BỘ GIÁO DỤC VÀ ĐÀO TẠO
VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

NGUYỄN THỊ KIM SƠN

NGHIÊN CỨU ỨNG DỤNG MỘT SỐ MÔ HÌNH SỬ DỤNG HỌC SÂU TRONG DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA NGƯỜI HỌC

TÓM TẮT LUẬN ÁN TIẾN SĨ MÁY TÍNH
Ngành: Hệ thống thông tin
Mã số: 9 48 01 04

Hà nội - 2025

Luận án được hoàn thành tại: Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Tập thể hướng dẫn khoa học

1. Người hướng dẫn 1: PGS.TS. Nguyễn Hữu Quỳnh
2. Người hướng dẫn 2: PGS. TS. Ngô Quốc Tạo

Phản biện 1: PGS. TS. Bùi Thu Lâm

Phản biện 2: TS. Nguyễn Như Sơn

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi 9 giờ 00, ngày 8 tháng 10 năm 2025

Luận án có thể được tìm thấy tại:

1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

GIỚI THIỆU

1. Giới thiệu chung

Sự bùng nổ của khoa học dữ liệu và trí tuệ nhân tạo (AI) trong giáo dục đã mở ra cơ hội mới để phân tích và khai thác thông tin từ dữ liệu học tập, từ đó cải thiện hiệu quả dạy và học trong bối cảnh chuyển đổi số. Trong đó, một ứng dụng nổi bật là dự đoán kết quả học tập của sinh viên dựa trên dữ liệu thu thập trong quá trình học, giúp phát hiện sớm các nguy cơ thất bại và triển khai các biện pháp can thiệp kịp thời. Phương pháp này trực tiếp hỗ trợ mục tiêu giáo dục hiện đại, bao gồm cá nhân hóa trải nghiệm học tập và nâng cao tỷ lệ tốt nghiệp.

Hiện nay, nhiều nghiên cứu vẫn sử dụng các mô hình học máy (machine learning) truyền thống như hồi quy tuyến tính, hồi quy logistic, SVM, cây quyết định, KNN hay Naive Bayes. Các mô hình này có ưu điểm là đơn giản, dễ triển khai và dễ giải thích, nhưng lại hạn chế trong việc xử lý các mối quan hệ phi tuyến và phụ thuộc theo thời gian trong dữ liệu học tập. Thực tế, dữ liệu học tập thường mang tính chuỗi, phản ánh sự tiến bộ của sinh viên qua thời gian, trong khi các mô hình học máy hoặc thống kê truyền thống chủ yếu dựa vào đặc trưng tĩnh như điểm cuối kỳ.

Bên cạnh đó, kết quả học tập của sinh viên chịu ảnh hưởng từ nhiều yếu tố đa chiều, bao gồm đặc điểm cá nhân (giới tính, thói quen học tập, việc làm thêm, khả năng chi trả học phí), nền tảng gia đình (trình độ học vấn của cha mẹ), cũng như kết quả đầu vào như điểm thi tốt nghiệp THPT, điểm thi các tổ hợp môn xét tuyển (Toán – Hóa – Sinh, Văn – Sử – Địa, v.v.), và điểm tiếng Anh. Bên cạnh đó, hình thức tuyển sinh (xét tuyển học bạ, điểm thi tốt nghiệp, ưu tiên xét tuyển...) cũng là một yếu tố quan trọng ảnh hưởng đến mức độ phù hợp và khả năng thích nghi của sinh viên với môi trường đại học. Ngoài ra, các yếu tố bối cảnh như điều kiện cơ sở vật chất, chính sách học bổng, chất lượng giảng dạy và mức độ hỗ trợ từ nhà trường cũng góp phần định hình kết quả học tập. Việc các yếu tố này có mối liên hệ phức tạp và phi tuyến khiến các mô hình truyền thống khó nắm bắt đầy đủ, đòi hỏi phải áp dụng các phương pháp phân tích hiện đại hơn như học máy, học sâu để dự đoán chính xác hơn.

Học sâu (Deep learning) nổi lên như một giải pháp đầy hứa hẹn nhờ khả năng tự động học biểu diễn dữ liệu và phát hiện các mẫu phức tạp mà không cần thiết kế đặc trưng thủ công. Các kiến trúc như LSTM và Transformer đặc biệt phù hợp với dữ liệu chuỗi, giúp phân tích quá trình học tập của sinh viên theo thời gian. Tuy

nhiên, thách thức đặt ra là các mô hình học sâu thường yêu cầu dữ liệu huấn luyện lớn, trong khi dữ liệu giáo dục lại nhỏ lẻ, phân tán và thiếu thống nhất giữa các hệ thống.

Một hướng tiếp cận tiềm năng là áp dụng các mô hình tiền huấn luyện (pre-trained models) hoặc kỹ thuật học truyền (transfer learning), vốn đã chứng minh hiệu quả vượt trội trong các lĩnh vực như thị giác máy tính và xử lý ngôn ngữ tự nhiên. Tuy nhiên, trong bối cảnh nghiên cứu giáo dục, rào cản lớn hiện nay là sự thiếu hụt các bộ dữ liệu chuẩn hóa và mô hình tiền huấn luyện được thiết kế chuyên biệt cho lĩnh vực này. Cho đến nay, cộng đồng nghiên cứu vẫn chưa xây dựng được một cơ sở dữ liệu chung hoặc hệ thống mô hình tiền huấn luyện có thể tái sử dụng rộng rãi trong các bài toán học thuật liên quan đến giáo dục.

Để giải quyết vấn đề này, luận án lựa chọn các mô hình học sâu làm nền tảng, kết hợp với các kỹ thuật như tăng cường dữ liệu, chọn lọc đặc trưng và tối ưu siêu tham số. Đồng thời, phát triển mô hình lai (kết hợp deep learning với machine learning truyền thống, hoặc kết hợp nhiều kiến trúc deep learning) là một hướng đi triển vọng, vừa tận dụng sức mạnh biểu diễn dữ liệu, vừa cải thiện tính giải thích của mô hình. Mục tiêu là xây dựng các mô hình có khả năng xử lý dữ liệu học tập theo chuỗi, tích hợp các yếu tố cá nhân, học thuật và xã hội, đồng thời duy trì hiệu quả dự đoán ngay cả khi dữ liệu bị giới hạn. Nghiên cứu này góp phần nâng cao phân tích học tập (Learning Analytics), hỗ trợ quá trình ra quyết định trong giáo dục đại học và thúc đẩy ứng dụng trí tuệ nhân tạo trong nghiên cứu giáo dục.

2. Mục tiêu nghiên cứu

Mục tiêu chung:

Nghiên cứu và phát triển các mô hình học máy và học sâu để phân tích dữ liệu giáo dục, nhằm dự đoán sớm kết quả học tập của sinh viên.

Mục tiêu cụ thể:

1) Đề xuất và so sánh hiệu năng của các mô hình học máy và học sâu hiện đại trong việc dự đoán kết quả học tập (ví dụ: điểm trung bình học kỳ (SGPA), phân loại tốt nghiệp), với trọng tâm là nâng cao độ chính xác và khả năng khái quát hóa.
2) Xây dựng các mô hình học sâu lai, thực hiện chọn lọc đặc trưng phù hợp và áp dụng các kỹ thuật tăng cường dữ liệu để khắc phục thách thức của bộ dữ liệu giáo dục quy mô nhỏ và không đồng nhất.

Việc đánh giá thực nghiệm sẽ được tiến hành trên các bộ dữ liệu huấn luyện thu thập từ các trường đại học và cao đẳng trong nước cũng như quốc tế.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Các bài toán dự đoán sớm kết quả học tập của sinh viên có thể được phân loại thành nhiều dạng cụ thể khác nhau, tùy theo mục tiêu và phạm vi phân tích. Cụ thể:

- Bài toán dự đoán sớm điểm số: bao gồm dự đoán sớm điểm trung bình học kỳ (semester GPA), điểm trung bình năm học, điểm trung bình tích lũy, điểm từng học phần, kết quả học phần ngắn hạn, điểm đánh giá thường xuyên, v.v.
- Bài toán phân loại sớm: bao gồm dự đoán sớm xếp loại học tập theo từng học phần, từng học kỳ, giai đoạn học tập, hoặc xếp loại tốt nghiệp.

Các nhiệm vụ dự đoán này đóng vai trò quan trọng trong hệ thống cảnh báo sớm học tập, giúp cơ sở đào tạo phát hiện sớm sinh viên có nguy cơ trượt học phần, phải học lại, hoặc không thể tốt nghiệp đúng hạn. Đồng thời, kết quả dự đoán hỗ trợ đề xuất các biện pháp can thiệp nhằm cải thiện kết quả học tập của sinh viên, cũng như cung cấp cơ sở dữ liệu định lượng giúp nhà quản lý giáo dục ra quyết định chính xác hơn.

Trong khuôn khổ luận án này, nghiên cứu tập trung vào hai dạng bài toán dự đoán cốt lõi: Dự đoán sớm điểm trung bình học kỳ (SGPA) và Dự đoán sớm xếp loại tốt nghiệp (graduation classification).

Từ đây về sau, thuật ngữ "kết quả học tập" trong luận án này được hiểu một cách cụ thể là "điểm trung bình học kỳ" hoặc "xếp loại tốt nghiệp".

Ngoài ra, luận án còn xem xét các đối tượng nghiên cứu ở cấp độ mô hình, bao gồm: các thuật toán học máy (KNN, DT, SVM, LR, RF) làm cơ sở; các kiến trúc học sâu (DNN, CNN, RNN, LSTM, Transformer, GNN/GCN/GAT) cho dữ liệu tuần tự và quan hệ; và các mô hình lai và nâng cao (NeutroDL, NeutroGNT, LATCGAd, AWG-GC) để giải quyết các tập dữ liệu nhỏ, mất cân bằng và không chắc chắn.

Phạm vi nghiên cứu: Các mô hình học máy và học sâu hiện đại, bao gồm cả các kiến trúc lai (hybrid models).

Các bộ dữ liệu được thu thập từ Trường Đại học Thủ đô Hà Nội (HNMU), Đại học Quốc gia Hà Nội (VNU), cùng với một số bộ dữ liệu quốc tế công khai nhằm mục đích tham khảo và đối sánh.

Dữ liệu sử dụng trong nghiên cứu bao gồm: Hồ sơ điểm số của sinh viên, thu thập từ hệ thống quản lý đào tạo của các trường đại học. Dữ liệu khảo sát về các yếu tố liên quan đến sinh viên, như thông tin cá nhân, sở thích, nền tảng học tập

trước khi vào đại học, hoàn cảnh gia đình, và các yếu tố xã hội-nghề nghiệp có thể ảnh hưởng đến kết quả học tập. Dữ liệu ở cấp độ cơ sở đào tạo đại học, bao gồm thông tin về cơ sở vật chất, chương trình đào tạo, giảng viên, v.v.

4. Phương pháp nghiên cứu

Nghiên cứu này kết hợp giữa nghiên cứu lý thuyết, tổng quan tài liệu, khảo sát điều tra và nghiên cứu thực nghiệm.

5. Những đóng góp chính của luận án

1. Đề xuất hai mô hình mới: mô hình NeutroDL và NeutroGNT, tích hợp quá trình neutrosophic vào các mô hình học sâu nhằm nâng cao hiệu quả dự đoán sớm điểm trung bình học kỳ (SGPA).
2. Đề xuất hai mô hình lai mới là LATCGAd và AWG-GC phục vụ dự đoán sớm xếp loại tốt nghiệp của sinh viên.
3. Xây dựng ba bộ dữ liệu đa thuộc tính từ nhiều nguồn khác nhau và đề xuất các khung phân tích phù hợp cho dữ liệu giáo dục.

Dưới góc độ hệ thống thông tin, theo nghĩa kiến trúc tích hợp của dữ liệu, phần mềm, phần cứng, con người và quy trình phối hợp để thu thập, xử lý và cung cấp thông tin cho việc ra quyết định, luận án có những đóng góp sau: (i) Phát triển và chuẩn hóa các bộ dữ liệu giáo dục nhằm hỗ trợ Khai phá dữ liệu giáo dục (EDM) và Phân tích học tập (LA). (ii) Thiết kế quy trình xử lý dữ liệu chặt chẽ để đảm bảo chất lượng dữ liệu và độ tin cậy của mô hình. (iii) Ứng dụng các khung học sâu tiên tiến để xây dựng và tối ưu hóa các mô hình dự đoán. (iv) Tận dụng hạ tầng CPU và GPU để xử lý dữ liệu và phân tích theo thời gian thực. (v) Đặt sinh viên ở vị trí trung tâm, đồng thời cung cấp các thông tin phân tích dựa trên dữ liệu nhằm hỗ trợ giảng viên và nhà quản lý trong việc nâng cao chất lượng giảng dạy và hoạch định chính sách. (vi) Tích hợp các thành phần của hệ thống thông tin để xây dựng một hệ thống phân tích giáo dục thông minh, thích ứng, hướng tới mô hình quản lý giáo dục đại học dựa trên dữ liệu.

6. Bố cục của luận án

Luận án được trình bày với cấu trúc bao gồm: phần mở đầu, ba chương chính, phần kết luận và hướng phát triển, danh mục các công trình đã công bố, cùng tài liệu tham khảo, cụ thể như sau:

Phần Mở đầu trình bày ý nghĩa khoa học và tính cấp thiết của đề tài, lý do lựa chọn đề tài nghiên cứu. Đồng thời nêu rõ mục tiêu, đối tượng, phạm vi, phương pháp nghiên cứu, những đóng góp chính và nội dung nghiên cứu của luận án.

Chương 1 cung cấp tổng quan về phân tích dữ liệu giáo dục, nhấn mạnh ứng dụng của học máy và học sâu trong dự đoán kết quả học tập của sinh viên. Chương này cũng tổng quan các công trình liên quan để xác định động cơ nghiên cứu, đồng thời giới thiệu ba bộ dữ liệu chính (HNMU1, HNMU2, VNU) thu thập từ Trường Đại học Thủ đô Hà Nội và Đại học Quốc gia Hà Nội, là cơ sở thực nghiệm cho các mô hình được phát triển ở các chương sau.

Chương 2 tập trung vào dự đoán điểm trung bình học kỳ (SGPA) bằng các mô hình học sâu kết hợp với lý thuyết Neutrosophy nhằm xử lý tính bất định của dữ liệu. Các mô hình như DNN, CNN, RNN, LSTM và Transformer được triển khai trong môi trường neutrosophic (Neutrosophic DLs) để dự đoán GPA của học kỳ kế tiếp từ dữ liệu học tập trong quá khứ. Để nâng cao hiệu quả, chương này đề xuất mô hình lai NeutroGNT, tích hợp quá trình neutrosophic hóa dữ liệu, sinh dữ liệu bằng CGAN, tiêm nhiễu và Transformer, giúp cải thiện độ chính xác và khả năng thích ứng trong điều kiện bất định.

Chương 3 chuyển hướng sang bài toán dự đoán xếp loại tốt nghiệp của sinh viên, một nhiệm vụ dài hạn và mang tính hệ thống. Chương này giới thiệu hai mô hình LATCGAd và AWG-GC, tận dụng các mô hình dựa trên đồ thị (Graphformer), GANs nâng cao (CGAN, WGAN) và Autoencoder, kết hợp với AdaLN để đảm bảo ổn định, nhằm xử lý dữ liệu ít và mất cân bằng. Các mô hình này vừa mở rộng dữ liệu, vừa nâng cao hiệu quả dự đoán, đem lại độ chính xác, độ tin cậy và khả năng mở rộng cao hơn cho các hệ thống phân tích giáo dục.

Phần Kết luận và Hướng phát triển tổng hợp các kết quả đã đạt được, rút ra những kết luận chính, đồng thời đề xuất các hướng nghiên cứu tiếp theo dựa trên các phát hiện của luận án.

Danh mục công trình đã công bố: Luận án liệt kê 08 bài báo của tác giả, cùng cộng sự, đã được công bố hoặc chấp nhận đăng trên các tạp chí và kỷ yếu hội thảo khoa học trong và ngoài nước.

Tài liệu tham khảo: Cuối luận án là danh mục các tài liệu tham khảo được sử dụng trong quá trình nghiên cứu.

7. Tổng quan nội dung chính

Ngoài chương 1 phân tích tổng quan và giới thiệu bài toán, tập dữ liệu, Chương 2 và Chương 3 của luận án tạo thành một cấu trúc thống nhất, trình bày hai cách tiếp cận bổ sung cho nhau trong dự đoán sớm kết quả học tập của sinh viên dựa trên dữ liệu học tập và các yếu tố khảo sát. Chương 2 giải quyết bài toán hồi quy nhằm

dự đoán điểm trung bình học kỳ (SGPA), một chỉ báo liên tục phản ánh kết quả học tập ngắn hạn. Chương 3 tập trung vào bài toán phân loại nhằm dự đoán sớm xếp loại tốt nghiệp, một kết quả rời rạc mang tính dài hạn. Hai nhiệm vụ này có mối liên hệ chặt chẽ, bởi kết quả GPA của nhiều học kỳ là đầu vào quan trọng cho mô hình dự đoán sớm xếp loại tốt nghiệp. Do đó, dự đoán GPA sớm giúp nâng cao độ chính xác cho bài toán phân loại sau này.

Về phương pháp luận, Chương 2 giới thiệu các mô hình học sâu (DNN, LSTM, Transformer) kết hợp với kỹ thuật xử lý bất định (Neutrosophy) và tăng cường dữ liệu (CGAN), đặt nền tảng cho Chương 3. Chương 3 kế thừa và phát triển các mô hình mở rộng như LATCGAd và AWG-GC, tích hợp CGAN, WGAN, Graphformer và Autoencoder để xử lý dữ liệu phân loại vốn phức tạp và mất cân bằng. Các chương này có sự gắn kết chặt chẽ cả về sự phụ thuộc dữ liệu lẫn quy trình mô hình hóa tiến triển, được thiết kế phù hợp với đặc thù của từng nhiệm vụ dự đoán.

8. Ý nghĩa của luận án

Ý nghĩa học thuật: Nghiên cứu này đóng góp vào sự phát triển của lĩnh vực Khai phá dữ liệu giáo dục (Educational Data Mining - EDM) thông qua việc tích hợp các mô hình học sâu vào hệ thống thông tin giáo dục. Các mô hình được đề xuất cho bài toán dự đoán điểm trung bình học kỳ (SGPA) và xếp loại tốt nghiệp, được huấn luyện trên dữ liệu thực tế với độ chính xác cao, tạo nền tảng khoa học vững chắc cho việc ứng dụng trí tuệ nhân tạo trong phân tích hành vi học tập của sinh viên.

Ứng dụng thực tiễn: Các kết quả nghiên cứu của luận án có khả năng ứng dụng cao trong quản lý giáo dục, đặc biệt là ở các khía cạnh sau:

- Cá nhân hóa học tập: hỗ trợ cố vấn học tập và xây dựng lộ trình học tập phù hợp cho từng sinh viên;
- Phát hiện sớm sinh viên có nguy cơ: cho phép cơ sở đào tạo triển khai các biện pháp can thiệp kịp thời;
- Ra quyết định dựa trên dữ liệu: hỗ trợ công tác quy hoạch, đánh giá và hoạch định chính sách giáo dục.

Đóng góp ở tầm hệ thống: Luận án minh chứng cho việc tích hợp công nghệ học sâu với các thành phần cốt lõi của hệ thống thông tin giáo dục (dữ liệu , phần cứng, phần mềm, con người, quy trình), hướng tới xây dựng một môi trường học tập thông minh, thích ứng và hiệu quả trong kỷ nguyên AI.

Kết quả của luận án đã được trình bày tại:

1. Seminar FS&IS, Trường Công nghệ Thông tin và Truyền thông, Trường Đại học Công nghiệp Hà Nội.
2. Hội nghị VNICT, 2024.
3. Hội nghị MCO, 2025.

CHƯƠNG 1. TỔNG QUAN VỀ DỰ ĐOÁN KẾT QUẢ HỌC TẬP DỰA TRÊN CÁC TIẾP CẬN HỌC MÁY VÀ HỌC SÂU

Chương này trình bày bối cảnh và động lực nghiên cứu (Mục 1.1), nhấn mạnh tầm quan trọng của việc dự đoán sớm kết quả học tập của sinh viên. Mục 1.2 tổng quan các nền tảng cơ bản của học máy và học sâu. Mục 1.3 tổng hợp các nghiên cứu trong và ngoài nước có liên quan, qua đó chỉ ra khoảng trống nghiên cứu. Mục 1.4 giới thiệu các bộ dữ liệu thực nghiệm, bao gồm ba bộ dữ liệu từ các trường đại học Việt Nam ([CT1], [CT3], và [CT4]) cùng một số bộ dữ liệu quốc tế dùng cho đối sánh. Cuối cùng, Mục 1.5 trình bày các thước đo đánh giá được sử dụng để kiểm định và so sánh hiệu năng của các mô hình trong các chương sau.

1.1. Bối cảnh và động lực nghiên cứu

Trong bối cảnh Cách mạng công nghiệp lần thứ tư, dữ liệu trở thành động lực cho việc cá nhân hóa học tập và ra quyết định dựa trên thông tin trong giáo dục. Cùng với sự phát triển của các công nghệ giáo dục, việc dự đoán kết quả học tập bằng các phương pháp học máy (ML) và học sâu (DL) ngày càng giữ vai trò trung tâm. Luận án này tập trung khai thác khai phá dữ liệu giáo dục dựa trên học sâu (DL-based EDM) nhằm nâng cao khả năng dự đoán và hỗ trợ công tác quản lý chiến lược trong giáo dục.

Khai phá dữ liệu giáo dục (Educational Data Mining - EDM) và Phân tích học tập (Learning Analytics - LA) sử dụng các phương pháp tính toán để phân tích dữ liệu học tập, từ đó cho phép can thiệp sớm, dự đoán kết quả học tập và cung cấp hỗ trợ cá nhân hóa. EDM tập trung vào việc hiểu hành vi học tập, trong khi LA theo dõi và báo cáo các quá trình học tập. Sự kết hợp của hai lĩnh vực này thúc đẩy cải tiến giáo dục dựa trên dữ liệu, mặc dù vẫn còn những thách thức như quyền riêng tư dữ liệu và tích hợp hệ thống.

1.2. Các phương pháp học máy và học sâu

Học máy được phân loại thành bốn nhóm chính: học có giám sát, học không giám sát, học bán giám sát và học tăng cường. Các thuật toán như KNN, SVM, LR,

DT và RF được giới thiệu ngắn gọn trong tiểu mục này. Trong khi đó, các mô hình học sâu như DNN, CNN, RNN, LSTM và Transformer được thiết kế để giải quyết các nhiệm vụ phức tạp.

1.3. Tổng quan các nghiên cứu liên quan

Các nghiên cứu quốc tế gần đây trong lĩnh vực khai phá dữ liệu giáo dục và phân tích học tập cho thấy sự gia tăng mạnh mẽ trong việc ứng dụng các kỹ thuật học máy và học sâu để dự đoán kết quả học tập và hỗ trợ quá trình học tập. Nhiều thuật toán ML đã được sử dụng để phân tích dữ liệu giáo dục từ hệ thống quản lý học tập (LMS) và các khóa học trực tuyến mở (MOOCs). Các phương pháp này tỏ ra hiệu quả trong việc dự đoán kết quả học tập của sinh viên, bao gồm điểm số, nguy cơ bỏ học và khả năng tốt nghiệp. Bên cạnh đó, các kỹ thuật DL ngày càng được chú ý nhờ khả năng nắm bắt các mối quan hệ phi tuyến, phức tạp và cải thiện độ chính xác trong dự đoán. Các mô hình lai (hybrid models) cũng được phát triển nhằm giải quyết những thách thức như mất cân bằng dữ liệu và dữ liệu không đầy đủ. Tuy nhiên, vẫn còn tồn tại những hạn chế, chẳng hạn như thiếu kiểm chứng trên các hệ thống giáo dục khác nhau, cỡ mẫu nhỏ, và sự cần thiết phải cải thiện khả năng xử lý dữ liệu tuần tự và mối quan hệ giữa các học phần.

Trong bối cảnh các nghiên cứu trong nước tại Việt Nam, việc ứng dụng ML trong giáo dục vẫn ở giai đoạn khởi đầu. Một số công trình đáng chú ý tập trung vào ứng dụng ML trong lựa chọn học phần, dự đoán kết quả học tập và phát hiện sớm sinh viên có nguy cơ. Tuy nhiên, nhiều nghiên cứu trong nước vẫn gặp thách thức liên quan đến cỡ mẫu nhỏ và sự cần thiết phải xây dựng các mô hình toàn diện hơn, có khả năng khai thác đồng thời cả dữ liệu có cấu trúc và phi cấu trúc.

Khoảng trống nghiên cứu: Các phương pháp hiện tại chủ yếu dựa vào dữ liệu tĩnh và các mô hình ML truyền thống, dẫn đến khó khăn trong việc nắm bắt tính tuần tự của quá trình học tập. Những thách thức bao gồm dữ liệu nhỏ, phân mảnh và thiếu ngữ cảnh thời gian. Nghiên cứu trong tương lai cần tập trung xây dựng các bộ dữ liệu tuần tự được chuẩn hóa và phát triển các mô hình DL lai phù hợp với đặc thù của dữ liệu giáo dục đa dạng.

1.4. Bộ dữ liệu

Bộ dữ liệu HNMU1: Thu thập từ Trường Đại học Thủ đô Hà Nội (HNMU) trong giai đoạn 2021–2022, bao gồm 2.763 bản ghi của sinh viên ngành Giáo dục Tiểu học; sau tiền xử lý, còn lại 933 bản ghi với 39 thuộc tính.

Bộ dữ liệu HNMU2: Thu thập tại HNMU trong giai đoạn 2023–2024, gồm 2.613 bản ghi của sinh viên ngành Sư phạm Toán và Sư phạm Vật lý; sau khi làm sạch dữ liệu, giữ lại 551 bản ghi của sinh viên ngành Toán với 88 thuộc tính.

Bộ dữ liệu VNU: Dữ liệu khảo sát từ sinh viên ngành Sư phạm Ngữ văn tại Đại học Quốc gia Hà Nội (VNU) năm 2023, gồm 521 bản ghi với 91 thuộc tính; trong đó 271 mẫu được gán nhãn.

Bộ dữ liệu quốc tế: Năm bộ dữ liệu từ nhiều cơ sở giáo dục trên thế giới cũng được sử dụng cho mục đích so sánh và đối sánh.

Thách thức về quyền riêng tư: Dữ liệu giáo dục thường chứa thông tin cá nhân và nhạy cảm. Các vấn đề như chương trình đào tạo không đồng nhất, thay đổi thường xuyên, cùng mức độ số hóa khác nhau gây khó khăn cho việc chuẩn hóa. Do đó, nghiên cứu cần đảm bảo nguyên tắc bảo mật dữ liệu, sự đồng thuận của người học, cũng như khả năng thích ứng của mô hình với các bộ dữ liệu có quy mô nhỏ và đa dạng.

1.5. Độ đo đánh giá mô hình dự đoán

Các mô hình phân loại được đánh giá bằng các độ đo như Accuracy, Precision, Recall và F1-Score. Đối với các mô hình hồi quy, các chỉ số chính bao gồm MSE, RMSE, MAE và R^2.

CHƯƠNG 2. DỰ ĐOÁN SỚM ĐIỂM TRUNG BÌNH HỌC KỲ DỰA TRÊN CÁC TIẾP CẬN HỌC SÂU

Trong giáo dục hiện đại, việc dự đoán điểm trung bình học kỳ (SGPA) của sinh viên có ý nghĩa quan trọng trong theo dõi kết quả học tập, phát hiện sớm sinh viên có nguy cơ, và định hướng các kế hoạch học tập cá nhân hóa. Tuy nhiên, SGPA không phải là một thước đo tuyệt đối hay ổn định. Giá trị này có thể thay đổi theo thời gian dưới tác động của nhiều yếu tố như phương pháp chấm điểm, cách thức giảng dạy, trạng thái tâm lý của sinh viên, cũng như sự khác biệt giữa các cơ sở đào tạo. Do đó, SGPA cần được xem xét như một chỉ báo linh hoạt, phản ánh cả tính bất định và sự biến động. Xuất phát từ quan điểm này, chương này trình bày các mô hình dự đoán áp dụng học sâu kết hợp với các phương pháp xử lý bất định nhằm nâng cao độ chính xác và phản ánh tốt hơn tính phức tạp của môi trường giáo dục thực tế.

Hai hướng tiếp cận mô hình được đề xuất gồm: NeutroDLs: tích hợp logic neutrosophic vào các mô hình học sâu tiêu chuẩn; NeutroGNT: mô hình lai kết hợp

Transformer, CGAN và biểu diễn neutrosophic để xử lý vấn đề mất cân bằng dữ liệu và bất định. Thực nghiệm trên bảy bộ dữ liệu thực cho thấy các mô hình này cải thiện đáng kể độ chính xác dự đoán, trong đó NeutroGNT đạt MSE = 0,018 và R^2 = 96,05%. Nội dung của chương này dựa trên các công bố [CT5] và [CT6].

2.1. Giới thiệu

Đánh giá kết quả học tập của sinh viên là một quá trình phức tạp do tồn tại nhiều yếu tố bất định, sự đa dạng trong chuẩn đánh giá, cùng ảnh hưởng từ phương pháp giảng dạy và yếu tố tâm lý của người học. Sự phát triển của học tập trực tuyến càng làm gia tăng biến động thông qua các chỉ số tương tác số. Những yếu tố này khiến dữ liệu giáo dục thường chứa nhiều nhiễu và khó chuẩn hóa, từ đó hạn chế hiệu quả của các mô hình ML/DL truyền thống.

Để khắc phục, luận án tích hợp logic mờ (fuzzy logic) và logic neutrosophic vào các mô hình học sâu. Trong đó, logic neutrosophic bổ sung thành phần bất định (indeterminacy), cho phép xử lý tốt hơn dữ liệu chưa hoàn chỉnh và chứa nhiều yếu tố không chắc chắn.

2.2. Tổng quan về lý thuyết Neutrosophy

Định nghĩa 2.1. Một tập neutrosophic (NS) A, xác định trên không gian X với phần tử x, được biểu diễn dưới dạng:

$$A = \{(x, T_A(x), I_A(x), F_A(x)) : x \in X\} \quad (2.1)$$

ở đó mỗi x thuộc X được đại diện bởi ba hàm thuộc: hàm đo độ thuộc, $T_A: X \rightarrow [0;1]$; hàm đo độ trung tính, $I_A: X \rightarrow [0;1]$; và $F_A: X \rightarrow [0;1]$: đo độ không thuộc của phần tử x vào tập A. Tổng của các hàm thuộc thỏa mãn điều kiện $0 \leq T_A(x) + I_A(x) + F_A(x) \leq 3$ với mọi $x \in X$.

2.3. Đặt bài toán

Cho X là tập con khác rỗng của R^m. $x = (x_1, x_2, \dots, x_m) \in X$. Trong chương này luận án xem xét 03 kịch bản sau:

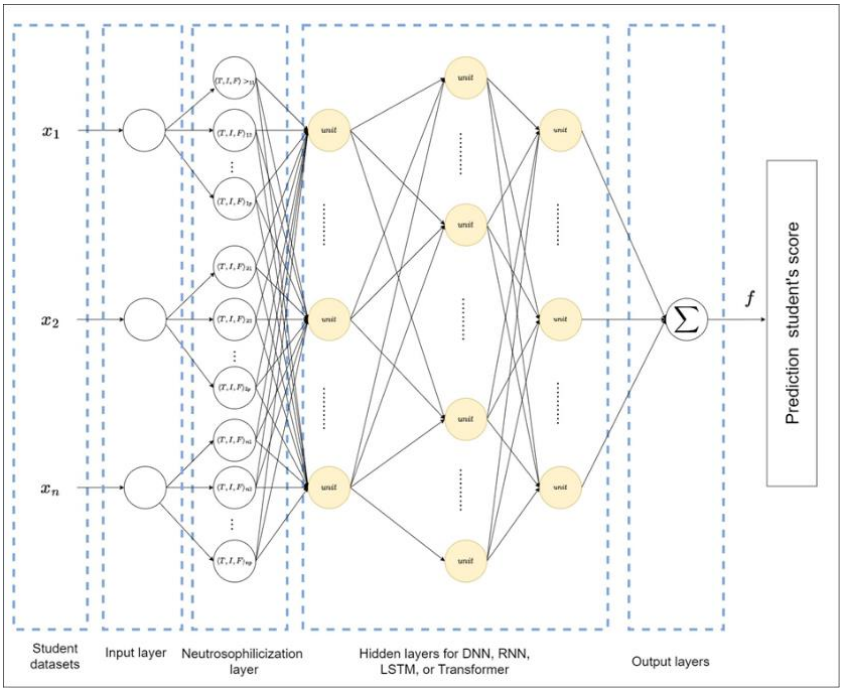
Kịch bản 1: Dự đoán sớm kết quả học tập học kỳ thứ n khi biết kết quả học tập của học kỳ thứ $n-1$. Tức là, cho biết giá trị của x_{n-1}, dự đoán giá trị của $x_n, 1 < n \leq m$.

Kịch bản 2: Dự đoán sớm kết quả học tập học kỳ thứ n khi biết kết quả học tập của hai học kỳ trước đó. Tức là, cho biết giá trị của x_{n-2}, x_{n-1}, dự đoán giá trị của $x_n, 2 < n \leq m$.

Kịch bản 3: Dự đoán sớm kết quả học tập học kỳ thứ n khi biết kết quả học tập của ba học kỳ trước đó. Tức là, cho biết giá trị của x_{n-3}, x_{n-2}, x_{n-1}, dự đoán giá trị của $x_n, 3 < n \leq m$.

2.4. Mô hình NeutroDL

Tập trung vào bài toán dự đoán SGPA, trong chương này, luận án đề xuất một cách tiếp cận mới tích hợp các lý thuyết về bất định vào các mô hình học sâu, nhằm nâng cao độ chính xác dự đoán, đặc biệt trong bối cảnh dữ liệu giáo dục không đầy đủ hoặc còn mơ hồ. Hình 2.4 minh họa kiến trúc tổng quát của các mạng NeutroDL (neutrosophic+DNN, CNN, RNN, LSTM và Transformer). Dữ liệu được xử lý thông qua các hàm neutrosophic để mô hình hóa tính bất định, từ đó nâng cao độ chính xác dự đoán. Mỗi mẫu dữ liệu được đặc trưng bởi một vector gồm 18 thuộc tính, với 6 nhân tố neutrosophic tương ứng với các mức độ kết quả học tập. Bằng cách xét đến các yếu tố không chắc chắn, các mô hình đề xuất hướng đến việc cung cấp những đánh giá linh hoạt và sát thực hơn về năng lực của sinh viên.



Hình 2.4. Mô hình NeutroDL

Thuật toán 2.1. NeutroDL dự đoán SGPA với Neutrosophic logic và Học sâu

Đầu vào: X: hồ sơ học tập lịch sử của sinh viên;
F_n: các hàm thuộc neutrosophic
Mô hình ∈ {DNN, CNN, RNN, LSTM, Transformer};
Siêu tham số: tốc độ học η, tỷ lệ dropout d, số epochs E
Đầu ra: $\hat{y}$ điểm dự đoán kết quả học tập của sinh viên
Tiền xử lý dữ liệu sinh viên thô: làm sạch, chuẩn hóa và sắp xếp theo thời gian.
Mã hóa neutrosophic: For $x_i \in X$ **do**
Với mỗi đầu vào x_i ánh xạ bằng hàm hình thang neutrosophic:

$$[\mathrm{T}(x_i), \mathrm{I}(x_i), \mathrm{F}(x_i)] \leftarrow F_n()$$

end for
Xây dựng mô hình gồm:
Input layer: vectơ neutrosophic $[T, I, F]$
Encoder (biến đổi neutrosophic.)
Hidden layers theo mô hình đã chọn (Mô hình ∈ {DNN, CNN, RNN, LSTM, Transformer})
Decoder: giải mờ (neutrosophic defuzzification)
Output layer: đầu ra hồi quy (regression head)
Huấn luyện: tối ưu bằng Adam với hàm mất mát MAE.
Chạy E epoch trên tập train.
Đánh giá trên tập test bằng các chỉ số RMSE, MAE, R^2
Trả về $\hat{y}$

Dữ liệu được sử dụng trong chương này là HNMU1. Sáu tập mờ neutrosophic được áp dụng, bao gồm: Excellent, Very Good, Good, Medium, Poor, và Very Poor. Các phương pháp học sâu (DL) dùng để phân tích dữ liệu gồm các mô hình cổ điển: DNN, CNN, RNN, LSTM, và Transformer. Kết quả so sánh được đánh giá dựa trên sai số (trung bình sau 10 lần chạy), như thể hiện trong Bảng 2.9.

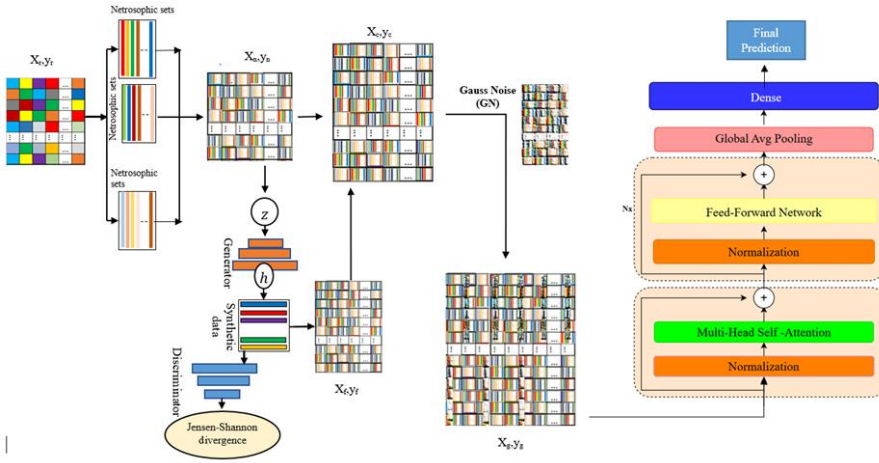
Bảng 2. 9. Sai số trung bình so sánh giữa các sai số trung bình của các kịch bản 1, 2, 3

Mô hình /Độ đo		RMSE		MAE		R^2 (%)	
		Real input	**NeutroDL**	**Real input**	**NeutroDL**	**Real input**	**NeutroDL**
Kịch bản 1	**DNN**	1.06 ± 0.33	**0.89** ± 0.09	1.07 ± 0.11	**0.75**± 0.06	48.26 ± 32.00	12.76 ± 4.90
	CNN	0.92 ± 0.06	**0.90**± 0.07	**0.73** ± 0.04	0.74± 0.04	8.52 ± 4.65	**11.40** ± 5.06
	RNN	**0.89**± 0.05	0.92 ± 0.07	**0.72**± 0.04	0.73 ± 0.04	12.6 ± 8.49	12.39 ± 6.06
	LSTM	**0.90** ± 0.05	0.91 ± 0.07	0.74 ± 0.03	**0.74** ± 0.04	9.72 ± 5.39	**12.73** ± 4.97
	Transformer	0.90 ± 0.04	**0.89** ± 0.08	0.74 ± 0.03	**0.74** ± 0.04	26± 5.40	**13.13** ± 7.65
Kịch bản 2	**DNN**	1.10 ± 0.11	**0.57** ± 0.05	1.13±0.12	**0.47** ± 0.05	186.07 ± 19.20	**48.42** ± 5.74
	CNN	**0.53** ± 0.05	0.58 ± 0.04	**0.41** ± 0.03	0.46 ± 0.02	46.39 ± 6.51	**47.03** ± 5.80
	RNN	**0.55** ± 0.07	0.60 ± 0.05	**0.42** ± 0.05	0.45 ± 0.03	37.12 ± 18.19	**46.16** ± 6.82
	LSTM	**0.52** ± 0.04	0.57 ± 0.04	**0.40** ± 0.03	0.45 ± 0.03	**52.42** ± 9.95	49.51 ± 5.40
	Transformer	0.63 ± 0.07	**0.59** ± 0.06	0.48 ± 0.04	**0.47** ± 0.06	26.83 ± 5.96	**45.54** ± 8.92
Kịch bản 3	**DNN**	2.44 ± 0.16	**0.87** ± 0.05	2.31 ± 0.16	**0.74** ± 0.04	-211.39 ± 75.45	**59.45** ± 4.00
	CNN	0.86 ± 0.08	**0.80** ± 0.08	0.67 ± 0.07	**0.62** ± 0.07	59.01 ± 7.00	**62.04** ± 5.36
	RNN	0.82 ± 0.13	**0.80** ± 0.08	0.62 ± 0.12	**0.60**± 0.06	60.69 ± 9.20	**62.14** ± 7.67
	LSTM	0.88 ± 0.13	**0.76** ± 0.11	0.71 ± 0.15	**0.59** ± 0.07	58.51 ± 1.54	**65.28** ± 8.93
	Transformer	0.93± 0.07	**0.79** ± 0.06	0.77 ± 0.06	**0.59** ± 0.05	53.05 ± 7.60	**65.95** ± 4.33

Các kết quả được in đậm thể hiện phương pháp dự đoán tương ứng đạt kết quả tốt hơn so với các phương pháp khác. Ba kịch bản nghiên cứu với bộ dữ liệu HNMU1 cho thấy cách tiếp cận đề xuất đã vượt trội hơn so với các mạng nơ-ron truyền thống khi làm việc với dữ liệu số thực. Trong các thí nghiệm, các mô hình RNN và Transformer được sử dụng trong luận án đã cho kết quả ổn định và tốt hơn so với các mô hình khác trong cùng điều kiện thử nghiệm.

2.5. Mô hình NeutroGNT

Phần này đề xuất một khung học sâu lai, tích hợp Transformer, Conditional GAN (CGAN) và biểu diễn đầu vào theo neutrosophic. Đồng thời, một chiến lược bổ sung nhiễu (noise injection) cũng được đưa vào nhằm nâng cao khả năng khái quát hóa của mô hình.



Hình 2. 10. Mô hình NeutroGNT

Cấu trúc tổng thể của mô hình được mô tả trong hình 2.10: Với tập dữ liệu thực (X_r, y_r), sử dụng các hàm thuộc neutrosophic dạng hình thang để mnetrốphic hóa dữ liệu và xây dưng tập dữ liệu mới (X_n, y_n). Để tăng tính hiệu quả của mô hình học sâu hồi quy, CGAN được tích hợp trong quá trình tền xử lý dữ liệu để sinh thêm dữ liệu (X_f,y_f). Hai tập dữ liệu (X_n, y_n) và(X_f,y_f) được kết hợp để tạ thành bộ dữ liệu tổng hợp (X_c, y_c). Sau đó, một chiến lược bổ sung nhiễu nhân tạo được tích hợp nhằm tăng cường tính bền vững (robustness) và khả năng khái quát hóa (generalization) của mô hình dự đoán, hình thành (X_g,y_g). Một mô hình Transformer được sử dụng để đưa ra kết quả dự đoán, hoạt động theo cách kết hợp để nắm bắt các mẫu phức tạp và các quan hệ phụ thuộc trong dữ liệu (X_g,y_g). Cuối cùng, thực hiện giải mờ (defuzzification) để chuyển

các giá trị neutrosophic về dạng giá trị thực và xuất ra kết quả dự đoán cuối cùng.đầu ra $\hat{Y}$.

Thuật toán 2.2. NeutroGNT – Dự đoán SGPA với Neutrosophic logic, CGAN, chiến lược bổ sung nhiễu nhân tạo và Transformer

1: Đầu vào: D_{real}: bộ dữ liệu thực gồm hồ sơ học tập của sinh viên và SGPA
2: Z : véc-tơ nhiễu tiềm ẩn cho CGAN
3: G : số lượng mẫu neutrosophic tổng hợp cần sinh
4: T_{Neutro}: mô hình Transformer với mã hóa neutrosophic và chiến lược bổ sung nhiễu (noise injection)
5: Đầu ra: $\hat{Y}$: các giá trị SGPA dự đoán cho tập kiểm thử

6: [X_r, y_r] ← Preprocess(D_{real}) ▷ Làm sạch, chuẩn hóa (scale), và sắp xếp theo học kỳ.
7: X_{Neutro} ← NeutrosophicTransform(X_r) Neutrosophic hóa bằng hàm membership hình thang
8: [G_{CGAN}, D_{CGAN}] ← Train CGAN([X_{Neutro}, y], Z)
9: for $i = 1$ **to** G **do**
10: z_i← Sample(Z)
11: y_i← SampleLabelDistribution(y_r)
12: $X_f[i] \leftarrow G_{CGAN}(z_i, y_i)$ ▷ Sinh đầu vào neutrosophic tổng hợp
13: $y_f[i] \leftarrow y_i$
14: end for
15: D_{aug} ← Concatenate([X_{Neutro} , y_r], [(X_f, y_f])
16: D_{aug} ← InjectNoise(D_{aug}) ▷ Gaussian noise injection
17: T_{Neutro} ← TrainTransformer(D_{aug})
18: $\hat{Y}$ ← Predict(T_{Neutro}, X_{test})
19: Trả về $\hat{Y}$

Kết quả và thảo luận: Trong phần này, chúng tôi sử dụng 06 bộ dữ liệu để tiến hành thực nghiệm và phân tích.

	Tập dữ liệu	**M**	**S**	**K**	**Kịch bản**	**Đặc trưng đầu vào**	**Đầu ra**
1	HNMU2	551	52	88	1	GPA Semester 1	GPA Semester 2
					2	GPA Semester 1, GPA Semester 2	GPA Semester 3
					3	GPA Semester 1, GPA Semester 2, GPA Semester 3	GPA Semester 4
2	VNU	271	43	91	2	GPA Semester 1, GPA Semester 2	GPA Semester 3

					3	GPA Semester 1, GPA Semester 2, GPA Semester 3	GPA Semester 4
3	Malaya-Stud	493	4	16	3	HSC, SSC, Last	Overall
4	Portugal-Math	395	3	33	2	G1, G2	G3
5	Portugal-Lang	649	3	33	2	G1, G2	G3
6	Covenant-Priv	1841	6	9	3	First Year GPA, Second Year GPA, Third Year GPA	Fourth Year GPA

Kích thước mẫu (M); Số lượng thuộc tính liên quan đến điểm số (S); Tổng số thuộc tính (k)

Trước khi tiến hành thực nghiệm, toàn bộ dữ liệu được tiền xử lý để loại bỏ các giá trị thiếu và loại bỏ các điểm số nằm ngoài khoảng 0–10. Sau đó, các bộ dữ liệu được chia thành 80% cho tập huấn luyện và 20% cho tập kiểm thử. Kết quả thực nghiệm (được lấy trung bình qua 10 lần chạy) cho thấy mô hình NeutroGNT luôn vượt trội so với tất cả các mô hình được đánh giá khác.

Bảng 2.7. Bảng sai số trung bình sau 10 lần chạy thực nghiệm-Kịch bản 1

	Real_T	**Neutro_T**	**NeutroCT**	**NeutroGNT**
MSE	0.519 ± 0.028	0.474 ± 0.040	0.469 ± 0.031	**0.458 ± 0.011**
MAE	0.576 ± 0.014	0.560 ± 0.029	0.558 ± 0.022	**0.548 ± 0.010**
R^2	-0.087±0.058	0.008 ± 0.085	0.017 ± 0.064	**0.041 ± 0.022**

Trong Bảng 2.7, NeutroGNT đạt MSE thấp nhất (0.458 ± 0.011) và cải thiện 12.8% R^2 so với Real_T, nhưng R^2 vẫn ở mức thấp (0.041 ± 0.022), cho thấy khả năng khái quát và giải thích dữ liệu còn hạn chế trong các tình huống thực tế nhiều nhiễu như HNMU2.

Bảng 2.8. Bảng sai số trung bình sau 10 lần chạy thực nghiệm-Kịch bản 2

Tập dữ liệu		Real_T	Neutro_T	NeutroCT	NeutroGNT
HNMU2	MSE	0.323 ± 0.101	0.183 ± 0.024	0.208 ± 0.052	0.181 ± 0.030
	MAE	0.459 ± 0.085	0.339 ± 0.025	0.363 ± 0.053	0.338 ± 0.035
	R^2	0.077 ± 0.288	0.478 ± 0.069	0.407 ± 0.147	**0.482 ± 0.084**
VNU	MSE	0.302 ± 0.031	0.320 ± 0.042	0.321 ± 0.044	**0.260 ± 0.04**6
	MAE	0.441 ± 0.032	0.453 ± 0.039	0.451 ± 0.042	**0.381 ± 0.054**
	R^2	0.201 ± 0.083	0.153 ± 0.112	0.150 ± 0.116	**0.202 ± 0.140**
Portugal-Math	MSE	2.536 ± 2.129	1.263 ± 0.080	1.409 ± 0.135	**1.197 ± 0.074**
	MAE	1.065 ± 0.567	0.770 ± 0.069	0.844 ± 0.077	0.725 ± 0.043
	R^2	0.505 ± 0.415	0.754 ± 0.016	0.725 ± 0.026	**0.767 ± 0.014**
Portugal-Lang	MSE	0.704 ± 0.550	0.423 ± 0.014	0.435 ± 0.032	0.440 ± 0.033
	MAE	0.528 ± 0.241	0.403 ± 0.004	0.413 ± 0.027	0.425 ± 0.027
	R^2	0.711 ± 0.225	0.826 ± 0.006	0.822 ± 0.013	0.820 ± 0.013

Trong Kịch bản 2, các mô hình đề xuất thể hiện hiệu năng vượt trội và ổn định trên cả 04 bộ dữ liệu chuẩn. Đặc biệt, mô hình NeutroGNT đạt kết quả xuất sắc về cả hai chỉ số MSE (MAE) và R^2.

Bảng 2.9. Bảng sai số trung bình sau 10 lần chạy thực nghiệm-Kịch bản 3

Tập dữ liệu		Real_T	Neutro_T	NeutroCT	NeutroGNT
HNMU2	MSE	0.212 ± 0.088	0.208 ± 0.081	0.175 ± 0.082	0.152 ± 0.025
	MAE	0.374 ± 0.078	0.382 ± 0.083	0.347 ± 0.081	0.322 ± 0.029
	R^2	0.047 ± 0.393	0.068 ± 0.364	0.216 ± 0.367	0.319 ± 0.111
VNU	MSE	0.119 ± 0.037	0.109 ± 0.041	0.121 ± 0.061	**0.088 ± 0.017**
	MAE	0.281 ± 0.039	0.271 ± 0.051	0.282 ± 0.074	**0.242 ± 0.026**
	R^2	0.549 ± 0.140	0.588 ± 0.154	0.541 ± 0.230	**0.666 ± 0.064**
Malaya – Stud	MSE	0.495 ± 0.563	**0.342 ± 0.03**8	0.412 ± 0.063	0.400 ± 0.055
	MAE	0.505 ± 0.249	**0.434 ± 0.025**	0.485 ± 0.048	0.473 ± 0.036
	R^2	0.788 ± 0.241	**0.854 ± 0.016**	0.824 ± 0.027	0.829 ± 0.024
Covenant - Priv	MSE	0.023 ± 0.001	0.022 ± 0.001	0.023 ± 0.003	**0.019 ± 0.00**2
	MAE	0.116 ± 0.003	0.114 ± 0.002	0.118 ± 0.008	**0.107 ± 0.005**
	R^2	0.949 ± 0.002	0.952 ± 0.001	0.950 ± 0.007	**0.958 ± 0.003**
	RMSE	0.152 ± 0.003	0.147 ± 0.002	0.150 ± 0.009	**0.138 ± 0.005**

Trong số các mô hình được đánh giá, NeutroGNT nổi bật nhờ đạt được sự cân bằng tối ưu giữa độ chính xác và tính ổn định. Trên bộ dữ liệu Covenant-PrivateEng, mô hình này ghi nhận giá trị R^2 trung bình cao nhất, vượt trội rõ rệt so với các mô hình khác. Đáng chú ý, RMSE trung bình của NeutroGNT thấp hơn 0.138 so với mô hình Real_T. Hơn nữa, mô hình đạt được RMSE nhỏ nhất là 0.1342, thấp hơn kết quả tốt nhất trước đây do Aderibigbe et al. (2019) báo cáo. Ngoài ra, MSE nhỏ nhất là 0.018 – thấp nhất trong toàn bộ nghiên cứu, trong khi R^2 lớn nhất đạt 96.05% đã vượt qua tất cả các mốc chuẩn trước đó. Các kết quả này khẳng định hiệu quả dự báo vượt trội và tính ưu việt của mô hình NeutroGNT.

2.4. Phụ lục Chương 2: Phần này tóm tắt các mô hình GAN, CGAN và Transformer cho bài toán dự đoán SGPA.

CHƯƠNG 3: NÂNG CAO HIỆU NĂNG CỦA CÁC MÔ HÌNH PHÂN LOẠI TỐT NGHIỆP SỚM

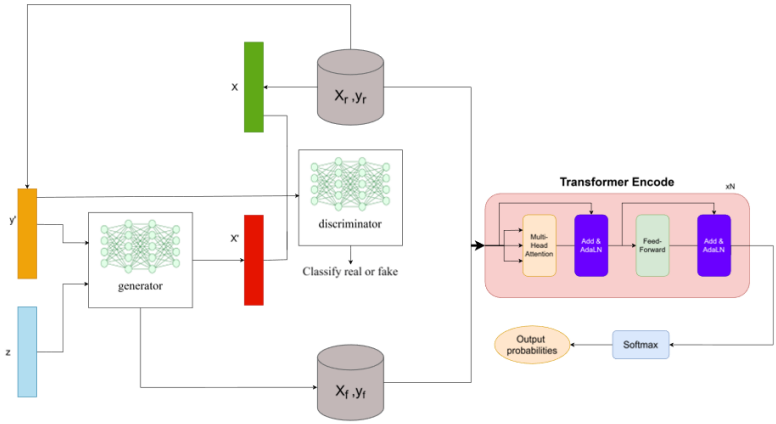
Chương này giới thiệu hai mô hình học sâu lai LATCGAd và AWG-GC nhằm nâng cao hiệu quả dự đoán tốt nghiệp sớm trong bối cảnh dữ liệu giáo dục nhỏ và mất cân bằng. LATCGAd kết hợp Transformer, CGAN và Adaptive Layer Normalization (AdaLN) để tăng cường dữ liệu, ổn định huấn luyện và giảm quá khớp, đạt 96.97% accuracy và 73.66% F1-score. AWG-GC tích hợp Autoencoder, Wasserstein GAN và Graphormer để đồng thời học biểu diễn, sinh dữ liệu và phân loại, đạt 98.54% accuracy và 99.25% F1-score, vượt trội so với các mô hình cơ sở. Nội dung chương 3 được phát triển dựa trên các kết quả trong [CT7] và [CT8].

3.1. Giới thiệu

Phương pháp LAGT [CT2] đã chứng minh ưu thế so với các mô hình truyền thống nhờ kết hợp GCN và Transformer trong khung bán giám sát. Trên nền tảng đó, chương này tiếp tục khai thác các tiến bộ của mô hình sinh dữ liệu (CGAN, WGAN), kiến trúc đồ thị (GAT, Graphormer) và Transformer để giải quyết bài toán dữ liệu nhỏ, mất cân bằng. Hai mô hình lai LATCGAd và AWG-GC được đề xuất, tích hợp sinh dữ liệu với học sâu, nâng cao độ chính xác và ổn định trong dự đoán tốt nghiệp sớm.

3.2. Mô hình LATCGAd

Mô hình này dùng CGAN để sinh mẫu tổng hợp cho các nhãn hiếm, khắc phục mất cân bằng dữ liệu. Tập dữ liệu mở rộng được xử lý bởi Transformer Encoder nhằm khai thác quan hệ đặc trưng phức tạp. AdaLN được tích hợp trong từng lớp Transformer để điều chỉnh chuẩn hóa theo đặc tính đầu vào, giúp giảm sai lệch, cải thiện hội tụ và hạn chế quá khớp trên dữ liệu nhỏ.



Hình 4. 1. Mô hình LATCGAd

Thuật toán 3.1. LATCGAd - Phân tích học tập với Transformer, CGAN, và Adaptive Layer Normalization

1: Đầu vào: D_{Real}: Dữ liệu thực có gán nhãn
2: Z : Véc–tơ nhiễu tiềm ẩn cho CGAN
3: G: Số lượng mẫu tổng hợp cần sinh
4: T_{AdaLN}: Mô hình Transformer với Adaptive Layer Normalization
5: Đầu ra: Ŷ: Các nhãn dự đoán

6: $[X_r, y_r] \leftarrow$ Preprocess(D_{real})
7: $[G_{\text{CGAN}}, D_{\text{CGAN}}] \leftarrow Train_{CGAN}([X_r, y_r], Z)$
8: for $i = 1$ **to** G **do**
9: $z_i \leftarrow$ Sample(Z)
10: $y_i \leftarrow$ Sample_Label_Distribution (y_r)
11: $X_f\ [i] \leftarrow G_{CGAN}(z_i, y_i)$ ▷ Sinh mẫu tổng hợp
12: $y_f\ [i] \leftarrow y_i$
13: end for
14: $D_{aug} \leftarrow$ Concatenate($[X_r, y_r], [X_f, y_f]$) ▷ Tập dữ liệu tăng cường
15: $T_{\text{AdaLN}} \leftarrow$ Train_Transformer(D_{aug})
16: Ŷ ← Predict(T_{AdaLN}, X_{test})
17: Trả về Ŷ

Thực nghiệm được thực hiện trên ba bộ dữ liệu: HNMU1, HNMU2 và VNU. Dữ liệu được chia thành ba phần: 60% cho huấn luyện, 15% cho kiểm định (validation) và 25% cho kiểm thử.

Trên HNMU1, mô hình LATCGAd đạt độ chính xác 95.56%, vượt qua tất cả các mô hình cơ sở. Đồng thời, các chỉ số Precision (72.50%), Recall (74.78%) và F1-score (73.61%) đều được cải thiện, cho thấy khả năng phân loại dương tính đúng khá mạnh.

Trên HNMU2, LATCGAd tiếp tục dẫn đầu về độ chính xác (96.97%), nhưng lại kém hơn so với Decision Tree (DT) ở chỉ số Precision và Recall, điều này phản ánh khả năng khái quát hóa tốt nhưng độ sắc nét trong việc nhận diện chính xác lớp mục tiêu còn hạn chế.

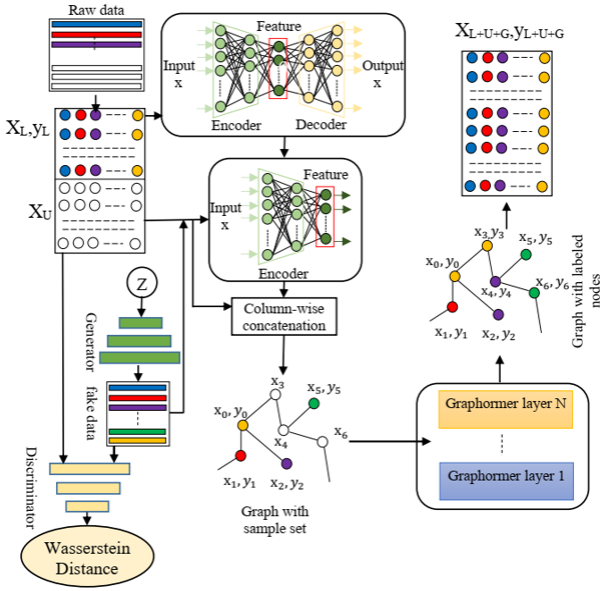
Bảng 3.8. Kết quả dự đoán trên bộ dữ liệu HNMU2

Method	**Accuracy**	**Precision**	**Recall**		**F1-Score**
DT	89.70	94.65	79.26		82.48
SVM	80.29	41.38	41.81		40.55
LR	71.74	64.57	62.25		60.46
DNN	87.05	69.32	60.92		63.75
GAT	89.05	53.52	57.95		55.16
Transformer	95.62	72.77	60.99		64.79
LATCGAd	**96.97**	**73.26**	**74.09**		**73.66**

Trên bộ dữ liệu VNU, mô hình LATCGAd đạt độ chính xác 87.65%, nhỉnh hơn so với Decision Tree (83.95%) và Transformer chuẩn (86.76%).

3.3. Mô hình AWG-GC

Mô hình AWG-GC là một mở rộng của mô hình LATCGAd, vừa kế thừa các ưu điểm trong sinh dữ liệu và huấn luyện, vừa bổ sung các thành phần để xử lý đặc trưng phức tạp, số lượng nhãn hạn chế và mối quan hệ đa chiều trong dữ liệu giáo dục. Sau giai đoạn tiền xử lý, dữ liệu thô được tổ chức thành tập mẫu ban đầu gồm $(L + U)$ mẫu, trong đó (X_L, y_L) là các mẫu có nhãn và X_U là các mẫu không nhãn. Mỗi mẫu trong tập này có kích thước n chiều. Tập dữ liệu này được sử dụng để huấn luyện mạng nơ-ron Autoencoder sâu nhằm học biểu diễn trong không gian tiềm ẩn. Đồng thời, tập $(L + U)$ cũng được sử dụng để huấn luyện WGAN, từ đó sinh thêm tập mẫu tổng hợp X_G, gồm G mẫu mới. Tập mở rộng $(L + U + G)$ sau đó được đưa vào bộ mã hóa (encoder) của Autoencoder để trích xuất đặc trưng và giảm chiều dữ liệu từ n xuống m. Như vậy, mỗi mẫu trong tập $L + U + G$ có hai dạng biểu diễn: một trong không gian gốc n chiều và một trong không gian đặc trưng m chiều. Dựa trên không gian đặc trưng kết hợp này, đồ thị lân cận của các mẫu trong tập $(L + U + G)$ được xây dựng bằng thuật toán KNN. Đồ thị thu được được đưa vào mô hình Graphormer. Với cơ chế global attention có trọng số theo khoảng cách đồ thị, Graphormer có khả năng học hiệu quả mối quan hệ giữa các nút, từ đó cải thiện độ chính xác trong bài toán dự đoán phân loại tốt nghiệp của sinh viên.



Hình 3. 1. Mô hình AWG-GC

Thuật toán 3.2: AWG-GC – Tích hợp Autoencoder, Wasserstein GAN và Graphormer cho bài toán phân loại tốt nghiệp.

1: Đầu vào: $\mathcal{D}_{\mathrm{L}}$ **:** Tập dữ liệu có nhãn
2: $\mathcal{D}_{\mathrm{U}}$: Tập dữ liệu chưa gán nhãn
3: m : Số lượng mẫu trong $\mathcal{D}_{\mathrm{U}}$
4: n : Số lượng mẫu trong $\mathcal{D}_{\mathrm{L}}$
5: z : Kích thước đặc trưng tiềm ẩn trích xuất từ Autoencoder
6: s : Số lượng mẫu tổng hợp được sinh bởi WGAN
7: Đầu ra: Ŷ ← Các nhãn dự đoán

8: $[X_L, y_L]$ ← Preprocess($\mathcal{D}_L$)
9: X_{U}← Preprocess($\mathcal{D}_{\mathrm{U}}$)
10: Train Autoencoder on $X_L \cup X_U$
11: Z_{L} ← Encode(X_L), Z_{U} ← Encode(X_{U})
12: $[\mathcal{G}, C]$ ← TrainWGAN(X_L, y_L)
13: **for** $i = 1$ **to** s **do**
14: z_i ← SampleNoise()
15: y_i← SampleLabel(y_L)
16: $x_i^{gen} \leftarrow \mathcal{G}(z_i, y_i)$
17: $\mathcal{D}_S \leftarrow \mathcal{D}_S \cup (x_i^{gen}, y_i)$
18: **end for**
19: $\mathcal{D}_{all} \leftarrow \mathcal{D}_L \cup \mathcal{D}_U \cup \mathcal{D}_S$
20: Z_{all} ← Encode($\mathcal{D}_{all}$)
21: $F_{combined}$ ← Concatenate(X_{all}, Z_{all})
22: G_{knn} ← ConstructGraph($F_{combined}$)
23: Train Graphormer on (G_{knn}, y_L)
24: Ŷ ← Predict(Graphormer, X_{test})
25: **Trả về** Ŷ

Luận án sử dụng ba bộ dữ liệu thực tế, gồm HNMU2, VNU, và SATDAP, để đánh giá hiệu năng của mô hình đề xuất. Bộ dữ liệu SATDAP (Bồ Đào Nha) bao gồm 4.424 bản ghi với 36 đặc trưng. Các bộ dữ liệu HNMU2 và VNU được chia thành ba tập: huấn luyện (60%), kiểm định (15%), và kiểm tra (25%). Bộ dữ liệu SATDAP cũng được chia thành ba tập: huấn luyện (65%), kiểm định (15%), và kiểm tra (20%).

Bảng 3.10: Kết quả dự đoán cho tập HNMU2

Mô hình	Accuracy	Precision	Recall	F1-Score
KNN	80.29	40.68	41.58	40.59
RF	95.62	47.79	48.31	48.34
Transformer	95.62	72.77	60.99	64.79
GAT	89.05	53.52	57.95	55.16
Graphomer	97.08	73.45	73.97	73.67

AutoGAT	93.43	59.84	59.74	59.74
AWG_GAT	**97.08**	**98.50**	**86.41**	**90.37**
AWG-GC	**98.54**	**99.25**	**99.25**	**99.25**

Kết quả dự đoán trên bộ dữ liệu VNU cho thấy mô hình AWG-GC liên tục vượt trội hơn tất cả các phương pháp khác trên toàn bộ các thước đo đánh giá.

Bảng 3.11: Kết quả dự đoán cho tập VNU

Mô hình	Accuracy	Precision	Recall	F1-Score
KNN	86.76	51.45	54.98	53.12
RF	82.35	54.03	46.91	49.39
Transformer	86.76	69.72	71.73	70.72
GAT	80.88	51.60	50.52	51.00
Graphomer	88.24	80.11	63.93	64.97
AutoGAT	85.29	74.50	58.59	53.96
AWG-GAT	**89.71**	**70.95**	**95.98**	**78.64**
AWG-GC	**94.12**	**81.67**	**97.70**	**88.17**

Kết quả trên bộ dữ liệu SATDAP cho thấy mô hình AWG-GC đạt hiệu năng tổng thể cao nhất trên tất cả các thước đo đánh giá. Cụ thể, AWG-GC vượt trội hơn mô hình XGBoost của *Martins et al. (2021)* với mức cải thiện 8.81% về độ chính xác và 9.21% về F1-score. So sánh này nhấn mạnh hơn nữa tính ưu việt của AWG-GC, không chỉ ở khả năng dự đoán chính xác mà còn ở hiệu năng phân loại cân bằng giữa các lớp.

Bảng 3.12: Kết quả dự đoán cho tập SATDAP

Mô hình	Accuracy	Precision	Recall	F1-Score
KNN	66.67	57.66	54.73	55.35
RF	79.32	70.78	68.65	69.37
Transformer	80.34	71.87	70.99	71.34
Graphomer	80.79	74.08	70.30	71.67
AWG-GC	**81.81**	**74.74**	**73.89**	**74.21**
XGBoost (Martin et al., 2021)	73.00	-	-	65.00

Kiến trúc Graphormer cho phép mô hình nắm bắt tốt hơn cấu trúc tiềm ẩn của dữ liệu giáo dục, đồng thời xử lý hiệu quả các thách thức như kích thước mẫu nhỏ và mất cân bằng lớp.

3.4. Phụ lục Chương 3: Phần này giới thiệu về Wasserstein GANs (WGAN) và Graphormer nhằm cung cấp cơ sở lý thuyết cho việc tích hợp chúng vào mô hình đề xuất.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

A. Những đóng góp chính của luận án

Luận án này đã giải quyết bài toán dự đoán kết quả học tập của sinh viên trong điều kiện không chắc chắn, thiếu hụt dữ liệu và mất cân bằng lớp, vốn đặc trưng cho các môi trường giáo dục thực tiễn.

Trước hết, đối với dự đoán SGPA ngắn hạn, luận án đã đề xuất hai mô hình NeutroDL và NeutroGNT, tích hợp học sâu với lý thuyết neutrosophic nhằm xử lý dữ liệu thiếu và không chắc chắn. Kết quả thực nghiệm khẳng định tính hiệu quả, trong đó NeutroGNT đạt MSE = 0.018 và $R^2 = 96.05\%$, vượt trội hơn so với các mô hình truyền thống, đồng thời hỗ trợ giám sát kịp thời, can thiệp sớm và cá nhân hóa học tập.

Tiếp nối, nghiên cứu mở rộng sang dự đoán phân loại tốt nghiệp dài hạn, với hai mô hình lai được đề xuất: LATCGAd, đạt độ chính xác 96.97% và F1-score 73.66%; và AWG-GC, đạt độ chính xác 98.54% và F1-score 99.25%. Các kết quả này vượt trội so với các mô hình đối sánh, chứng minh ưu thế của việc kết hợp kiến trúc sinh dữ liệu và mạng dựa trên đồ thị trong bối cảnh dữ liệu giáo dục mất cân bằng.

Tóm lại, luận án đóng góp ở ba khía cạnh chính: (i) Xây dựng các khung mô hình học sâu tích hợp yếu tố bất định cho dự đoán SGPA. (ii) Thiết kế các mô hình lai cho bài toán phân loại tốt nghiệp trong điều kiện dữ liệu nhỏ và mất cân bằng. (iii) Phát triển bộ dữ liệu mở rộng và quy trình phân tích phục vụ ứng dụng trong giáo dục.

Những đóng góp này cung cấp công cụ thực tiễn nhằm hỗ trợ ra quyết định thích ứng, thông minh và dựa trên dữ liệu trong giáo dục đại học.

B. Hướng nghiên cứu tiếp theo

Dựa trên các kết quả đạt được, luận án đề xuất một số hướng nghiên cứu tiềm năng:

1. Mở rộng mục tiêu dự đoán sang các khía cạnh khác như: nguy cơ bỏ học, khả năng hoàn thành chương trình, mức độ hài lòng môn học, và định hướng nghề nghiệp, nhằm cung cấp bức tranh toàn diện hơn về hành trình học tập của sinh viên.
2. Ứng dụng học tăng cường và học không giám sát, kết hợp với các kỹ thuật AI có khả năng giải thích (XAI), để vừa cá nhân hóa lộ trình học tập, vừa cung cấp căn cứ minh bạch giúp giảng viên và nhà quản lý tin tưởng hơn vào quyết định can thiệp sớm.
3. Khai thác học liên kết (Federated Learning) và học chuyển giao (Transfer Learning) để xây dựng các mô hình vừa đảm bảo hiệu quả, khả năng tổng quát hóa, vừa duy trì quyền riêng tư dữ liệu giữa các cơ sở đào tạo.
4. Phát triển hệ thống Phân tích học tập (Learning Analytics) trực tuyến dựa trên các mô hình đề xuất, tích hợp XAI để hỗ trợ giải thích trực quan và cung cấp các khuyến nghị theo thời gian thực cho cả sinh viên và giảng viên.

DANH MỤC CÔNG TRÌNH CÔNG BỐ

[**CT1**]. **Son, N. T. K.**, Bien, N. V., Quynh, N. H., and Tho, C. C. (2022). Machine learning-based admission data processing for early forecasting of students' learning outcomes. *International Journal of Data Warehousing and Mining, 18*(1). (SCIE). https://www.igi-global.com/gateway/article/313585

[**CT2**]. **Son, N. T. K.,** Quynh, N. H., and Minh, B. T. (2024). Early prediction of students' graduation rank using LAGT: Enhancing accuracy with GCN and Transformer. *Journal of Computer Science and Cybernetics, 40*(4), 299-314. https://doi.org/10.15625/1813-9663/21095

[**CT3**]. **Son, N. T. K.**, Hoa, N. H., Trang, H. T. T., and Ngan, T. Q. (2024). Introducing a cutting-edge dataset: Revealing key factors in student academic outcomes and learning processes. *VNU Journal of Science: Education Research, 40*(4). https://js.vnu.edu.vn/ER/article/view/5157.

[**CT4**]. Vinh, T. D., **Son, N. T. K.,** and Diep, L. N. (2025). Dataset of factors affecting learning outcomes of students at the University of Education, Vietnam National University, Hanoi. *Data in Brief, 59*, 111438. (ESCI). https://doi.org/10.1016/j.dib.2025.111438

[**CT5**]. **Son, N. T. K.**, Dat, B. V., Hoa, N. H., and Trang, H. T. T. (2025). Hybrid artificial intelligence models incorporating neutrosophy for predicting student outcomes. *Lecture Notes in Networks and Systems.* (Accepted; Scopus).

[**CT6**]. **Son, N. T. K.**, Thong, N. T., and Quynh, N. H. (2025). Neutrosophy-driven deep learning for predicting student performance. *Neutrosophic Sets and Systems, Volume 87 (2025).* (Scopus).

[**CT7**]. **Son, N. T. K**., Quynh, N. H., and Minh, B. T. (2025). Refining graduation classification accuracy with synergistic deep learning models. *Cybernetics and Information Technologies, Volume **25**, No 2, 2025.* (Scopus, ESCI).

[**CT8**]. **Son, N. T. K.,** Vinh, T. D., Quynh, N. H., Tao, N. Q., Hop, D. T., and Minh, B. T. (2025). Enhancing student graduation prediction with integrated deep learning under data constraints. *Data Intelligence.* (Accepted; In proof, Scopus, ESCI).