MỘT PHƯƠNG PHÁP XÂY DỰNG DỮ LIỆU CHO HỆ THỐNG HỌC SÂU TRONG CHẨN ĐOÁN MỘT SỐ BỆNH THÔNG THƯỜNG Ở TRẺ EM

Huỳnh Trung Trụ*, Tân Hạnh*
* Học Viện Công Nghệ Bưu Chính Viễn Thông cơ sở tại TP.HCM

Tóm tắt— Chẩn đoán ban đầu có vai trò quan trọng trong quá trình khám chữa bệnh. Nếu xác định được sớm trường hợp khám là có dấu hiệu bệnh nặng thì việc chữa trị sẽ gặp thuận lợi. Ngược lại, người khám sẽ không còn lo lắng hoặc chỉ cần khám tại các cơ sở ý tế nhỏ tại địa phương, tránh được sự lãng phí và cũng góp phần giảm tải cho bệnh viện trung tâm. Bài báo này đề xuất phương pháp dùng các mô hình học sâu cho việc chẩn đoán ban đầu giúp nhận định bệnh. Phương pháp mà bài báo đề xuất ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên đối với tiếng Việt trong việc xây dựng kho dữ liệu huấn luyện hệ thống học sâu từ các bệnh án điện tử. Kết quả thử nghiệm với mô hình CNN, LSTM và CNN-LSTM kết hợp là khá tốt khi nhận định 3 loại bệnh phổi, tiêu hóa, da liễu.

Từ khóa- Kho ngữ liệu, Học sâu, phân lớp, CNN, Mạng Neural Network, y tế, khám bệnh.

I. GIỚI THIỆU

Xây dựng một hệ thống hỗ trợ chăm sóc sức khỏe hoặc thăm khám bệnh tự động là mong muốn đã có từ lâu. Với sự phát triển của khoa học máy tính, và nhất là sự phát triển của các phương pháp học sâu, nhu cầu này càng trở nên được quan tâm hơn và cũng càng có cơ sở thành công hơn. Trên thế giới đã có nhiều công trình nghiên cứu về lĩnh vực này [1]. Các công trình này nghiên cứu ứng dụng từ nhiều lĩnh vực khác nhau của khoa học máy tính như thị giác máy tính, nhận dạng giọng nói cũng như xử lý ngôn ngữ tự nhiên cho tiếng Anh. Những công trình xử lý bài toán chuyên sâu theo chuyên ngành hẹp như [5] và [6] đòi hỏi công sức rất lớn và cũng thu được những kết quả rất tốt.

Việc thu thập kiến thức và hiểu biết từ dữ liệu y sinh phức tạp, nhiều chiều và không đồng nhất vẫn là một thách thức quan trọng trong việc xây dựng kho dữ liệu để huấn luyện các hệ thống deep learning. Nhiều loại dữ liệu khác nhau đã và đang xuất hiện trong nghiên cứu y sinh hiện đại, bao gồm hồ sơ sức khỏe điện tử, hình ảnh, dữ liệu cảm biến … Đặc điểm chung của các loại dữ liệu này là phức tạp, không đồng nhất, chú thích kém và nói chung là không có cấu trúc. Việc xử lý các dữ liệu này đòi hỏi nền tảng kiến thức miền đầy đủ.

Tác giả liên hệ: Huỳnh Trung Trụ,
Email: truht@ptithcm.edu.vn
Đến toà soạn: 10/2020, Chỉnh sửa:11/2020, Chấp nhận đăng:12/2020

Nhiều khái niệm và mối quan hệ đang nằm trong các dữ liệu y tế như: các tóm tắt xuất viện, các kết quả xét nghiệm, các công trình nghiên cứu khoa học… Những dữ liệu này được tạo ra liên tục hằng ngày và đang lưu trữ với nhiều dạng khác nhau như: âm thanh, hình ảnh và văn bản. Cụ thể, văn bản tường thuật (clinical arratives) chứa nhiều khái niệm đề cập đến các điều kiện lâm sàng, các vị trí giải phẫu trên cơ thể, các loại thuốc được sử dụng trong quá trình điều trị và những thủ tục (thủ thuật). Việc rút trích các khái niệm và mối quan hệ giữa chúng là cơ sở nền tảng để phát triển các ứng dụng như: tìm kiếm thông tin, hỏi đáp, tóm tắt văn bản và hệ thống hỗ trợ ra quyết định. Nhiều hình thức mặt chữ (surface forms) biểu diễn cùng khái niệm, cho nên việc rút trích và ánh xạ những khái niệm xuất hiện trong tài liệu văn bản đến những thuật ngữ đã được định nghĩa trong các từ vựng hoặc ontology (hay gọi là chuẩn hóa) nhằm giúp cho người dùng dễ dàng nhận biết và hiểu được các khái niệm và mối quan hệ một cách dễ dàng.

Trong lĩnh vực y học có nhiều nguồn tài nguyên từ vựng và ontology phong phú, có thể được tận dụng để nhận diện các khái niệm và liên kết các khái niệm hoặc chuẩn hóa. Một trong những nguồn tài nguyên đó là UMLS (Unified Medical Language System), nó chứa trên 130 từ vựng (lexicons/thesauri) với các thuật ngữ từ nhiều ngôn ngữ khác nhau, trong đó UMLS Metathesaurus tích hợp những nguồn tài nguyên chuẩn như: SNOMED-CT, ICD9 và RxNORMđược sử dụng rộng rãi trên thế giới trong chăm sóc lâm sàng, y tế cộng đồng và dịch tễ học. Ngoài ra, UMLS cũng cung cấp một mạng ngữ nghĩa, trong đó mỗi khái niệm trong Metathesaurus được biểu diễn bởi một ký hiệu nhận dạng duy nhất khái niệm (CUI - Concept Unique Identifier) và được phân loại ngữ nghĩa [16].

Trong phần tiếp theo của bài báo, các tác giả sẽ trình bày một số công trình liên quan ở mục 2. Mục 3 sẽ trình bày về phương pháp thực hiện từ quá trình xử lý dữ liệu đến các cấu hình của một số giải thuật học sâu dùng trong thử nghiệm của bài báo. Mục 4 các tác giả sẽ trình bày kết quả đạt được và các ý kiến thảo luận. Các tác giả sẽ trình bày những ý kiến kết quận và hướng phát triển tiếp dựa trên kết quả đạt được từ bài báo này trong mục 5.

II. CÁC CÔNG TRÌNH LIÊN QUAN

Trong lĩnh vực y khoa, việc ứng dụng trí tuệ nhân tạo đã được phát triển từ lâu. Với sự phát triển của các giải

thuật học sâu thì lĩnh vực này càng đó điều kiện phát triển, nhất là với các bài toán thuộc lĩnh vực thị giác máy tính (computer vision).

Ở công trình [10] các tác giả giới thiệu một mô hình học sâu phân loại trẻ em khỏe mạnh hoặc có khả năng mắc chứng tự kỷ. Mô hình các tác giả sử dụng là CNN kết hợp với mô hình MobileNet. Kết quả đạt được rất tốt, độ chính xác đạt 94,6%. Trong khi đó, Amjad Rehman [11] và các cộng sự phân loại bệnh bạch cầu mãn tính dòng tế bào lympho sử dụng mô hình CNN phân loại ảnh chụp tế bào đạt độ chính xác 97.78%. Ở bài báo [12] các tác giả sử dụng mô hình học sâu trong chẩn đoán ký sinh trùng đường ruột ở người, tác giả sử dụng mạng nơ-ron tính chập ConvNet với độ chính xác 96.49%. Trong bài báo [13] tác giả phát hiện và chẩn đoán sâu răng bằng cách sử dụng thuật toán mạng nơ-ron CNNs dựa trên mô hình học sâu, với độ chính xác 95%.

Các công trên đạt được độ chính xác rất cao khi giải quyết bài toán xác định một loại bệnh cụ thể.

Các giải thuật học sâu trong lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt cũng được dùng trong nhiều công trình, nhất là cho lĩnh vực phân loại ý kiến đánh giá sản phẩm như [14] và [15]. Kết quả đạt được ở các công trình này cũng khá tốt, độ chính xác trên 80%. Ở công trình [14] các tác giả đã thử nghiệm phân loại ý kiến cho tiếng Anh và tiếng Việt để nhận thấy các giải thuật học sâu không phụ thuộc vào ngôn ngữ trong lĩnh vực xử lý ngôn ngự tự nhiên. Vấn đề sử dụng các giải thuật học sâu cho lĩnh vực xử lý ngôn ngữ tự nhiên là xây dựng kho ngữ liệu đầy đủ và chất lượng để các giải thuật học sâu học tốt tri thức của lĩnh vực cần xử lý.

III. PHƯƠNG PHÁP TIẾP CẬN

Các bệnh án điện tử có nhiều yếu tố như số đo huyết áp, thân nhiệt, hoặc các chỉ trong xét nghiệm … là những giá trị có ý nghĩa quan trọng thuộc về chuyên ngành khoa học sức khỏe. Nếu chỉ xem các giá trị này như một từ hoặc cụm từ thông thường sẽ dẫn đến chẩn đoán hoặc nhận định sai trong khám chữa bệnh. Tuy nhiên, ngoài những chỉ số kết quả của quá trình khám cận lâm sàng có tính chuyên môn sâu về y khoa, các thông số của quá trình khám tổng quát như chiều cao, cân nặng, huyết áp … không đòi độ chính xác cao. Đây là các thông số cơ bản góp phần vào nhận định phân loại bệnh trong giai đoạn đầu của quá trình khám chữa bệnh. Trong bài báo này các tác giả trình bày phương pháp tiếp cận xử lý các thông tin ban đầu này thành cơ sở tri thức nhằm khai thác khả năng của các hệ thống học sâu cho mục đích hỗ trợ phân loại một số bệnh ở giai đoạn đầu của quá trình khám chữa bệnh.

3.1 Tiền xử lý dữ liệu

3.1.1 Xây dựng kho dữ liệu

Dữ liệu mà các tác giả thu thập là các bệnh án điện tử một số bệnh viện và phòng khám tư nhân. Quá trình xử lý tạo kho dữ liệu được thực hiện theo các bước:

Bước 1: Rút trích dữ liệu theo từng ca khám và kết luận của các bác sỹ.

Bước 2: Tạo văn bản cho mỗi ca khám bệnh. Mỗi ca tạo thành một văn bản. Mỗi câu trong văn bản là một thông tin theo khía cạnh như tiền sử bệnh, chẩn đoán, kết luận.

Điều quan trọng trong thông tin bệnh án là chẩn đoán và kết luận của bác sĩ. Vì đây là thông tin gán nhãn của mẫu dữ liệu. Các bệnh án không có thông tin chẩn đoán và kết luận của bác sĩ sẽ bị loại bỏ. Các thông tin khác có thể bị khuyết.

Ví dụ:

"*17 tháng, cao 120 cm, nặng 16 kg, biểu hiện lâm sàng sốt, ho, ngủ ly bì, thở rít khi năm yên. Chẩn đoán khả năng viêm phôi. Kết luận viêm phổi nặng*"

Nhãn của dữ liệu này: *viem_phoi* (viêm phổi)

Như vậy, cấu trúc kho dữ liệu bệnh án này gồm:

- Thuộc tính xác định mỗi mẫu dữ liệu.
- Văn bản nội dung các mẫu dữ liệu bệnh án
- Nhãn bệnh cho mỗi bệnh án.

Sau quá trình xử lý như trên tác giả thu được một kho dữ liệu với số liệu như bảng 1.

Các loại bệnh được thu thập thử nghiệm trong bài bài báo này là: da liễu, tiêu hóa và bệnh liên quan đến phổi. Đây là các bệnh rất thường gặp ở trẻ em. Trong đó, bệnh da liễu là loại bệnh có triệu chứng thuộc dạng đa dạng và phức tạp nhất.

Bảng 3.1: Số liệu kho dữ liệu bệnh án bằng tiếng việt

Đặc tính	Số lượng
Số bệnh nhân	4027
Số văn bản	8791
Số loại nhãn (loại bệnh)	3 (da liễu, tiêu hóa, phổi)

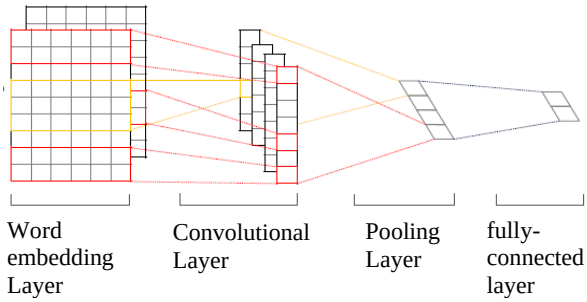
3.1.2 Tạo dữ liệu cho mô hình học sâu

Dữ liệu văn bản được chuyển đổi về dạng ma trận trọng số để sử dụng huấn luyện các mô hình học sâu. Bài báo này sử dụng công cụ word2vec [8] cho việc chuyển đổi này. Word2vec chứa mô hình Continuous Bag-of-Words (CBOW) và mô hình Skip-Gram [9]. Mô hình CBOW dự đoán từ mục tiêu (ví dụ: từ "mặc" có thể tìm ra khi dùng từ "kệ" nếu trong kho ngữ liệu hai từ này có mối quan hệ) từ các từ cùng ngữ cảnh với nó, trong khi mô hình Skip-Gram thực hiện ngược lại, dự đoán các từ ngữ cảnh được đưa ra từ mục tiêu.

3.2 Sơ lược về phương pháp học sau CNN và LSTM

3.2.1 CNN

CNN là một trong những mô hình học sâu tiên tiến giúp cho chúng ta xây dựng được những hệ thống xử lý thông minh, cho kết quả có độ chính xác cao. Mô hình CNN như hình 1 có các layer liên kết được với nhau thông qua cơ chế tích chập (convolution). Layer tiếp theo là kết quả tích chập từ layer trước đó. Nhờ vậy, ta có được các kết nối cục bộ. Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua tích chập (convolution) từ các bộ lọc.



Hình 1: Mô hình Convolutional Neural Network chuẩn [2]

Với bài toán xử lý ngôn ngữ tự nhiên, tầng Word embedding có thể được tạo từ công cụ word2vec. Tầng này gồm các ma trận kích thước *n* x *k*, biểu diễn câu có *n* từ, mỗi từ biểu diễn một vector k chiều. Lớp này mã hóa mỗi từ trong câu được chọn thành một vector từ. Đặt l ∈ R là chiều dài câu, $|D| \in R$ là kích thước từ vựng và $W^{(l)} \in R^{k \times |D|}$ là ma trận nhúng các vector từ k chiều. Từ thứ *i* trong câu được chuyển thành một vector *k* chiều w_i bằng công thức (1):

$$w_i = W^{(l)} x_i \quad (1)$$

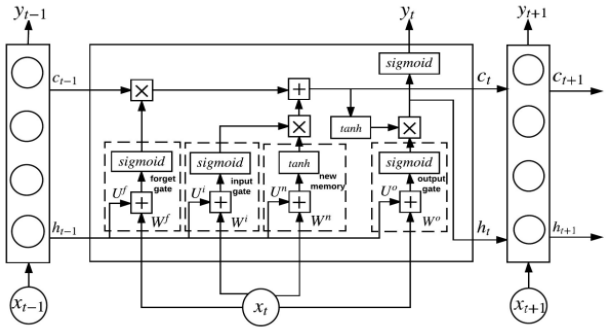
trong đó x_i là một biểu diễn one-hot vector cho từ thứ *i*.

Tầng Convolution sử dụng phép tích chập để xử lý dữ liệu bằng cách trượt cửa sổ trượt (slide windows) có kích thước cố định (còn gọi là kernel) trên ma trận dữ liệu đầu vào để thu được kết quả đã được tinh chỉnh. Trong khi đó, tầng Pooling tổng hợp các vector kết quả của tầng Convolution và giữ lại những vector quan trọng nhất.

Tầng full-connected đơn giản là một neural network truyền thống sử dụng những vector còn lại ở các lớp trên làm đầu vào để tạo ra kết quả cuối cùng thông qua quá trình huấn luyện.

3.2.2. LSTM

Mạng LSTM [7] thuộc nhóm phương pháp học sâu hồi quy (Recurrent Neural Networks – RNN). Mô hình mạng LSTM như ở hình 2. LSTM có các kết nối giữa các neural tạo thành dạng có hướng có tính chu kỳ và có khả năng học các phụ thuộc dài. Tất cả các RNN có dạng một chuỗi các module lặp lại. Trong các RNN tiêu chuẩn, mô đun lặp này thường có cấu trúc đơn giản. Tuy nhiên, module lặp trong LSTM thì phức tạp hơn. Thay vì có một tầng neural thì có bốn lớp tương tác theo một cách đặc biệt. Bên cạnh đó, nó có hai trạng thái: trạng thái ẩn và trạng thái tế bào (cell state). Hình 2 minh họa mô hình LSTM.



Hình 2: Mô hình Long Short Term Memory network [7]

Tại thời điểm bước t, LSTM trước tiên quyết định thông tin nào sẽ được đổ vào trạng thái tế bào. Quyết định này được đưa ra bởi một hàm sigmoid hoặc tầng σ, được gọi là cổng quên (forget gate). Hàm lấy h_{t-1} (đầu ra từ lớp ẩn trước đó) và x_t (đầu vào hiện tại) và xuất ra một số trong [0, 1], trong đó 1 có nghĩa là *giữ hoàn toàn* và 0 có nghĩa là *bỏ qua hoàn toàn* trong công thức (2)

$$f_t = \sigma(W^f x_t + U^f h_{t-1}) \quad (2)$$

Sau đó LSTM quyết định những thông tin mới sẽ lưu trữ trong trạng thái tế bào. Việc này gồm hai bước. Đầu tiên, một hàm hay lớp sigmoid, được gọi là cổng đầu vào như ở công thức (3), quyết định giá trị nào LSTM sẽ cập nhật. Tiếp theo, một hàm hoặc lớp *tanh* tạo ra một vectơ các giá trị ứng viên mới $\tilde{C}$.

$$i_t = \sigma(W^i x_t + U^i h_{t-1}) \quad (3)$$

$$\tilde{C} = \tanh(W^n x_t + U^n h_{t-1}) \quad (4)$$

Tiếp theo, cập nhật trạng thái tế bào cũ C_{t-1} vào trạng thái tế bào mới C_t như công thức (5). Cổng quên f_t có thể kiểm soát độ dốc đi qua nó và cho phép xóa và cập nhật bộ nhớ một cách tường minh, giúp giảm bớt sự hao hụt của độ dốc hoặc làm bùng nổ về độ dốc trong RNN tiêu chuẩn.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

LSTM quyết định đầu ra dựa trên trạng thái tế bào. Trước tiên, LSTM chạy một lớp sigmoid, quyết định phần nào của trạng thái tế bào sẽ xuất ra trong công thức (6), được gọi là ngõ ra (output gate). Sau đó, LSTM đặt trạng thái tế bào vào hàm *tanh* và nhân nó với đầu ra của cổng sigmoid, để LSTM chỉ xuất ra các phần mà nó quyết định như công thức (7).

$$o_t = \sigma(W^0 x_t + U^0 h_{t-1}) \quad (6)$$

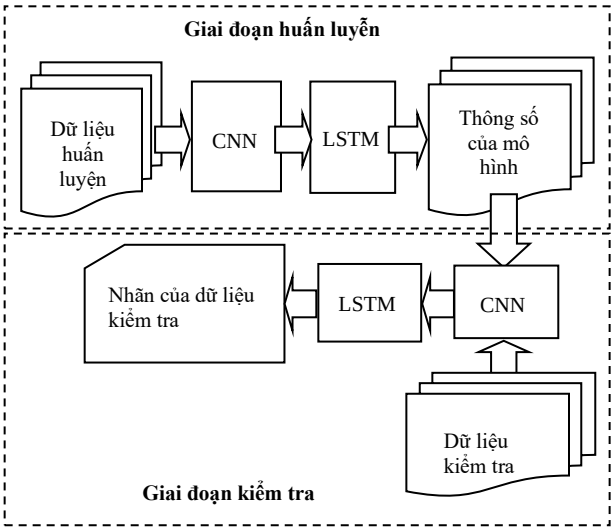
$$h_t = o_t * \tanh(C_t) \quad (7)$$

LSTM thường được áp dụng cho dữ liệu tuần tự nhưng cũng có thể được sử dụng cho dữ liệu có cấu trúc cây.

3.2.3 Mô hình CNN-LSTM

Phương pháp LSTM có thể làm việc hiệu quả với loại dữ liệu tuần tự có kích thước lớn. Với đặc trưng của loại dữ liệu bệnh án có các chỉ số có tính chuyên môn sâu. Đặc trưng này của dữ liệu sẽ phù hợp với mô hình tích chập của CNN như đã trình bày ở mục 3.2.1. Bài báo thử nghiệm kết hợp mô hình CNN và LSTM nhằm khai thác ưu điểm của mỗi mô hình trong vấn đề học đặc trưng của dữ liệu y tế. Mô hình kết hợp này được thể hiện trong Hình 3.

Tầng tích chập (Convolutional layer) của mạng CNN tạo ra một tập các vector đặc trưng của đối tượng. Số vectơ đặc trưng bằng số bộ lọc được sử dụng trong quá trình tích chập. Trong lớp tổng hợp số chiều (Pooling layer), các giá trị giá trị đặc trưng tốt nhất từ mỗi lớp sẽ được chọn để thu được đặc trưng quan trọng nhất của văn bản. Các vectơ đặc trưng qua mạng CNN được kết nối đầy đủ (Full connected layer) tạo ra một tập hợp các tham số ở đầu ra của mạng CNN. Bộ LSTM sử dụng các tham số đầu ra của CNN để thực hiện quá trình phân loại các văn bản.



Hình 3: Mô hình kết hợp CNN và LSTM [14]

IV. THỰC NGHIỆM

4.1. Cấu hình các mô hình học sâu

a) LSTM

Dựa trên thư viện Keras. Các thông số được chọn để thử nghiệm như liệt kê ở bảng 4.1.

Bảng 4.1: Thông số thử nghiệm mô hình LSTM

Đặc tính	Giá trị
Số neural ẩn	100, 200
dropout	0.2
Recurrent_dropout	0.2
Epoch	500
Kích thước embedding w	300
Hàm activation	sigmoid

b) CNN

Dựa trên thư viện Tensorflow. Các thông số được chọn để thử nghiệm như liệt kê ở bảng 4.2

Bảng 4.2: Thông số thử nghiệm mô hình CNN

Đặc tính	Giá trị
Kích thước embedding word	300
Số bộ lọc	300
Dropout	0.5
Epoch	500
L2	0.0008
Hàm activation	Sigmoid
Kích thước bộ lọc	3,4,5

c) CNN – LSTM kết hợp

Dựa trên thư viện Keras. Các thông số được chọn để thử nghiệm như liệt kê ở bảng 4.3.

Bảng 4.3: Thông số thử nghiệm mô hình CNN + LSTM

Đặc tính	Giá trị
Epoch	500
LSTM	
Số bộ lọc	200
Hàm activation	softmax
CNN	
Kích thước embedding word	
Số bộ lọc	300
Kích thước bộ lọc	3
Pool size	2
Hàm activation	sigmoid

4.2 Kết quả thử nghiệm

Kết quả thử nghiệm trên bộ dữ liệu trình bày ở phần 3.1

Bảng 4.4: Độ chính xác (accuracy - %) của các thử nghiệm

Phương pháp	Da liễu	Tiêu hóa	Phổi	Tổng
CNN	61.57	67.43	66.99	65.42
LSTM	60.64	67.57	66.66	65.06
CNN-LSTM	68.73	73.60	71.64	71.38

Từ kết thu được về độ chính xác của các phương dùng trong thử nghiệm của bài báo này có thể rút ra một số nhận xét sau:

- Sự kết hợp giữa bộ CNN và bộ LSTM có sự cải thiện đáng kể về hiệu năng khi so với khi thực thi riêng từng giải thuật. Mức chênh lệch cao nhất lên đến trên 8% đối với loại nhãn bệnh da liễu. Như vậy, sự phức tạp của triệu chứng của bệnh da liễu, khi được chuyển qua mô hình ngôn ngữ, khiến cho giải thuật CNN và LSTM học không hiệu quả. Khi kết hợp hai mô hình này thì những ưu điểm của mỗi mô hình sẽ bổ sung cho nhau làm tăng khả năng học tri thức từ dữ liệu, như đã đề cập ở phần 3.2.3.
- Đối với kho dữ liệu thử nghệm trong bài báo này, kết quả thu được về độ chính xác của phương pháp CNN và LSTM tương đương nhau trong khả năng phân biệt cả ba nhãn bệnh cũng như trong đánh giá chung. Chiều dài lớn nhất của một mẫu dữ liệu trong thử nghiệm của bài báo này là 157 từ. Đây là kích thước không quá lớn để giải thuật LSTM thể hiện ưu điểm trong phân tích chuỗi dữ liệu dài. Tương tự, giải thuật CNN có thể chưa thể hiện được ưu điểm do kích thước bộ dữ liệu chưa đủ lớn, như bảng 3.1.
- Nhãn bệnh da liễu có kết quả thấp nhất. Điều này có thể lý giải là do các triệu chứng về da là rất đa dạng, khó phân biệt nếu không có sự hỗ trợ của quá trình khám cận lâm sàng. Một yếu tố có thể làm hạn chế độ chính xác của loại bện này là kích thước bộ dữ liệu. Với sự đa dạng về triệu chứng, các dệnh da liễu sẽ cần một lượng mẫu huấn luyện lớn hơn để thể hiện lượng tri thức của lĩnh vực phong phú hơn.
- Các kết quả đạt được tuy không cao, nhưng có thể nói là có nhiều triển vọng về việc ứng dụng các phương

pháp học sâu vào việc hỗ trợ phân loại ban đầu các bệnh nhân. Các kết quả có thể sẽ được cải thiện nếu lượng tri thức lĩnh vực được bổ sung cho phong phú hơn.

V. KẾT LUẬN

Kết quả thu được của bài báo này cho thấy phương pháp tiếp cận của bài báo là khá triển vọng. Mô hình nhận định bệnh của bài báo có ưu điểm là linh hoạt, dễ tiếp cận và sử dụng với nhiều đối tượng người nếu triển khai dưới dạng website hoặc ứng dụng di động. Tuy vậy, để có thể đánh giá đầy đủ sự hiệu quả của phương phái đề xuất của bài báo, cũng như có thể ứng dụng phương pháp này vào thực tế, thời gian tới các tác giả sẽ thu thập thêm dữ liệu cho nhiều loại bệnh hơn và thử nghiệm với nhiều mô hình học sâu khác.

TÀI LIỆU THAM KHẢO

[1] MIOTTO, Riccardo, et al. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, vol 19, isuue 6, 2018, pages 1236-1246.

[2] Yoon Kim, "Convolutional neural networks for sentence classification", in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, pages 1746-1751.

[3] FAUST, Oliver, et al. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, vol 161, 2018, pages 1-13.

[4] BEAM, Andrew L.; KOHANE, Isaac S. "Big data and machine learning in health care". *Jama*, vol 319, isuue 13, 2018, pages 1317-1318.

[5] WANG, Dayong, et al. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[6] LIU, Saifeng, et al. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. In: *Medical imaging 2017: computer-aided diagnosis*. International Society for Optics and Photonics, 2017. pages 1013428.

[7] Lei Zhang, Suai Wang, and Bing Liu (2018), "Deep learning for sentiment analysis: A survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol 8, Issue 4, 2018, page e1253.

[8] Xin Rong, "word2vec parameter learning explained", In *arXiv preprint arXiv:1411.2738*, 2014.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed representations of words and phrases and their compositionality". In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems* (NIPS 2013), 2013.

[10] Madison Beary, Alex Hadsell, Ryan Messersmith, Mohammad-Parsa Hosseini, "Diagnosis of Autism in Children using Facial Analysis and Deep Learning". *arXiv preprint arXiv:2008.02890*, 2020.

[11] Amjad Rehman, Naveed Abbas, Tanzila Saba, Syed Ijaz ur Rahman, Zahid Mehmood, HoshangKolivand. "Classification of acute lymphoblastic leukemia using deep learning". *Microscopy Research and Technique*, cil 81, isuue 11, 2018, pages 1310-1317.

[12] A.Z. Peixinho, S.B. Martins, J.E. Vargas and A.X. Falc˜ao, J.F. Gomes, C.T.N. Suzuki, "Diagnosis of Human Intestinal Parasites by Deep Learning". In: *Computational Vision and Medical Image Processing V: Proceedings of the 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing (VipIMAGE 2015, Tenerife, Spain*. 2015. pages 107.

[13] Jae-Hong Leea, Do-Hyung Kima, Seong-Nyum Jeonga, Seong-Ho Choib, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm". *Journal of dentistry*, vol 77, 2018, pages 106-111.

[14] Duy Nguyen Ngoc, Tuoi Phan Thi and Phuc Do, "Preprocessing Improves CNN and LSTM in Aspect-Based Sentiment Analysis for Vietnamese". In *Proceedings of Fifth International Congress on Information and Communication Technology. ICICT 2020. Springer, Singapore*, 2020. pages. 175-185

[15] Duy Nguyen Ngoc, Tuoi Phan Thi and Phuc Do, "A Data Preprocessing Method to Classify and Summarize Aspect-Based Opinions using Deep Learning", *Asian Conference on Intelligent Information and Database Systems. Springer, Cham*, 2019. pages 115-127

[16] BODENREIDER, Olivier; MCCRAY, Alexa T. "Exploring semantic groups through visual approaches". *Journal of biomedical informatics*, vol 36 issue 6, 2003, pages 414-432.

A METHOD OF BUILDING DATA FOR THE FOLLOWING SYSTEMS IN MEASURING A NUMBER OF COMMON DISEASES IN CHILDREN

Abstract— Initial diagnosis has an important role in the medical examination and treatment process. If the examination case is identified early as having serious illness, the treatment will be favorable. On the contrary, the examiners will no longer worry or just need to examine the small local medical facilities, avoiding waste and also helping to reduce the load of the central hospital. This paper proposes a method to use deep learning models for primary diagnosis to help identify diseases. The method that the paper proposes to apply natural language processing techniques to Vietnamese in building a database for training deep learning systems from electronic medical records. The test results with the model CNN, LSTM and CNN-LSTM combined are quite good when identifying 3 types of pneumonia, digestive, and dermatological diseases**.**

Keywords - Corpus, Deep Learning, classification, CNN, Convolution Neural Network, Healthcare, Medicine, Physical exam, Examination

LỜI CẢM ƠN

Trong quá trình thực hiện nghiên cứu tác giả cảm ơn NCS Nguyễn Ngọc Duy, công tác tại khoa Công nghệ thông tin 2, Học viện Công nghệ Bưu chính Viễn Thông cơ sở tại TP.HCM đã hỗ trợ. Bác sỉ chuyên khoa I Huỳnh Trung Quân, công tác tại bênh viện Đa Khoa Phúc Hưng Quảng Ngãi đã hỗ trợ.

SƠ LƯỢC TÁC GIẢ

Huỳnh Trung Trụ, Nhận học vị Thạc sỹ năm 2016. Hiện nay đang công tác tại khoa Công nghệ thông tin 2, Học viện Công nghệ Bưu chính Viễn thông cơ sở tại TP.HCM. Lĩnh vực nghiên cứu, học máy, khoa học dữ liệu, xử lý ngôn ngữ tự nhiên.

Tân Hạnh, Phó giám đốc Học Viện Công Nghệ Bưu Chính Viễn Thông cơ sở tại TP.HCM. Lĩnh vực nghiên cứu, học máy, truy xuất thông tin, khai phá dữ liệu.