

Áp dụng mạng CNN trong phân loại quang phổ Raman của dung dịch glucose tinh khiết với bộ dữ liệu tăng cường

Nguyễn Thanh Tùng^{1,2*}, Ngô Quang Trí³

¹Trường Quốc tế, Đại học Quốc gia Hà Nội, 144 Xuân Thủy, phường Cầu Giấy, Hà Nội, Việt Nam

²Khoa Công nghệ thông tin, Trường Đại học Nguyễn Tất Thành, 300A Nguyễn Tất Thành, phường Xóm Chiếu, TP. Hồ Chí Minh, Việt Nam

³Trường Đại học Kinh tế - Kỹ thuật Công nghiệp, 456 Minh Khai, phường Vĩnh Tuy, Hà Nội, Việt Nam

Ngày nhận bài 16/5/2025; ngày gửi phản biện 19/5/2025; ngày nhận phản biện 9/6/2025; ngày chấp nhận đăng 13/6/2025

Tóm tắt:

Trong những năm gần đây, ứng dụng công nghệ học máy trong chẩn đoán tiểu đường theo phương pháp xâm lấn trở thành một lĩnh vực nghiên cứu quan trọng trong y học. Trong giai đoạn nghiên cứu hiện tại, chúng tôi tạo ra các mẫu dung dịch glucose với 14 mức khác nhau dựa trên nồng độ mol của glucose trong dung dịch, trước khi chiếu tia laser để thu thập quang phổ Raman. Dựa trên kết quả tích cực từ nghiên cứu trước, chúng tôi tái sử dụng phương pháp tăng quy mô bộ dữ liệu bằng kỹ thuật tạo dữ liệu ảo thêm nhiễu jitter. Từ đó một bộ dữ liệu mới được tạo ra, gồm dữ liệu thật và dữ liệu ảo có số mẫu lớn hơn gấp 100 lần. Mô hình mạng nơ-ron được sử dụng là mô hình tích chập một chiều với mục đích xác định mức glucose trong các mẫu. Kết quả cho thấy, kỹ thuật tạo mẫu giả thông qua phương pháp thêm nhiễu jitter nâng cao độ chính xác rõ rệt trong việc giảm thiểu hiện tượng quá khớp của mô hình học sâu, đồng thời đạt độ chính xác cao trong việc phân loại tín hiệu tán xạ Raman.

Từ khóa: chẩn đoán tiểu đường, mạng nơ-ron tích chập, nhân bản dữ liệu, nồng độ glucose, quang phổ Raman.

Application of jitter augmentation in classification of Raman scattering from glucose liquid with several machine learning models

Thanh Tung Nguyen^{1,2*}, Quang Tri Ngo³

¹International School, Vietnam National University - Hanoi, 144 Xuan Thuy Street, Cau Giay Ward, Hanoi, Vietnam

²300A Nguyen Tat Thanh Street, Xom Chieu Ward, Ho Chi Minh City, Vietnam

³University of Economics - Technology for Industrial, 456 Minh Khai Street, Vinh Tuy Ward, Hanoi, Vietnam

Received 16 May 2025; revised 9 June 2025; accepted 13 June 2025

Abstract:

The utilisation of machine learning technology for invasive diabetes diagnosis has emerged as a prominent research area in the medical field in recent years. The study employs Raman spectroscopy of a glucose fluid sample to ascertain glucose levels. We produce glucose-liquid samples with 14 varying ratios of glucose to pure water to replicate the 14 glucose concentrations

*Tác giả liên hệ: Email: tung_nt@vnu.edu.vn

found in human blood and acquire the corresponding Raman spectra. With the optimistic result of the previous study, we also design the jittering augmentation approach to enhance the dataset and the dataset size was increased by a factor of 100. We employ a 1-D Convolutional Neural Network to ascertain glucose levels in samples. The result proves the positive effect of jitter augmentation for improving the performance of the deep learning model as well as high accuracy of these models in classification of Raman scattering.

Keywords: artificial intelligence, convolutional neural network, glucose concentration, jitter augmentation, Raman scattering.

1. Đặt vấn đề

Việc phát hiện kịp thời bệnh tiểu đường có thể được thực hiện thông qua các phương pháp theo dõi glucose máu chỉ phí thấp. Hiện nay, phương pháp phát hiện đường huyết xâm lấn vẫn phổ biến do tính đơn giản, liên quan đến việc thu mẫu máu để phân tích in vitro. Các thiết bị theo dõi glucose truyền thống sử dụng phương pháp điện hóa, yêu cầu lấy một lượng máu xác định từ bệnh nhân thông qua việc chích ngón tay hoặc sử dụng kim cấy dưới da. Hầu hết những người mắc bệnh tiểu đường cảm thấy không thoải mái khi phải kiểm tra mức đường huyết thường xuyên, do cần lấy máu nhiều lần mỗi ngày để xét nghiệm.

Nghiên cứu này áp dụng chiến lược học sâu nhằm giải quyết khó khăn đó, kế thừa từ các nghiên cứu trước của chúng tôi. Mạng nơ-ron tích chập một chiều (viết tắt là 1-D CNN hoặc CNN) được sử dụng để phân tích mối tương quan giữa dữ liệu quang phổ Raman và mức glucose. Để giảm thiểu hiện tượng quá khớp (overfitting) trong quá trình huấn luyện, chúng tôi áp dụng chiến lược tăng cường dữ liệu bằng cách thêm nhiễu jitter nhằm tạo ra các mẫu tương đương, góp phần mở rộng tập dữ liệu. Kết quả thu được cho thấy hiệu quả rõ rệt, độ chính xác đạt trên 96%.

Bài báo này được tổ chức thành bốn phần chính: Giới thiệu, phương pháp, kết quả thực nghiệm và kết luận. Phần giới thiệu trình bày tổng quan về vấn đề nghiên cứu và tầm quan trọng của nó. Phần phương pháp mô tả các chiến lược và quy trình được sử dụng để xây dựng mô hình 1-D CNN và xử lý tập dữ liệu. Trong phần kết quả thực nghiệm, chúng tôi phân tích và thảo luận các

phát hiện của nghiên cứu, bao gồm hiệu quả của mô hình CNN. Phần kết luận tổng hợp các kết quả chính, đánh giá tác động của chúng và đề xuất các hướng nghiên cứu trong tương lai.

2. Nghiên cứu liên quan

Do những hạn chế của các phương pháp xét nghiệm đường huyết truyền thống, nhiều phương pháp không xâm lấn đã được phát triển nhằm cung cấp đánh giá nồng độ glucose một cách chính xác. Gần đây, quang phổ Raman đã được nghiên cứu như một công cụ để đánh giá mức glucose. Quang phổ Raman, dựa trên hiện tượng Raman được phát hiện vào năm 1928 [1], hoạt động như một kỹ thuật tán xạ. Tán xạ ánh sáng xảy ra do sự tương tác giữa các phân tử trong mẫu với tia laser đơn sắc. Ánh sáng bị tán xạ ở một tần số khác với tán xạ đàn hồi sẽ tạo nên phổ Raman, phản ánh sự tương tác không đàn hồi giữa phân tử mẫu và ánh sáng đơn sắc. Thiết bị sử dụng trong quang phổ Raman, gọi là máy quang phổ Raman, thường bao gồm bốn thành phần chính: nguồn phát laser, buồng chứa mẫu, buồng quang phổ sử dụng cách tử nhiễu xạ và bộ phận thu nhận tín hiệu. Quang phổ Raman có thể được ứng dụng để đo nhiều loại hợp chất hóa học khác nhau [2]. Một phát hiện đáng chú ý về mối liên hệ giữa dữ liệu phổ Raman thu được từ cơ thể người và mức đường huyết đã được trình bày trong [3]. Một số nghiên cứu trong những năm gần đây đã xây dựng chiến lược hoặc phương thức đo glucose không xâm lấn bằng phổ Raman trên người đã cho thấy tiềm năng ứng dụng. Nghiên cứu của I. Rehan và cs (2024) [4] đã áp dụng kỹ thuật học máy cơ bản là máy vector hỗ trợ (viết tắt là SVM) để phân tích dữ liệu phổ Raman từ nước tiểu để đánh giá xem người đó có

bị tiểu đường không, kết quả đạt được đáng kể là tỷ lệ chính xác lên tới 94%. Trong nghiên cứu của L.A. Duc và cs (2023) [5], bệnh tiểu đường tuýp 2 cũng đã được phát hiện thông qua sự kết hợp giữa phổ Raman và các kỹ thuật học máy. Tuy nhiên, thách thức trong việc sử dụng phổ Raman là đảm bảo hiệu quả dự đoán mức glucose một cách ổn định. Nhiều yếu tố có thể ảnh hưởng đến kết quả, chẳng hạn như cường độ tín hiệu Raman yếu và đặc điểm phức tạp của cơ thể người khi có cấu trúc dày đặc gồm các mô và chất lỏng [6]. Trước tình hình đó, một số nghiên cứu đã có những cách tiếp cận đơn giản hơn khi thu thập Raman từ dung dịch glucose thay vì thu trực tiếp trên da người, nhằm đánh giá hiệu quả học máy khi trích xuất glucose. Nghiên cứu của González-Viveros đề xuất giải pháp áp dụng Mạng nơ-ron tích chập để phân loại nồng độ đường trong dung dịch nước đường với độ chính xác cao [7].

Từ các nghiên cứu trước, nhóm nghiên cứu đã thực hiện toàn bộ các bước, từ thu thập dữ liệu đến đề xuất các mô hình học máy để giải quyết bài toán phân loại dựa trên dữ liệu phổ Raman có đặc trưng nồng độ glucose. Theo T.N. Quang và cs (2024) [8], các mẫu giả lập được tạo ra với giá trị mong muốn thay vì sử dụng mẫu sinh học, do gặp nhiều hạn chế trong việc thu thập mẫu theo cách không xâm lấn hoặc xâm lấn ở các vùng khác nhau trên cơ thể người [9], cũng như các vấn đề đạo đức liên quan đến thử nghiệm trên người.

Trong nghiên cứu trước, chúng tôi đã tiến hành toàn bộ quá trình từ thu thập dữ liệu đến đề xuất mô hình học máy để phân loại dữ liệu phổ Raman với đặc trưng nồng độ glucose [8]. Một máy quang phổ Raman kích thước lớn được sử dụng để thu thập dữ liệu từ vật liệu. Các mẫu được tổng hợp bằng cách pha trộn glucose tinh khiết với nước khử ion theo tỷ lệ chính xác nhằm mô phỏng mẫu máu có các mức glucose cụ thể trong môi trường phòng thí nghiệm. Chúng tôi thiết lập công suất laser ở mức 115 mW và cấu hình phạm vi đo để bao phủ bước sóng 792,3 nm trong 300 giây. Sau đó, chúng tôi nhỏ 3 ml dung dịch glucose vào vị trí cần đo để thu nhận phổ Raman. Tín hiệu tán xạ này được số hóa thành dữ liệu có thể tiền xử

lý và đưa vào mô hình học máy. Độ chính xác ở cấp mẫu đóng vai trò then chốt trong việc đánh giá khả năng dự đoán mức glucose của mô hình trong nghiên cứu này. Chúng tôi cũng sử dụng mô hình CNN 4 lớp cùng với các mô hình học máy đơn giản như SVM, Random Forest và Extra Trees. Tuy nhiên, do số lượng dữ liệu vẫn còn hạn chế, chúng tôi đã áp dụng kỹ thuật tăng cường dữ liệu jitter để tăng số lượng mẫu, qua đó tránh hiện tượng quá khớp trong các mô hình học máy phức tạp như CNN [10]. Kết quả cho thấy hiệu quả rõ rệt với độ chính xác cao trên tất cả các mô hình học máy, đặc biệt là mô hình CNN. Đáng chú ý, hiệu suất của mô hình khi sử dụng tập dữ liệu đã tăng cường cao hơn so với tập gốc, điều này cho thấy hiệu quả của kỹ thuật jitter augmentation đối với dữ liệu Raman từ mẫu tinh khiết. Do đó, trong bài báo này, chúng tôi áp dụng kỹ thuật jitter augmentation với quy mô lớn hơn và đề xuất một mô hình CNN có kích thước lớn.

3. Tạo dữ liệu ảo với thêm nhiễu jitter

Chúng tôi sử dụng tập dữ liệu Raman từ nghiên cứu trước được thu thập bằng thiết bị phổ kế quy mô lớn [8]. Tập dữ liệu gốc, được tạo ra từ các phép đo thực nghiệm, bao gồm 14 nhãn tương ứng với 14 mức nồng độ glucose từ 5,0 đến 7,5 mmol/l. Mặc dù khoảng giá trị glucose này không bao phủ toàn bộ dải chỉ số glucose của bệnh tiểu đường, mục tiêu của bài báo này cũng như nghiên cứu trước [8] là nhằm chứng minh tính khả thi của mô hình CNN và việc sử dụng các nguồn dữ liệu không lấy từ người. Do đó, phạm vi glucose giới hạn này không làm ảnh hưởng đến tính toàn vẹn của nghiên cứu. Các khoảng nồng độ glucose được gán nhãn một cách đều đặn, trung bình từ 0,1 đến 0,5 mmol/l. Mỗi mẫu chứa nhiều cặp dữ liệu bước sóng - cường độ trong phổ Raman, các cặp này được đánh chỉ số và sắp xếp theo thứ tự tăng dần của độ dịch bước sóng.

Trong nghiên cứu trước của T.N. Quang và cs (2024) [8], thuật toán tăng cường jitter được sử dụng để tạo ra các mẫu dữ liệu mới, dựa trên sự tương đồng giữa phổ tán xạ Raman và dữ liệu chuỗi thời gian mà đặc trưng của các nhãn glucose

được xác định bởi mối tương quan tỷ lệ giữa các giá trị cường độ liên kề [11]. Trong bài báo này, công thức tổng quát để tạo dữ liệu chuỗi thời gian mới bằng phương pháp jitter [12] được áp dụng cho phổ Raman như công thức (1), với một mẫu Raman thực $I = \{I_1, I_2, \dots, I_j, \dots, I_n\}$:

$$I(p_i) = \{I_1 + \epsilon_1, I_2 + \epsilon_2, \dots, I_j + \epsilon_j, \dots, I_n + \epsilon_n\} \quad (1)$$

trong đó: $I(p_i)$: mẫu ảo thứ i được tạo từ mẫu thật I ; I_j : Cường độ ở bước sóng thứ j ; n : Số bước sóng trong quang phổ Raman; ϵ_j : Giá trị nhiễu jitter ngẫu nhiên ở cường độ được thêm vào bước sóng thứ j .

Trong nghiên cứu của T.N. Quang và cs (2024) [8], kích thước dữ liệu sau khi sao chép bằng kỹ thuật tăng cường jitter đã tăng lên gấp 20 lần, từ 70 mẫu ban đầu lên thành 1400 mẫu. Tương ứng, mỗi mức glucose trong tập dữ liệu gốc bao gồm 5 mẫu, trong khi tập dữ liệu đã tăng cường chứa 100 mẫu gồm 5 mẫu gốc và 95 mẫu giả lập. Trong nghiên cứu này, kích thước dữ liệu được nhân lên 100 lần (từ 70 lên 7000 mẫu). Cụ thể, mỗi mức glucose trong tập dữ liệu tăng cường bao gồm 500 mẫu,

Bảng 1. Số lượng mẫu tương ứng với mỗi mức glucose với mỗi bộ dữ liệu.

Mức đường huyết	Số mẫu mỗi nhãn mức glucose		
	Bộ dữ liệu gốc	Bộ dữ liệu tăng cường	Bộ dữ liệu tăng cường ở nghiên cứu trước [8]
5,0	5	500	100
5,2	5	500	100
5,4	5	500	100
5,5	5	500	100
5,6	5	500	100
5,8	5	500	100
6,0	5	500	100
6,2	5	500	100
6,4	5	500	100
6,5	5	500	100
6,6	5	500	100
6,8	5	500	100
7,0	5	500	100
7,5	5	500	100

trong đó có 5 mẫu gốc và 495 mẫu giả lập. Bảng 1 trình bày số lượng mẫu tương ứng ở mỗi mức glucose trong từng tập dữ liệu.

Chúng tôi cũng lưu ý trong bảng 1 rằng, mã nhãn được sử dụng như một định danh trong quá trình huấn luyện và kiểm thử của mô hình học sâu. Do đặc trưng của bài toán phân loại, mức glucose của mỗi mẫu được xác định dựa trên mã định danh này thay vì giá trị thực tế của nó.

4. Thử nghiệm với mô hình học máy

4.1. Đề xuất mô hình CNN

Trong nghiên cứu trước [8], chúng tôi đã đề xuất một mô hình CNN gồm 4 lớp và kết quả thu được khá khả quan với tập dữ liệu đã được tăng cường. Trong nghiên cứu này, kích thước tập dữ liệu được tăng lên gấp 100 lần và dữ liệu của mỗi mẫu có kích thước rất lớn với 2030 giá trị, do đó, một mô hình CNN lớn hơn cũng được đề xuất. Cụ thể, một mạng 1-D CNN với 7 lớp nơ-ron được xây dựng, bao gồm 5 lớp tích chập, 1 lớp đầu vào và 1 lớp đầu ra.

Lớp đầu vào có nhiệm vụ tiếp nhận giá trị của mẫu dưới dạng mảng gồm 2030 phần tử, tương ứng với số lượng giá trị cường độ của phổ Raman. Lớp đầu ra có nhiệm vụ làm phẳng và trả về nhãn thuộc tập 14 nhãn. Các đặc tính của từng lớp nơ-ron được mô tả trong bảng 2.

Bảng 2. Mô hình CNN đề xuất.

Thứ tự lớp	Kích thước nhân	Số bộ lọc
Lớp 1	48	1152
Lớp 2	24	576
Lớp 3	12	144
Lớp 4	6	36
Lớp 5	3	9

Tương tự như trong nghiên cứu trước [8], lớp liên kế lớp đầu ra có kích thước kernel nhỏ hơn một cách tỷ lệ so với các lớp ở xa hơn, điều này là do cơ chế pooling vốn có trong mạng 1-D CNN. Giá trị của lớp trước đó (gần lớp đầu vào hơn) có kích thước kernel gấp đôi so với lớp liên kế phía sau, tức lớp ở gần lớp đầu ra hơn. Ngoài ra, để

nâng cao hiệu suất của mô hình CNN, số lượng bộ lọc tại mỗi lớp được xác định bằng bình phương của giá trị kích thước kernel tương ứng. Chúng tôi sử dụng hàm kích hoạt ReLU và áp dụng phương pháp chuẩn hóa L2 với hệ số điều chỉnh là 0,05 ở tất cả các lớp.

Chương này trình bày cấu trúc và nội dung của tập dữ liệu được sử dụng cho mô hình CNN, cùng với kiến trúc của chính mô hình CNN. Trong chương tiếp theo, chúng tôi sẽ trình bày quá trình triển khai mô hình CNN, quy trình huấn luyện và mô tả thiết lập thực nghiệm.

4.2. Quá trình huấn luyện và kiểm thử

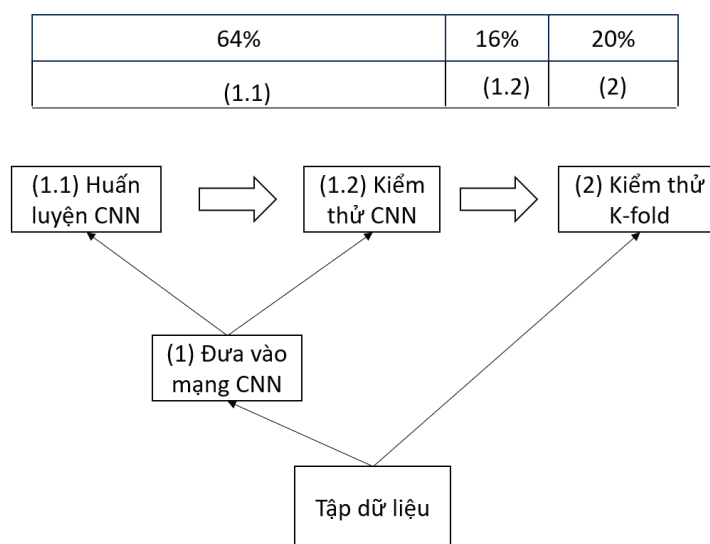
Yếu tố cốt lõi trong việc đánh giá mô hình CNN là xác thực cách mô hình học trong quá trình huấn luyện và hoạt động như thế nào trong quá trình kiểm thử. Ngoài ra, chúng tôi cũng so sánh hiệu suất của mô hình CNN. Do đó, một loạt thí nghiệm được thiết kế với phương pháp tạo tập dữ liệu riêng biệt cho huấn luyện và kiểm thử, lịch trình huấn luyện và kết quả từ quá trình kiểm thử.

Trong nghiên cứu trước [8], kỹ thuật kiểm định chéo K-fold được sử dụng, theo đó tập dữ liệu được chia thành 5 phần bằng nhau, trong đó một phần được chọn làm tập kiểm thử, bốn phần còn lại được dùng cho huấn luyện.

Với cách phân chia này, tập kiểm thử bao gồm 1330 mẫu được chia đều cho 14 mã nhãn, tức mỗi nhãn có 95 mẫu. Tương tự, tập huấn luyện có tổng

cộng 5320 mẫu, trong đó mỗi nhãn có 380 mẫu. Số lần lặp lại quá trình là 50. Để giảm thiểu hiện tượng quá khớp, chúng tôi áp dụng kiểm định chéo K-fold với thư viện Sklearn, một kỹ thuật trong đó mỗi phần được luân phiên làm tập kiểm thử, trong khi các phần còn lại được sử dụng cho huấn luyện. Ngoài ra, sau khi đưa vào mô hình học máy, quá trình chia dữ liệu cũng tự động tiến hành với thư viện Tensorflow. Hình 1 mô tả các bước thực hiện thí nghiệm gồm hai bước là: bước (1) đưa dữ liệu vào mạng 1-D CNN phục vụ việc huấn luyện, điều chỉnh trọng số các lớp tích chập và bước (2) là cho mạng CNN đã “được học” thông qua tối ưu hóa trọng số các lớp tích chập thí nghiệm với dữ liệu độc lập. Bản thân bước (1) cũng bao gồm nhiều đợt với mỗi đợt được thực hiện theo chu trình tuần hoàn là (1.1) huấn luyện và (1.2) kiểm thử. Lưu ý rằng kiểm thử ở bước (1.2) khác với kiểm thử K-fold ở bước (2). Kiểm thử bước (1.2) được thực hiện sau mỗi lần huấn luyện ở bước (1.1) và thực hiện nhiều lần theo từng đợt với mục đích tối ưu hóa trọng số lớp tích chập trong khi kiểm thử K-fold được thực hiện duy nhất một lần sau khi mô hình đã được tối ưu.

Như vậy, có thể thấy việc phân chia tập dữ liệu diễn ra hai lần, một lần thực hiện để phục vụ việc đánh giá hiệu năng theo cơ chế K-fold, một lần để mạng 1-D CNN điều chỉnh tham số sau huấn luyện. Với cơ chế K-fold, tập dữ liệu được chia thành 5 phần bằng nhau, trong đó một phần được



Hình 1. Phân bổ tập dữ liệu vào các bước trong mỗi lần vận hành mạng 1-D CNN.

chọn làm tập kiểm thử và bốn phần còn lại dùng cho huấn luyện. Như vậy dữ liệu phục vụ mạng 1-D CNN là 80%. Một lần nữa, tỷ lệ chia trong mạng 1-D CNN được thiết lập theo chuẩn là 0,2, tương tự tỷ lệ với cơ chế K-fold. Việc phân chia tập dữ liệu sẽ được trực quan hóa ở hình 1, với tỷ lệ 64% phục vụ huấn luyện CNN, 16% phục vụ kiểm thử CNN và 20% phục vụ kiểm thử K-fold. Lưu ý rằng, do kiểm thử và huấn luyện CNN được thực hiện nhiều lần trong một lần vận hành, dữ liệu phân chia vào hai tập kiểm thử và huấn luyện trong CNN sẽ được tự động sắp xếp, phân chia liên tục trong tập dữ liệu đưa vào mạng CNN (chiếm 80%) còn dữ liệu phục vụ kiểm thử K-fold sẽ là cố định.

4.3. Tiêu chí đo đạc

Trong phạm vi nghiên cứu, một số tiêu chí phổ biến trong lĩnh vực học máy được sử dụng để đánh giá như: độ chính xác (accuracy) và độ sai số. Trong các chỉ số này, độ chính xác là giá trị quan trọng nhất phản ánh hiệu suất của mô hình học máy với một tập dữ liệu cụ thể.

Trước khi tính toán độ chính xác, chúng tôi xác định cách thức đánh giá hiệu suất của mô hình. Trong bài toán phân loại, nhiệm vụ chính là mô hình sẽ dự đoán một nhãn cho một mẫu phổ Raman, sau đó kiểm tra kết quả bằng cách so sánh với mã nhãn thực tế. Do đó, kết quả đầu ra của mô hình sẽ có hai trạng thái: đúng hoặc sai.

Độ chính xác là tỷ lệ giữa số lần dự đoán đúng trên tổng số lần phân loại, được tính theo công thức (2):

$$Accuracy (\%) = \frac{N_T \times 100\%}{N_T + N_F} \quad (2)$$

trong đó: N_T : Số lần phân loại đúng; N_F : Số lần phân loại sai.

Với tiêu chí hàm sai số, chúng tôi sử dụng hàm mất mát entropy dạng Categorical để đo độ sai khác giữa phân phối xác suất dự đoán của mô hình đối với các lớp và phân phối thực tế của nhãn. Hàm mất mát này được tính theo công thức (3), nhằm đo lường mức độ khác biệt giữa hai phân phối p và q [13]:

$$H(p, q) = - \sum_i p_i \log(q_i) \quad (3)$$

Trong phạm vi nghiên cứu, độ chính xác được tính theo công thức (2) và (3) cho từng vòng lặp trong tổng số 50 vòng lặp, trước khi tính giá trị trung bình cuối cùng.

4.4. Kết quả

Bảng 3 trình bày giá trị trung bình của hai tiêu chí đánh giá: tỷ lệ chính xác tính theo công thức (2) và độ sai số tính theo công thức (3) của mô hình CNN với 5 lớp tích chập trên ba loại tập dữ liệu: tập dữ liệu gốc, tập dữ liệu đã tăng cường trong nghiên cứu trước [8] và tập dữ liệu đã tăng cường trong nghiên cứu này.

Bảng 3. Kết quả đo đạc và đối sánh.

CNN	Tập dữ liệu	Độ chính xác (%)	Độ mất mát
4 lớp tích chập [8]	Raw [8]	90,08	9,53
	Modest Enriched [8]	95,03	8,51
5 lớp tích chập	Tập dữ liệu thô	88,73	11,02
	Modest Enriched [8]	95,92	8,16
	Large Enriched	96,81	7,87

Kết quả bảng 3 cho thấy, mô hình CNN với 5 lớp tích chập gặp phải hiện tượng underfitting khi xử lý dữ liệu quy mô rất nhỏ như tập dữ liệu gốc. Khi áp dụng mô hình CNN lớn hơn lên tập dữ liệu có quy mô trung bình được sử dụng trong nghiên cứu trước [8], độ chính xác của mô hình có cải thiện nhẹ so với mô hình CNN nhỏ hơn trong nghiên cứu đó, song song với đó là độ sai số thấp hơn. Đối với tập dữ liệu đã được tăng cường ở quy mô lớn trong nghiên cứu này, kết quả thu được rất tích cực, với độ chính xác ở mức cao nhất là 96,81% và độ sai số ở mức thấp nhất là 7,87.

5. Kết luận

Chúng tôi đánh giá tính khả thi của mô hình mạng nơ-ron tích chập một chiều (1-D CNN) trong việc phân loại các mức nồng độ glucose cụ thể dựa trên dữ liệu phổ Raman thu được từ dung dịch glucose tinh khiết.

Trong quá trình thu thập dữ liệu, chúng tôi tạo ra một tập gồm năm mẫu, mỗi mẫu chứa dữ liệu của 14 mức nồng độ glucose, dao động từ 5,0 đến 7,5 mmol/l. Tuy nhiên, kích thước mẫu như vậy

là chưa đủ để giảm thiểu hiện tượng quá khớp thường gặp trong các mô hình học sâu. Vì vậy, chúng tôi áp dụng kỹ thuật tăng cường dữ liệu bằng jitter để tạo ra các mẫu mô phỏng. Tổng số mẫu được tạo ra sau quá trình tăng cường dữ liệu là 7000 mẫu, với mỗi mức glucose được đại diện bởi 500 mẫu.

Trong nghiên cứu này, chúng tôi hoàn toàn kiểm soát toàn bộ quy trình tạo tập dữ liệu bao gồm việc tạo mẫu vật liệu, thu thập phổ Raman thông qua tương tác giữa ánh sáng và vật liệu, cũng như thiết kế và triển khai thuật toán tăng cường dữ liệu jitter.

Về thiết kế mô hình học sâu, chúng tôi triển khai mạng CNN một chiều với bốn lớp và sử dụng hàm mất mát dạng entropy phân loại (categorical entropy). Kỹ thuật kiểm định chéo K-fold được sử dụng để chia tập dữ liệu đã tăng cường thành các phần huấn luyện và kiểm thử. Mô hình được huấn luyện trên tập huấn luyện và đánh giá trên tập kiểm thử qua 50 lần lặp lại.

Chúng tôi sử dụng thư viện TensorFlow với ngôn ngữ Python để xây dựng mô hình CNN, và sử dụng thư viện scikit-learn (sklearn) với Python để triển khai kiểm định chéo K-fold. Việc đánh giá hiệu suất mô hình được thực hiện dựa trên bốn tiêu chí, trong đó có tỷ lệ chính xác và giá trị hàm mất mát. Mặc dù có một vài bất ổn trong một số vòng lặp, chúng tôi vẫn tiếp tục thử nghiệm mô hình CNN với tập dữ liệu thu được từ người thật vốn đặc trưng bởi mức nhiễu cao và không có giá trị glucose xác định rõ ràng.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Bộ Khoa học và Công nghệ theo mã số đề tài DTDL.CN-25/23. Các tác giả xin chân thành cảm ơn.

TÀI LIỆU THAM KHẢO

[1] C.V. Raman, K.S. Krishnan (1928), “A new type of secondary radiation”, *Nature*, **121**(3048), pp.501-502, DOI: 10.1038/121501c0.

[2] J. Xu, Q. He, Z. Xiong, et al. (2021), “Raman spectroscopy as a versatile tool for investigating thermochemical processing of coal, biomass, and wastes: Recent advances and future perspectives”, *Energy & Fuels*, **35**(4), pp.2870-2913, DOI: 10.1021/acs.energyfuels.0c03975.

[3] J. Shao, M. Lin, Y. Li, et al. (2012), “In vivo blood glucose quantification using Raman spectroscopy”, *PLOS ONE*, **7**(10), DOI: 10.1371/journal.pone.0048127.

[4] I. Rehan, S. Khan, R. Ullah (2024), “Raman spectroscopy assisted support vector machine: A steadfast tool for noninvasive classification of urinary glucose of diabetes mellitus”, *Physica Scripta*, **99**(2), DOI: 10.1088/1402-4896/ad1da8.

[5] L.A. Duc, N.T. Tung, T.T. Oanh, et al. (2023), “Non-invasive in vivo type 2 diabetes mellitus diagnosis using Raman spectroscopy in combination with machine learning”, *Mobile Networks and Applications*, DOI: 10.1007/s11036-023-02220-9.

[6] J. Smulko, M.S. Wróbel, I. Barman (2015), *Noise in Biological Raman Spectroscopy*, Book Noise in Biological Raman Spectroscopy, pp.1-6.

[7] N.G. Viveros, P.G. Gil, J.C. Ramos, et al. (2021), “On the estimation of sugars concentrations using Raman spectroscopy and artificial neural networks”, *Food Chemistry*, **352**, DOI: 10.1016/j.foodchem.2021.129375.

[8] T.N. Quang, T.T. Nguyen, H.P.T. Viet (2024), “Machine learning approach for early detection of diabetes using Raman spectroscopy”, *Mobile Networks and Applications*, **29**, pp.294-305, DOI: 10.1007/s11036-024-02340-w.

[9] M. Shokrehodaie, D.P. Cistola, R.C. Roberts, et al. (2021), “Non-invasive glucose monitoring using optical sensor and machine learning techniques for diabetes applications”, *Ieee Access: Practical Innovations, Open Solutions*, **9**, pp.73029-73045, DOI: 10.1109/access.2021.3079182.

[10] A. Gavrilov, A. Jordache, M. Vasdani, et al. (2018), “Preventing model overfitting and underfitting in convolutional neural networks”, *International Journal of Software Science and Computational Intelligence*, **10**(4), pp.19-28, DOI: 10.4018/IJSSCI.2018100102.

[11] P. Esling, C. Agon (2012), “Time-series data mining”, *ACM Computing Surveys (CSUR)*, **45**(1), DOI: 10.1145/2379776.2379788.

[12] G. Iglesias, E. Talavera, A.G. Prieto, et al. (2023), “Data augmentation techniques in time series domain: A survey and taxonomy”, *Neural Computing and Applications*, **35**(14), pp.10123-10145, DOI: 10.1007/s00521-022-08122-y.

[13] C.H. Chen, P.H. Lin, J.G. Hsieh, et al. (2020), “Robust multi-class classification using linearly scored categorical cross-entropy”, *2020 International Joint Conference on Neural Networks (IJCNN)*, DOI: 10.1109/IJCNN48605.2020.9207212.