ỨNG DỤNG CÔNG NGHỆ AI TRONG DỰ ĐOÁN NGUY CƠ MẮC UNG THƯ PHỔI Ở NGƯỜI

Vũ Đình Tuấn*, Nguyễn Đức Quân, Nguyễn Công Minh
Vũ Thu Phương, Phạm Thị Quỳnh Trang
Trường Đại học Công nghiệp Hà Nội

Tóm tắt

Trong những năm gần đây, trí tuệ nhân tạo (AI) đã trở thành một phần không thể thiếu trong cuộc cách mạng công nghiệp thứ tư. Khả năng của AI đã lan rộng vào nhiều lĩnh vực của cuộc sống con người, từ kinh tế, giáo dục, y học đến công việc hàng ngày, giải trí và thậm chí trong lĩnh vực quân sự. Bài báo này trình bày một ứng dụng cụ thể của trí tuệ nhân tạo trong việc dự đoán nguy cơ mắc bệnh ung thư phổi dựa trên việc phân tích ảnh chụp CT phổi. Trong nghiên cứu này, nhóm tác giả đề xuất một mô hình học máy sử dụng thuật toán mạng nơ ron tích chập (CNN) để huấn luyện trên tập dữ liệu về ung thư phổi từ IQ-OTH/NCCD. Kết quả mô phỏng cho thấy độ chính xác của dự đoán được cải thiện đạt đến 97,815 % và thiết kế ra một ứng dụng dự đoán nguy cơ mắc ung thư phổi ở người. Điều này chứng tỏ sức mạnh của AI không chỉ là một xu hướng mà còn là một công cụ mạnh mẽ trong việc cải thiện chẩn đoán và dự đoán bệnh tật, từ đó giúp nâng cao chất lượng cuộc sống và sức khỏe của con người. Việc áp dụng AI trong lĩnh vực y học không chỉ giúp tăng cường khả năng phát hiện sớm bệnh, mà còn giúp tối ưu hóa quá trình điều trị và giảm thiểu tác động của bệnh tật lên cơ thể con người.

Từ khoá: Trí tuệ nhân tạo (AI); Ung thư phổi; Mạng nơ ron tích chập (CNN).

Abstract

Applying AI technology to predict the risk of Lung cancer in humans

In recent years, Artificial Intelligence (AI) has emerged as a representative of the fourth industrial revolution. AI appears in many areas of human life such as economics, education, medicine, housework, and even military operations. This article presents an application of artificial intelligence in predicting the risk of Lung cancer in humans based on Lung CT scans. In this study, we propose a machine learning model using the Convolutional Neural Network (CNN) algorithm to train the Lung cancer dataset of IQ-OTH/NCCD. The simulation results show that the prediction accuracy is improved to 97.815 % and an application is designed to predict the risk of Lung cancer in humans. This proves that the power of AI is not just a trend but also a powerful tool in improving disease diagnosis and prediction, thereby helping to improve the quality of human life and health. Applying AI in the field of medicine not only enhances the ability to detect diseases early but also helps optimize the treatment process and minimize the impact of diseases on the human body.

Keywords: Artificial Intelligence (AI); Lung cancer; Convolutional Neural Network (CNN).

***Tác giả liên hệ, Email:** vdt.12012002@gmail.com
DOI: https://doi.org/10.63064/khtnmt.2024.562

1. Giới thiệu chung

Với mức độ ô nhiễm đang ngày càng tăng do tác nhân chính đó là con người. Môi trường đang bị ảnh hưởng nghiêm trọng chính bởi chất thải của con người và các nhà máy, xí nghiệp. Điều đó dẫn đến hàng loạt những hệ lụy xấu tới sức khỏe của con người. Thực tế nhất là ô nhiễm không khí, vấn đề đang ngày càng trở nên nghiêm trọng, đây chính là nguyên nhân chính dẫn tới những căn bệnh quái ác về hệ hô hấp của con người. Căn bệnh ung thư phổi là một mối đe dọa lớn đến sức khỏe của con người, nhất là hệ quả sau đại dịch COVID-19 [1].

Ung thư phổi đứng thứ 2 về tỉ lệ mắc mới và là nguyên nhân gây tử vong nhiều nhất trong số các bệnh ung thư trên toàn cầu. Tại Việt Nam, mỗi năm ghi nhận 26.262 ca mắc mới ung thư phổi và 23.797 ca tử vong vì căn bệnh này [1]. Trong thời đại công nghiệp 4.0, ứng dụng công nghệ trí tuệ nhân tạo (AI) trong các lĩnh vực như một bước cải tiến lớn trong khoa học công nghệ. Trí tuệ nhân tạo là lĩnh vực khoa học máy tính chuyên giải quyết các vấn đề nhận thức thường liên quan đến trí tuệ con người như học tập, sáng tạo và nhận diện hình ảnh. Các tổ chức hiện đại thu thập vô số dữ liệu từ nhiều nguồn khác nhau như cảm biến thông minh, nội dung do con người tạo ra, công cụ giám sát và nhật ký hệ thống. Mục tiêu của AI là tạo ra các hệ thống tự học có thể tìm ra ý nghĩa của dữ liệu. Sau đó, AI áp dụng kiến thức thu được để giải quyết các vấn đề mới theo cách giống như con người [12].

Trong y tế: AI hỗ trợ dự đoán bệnh tật, chẩn đoán nhanh chóng và chính xác hơn, hỗ trợ trong quản lý dữ liệu y tế và phát triển thuốc mới. Trong lĩnh vực tài chính: Việc tối ưu hóa quy trình giao dịch, phân tích dữ liệu tài chính phức tạp và dự đoán xu hướng thị trường cũng được thực hiện một cách nhanh chóng nhờ sự hỗ trợ từ trí tuệ nhân tạo. Trong sản xuất: AI giúp tăng cường tự động hóa và tối ưu hóa quy trình sản xuất, giảm thiểu lỗi lầm và tăng năng suất. Với lĩnh vực vận tải: AI cho phép thực hiện tối ưu hóa lộ trình vận chuyển, dự đoán và phòng tránh tai nạn, cải thiện quản lý hệ thống giao thông. Dự đoán mùa màng, tối ưu hóa sử dụng nguồn tài nguyên như nước và phân bón và giảm thiểu thiệt hại do thảm họa thiên nhiên cũng có những kết quả nhanh chóng và chính xác hơn nhờ việc ứng dụng AI trong nông nghiệp. Trong giáo dục: AI giúp tùy chỉnh hóa quá trình học tập, cung cấp phản hồi và hỗ trợ cá nhân hóa cho sinh viên, giáo viên và phụ huynh. Những ứng dụng này mang lại hiệu quả và tiết kiệm thời gian, tài nguyên, cũng như cải thiện chất lượng và hiệu suất làm việc. Nhóm chuyên gia tại Đại học Canterbury (New Zealand) công bố nghiên cứu cho thấy việc sử dụng AI có thể giúp các chuyên gia y tế xây dựng chiến lược điều trị ung thư hiệu quả hơn, từ đó giúp tăng khả năng cứu sống người bệnh. Đây là kết quả nghiên cứu trong vòng bốn năm do Phó Giáo sư Alex Gavryushkin tại Trung tâm Nghiên cứu toán sinh học thuộc Đại học Canterbury dẫn đầu [2].

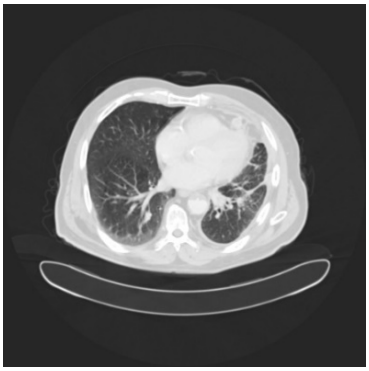
Trong bối cảnh tình trạng quá tải của các bệnh viện ở Việt Nam do số lượng bệnh nhân tăng lên không ngừng, sự cần thiết của việc áp dụng AI để hỗ trợ chuẩn đoán trở nên khẩn cấp và quan trọng hơn bao giờ hết. Nhằm mục đích phát hiện căn bệnh sớm và tìm ra các phương pháp điều trị phù hợp, nhóm nghiên cứu đã quyết

định tiến hành nghiên cứu và phát triển một ứng dụng AI dựa trên công nghệ tiên tiến, nhằm dự đoán nguy cơ mắc bệnh ung thư phổi ở người. Điều này sẽ giúp tăng cơ hội phát hiện sớm và cung cấp các biện pháp can thiệp chính xác và hiệu quả nhất để cứu sống và cải thiện chất lượng cuộc sống cho các bệnh nhân.

Nhóm tác giả đề xuất mô hình và phát triển thành một phần mềm dự đoán nguy cơ mắc ung thư phổi. Sử dụng mô hình Mạng nơ ron tích chập (CNN) để thực hiện các phép tích chập trên đầu vào từ ảnh chụp CT - cắt lớp với 96 bộ lọc có kích thước 3 × 3 và hàm kích hoạt ReLU. Phương pháp này đã cải thiện độ chính xác và nâng cao độ tin cậy trong việc dự đoán nguy cơ mắc ung thư phổi. Dựa vào mô hình này, nhóm nghiên cứu đã phát triển một phần mềm dự đoán nguy cơ mắc ung thư phổi từ ảnh chụp CT. Phần mềm này có khả năng phân loại ảnh thành một trong bốn kết quả: Bình thường, Lành tính, Ác tính và Không xác định.

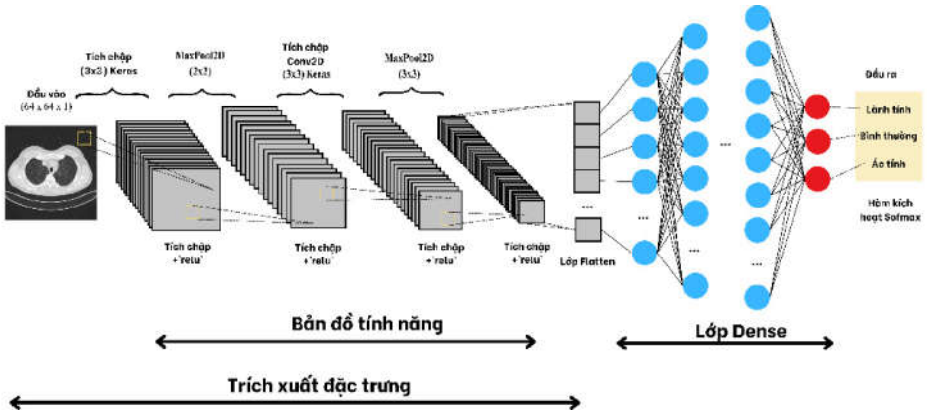
2. Bộ dữ liệu về ung thu phổi của IQ-OTH/NCCD

Bộ dữ liệu về ung thư phổi của Bệnh viện Giảng dạy Ung thư Iraq/Trung tâm Bệnh Ung thư Quốc gia Iraq (IQ-OTH/NCCD) đã được thu thập tại các bệnh viện chuyên khoa nói trên trong khoảng thời gian ba tháng vào mùa thu năm 2019. Bộ dữ liệu bao gồm ảnh chụp CT của bệnh nhân được chẩn đoán mắc bệnh ung thư phổi ở các giai đoạn khác nhau cũng như đối tượng khỏe mạnh. Các slide IQ-OTH/NCCD được đánh dấu bởi các bác sĩ ung thư và bác sĩ X-quang ở hai trung tâm này. Bộ dữ liệu chứa tổng cộng 1.190 hình ảnh đại diện cho các lát cắt CT của 110 trường hợp. Những trường hợp này được chia thành 3 loại: Bình thường, lành tính và ác tính. Trong đó có 40 trường hợp được chẩn đoán là ác tính; 15 trường hợp được chẩn đoán lành tính; 55 trường hợp được phân loại là trường hợp bình thường. Quét CT ban đầu được thu thập ở định dạng DICOM. Máy quét được sử dụng là SOMATOM của Siemens. Giao thức CT bao gồm: 120 kV, độ dày lát cắt 1 mm, với chiều rộng cửa sổ từ 350 đến 1.200 HU và tâm cửa sổ từ 50 đến 600 được sử dụng để đọc. Tất cả các hình ảnh đã được xác định lại trước khi thực hiện phân tích. Sự cho phép bằng văn bản đã được miễn trừ bởi hội đồng đánh giá giám sát. Nghiên cứu đã được phê duyệt bởi hội đồng đánh giá thể chế của các trung tâm y tế tham gia. Mỗi lần quét có chứa một số lát. Số lượng các lát này dao động từ 80 đến 200 lát, mỗi lát tượng trưng cho hình ảnh bộ ngực của con người với các cạnh và góc khác nhau. 110 trường hợp khác nhau về giới tính, độ tuổi, trình độ học vấn, khu vực cư trú và tình trạng sống. Một số người trong số họ là nhân viên của Bộ Giao thông vận tải và Dầu mỏ Iraq, những đối tượng khác như nông dân,… Hầu hết họ đến từ những nơi ở khu vực miền Trung Iraq, đặc biệt là các tỉnh Baghdad, Wasit, Diyala, Salahuddin và Babylon [6, 7, 10].



Hình 1: Hình ảnh quét CT mẫu

3. Mô hình học máy dự đoán nguy cơ mắc ung thư phổi



Hình 2: Mô hình học máy CNN

Các công thức ứng với các lớp:

Lớp Convolutional:

$$O_{i,j,k} = \sigma\left(\sum_{l=0}^{0}\sum_{m=0}^{2}\sum_{n=0}^{2} I_{i+m,j+n,l} \cdot K_{m,n,l,k} + b_k\right)$$

trong đó:

- $O_{i,j,k}$ là giá trị đầu ra tại vị trí (i, j) của kênh k.

- σ là hàm kích hoạt ReLU.

- $I_{i+m,j+n,l}$ là giá trị pixel tại vị trí (i+m, j+n) của kênh l trong ảnh đầu vào.

- $K_{m,n,l,k}$ là giá trị của bộ lọc tại vị trí (m, n) của kênh l và kênh k.

- b_k là độ lệch của kênh k.

Lớp Max Pooling:

$$O'_{i,j,k} = max(O_{2i:2i+2,2j:2j+2,k})$$

- Là giá trị tại vị trí (i, j) của kênh k trong đầu ra sau Max Pooling.

Lớp Dense:

$$z = \sigma(W.x + b)$$

trong đó:

- W là ma trận trọng số.

- x là đầu vào.

- b là độ lệch.

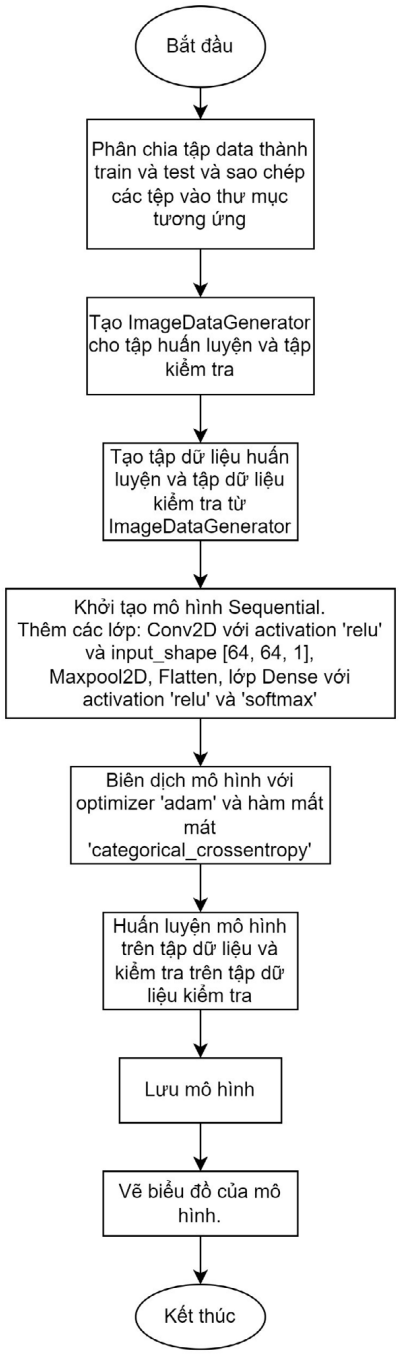
- σ là hàm kích hoạt ReLU.

Lớp Output:

$$z = \text{softmax}(W.x + b)$$

Lưu đồ thuật toán mô hình CNN được thể hiện trong Hình 3. Khi bắt đầu quá trình huấn luyện, chương trình sẽ chia tập dữ liệu thành 2 thư mục "train" và "test". Sau đó lớp ImageDataGenerator trong thư viện Keras được sử dụng để tạo ra các biến thể từ ảnh gốc. Lúc này mô hình Squential được khởi tạo nhằm thêm các lớp và sử dụng các hàm để thực hiện quá trình huấn luyện. Cuối cùng mô hình sẽ được lưu lại và được biểu diễn dưới dạng biểu đồ.

Sử dụng lớp ImageDataGenerator trong thư viện Keras để tạo ra các biến thể từ ảnh gốc. Lớp Conv2D: Thực hiện các phép tích chập trên đầu vào với 96 bộ lọc kích thước 3 × 3 và hàm kích hoạt ReLU. Mục đích là học các đặc trưng cấp cao từ dữ liệu đầu vào.



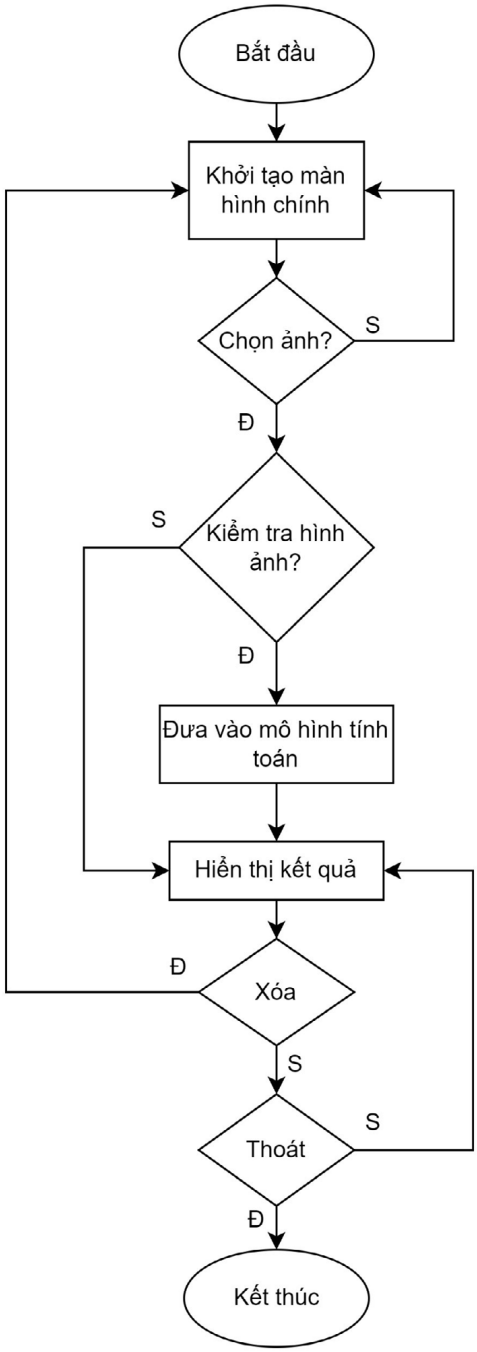
Hình 3: Lưu đồ thuật toán mô hình CNN

Lớp MaxPooling2D: Giảm kích thước của dữ liệu bằng cách giữ lại giá trị lớn nhất từ mỗi vùng, giúp giảm số lượng tham số và tăng tốc quá trình tính toán.

Optimizer (‘adam’): Là thuật toán để điều chỉnh trọng số của mô hình.

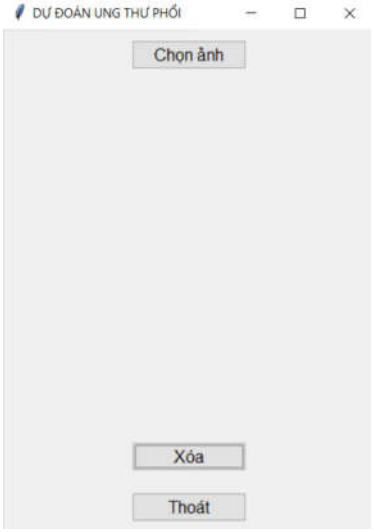
Loss Function (‘categorical_crossentropy’): Là hàm mất mát để đánh giá sự sai lệch giữa dự đoán của mô hình và nhãn thực tế.

Việc sử dụng ‘adam’ và ‘categorical_crossentropy’ giúp quá trình huấn luyện hiệu quả thông qua quá trình tối ưu hóa và đánh giá mất mát.



Hình 4: Lưu đồ thuật toán chương trình chính

Hình 5 là giao diện chính của chương trình ứng dụng dự đoán nguy cơ ung thư phổi được viết bằng ngôn ngữ Python, chạy trên máy tính dùng hệ điều hành Windows. Chương trình cho phép người dùng dễ dàng tiến hành dự đoán bằng cách chọn một bức ảnh từ máy tính cá nhân. Sau khi thực hiện dự đoán, kết quả được hiển thị một cách rõ ràng với các lớp quan trọng mà người dùng quan tâm: "Bình thường" cho những trường hợp không có dấu hiệu ung thư, "Lành tính" để những trường hợp có khả năng ung thư đang phát triển, "Ác tính" để những trường hợp nghiêm trọng và "Không xác định" cho các đầu vào không xác định.

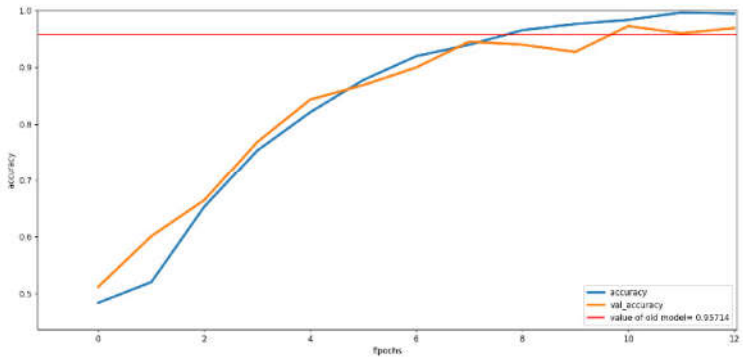


Hình 5: Giao diện màn hình chính

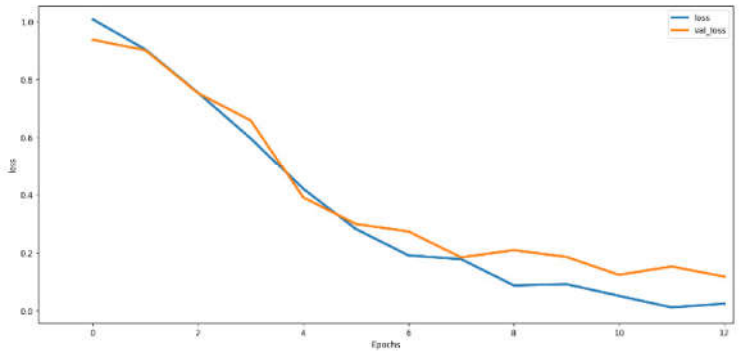
Người dùng bấm nút "Chọn ảnh" để được dự đoán về tình trạng phổi với 1 trong 4 kết quả: Bình thường, Lành tính, Ác tính và Không xác định như mô tả trong Hình 6.

Trong nghiên cứu, một mô hình AI sử dụng mạng nơ ron tích chập (CNN) được xây dựng bằng cách sử dụng thư viện TensorFlow và Keras trong Python để hình thành mô hình đề xuất. Bộ dữ liệu bao gồm 1.100 bản chụp CT ung thư phổi chia làm 3 loại: Trường hợp bình thường, trường hợp lành tính và trường hợp ác tính. Trong quá trình đào tạo, chia tập dữ liệu thành 2 nhóm, 50 % cho giai đoạn huấn luyện, 50 % cho giai đoạn thử nghiệm làm tăng tính khách quan của quá trình đánh giá hiệu suất của mô hình. Việc chia dữ liệu cho giai đoạn thử nghiệm giúp đảm bảo rằng mô hình không chỉ học thuộc lòng dữ liệu huấn luyện mà còn có khả năng tổng quát hóa tốt trên dữ liệu mới (thử nghiệm). Nếu thay đổi tỷ lệ chia dữ liệu có thể xảy ra một số vấn đề như mô hình không được huấn luyện đủ với đủ loại dữ liệu, có thể dẫn đến độ chệch trong việc dự đoán các lớp cụ thể hoặc làm thay đổi đột ngột hiệu suất của mô hình trên dữ liệu thử nghiệm. Sau khi kết thúc khóa đào tạo mà được thực hiện ngẫu nhiên, mô hình này cho kết quả sau 12 trong số 13 vòng lặp Epoch huấn luyện, độ chính xác tổng thể là 97,815 %.

Hình 6: Kết quả bình thường



Hình 7: Độ chính xác



Hình 8: Độ mất mát

Độ chính xác trên tập huấn luyện (training accuracy) là 99,64 %. Độ chính xác trên tập kiểm tra (validation accuracy) là 95,99 %. Độ chính xác cao một chỉ số tích cực, đồng nghĩa với việc mô hình này hiệu quả trong quá trình phân loại dữ liệu huấn luyện.

Độ mất mát của mô hình được thể hiện trên Hình 8. Mất mát trên tập huấn luyện (training loss) là 0,0114. Mất mát trên tập kiểm tra (validation loss) là 0,1525. Mất mát trên tập huấn luyện ở mức rất thấp, cho thấy mô hình đang thể hiện khả năng xuất sắc trong việc học từ dữ liệu huấn luyện. Mức mất mát này chứng minh rằng mô hình có khả năng dự đoán chính xác nhãn của dữ liệu huấn luyện với độ chính xác cao.

Mô hình Mạng nơ ron tích chập (CNN) này đã đạt được kết quả đáng chú ý sau khi được huấn luyện trên tập dữ liệu. Độ chính xác trên tập huấn luyện của mô hình đạt đến 99,64%, tăng cao đáng kể so với mô hình trước đó với độ chính xác chỉ là 93,548 %. Điều này cho thấy mô hình mới không chỉ có khả năng học tốt từ dữ liệu huấn luyện mà còn có khả năng dự đoán chính xác cao.

Với mức mất mát trên tập huấn luyện chỉ là 0,0114 và trên tập kiểm tra là 0,1525, mô hình đã cho thấy khả năng học tốt và khả năng tổng quát hóa tương đối tốt. Các kết quả này đều là những tiêu chí quan trọng trong việc đánh giá hiệu suất của mô hình và đồng thời cung cấp cái nhìn sâu sắc về khả năng dự báo của nó.

Nhìn chung, mô hình CNN mới đã mang lại sự cải thiện đáng kể so với mô hình trước đó và nhóm tác giả sẽ tiếp tục nghiên cứu và tối ưu hóa để đảm bảo hiệu suất cao và tính ổn định trong ứng dụng thực tế.

Bảng 1. So sánh với các nghiên cứu liên quan

Tác giả	Mô hình	Hiệu suất
Taher và cộng sự [9]	FCM & HNN	Độ nhạy: 83 % Độ đặc hiệu: 99 % Độ chính xác: 98 %
Dandil và cộng sự [4]	ANN	Độ nhạy: 92,3 % Độ đặc hiệu: 89,47 % Độ chính xác: 90,63 %
Diaz và cộng sự [5]	GA, SVM và ANN	Độ chính xác lần lượt: 95,87 % và 93,66 %
J. Lee và cộng sự [8]	CNN	Độ chính xác: 94,61%
Nhóm nghiên cứu	CNN	Độ chính xác: 97,815 %

Nhóm nghiên cứu đã đạt được mức độ chính xác cao hơn so với các nhóm Dandil, Diaz và Lee [4, 5, 8], lần lượt là khoảng 7,185 %; 4,155 % và 3,205 %. Điều này có ý nghĩa quan trọng trong việc dự đoán nguy cơ và tình trạng bệnh, đặc biệt khi sự chính xác trong dự đoán có thể có ảnh hưởng lớn đến quyết định điều trị và chăm sóc sức khỏe của bệnh nhân.

4. Kết luận

Bài báo đã trình bày một mô hình sử dụng thuật toán học máy mạng nơ ron tích chập (CNN) áp dụng cho dự đoán nguy cơ mắc ung thư phổi ở người dựa trên ảnh chụp CT phổi. Một chương trình ứng dụng đơn giản dự đoán ung thư phổi cũng được xây dựng cho phép người dùng đưa ảnh chụp CT phổi của mình vào để có được cảnh báo nguy cơ mắc bệnh của mình một cách nhanh chóng. Với độ chính xác 97,815 %. Khi kiểm tra (test) trên tập dữ liệu của IQ-OTH/NCCD cho thấy mô hình học máy được đề xuất phù hợp với ứng dụng này.

Nhóm nghiên cứu sẽ tiếp tục phát triển để cải thiện ứng dụng, đề xuất một số ý tưởng như tăng cường bộ dữ liệu để đạt được sự đa dạng và độ chính xác cao hơn, tối ưu hóa mô hình, giảm thiểu sự quá mức của mô hình đối với dữ liệu huấn luyện và tạo ra một mô hình mà tổng quát hóa tốt hơn cho dữ liệu mới và cải thiện giao diện người dùng để tạo trải nghiệm tốt nhất cho người sử dụng. Đồng thời, việc tích hợp các công nghệ mới như học máy giải thích và học máy tăng cường sẽ làm cho hệ thống trở nên mạnh mẽ và hiệu quả hơn.

TÀI LIỆU THAM KHẢO

[1]. Báo Điện tử Chính phủ (2023). *Tỉ lệ mắc mới ung thư phổi có xu hướng gia tăng*. Đăng tải ngày 13/8/2023. https://baochinhphu.vn/ti-le-mac-moi-ung-thu-phoi-co-xu-huong-gia-tang 102230813154337568.htm.

[2]. Hà Dung (2023). *Tiềm năng của AI trong điều trị y tế*. Báo nhân dân. Đăng tải ngày 09/9/2023. https://nhandan.vn/tiem-nang-cua-ai-trong-dieu-tri-y-te-post771481.html.

[3]. AL-Huseiny, Muayed (2023). *The IQ-OTH/NCCD Lung cancer dataset.* Mendeley data.

[4]. Dandıl, M. E. Emre Çakiroğlu, Ziya Özkan, Murat Kurt, Özlem Kar Canan and Arzu (2014). *Artificial neural networkbased classification system for Lung nodules on computed tomography scans.* 6th International Conference of soft computing and pattern recognition (SoCPaR), p. 382 - 386: IEEE.

[5]. Diaz, Joey Mark Pinon, Raymond Christopher Solano and Geoffrey (2014). *Lung cancer classification using genetic algorithm to optimize prediction models.* The 5th International Conference on Information,

Intelligence, Systems and Applications, p. 1 - 6, IEEE.

[6]. H. F. Al-Yasriy, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil and Z. S. Hassan (2020). *Diagnosis of Lung cancer based on CT scans using CNN.* IOP Conference Series: Materials Science and Engineering.

[7]. H. F. Kareem, M. S. A. Husieny, F. Y. Mohsen, E. A. Khalil and Z. S. Hassan (2021). *Evaluation of SVM performance in the detection of Lung cancer in marked CT scan dataset.* Indonesian Journal of Electrical Engineering and Computer Science.

[8]. J. Lee et al., *Lung cancer detection method using Convolution Neural Network.* University of Kyung Hee.

[9]. Taher, Fatma Sammouda and Rachid (2011). *Lung cancer detection by using artificial neural network and fuzzy clustering methods*. IEEE GCC Conference and Exhibition (GCC), p. 295 - 298.

[10]. https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset.

[11]. https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/.

[12]. https://aws.amazon.com/vi/what-is/artificial-intelligence/.

BBT nhận bài: 24/02/2024; Phản biện xong: 05/3/2024; Chấp nhận đăng: 28/3/2024