
Tất nhiên ngày nay các công cụ chẩn đoán chính xác đã có sẵn (ví dụ: phân loại miễn dịch học) nhưng chúng yêu cầu mẫu máu, và vì phương pháp không dựa trên hình ảnh nên khả năng sử dụng những công cụ này trong các ứng dụng y tế từ xa là khá hạn chế. Theo phân tích thị giác trong chẩn đoán ALL (phương pháp FAB), những đặc trưng mà các kỹ thuật viên phòng lab xem xét trong quá trình quan sát hình ảnh là:

- L1: Các phôi bào ALL nhỏ và đồng nhất. Các hạt nhân trò, đều đặn với ít khe hở và không dễ thấy. Ít tế bào chất và thường không có không bào.
- L2: Các phôi bào ALL đều lớn và không đồng nhất. Các hạt nhân không đều và thường bị tách ra. Một hoặc nhiều có nhân con lớn. Thể tích tế bào chất có thể thay đổi, nhưng thường lớn và có thể chứa không bào.
- L3: Các phôi bào ALL có kích thước lớn vừa phải và đồng nhất. Hạt nhân đều đặn và có hình bầu dục tròn. Có một hoặc nhiều nhân con nổi bật. Thể tích tế bào chất vừa phải và chứa các không bào nổi bật.



Hình 5.1: Sự biến đổi hình thái phôi bào theo phương pháp phân loại FAB: (a) tế bào lympho ở người bình thường, (b-d) nguyên bào lympho từ các bệnh nhân theo thứ tự lần lượt là L1, L2 và L3.

5.1.2 ALL-IDB

Bộ dữ liệu được cung cấp bởi Khoa Công nghệ thông tin của trường Đại học Milano. Hình ảnh tế bào bạch cầu được chụp bằng kính hiển vi quang học kết hợp với máy ảnh kỹ thuật số Canon PowerShot G5. Tập ảnh có định dạng JPG với độ sâu màu là 24 bit. Độ phóng đại của kính hiển vi là từ 300 đến 500 lần. Cơ sở dữ liệu ALL-IDB bao gồm hai tập dữ liệu khác nhau IDB1 và IDB2. Ta sẽ kiểm tra thuật toán trên tập ALL-IDB2 vì nó được xây dựng để kiểm tra hiệu suất của các hệ thống phân lớp. Tập dữ liệu này chứa những vùng ảnh các tế bào bạch cầu lành tính và ác tính, được cắt ra từ tập ALL-IDB1. ALL-IDB2 được dùng để phát hiện phân đoạn và bài toán phân lớp.

Tập dữ liệu chứa 260 hình ảnh, 50% lành tính và 50% ác tính. Phương pháp phân lớp được đề xuất sẽ xây dựng một bộ phân lớp nhị phân vì tập dữ liệu chứa hai lớp (tế bào lành tính và tế bào ác tính).

Chú thích về tên ảnh của tập dữ liệu: Các tệp hình ảnh ALL-IDB2 được đặt tên với ký hiệu ImXXX_Y.jpg trong đó XXX là số nguyên có 3 chữ số (số thứ tự của bức ảnh) và Y là chữ số boolean bằng 0 nếu tế bào được đặt ở giữa ảnh là một tế bào bình thường và bằng 1 nếu tế bào đó là phôi bào. Tất cả hình ảnh có nhãn Y = 0 là của những người khỏe mạnh và tất cả hình ảnh có nhãn Y = 1 là của bệnh nhân ALL.

5.1.3 C-NMC

Để khắc phục hạn chế của việc sử dụng một tập dữ liệu duy nhất cũng như để mở rộng phạm vi công việc, chúng tôi sẽ mở rộng nghiên cứu thêm tập dữ liệu thứ hai, độc lập và gần đây hơn, C-NMC. Bộ dữ liệu này được dùng để phân loại tế bào bình thường và tế bào ác tính B-ALL tại IEEE ISBI-2019, bao gồm một lượng lớn hình ảnh được dán nhãn của các tế bào bình thường và ác tính. Hình ảnh tế bào được trích xuất từ hình ảnh hiển vi phết máu sau khi bình thường hóa vết bẩn. Kích thước của tập dữ liệu huấn luyện là 10.661 hình ảnh từ 76 đối tượng, bao gồm 7.272 hình ảnh tế bào của 47 bệnh nhân ALL và 3.389 hình ảnh của 29 đối tượng có tế bào bạch cầu khỏe mạnh.

5.2 Tiêu chuẩn đánh giá

Để kiểm tra hiệu suất của phương pháp được đề xuất, ta sử dụng độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc trưng (specificity), precision, recall, độ đo F1 (F-measure), root mean square error (RMSE) và hệ số xác định (R2) cũng như thời gian tính toán để chọn lọc đặc trưng. Công thức của các độ đo:

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

$$Sensitivity = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$Specificity = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

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

$$F_1 = 2 \times \frac{\text{Specificity} \times \text{Sensitivity}}{\text{Specificity}+\text{Sensitivity}}$$

trong đó: “TP” (true positives) là số mẫu tế bào ác tính được gán nhãn chính xác bởi bộ phân lớp, “TN” (true negatives) là số mẫu lành tính được gán nhãn chính xác bởi bộ phân lớp. “FP” (false positives) là số mẫu ác tính nhưng lại bị gán nhãn sai là lành tính, “FN” (false negatives) là số mẫu lành tính nhưng bị gán nhãn là ác tính.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)}{(y_i - \bar{y}_i)}$$

trong đó: biểu thị giá trị đầu ra, là giá trị mục tiêu, n là số mẫu, và là trung bình giá trị đầu ra.

Tập dữ liệu được chia thành tập huấn luyện và tập kiểm thử như sau: 80% để huấn luyện (tách 80% để huấn luyện và 20% để kiểm tra nội sử dụng kiểm tra chéo 5-fold) và 20% để kiểm tra sau cùng. Hai tập tách biệt hoàn toàn với nhau. Đối với tập ALL-IDB2, tỷ lệ này tương ứng với 208 và 52 ảnh, trong khi tập C-NMC lần lượt là 8529 và 2132 ảnh.

5.3 Các kết quả ban đầu

6

TỔNG KẾT

Trong chương này, chúng tôi xin trình bày các công việc đã thực hiện trong quá trình làm đề cương luận văn và các công việc sẽ thực hiện khi làm luận văn.

Mục lục

6.1 Các công việc đã hoàn thành

- Hiểu về bệnh ung thư bạch cầu cấp dòng lympho (Acute Lymphoblastic Leukemia) và tầm quan trọng của việc phát hiện sớm và phân lớp tế bào bạch cầu (lành tính và ác tính) trong y học hiện nay.
- Nghiên cứu bài toán phân lớp tế bào bạch cầu, tìm hiểu các cách tiếp cận bài toán phân lớp dựa trên Machine Learning và Deep Learning hiện nay, nắm được thách thức phổ biến của những phương pháp này.
- Đề xuất một phương pháp xây dựng mô hình phân loại dựa vào các đặc trưng trích xuất từ tập dữ liệu hình ảnh tế bào bạch cầu và tối ưu nó bằng kỹ thuật lựa chọn đặc trưng.
- Tìm hiểu về đặc trưng sâu, phương pháp trích xuất đặc trưng sâu dựa vào mạng nơ-ron tích chập (CNN), cụ thể là những kiến trúc mạng CNN hiện đại như AlexNet, VGGNet, ResNet,...
- Tìm hiểu lý thuyết về trí thông minh bầy đàn, các thuật toán nổi bật trong tối ưu hóa bầy đàn và phương pháp sử dụng chúng để lựa chọn đặc trưng trong các bài toán phân lớp.
- Nghiên cứu một phương pháp để cải thiện giải thuật Salp Swarm Algorithm (SSA).
- Chuẩn bị các tập dữ liệu có dán nhãn từ những nguồn đáng tin cậy.
- Hiện thực trích xuất đặc trưng từ một tập dữ liệu nhỏ (ALL-IDB2) và huấn luyện một mô hình phân lớp.

6.2 Kế hoạch thực hiện luận văn

- Sử dụng thuật toán SSA để lựa chọn đặc trưng.
- Cải thiện SSA bằng mô hình đã đề xuất để chọn lọc được những đặc trưng tốt nhất.
- Dùng mô hình đã xây dựng và một giải thuật phân lớp để phân loại tế bào bạch cầu từ các tập dữ liệu đã chuẩn bị sẵn.
- Phân tích kết quả thực nghiệm của mô hình, tiếp tục cải thiện mô hình để thu được kết quả tốt nhất.
- Đưa ra một ứng dụng cho việc chẩn một số bệnh liên quan tới số lượng bạch cầu trong máu.
- Tổng hợp và đánh giá kết quả thực hiện luận văn.
- Hoàn thành báo cáo luận văn từ đề cương luận văn.

BIBLIOGRAPHY

[1] J. Laosai and K. Chamnongthai, "Acute leukemia classification by using svm and k-means clustering," in Proceedings of the 2014 International Electrical Engineering Congress, iEECON 2014, Thailand, March 2014.

[2] P. M. Gumble and S. V. Rode, "Analysis classification of acute lymphoblastic leukemia using knn algorithm," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 2, pp. 94–98, 2017.

[3] M. Saraswat and K. V. Arya, "Feature selection and classification of leukocytes using random forest," Medical Biological Engineering Computing, vol. 52, no. 12, pp. 1041–1052, 2014.

[4] M. S. Hosseini and M. Zekri, "Review of medical image classification using the adaptive neuro-fuzzy inference system," Journal of Medical Signals and Sensors, vol. 2, no. 1, pp. 49–60, 2012.

[5] J. Prinyakupt and C. Pluempitiwiriyawej, "Segmentation of white blood cells and comparison of cell morphology by linear and naïve bayes classifiers," Biomedical Engineering Online, vol. 14, no. 1, article 63, 2015.

[6] M. Y. M. A. S. Abdul Nasir and H. Rosline, "Detection of acute leukaemia cells using variety of features and neural networks," June 2011.

[7] I. U. R. Khan, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. pattern recognition letters," 2019.

[8] R. E. J. Kennedy, "Particle swarm optimization, in: The proceedings of the ieee international conference on neural networks," Washington, DC,1995, pp. 1942–1948.

[9] S. Banka Dara, "A hamming distance based binary particle swarm optimization (hdbpso) lgorithm for high dimensional feature selection, classification and validation," 2015.

[10] Y. e. a. Chen, "An effective feature selection scheme for healthcare data classification using binary particle swarm optimization," 2018 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE, 2018.

[11] M. Tran Zhang and Xue, "A pso based hybrid feature selection algorithm for highdimensional classification," In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 3801-3808).

[12] B. Gupta S. and Iqbal, "Threshold controlled binary particle swarm optimization for high dimensional feature selection, international journal of intelligent systems and applications(ijisa)," 2018.

[13] S. S. Kannan and N. Ramaraj, "A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm.," Knowledge-Based Systems 23.6 (2010): 580-585.

[14] e. a. GaneshKumar Pugalendhi, "Hybrid ant bee algorithm for fuzzy expert system based sample classification," IEEE/ACM transactions on computational biology and bioinformatics 11.2 (2014): 347-360.

[15] A. Rouhi and H. Nezamabadi-pour, "A hybrid method for dimensionality reduction in microarray data based on advanced binary ant colony algorithm," 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC). IEEE, 2016.