Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00027

HỆ THỐNG PHÂN LOẠI ẢNH XUẤT HUYẾT NÃO THEO HƯỚNG TIẾP CẬN XỬ LÝ DỮ LIỆU LỚN

Phan Anh Cang [1], Phan Thượng Cang[2], Phạm Duy Khang[2], La Ngọc Nguyên[2], Trần Hồ Đạt[1]

[1]Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật Vĩnh Long
[2]Khoa Công nghệ thông tin, Trường Đại học Cần Thơ

cangpa@vlute.edu.vn, ptcang@cit.ctu.edu.vn, pdkhang@cit.ctu.edu.vn, lnnguyen@cit.ctu.edu.vn, datth@vlute.edu.vn

TÓM TẮT: *Tai biến mạch máu não hay đột quỵ là căn bệnh gây tử vong đứng hàng thứ ba sau ung thư và tim mạch ở các nước công nghiệp phát triển, riêng đối với Việt Nam đứng hàng thứ nhất. Đây là bệnh lý không những gây tử vong cao mà còn là nguy cơ tàn phế hàng đầu trong các loại bệnh. Vì tính chất nguy hiểm của xuất huyết não nên việc chẩn đoán đòi hỏi phải thật nhanh chóng và chính xác. Hơn nữa do số lượng ca xuất huyết não nhập viện ngày càng tăng nên yêu cầu đặt ra là cần phải xử lý một lượng lớn dữ liệu từ nhiều bệnh viện khác nhau. Do vậy, việc xây dựng một hệ thống có khả năng phân loại tự động các dạng xuất huyết với tập dữ liệu lớn, thời gian xử lý nhanh và độ chính xác cao là điều hết sức cần thiết. Trong nội dung bài báo, chúng tôi đề xuất phương pháp sử dụng chỉ số Hounsfield và áp dụng các thuật toán máy học để nhận dạng và phân loại xuất huyết não từ các hình ảnh CT/MRI. Phương pháp dựa trên nền tảng xử lý dữ liệu lớn. Kết quả nghiên cứu cho thấy, phương pháp này giúp rút ngắn khá nhiều thời gian cho bác sĩ trong việc chẩn đoán xuất huyết não, từ đó phát hiện sớm căn bệnh và có hướng điều trị kịp thời cho bệnh nhân. Kết quả thực nghiệm của phương pháp đề xuất đạt độ chính xác 98% với tốc độ nhận dạng cận thời gian thực.*

Từ khóa: *Dữ liệu lớn, xuất huyết não, Spark, Hounsfield, xử lý dữ liệu lớn.*

I. GIỚI THIỆU

Bệnh xuất huyết não xảy ra khi máu đột nhiên tràn vào mô não, tập trung thành một khối (tụ máu), tạo áp lực lên các mô xung quanh và giết chết các tế bào não. Xuất huyết có thể xảy ra bên trong, giữa và màng bao bọc não, giữa các lớp màng não hoặc giữa hộp sọ và phần bao ngoài của não. Tỉ lệ tử vong của xuất huyết não trong 30 ngày là 50%. Trong đó, một nửa số bệnh nhân tử vong trong hai ngày đầu tiên sau xuất huyết. Những bệnh nhân còn sống sau xuất huyết não nặng thường bị di chứng nặng nề. Nhiều trường hợp ở trong tình trạng sống thực vật và sẽ chết do bội nhiễm, suy kiệt. Vì tính chất nguy hiểm của xuất huyết não nên việc chẩn đoán cần phải chính xác và nhanh chóng. Một trong những biện pháp cận lâm sàng để phát hiện căn bệnh này là chẩn đoán dựa trên hình ảnh chụp cộng hưởng từ (MRI) hoặc chụp cắt lớp (CT). Tuy nhiên, cấu trúc não có độ phức tạp lớn và khó khăn trong việc chẩn đoán. Cùng với đó, ở các tuyến bệnh viện tại Đồng bằng sông Cửu Long, chuyên môn và kinh nghiệm của các bác sĩ là khác nhau. Chính vì vậy điều này dẫn đến trường hợp kết quả chẩn đoán có thể sai sót và gây ra hậu quả rất nghiêm trọng hoặc những bệnh nhân ở các tuyến vùng sâu vùng xa có xu hướng dồn về các bệnh viện tuyến trên để chẩn đoán và điều trị gây áp lực lớn cho các bác sĩ và cơ sở vật chất ở các bệnh viện này. Hơn nữa, bệnh nhân phải di chuyển một quãng đường xa để tới các bệnh viện này dẫn tới bệnh trở nặng hoặc tăng nguy cơ tử vong. Vì những lí do trên, bài toán đặt ra là làm sao tăng độ chính xác của kết quả chẩn đoán tại các tuyến bệnh viện, đồng thời giảm áp lực cho các bác sĩ và tiết kiệm thời gian chẩn đoán, điều trị bệnh cho bệnh nhân. Nội dung trình bày trong bài báo gồm giới thiệu các công việc liên quan; các phương pháp sử dụng: thu thập dữ liệu và rút trích đặc trưng dựa trên chỉ số Hounsfield, sử dụng Apache Spark trên nền tảng Streaming để xử lý phân tán dữ liệu lớn; mô hình đề xuất nhận dạng và phân loại ảnh xuất huyết não cận thời gian thực dựa trên nền tảng xử lý dữ liệu lớn; một số kết quả thực nghiệm đạt được.

II. CÔNG VIỆC LIÊN QUAN

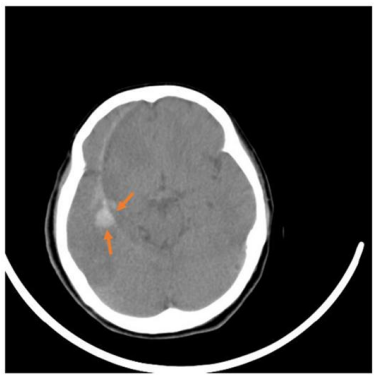
Bài toán nhận dạng và phân loại ảnh xuất huyết não đang trở thành nhu cầu cấp thiết và quan trọng trong lĩnh vực y tế hiện nay và việc áp dụng công nghệ thông tin vào các hệ thống nhận dạng đem lại những hiệu quả thiết thực góp phần nâng cao khả năng chẩn đoán và cứu sống thêm nhiều bệnh nhân đột quỵ kịp thời. Có rất nhiều công trình nghiên cứu của các tác giả trong và ngoài nước liên quan đến hệ thống trích xuất đặc trưng và nhận dạng ảnh xuất huyết não. Tác giả bài báo [1] tập trung vào việc phân đoạn (segment) ảnh y khoa để tìm ra vùng bệnh và cô lập chúng, bằng cách sử dụng thuật toán phân cụm K-means để phân đoạn dựa trên đồ thị histogram của ảnh. Trong bài báo [2], nhóm nghiên cứu đã sử dụng phương pháp phát hiện cung (Edge Detection) trong xử lý hình ảnh giúp cô lập vùng xuất huyết. Một bài nghiên cứu khác như [3] đề cập đến vấn đề lựa chọn được đúng lát cắt ở vùng sọ não và loại bỏ các lát cắt tại vùng mũi của bệnh nhân và sử dụng phương pháp "*Wavelet and Haralick texture features*" để phân loại hình ảnh. Ngoài ra, trong nghiên cứu [4] đã cung cấp những thông tin mới về chỉ số HU giúp nhận dạng vùng xuất huyết chính xác hơn. Bên cạnh đó, những đặc trưng của loại xuất huyết cần tính toán và cách rút trích đặc trưng từ hình ảnh cũng được đề xuất trong nghiên cứu [5].

2.1. Xuất huyết não

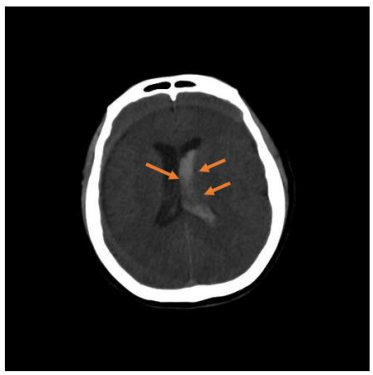
Xuất huyết não là một dạng đột quỵ, xảy ra khi một động mạch trong não bị vỡ gây chảy máu cục bộ trong các mô xung quanh. Khi não nhận một tổn thương, mạch máu vỡ bị tụ lại một chỗ, gây chèn ép lên mô não gần đó, làm giảm lưu lượng máu và gây chết các mô não. Xuất huyết có thể xảy ra bất kỳ đâu, dựa vào vị trí hình dạng của vùng

xuất huyết ta có thể phân loại thành 4 dạng xuất huyết sau: Xuất huyết nội sọ, xuất huyết dưới nhện, máu tụ dưới màng cứng và máu tụ ngoài màng cứng. Nguyên nhân gây ra xuất huyết não bắt nguồn từ chấn thương vùng đầu, huyết áp cao, dị tật động mạch máu, các bệnh lý về gan,… Ngoài các biểu hiện lâm sàng, để xác định xuất huyết não, các bác sĩ sử dụng phương pháp chụp cắt lớp vi tính CT hoặc cộng hưởng từ MRI sọ não. Trong thực tế, xuất huyết não có bốn dạng thường gặp như sau:

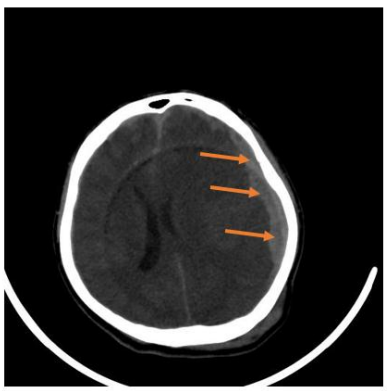
a. Xuất huyết nội sọ (Hình 1) với các đặc trưng là vùng xuất huyết não có hình tròn và không có các cung lồi hoặc lõm cùng với thể tích khá lớn, khoảng cách đối với tâm sọ não với vùng bị xuất huyết là tương đối không quá gần cũng không quá xa.



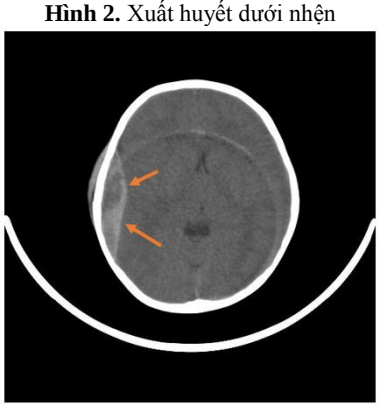
Hình 1. Xuất huyết nội sọ



Hình 2. Xuất huyết dưới nhện



Hình 3. Máu tụ dưới màng cứng



Hình 4. Máu tụ ngoài màng cứng

b. Xuất huyết dưới nhện (Hình 2) thường xảy ra ở vùng khoang nhện, vị trí nằm rất gần với tâm sọ não, có độ nghiêng ở khoảng 45^o và nằm trong khoang nhện của não.

c. Xuất huyết dưới màng cứng (Hình 3) và ngoài màng cứng (Hình 4) đều có đặc trưng là nằm rất gần với sọ não, sự khác nhau là thể tích xuất huyết của hai loại này khác nhau, với máu tụ dưới màng cứng thì thể tích xuất huyết nhỏ hơn, và nằm dọc theo não, có độ dài trục chính lớn hơn nhiều so với độ dài trục phụ của hình elip bao quanh đó.

Như ta thấy trong các đặc trưng phân loại thì có một vài đặc trưng rất quan trọng để giúp phân loại một cách hiệu quả như là đặc trưng về độ dài của trục chính và trục phụ, nó có thể giúp phân biệt giữa máu tụ dưới màng cứng và ngoài màng cứng. Đặc trưng về chu vi có thể giúp phân loại các dạng xuất huyết bằng kích thước của nó; khoảng cách so với tâm sọ não có thể phân biệt dựa vào vị trí của các vùng xuất huyết.

2.2. Chỉ số Hounsfield

Năm 1970 Sir Godfrey Newbold Hounsfield phát hiện ra phương pháp chụp cắt lớp CT, cùng với đó tên của ông được dùng để đặt cho một chỉ số quan trọng trong hình ảnh chụp cắt lớp là Hounsfield Unit. Chỉ số Hounsfield đại diện cho khả năng hấp thụ tia X của một mô hoặc bộ phận nào đó trong cơ thể con người và có thể được tính thông qua các đại lượng được lưu trong tập tin hình ảnh chuẩn y khoa Dicom. Sử dụng công thức (1) để tính toán chỉ số HU tại một điểm ảnh [6].

$$HU = pixelvalue \times RescaleSlope + RescaleIntercept \quad (1)$$

- $RescaleSlope$ và $RescaleIntercept$ là hai giá trị được lưu trực tiếp trong hình ảnh chụp cắt lớp.
- $pixelvalue$ là giá trị pixel tại từng điểm ảnh của ảnh Dicom.

Dựa vào chỉ số HU tính được, ta có thể xác định tại một điểm trên hình ảnh đó thuộc về bộ phận nào của cơ thể dựa vào một bảng được gọi là thang Hounsfield.

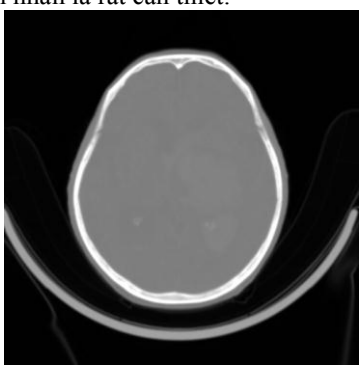
Bảng 1. Thang đo Hounsfield *[7]*

Nước	0 HU
Xương	1.000 HU
Không khí	-1.000 HU
Chất xám	35 đến 40 HU
Chất trắng	20 HU
Xuất huyết	40 đến 90 HU
Phổi	-700 đến -600
Thận	20 đến 45

III. MÔ HÌNH ĐỀ XUẤT

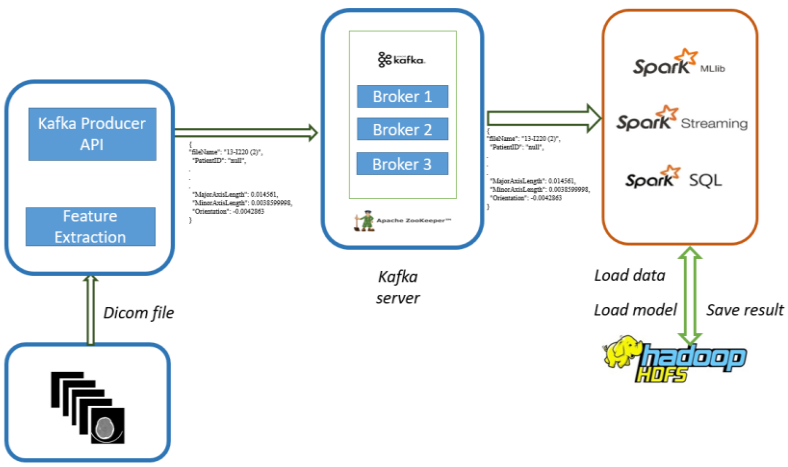
3.1. Phương pháp tổng quát

Với các nghiên cứu trước đây, việc rút trích các đặc trưng hình ảnh CT/MRI đa số sử dụng phương pháp dựa trên thông tin mức xám, nghĩa là chúng ta phải chuyển đổi hình ảnh Dicom trở thành dạng ảnh thông thường như jpg, png,... Điều này làm mất một số thông tin quan trọng cho việc chẩn đoán bệnh của bệnh nhân, gây khó khăn cho việc nhận dạng và phân loại đối với các hình ảnh này. Trong nghiên cứu này, chúng tôi đề xuất sử dụng chỉ số HU của các pixel trong ảnh CT/MRI để nhận biết khu vực xuất huyết dựa vào Bảng 1. Các phương pháp xử lý ảnh được áp dụng để khử bỏ các vùng ảnh không phải xuất huyết từ đó chúng tôi tính toán các đặc trưng. Hình 5 là ảnh chụp cộng hưởng từ chứa thông tin về não của bệnh nhân để các bác sĩ chẩn đoán bằng mắt thường mà không cần tăng độ tương phản hoặc làm rõ nét ảnh. Vì vậy, việc rút trích đặc trưng về thông tin não của bệnh nhân phục vụ nhận dạng và phân lớp giúp bác sĩ chẩn đoán, điều trị và quản lý hồ sơ bệnh nhân là rất cần thiết.



Hình 5. Ảnh CT/MRI não được chụp cộng hưởng từ

Mô hình tổng quan đề xuất theo hướng tiếp cận xử lý dữ liệu lớn được mô tả trong Hình 6:



Hình 6. Mô hình đề xuất tổng quát về phân loại xuất huyết não

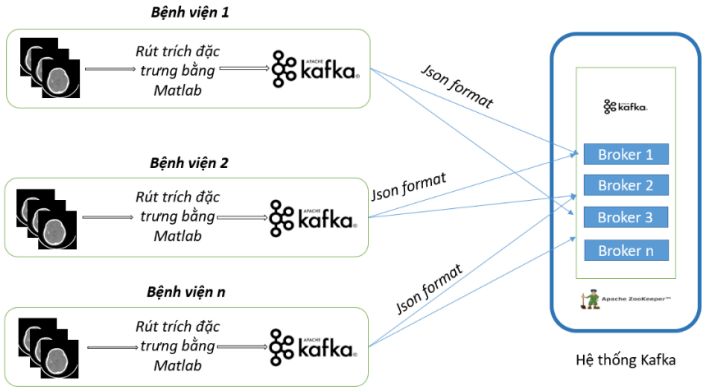
Quá trình xử lý được chia thành 2 giai đoạn:

+ ***Giai đoạn 1:*** dữ liệu hình ảnh được thu thập từ nhiều bệnh viện thông qua bộ thu thập dữ liệu phân tán Kafka có khả năng chịu lỗi cao và tiến hành rút trích đặc trưng trong giai đoạn này với mục tiêu giảm lưu lượng thông tin truyền tải trên mạng, kết quả đầu ra là tập dữ liệu đặc trưng được gửi tới các Broker đóng vai trò là bộ nhớ đệm của quá trình xử lý.

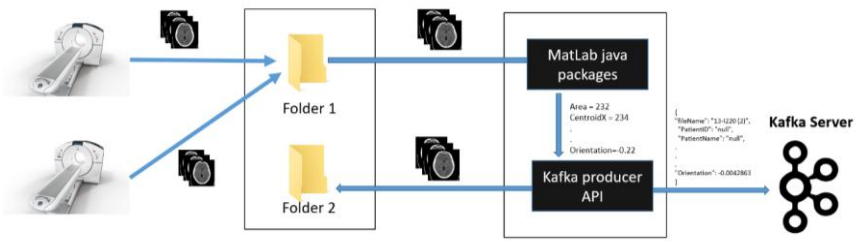
+ ***Giai đoạn 2:*** tập dữ liệu đặc trưng từ các Broker được gửi đến bộ phân loại song song trong môi trường Spark. Spark Streaming tiếp nhận dữ liệu (tập đặc trưng) từ các bệnh viện và sau đó chúng được xử lý và lưu tại hệ thống HDFS để huấn luyện mô hình máy học. Sau khi có được mô hình máy học tốt nhất, các dữ liệu kiểm tra sẽ được phân loại bởi mô hình vừa huấn luyện, các kết quả cũng được lưu tại hệ thống HDFS.

3.2. Xây dựng bộ thu thập dữ liệu và rút trích đặc trưng

Hệ thống thu thập dữ liệu làm nhiệm vụ tiếp nhận hình ảnh Dicom ngay khi xuất từ máy chụp cắt lớp, tiến hành rút trích các đặc trưng của ảnh chụp CT/MRI và sử dụng Kafka API để gửi dữ liệu vào hệ thống Kafka. Với mỗi hình ảnh từ bệnh viện được gửi vào một Folder tương ứng để quản lý được dễ dàng.

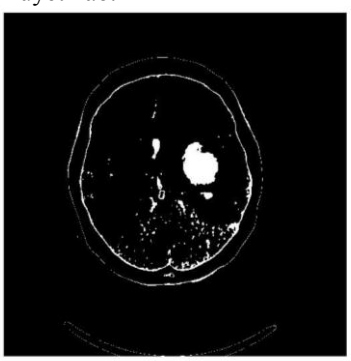


Hình 7. Mô hình rút trích đặc trưng song song



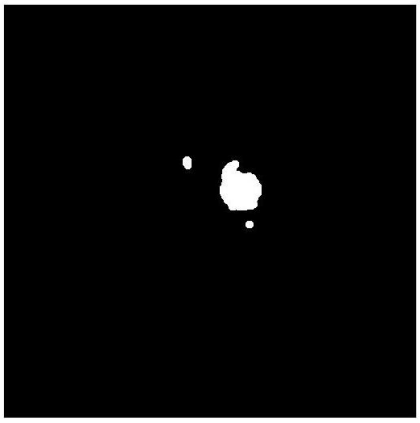
Hình 8. Mô hình xử lý phân loại song song

Hình 7 và Hình 8 mô tả quá trình thu thập và xử lý dữ liệu. Hình ảnh CT/MRI sau khi được xuất ra từ máy chụp CT/MRI được hệ thống tiếp nhận xử lý, các đặc trưng được xác định bởi các công cụ và ngôn ngữ lập trình MATLAB. Kết quả đầu vào giai đoạn này là tên một tập tin dicom và đầu ra là một lớp Bean tên là HemorrhageFeature, chứa các thông tin cần thiết. Các dữ liệu được gửi đến Server Kafka, chúng tôi sử dụng một Producer API của Kafka cung cấp để giúp gửi dữ liệu được dễ dàng. Tiếp theo, chúng tôi thực hiện tính toán chỉ số HU tại tất cả điểm ảnh của ảnh cắt lớp dựa vào thông tin chỉ số RescaleSlope và RescaleIntercept được lưu trữ trên ảnh Dicom. Dựa vào chỉ số HU trong Bảng 1, chúng ta có thể xác định vùng xuất huyết não.

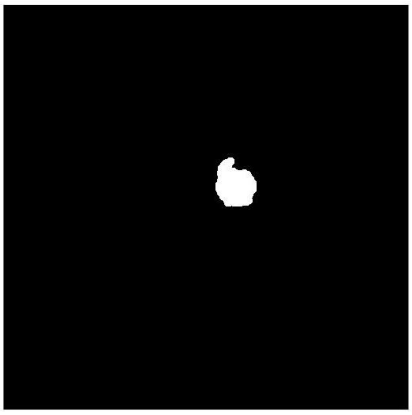


Hình 9. Ảnh kết quả sau khi phân đoạn dựa trên chỉ số HU

Hình 9 mô tả quá trình phân đoạn dựa trên chỉ số HU. Sau đó, chúng tôi áp dụng phương pháp hình thái học trong việc xử lý cấu trúc hình học của ảnh nhị phân với mục đích giảm lỗi trong quá trình nhận dạng. Kết quả vùng hộp sọ được loại bỏ như mô tả trong Hình 10. Cuối cùng, chúng tôi thực hiện tính toán và loại bỏ các vùng nhỏ hơn một ngưỡng xác định sẵn. Kết quả vùng xuất huyết được xác định với các hình dạng và vị trí khác nhau tuỳ vào loại xuất huyết (Hình 11).



Hình 10. Ảnh kết quả sau khi áp dụng phương pháp hình thái học



Hình 11. Kết quả sau khi loại bỏ các vùng nhỏ hơn một ngưỡng xác định

Rút trích đặc trưng: Việc rút trích tốt các đặc trưng quan trọng có thể tăng độ chính xác của mô hình máy học. Trong bài báo này, chúng tôi rút trích các đặc trưng được mô tả như sau [5]:

1. Kích thước của vùng xuất huyết

2. Toạ độ trọng tâm của vùng xuất huyết

3. Chu vi của vùng xuất huyết

4. Khoảng cách của vùng xuất huyết và trọng tâm của sọ não bệnh nhân, đặc trưng này giúp phân biệt giữa tụ máu nội sọ hoặc dưới nhện và máu tụ dưới màng cứng và ngoài màng cứng. Tụ máu nội sọ và dưới nhện có khoảng cách với trọng tâm não thấp hơn là máu tụ dưới màng cứng và máu tụ ngoài màng cứng.

5. Đường kính của vòng tròn có cùng kích thước với vùng xuất huyết não.

6. Tỉ lệ pixel trong vùng xuất huyết so với vùng lồi của hình phần xuất huyết, điều này giúp phân biệt giữa các loại xuất huyết có hình dạng lồi như là máu tụ ngoài màng cứng và xuất huyết nội sọ so với các loại xuất huyết có hình dạng lõm như là máu tụ dưới màng cứng.

7. Vùng giới hạn hình ellipse nhỏ nhất chứa vùng xuất huyết cũng được sử dụng để tính toán các đặc trưng gồm:

+ Số lượng pixel của vùng giới hạn chứa vùng xuất huyết

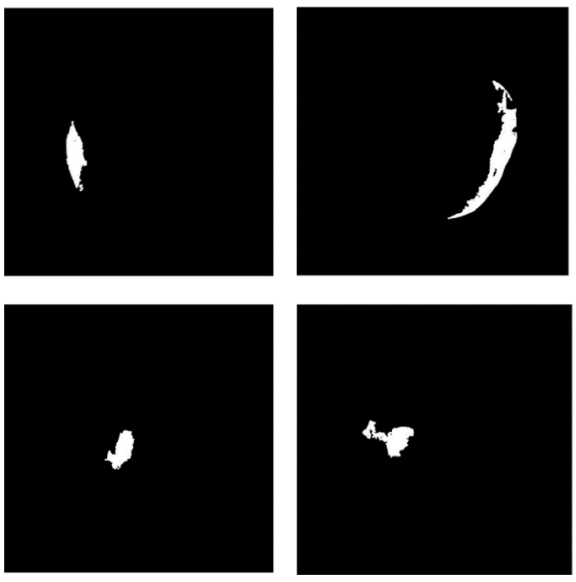
+ Tỉ lệ pixel của vùng giới hạn so với vùng xuất huyết

8. Các đặc trưng khác liên quan đến vùng chứa vùng xuất huyết bao gồm:

+ Độ lệch tâm của hình ellipse

+ Độ dài trục chính và trục phụ của hình ellipse

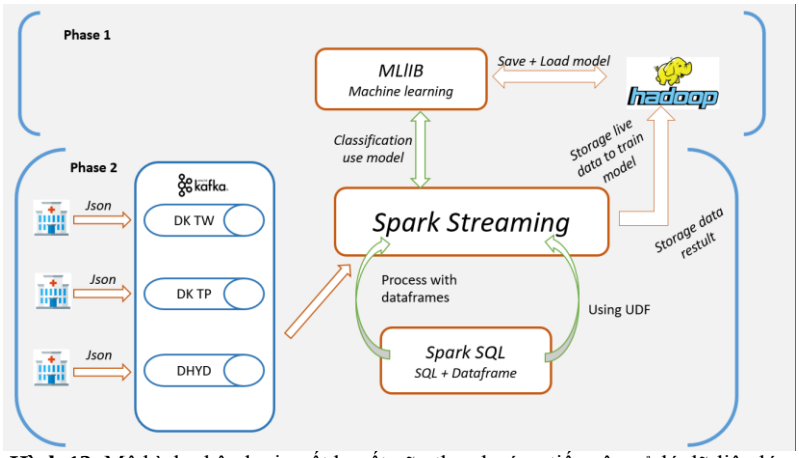
+ Góc nghiêng của trục chính hình ellipse so với trục hoành.



Hình 12. Kết quả sau khi nhận dạng của bốn loại xuất huyết não

Hình 12 mô tả kết quả phân loại trên các ảnh có xuất huyết não. Đối với những hình ảnh không có xuất huyết, vùng xuất huyết có giá trị HU nhỏ hơn 40. Sau giai đoạn 1, dữ liệu đầu ra là hệ thống tập tin chứa các đặc trưng được lưu trữ dưới dạng tập tin json. Trong giai đoạn 2, hệ thống sẽ thực hiện nhận dạng và phân loại ảnh xuất huyết não.

3.3. Nhận dạng và phân loại ảnh xuất huyết não



Hình 13. Mô hình phân loại xuất huyết não theo hướng tiếp cận xử lý dữ liệu lớn

Hình 13 mô tả quá trình nhận dạng và phân loại xuất huyết não theo hướng tiếp cận xử lý dữ liệu lớn. Tập dữ liệu đặc trưng được tiếp nhận về Spark Streaming sẽ được xử lý một lần. Kết thúc việc này là một mô hình máy học có độ chính xác cao được huấn luyện và sẽ được dùng để phân loại dữ liệu đầu vào. Kết quả phân loại sẽ được lưu vào hệ cơ sở dữ liệu phân tán HDFS.

fileName	PatientID	PatientName	PatientAge	PatientSex	InstitutionName	institutionAddress	AccessionNumber	Manufacturer	Modality	Area	CentroidX	CentroidY
65-I160 (9)	8296	DNH BBB MMM	null	F	BVDK TP CAN THO	CAN THO, VIETNAM	CCTH	Philips	CT	0.2858	0.028057998	0.02535
8-I140	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.2811	0.029801002	0.020244
8-I140_4	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.2811	0.029801002	0.020244
8-I150	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.5254	0.030565	0.020616999
8-I150_4	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.5254	0.030565	0.020616999
8-I160	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.7749	0.031463	0.021708999
8-I160_4	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.7749	0.031463	0.021708999
8-I170	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.7339	0.030883	0.022510001
8-I170_4	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.7339	0.030883	0.022510001
8-I180	8115	HUYNH XXX YYYY	null	M	BVDK TP CAN THO	CAN THO, VIETNAM	NOI TK	Philips	CT	0.7444	0.030764999	0.023035001
37-I280 (4)	8344	NGUYEN CCCC YY	null	F	BVDK TP CAN THO	CAN THO, VIETNAM	CCTH	Philips	CT	0.6447	0.037698	0.028445002

Hình 14a. Kết quả rút trích tập đặc trưng

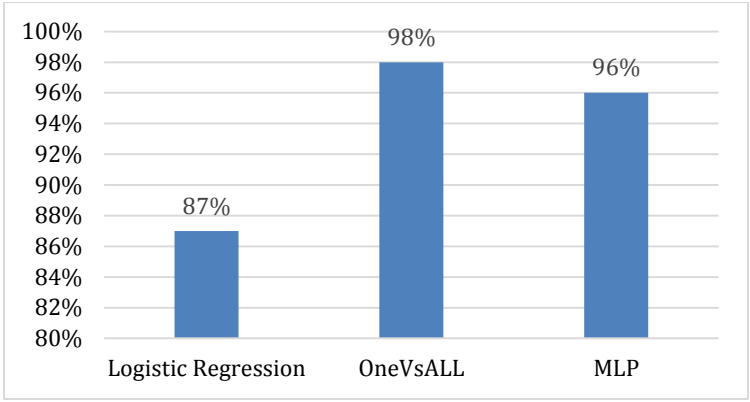
stanceWithSkull	Diameter	Solidity	BBULX	BBULY	BBWith	BBHeight	FilledArea	Extent	Eccentricity	MajorAxisLength	MinorAxisLength	Orientation	prediction
0.0063046003	0.0060324003	5.1182004E-5	0.02385	0.01935	0.009	0.0092	0.2858	3.4516997E-5	7.0511E-5	0.0089085	0.006317	0.0026695	3.0
0.0059043	0.0059826	7.6386E-5	0.02575	0.01705	0.0085	0.0065	0.2812	5.0878E-5	7.6918E-5	0.0081591	0.0052139	0.0027819	3.0
0.0059043	0.0059826	7.6386E-5	0.02575	0.01705	0.0085	0.0065	0.2812	5.0878E-5	7.6918E-5	0.0081591	0.0052139	0.0027819	3.0
0.006337	0.008179	8.1800004E-5	0.02565	0.01575	0.0101	0.0092	0.526	5.6543E-5	7.1721E-5	0.010187999	0.0070995004	0.0034556	3.0
0.006337	0.008179	8.1800004E-5	0.02565	0.01575	0.0101	0.0092	0.526	5.6543E-5	7.1721E-5	0.010187999	0.0070995004	0.0034556	3.0
0.0065877	0.009933	8.2252E-5	0.02595	0.01595	0.011	0.0112	0.8463	6.2898E-5	3.0239999E-5	0.010919999	0.010408999	-8.2968996E-4	2.0
0.0065877	0.009933	8.2252E-5	0.02595	0.01595	0.011	0.0112	0.8463	6.2898E-5	3.0239999E-5	0.010919999	0.010408999	-8.2968996E-4	2.0
0.0057801	0.0096666	8.0799E-5	0.02615	0.01685	0.0102	0.0109	0.7388	6.601E-5	5.9772003E-5	0.011252	0.0090211	-0.0055178003	3.0
0.0057801	0.0096666	8.0799E-5	0.02615	0.01685	0.0102	0.0109	0.7388	6.601E-5	5.9772003E-5	0.011252	0.0090211	-0.0055178003	3.0
0.0055469	0.0097354	8.4142E-5	0.02605	0.01695	0.0101	0.0113	0.753	6.5224E-5	7.0949E-5	0.011895	0.008382901	-0.0055846	3.0
0.011904	0.009060199	3.8955E-5	0.02855	0.01425	0.013	0.0263	0.6507	1.8855999E-5	9.7592005E-5	0.02736	0.0059677	0.0071583	1.0

Hình 14b. Kết quả phân loại

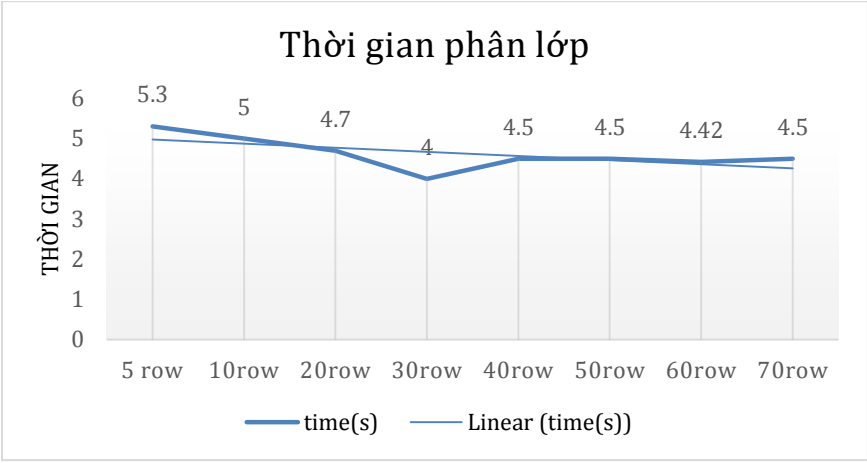
Hình 14a và 14b mô tả kết quả đạt được khi tiến hành rút trích đặc trưng và phân loại xuất huyết não cận thời gian thực dựa trên phương pháp đề xuất. Các kết quả phân lớp được hiển thị dưới dạng các nhãn với thông số như sau: 0- Chảy máu dưới nhện; 1- Máu tụ dưới màng cứng; 2- Máu tụ ngoài màng cứng; 3- Chảy máu nội sọ. Kết quả được lưu vào hệ thống HDFS ở dạng parquet để tiết kiệm bộ nhớ trên hệ thống HDFS và có khả năng truy cập nhanh hơn là việc lưu ở dạng text.

IV. KẾT QUẢ THỰC NGHIỆM

Chúng tôi thu thập tập ảnh đầu vào từ bệnh viện Đại học Y Dược Cần Thơ và Bệnh viện Đa khoa thành phố Cần Thơ bao gồm 394 hình ảnh CT/MRI của não người thuộc một trong bốn loại xuất huyết: 41 hình ảnh của bệnh xuất huyết dưới nhện, 52 hình ảnh của bệnh xuất huyết dưới màng cứng, 100 hình ảnh xuất huyết ngoài màng cứng và 201 hình ảnh xuất huyết nội sọ. Dữ liệu đầu vào là một lô các tập tin json có chứa các thông tin đặc trưng được rút trích ở giai đoạn 1. Dữ liệu đầu vào được chia thành hai tập: 80% dữ liệu được dùng huấn luyện mô hình máy học và 20% dữ liệu còn lại là dùng để đánh giá độ chính xác của mô hình. Chúng tôi lần lượt huấn luyện 3 mô hình phân lớp với nhiều thông số thực nghiệm để đạt được mô hình tốt nhất bao gồm: mô hình hồi quy logistic, OneVsAll, Multilayer Perceptron (MLP).



Hình 15. Biểu đồ đánh giá độ chính xác phân loại



Hình 16. Thời gian thực hiện với việc thay đổi tập dữ liệu dùng mô hình phân lớp OneVsAll

Hình 15 cho thấy độ chính xác của thuật toán OneVsAll cao nhất 98% và thuật toán này sẽ được áp dụng vào trong quá trình nhận dạng nhằm nâng cao độ tin cậy của hệ thống. Bên cạnh đó, một đặc điểm dễ thấy ở các hệ thống xử lý dữ liệu lớn đặc biệt trong môi trường Spark đó chính là thời gian xử lý chậm khi thực hiện với tập dữ liệu nhỏ và rất nhanh với tập dữ liệu lớn. Ta có thể thấy thời gian thực hiện công việc có xu hướng giảm xuống theo chiều tăng dần của kích thước tập dữ liệu (Hình 16) do các hệ thống nền tảng dữ liệu lớn nói chung và Spark nói riêng tốn nhiều tài nguyên để khởi chạy hệ thống và các tính toán. Trong nghiên cứu này, chúng tôi sử dụng cấu hình máy tính với tốc độ xử lý của CPU Core i5 (2 nhân, 4 luồng) và 4GB bộ nhớ nên thời gian xử lý cận thời gian thực. Tuy nhiên, khi hệ thống được triển khai trên các cụm của Apache Spark sẽ rút ngắn hơn nữa thời gian xử lý của hệ thống và tiến gần hơn đến xử lý thời gian thực.

V. KẾT LUẬN

Trong bài báo này chúng tôi đề xuất phương pháp hiệu quả nhằm xây dựng một hệ thống nhận dạng và phân loại xuất huyết não theo hướng tiếp cận xử lý dữ liệu lớn cận thời gian thực. Phương pháp đề xuất xác định vùng xuất huyết não sử dụng chỉ số Hounsfield; rút trích các đặc trưng để phân loại xuất huyết bằng cách ứng dụng khả năng xử lý dữ liệu lớn của Apache Spark để xây dựng hệ thống cận thời gian thực đem lại kết quả nhận dạng với độ chính xác cao. Nó có thể phân lớp các hình ảnh CT/MRI nhanh chóng và chính xác, hỗ trợ các bác sĩ trong việc chẩn đoán và điều trị kịp thời căn bệnh nguy hiểm xuất huyết não. Điều này không những giúp tăng cơ hội sống sót của bệnh nhân xuất huyết não mà còn giảm tải cho các bệnh viện tuyến trung ương và các bệnh viện tuyến tỉnh. Trong hướng phát triển của nghiên cứu, chúng tôi tiếp tục triển khai trên các cụm của Apache Spark để rút ngắn hơn nữa thời gian xử lý của hệ thống và tiến gần hơn đến xử lý thời gian thực nhằm đáp ứng hiệu quả về thời gian phân loại đồng thời mở rộng mô hình trong việc nhận dạng và phân loại nhiều loại bệnh khác như u não, ung thư phổi, và các bệnh lý về gan.

TÀI LIỆU THAM KHẢO

[1] A. D. J. B. a. T. B. S. Loncaric, "3-d image analysis of intra-cerebral brain hemorrhage from digitized ct films," *Computer Methods and Programs in Biomedicine,* p. 207–216, 1995.

[2] K. T. J. Santosh H. Suryawanshi, "Smart Brain Hemorrhage Diagnosis Using Artificial Neural Networks," *International journal of scientific & technology research,* pp. 267-271, 2015.

[3] C. L. T. T. Y. L. C. K. L. B. C. P. C. L. Q. T. S. T. a. Z. Z. R. Liu, "Hemorrhage slices detection in brain ct images," *In 19th International Conference on Pattern Recognition (ICPR 2008),* p. 1-4, 2008.

[4] N. T. M. N. P. C. Phan Anh Cang, "Chẩn đoán xuất huyết não dựa trên chỉ số Hounsfield và kỹ thuật mạng nơ-ron tích chập," *Hội nghị khoa học quốc gia lần thứ XI (FAIR XI),* 2018.

[5] D. A. K. A.-D. I. A. Mahmoud Al-Ayyoub, "Automatic Detection and Classification of Brain Hemorrhages," *WSEAS transactions on computers 12.10,* pp. 395-405, 2013.

[6] P. T. C. V. V. Q. L. T. H. Y. Phan Anh Cang, "Phát hiện và phân loại tự động xuất huyết não trên các ảnh CT/MRI," *Hội thảo quốc gia lần thứ XX: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông,* 2017.

[7] https://en.wikipedia.org/wiki, "Hounsfield_scale".

[8] M. C. M. F. M. J. S. S. &. S. I. Zaharia, "Spark: Cluster computing with working sets," *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing,* vol. 10, pp. 10-10, 2010.

[9] A. P. a. D. M. S. Ranjani, "Spark--an efficient framework for large scale data analytics," *International Journal of Scientific & Engineering Research,* 2016.

[10] H. K. A. W. P. &. Z. M. Karau, "Learning spark: lightning-fast big data analysis," *O'Reilly Media, Inc.*.

[11] M. C. M. J. F. S. S. a. I. S. Matei Zaharia, "Spark: cluster computing with working sets," *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing,* p. 10, 2010.

[12] A. P. a. M. S. Ranjani, "Spark-An Efficient Framework for Large Scale Data Analytics," *International Journal of Scientific & Engineering Research,* vol. 7, 2016.

[13] [Online]. Available: https://spark.apache.org/.

THE SYSTEM IDENTIFIES AND CLASSIFIES CEREBRAL HEMORRHAGE ON A LARGE DATA PROCESSING PLATFORM

Phan Anh Cang, Phan Thuong Cang, Pham Duy Khang, La Ngoc Nguyen, Tran Ho Dat

ABSTRACT: *Stroke is the third leading cause of death after cancer and cardiovascular disease in industrialized countries, especially in Vietnam. This is a disease that not only causes high mortality, but also is a leading risk of disability in all kinds of diseases. Because of the dangerous nature of cerebral hemorrhage, the diagnosis requires a quick and accurate diagnosis. Moreover, due to the increasing number of cases of cerebral hemorrhage, the need to deal with a large amount of data from many different hospitals is required. Therefore, building a system of automatically identifying brain hemorrhages with large data sets, fast processing times and high accuracy is essential. In this paper, we propose a method to use Hounsfield Unit; apply machine learning algorithms to identify and classify brain hemorrhage from CT / MRI images. Our proposed method is based on large data processing platform. Research results show that this method helps shorten a lot of time for doctors in diagnosing brain hemorrhage, thereby discovering the disease early and having timely treatment for patients. Experimental results of the proposed method achieve 98% accuracy with real-time proximity recognition speed.*