

Ứng dụng học máy trong dự báo chậm tiến độ dự án xây dựng: xu hướng và thách thức

Machine learning applications in construction delay prediction: trends and challenges

Nguyễn Thế Minh^{1,*}

¹ Phòng Kinh tế, Hạ tầng và Đô thị phường Long Châu, tỉnh Vĩnh Long;

*Tác giả liên hệ: theminhpltdt@gmail.com

■ Nhận bài: 31/01/2026 ■ Sửa bài: 26/02/2026 ■ Duyệt đăng: 10/03/2026

DOI: <https://doi.org/10.66195/mtu.2026.16.162>

TÓM TẮT

Chậm tiến độ là rủi ro phổ biến trong các dự án xây dựng, làm tăng chi phí, kéo dài thời gian và ảnh hưởng hiệu quả đầu tư cũng như phối hợp giữa các bên; tuy nhiên, kiểm soát tiến độ trong thực tiễn vẫn chủ yếu dựa trên so sánh kế hoạch - thực tế và kinh nghiệm nên khó phát hiện sớm dấu hiệu trễ và thiếu cơ sở định lượng để cảnh báo kịp thời. Trước xu hướng số hóa, dữ liệu dự án ngày càng sẵn có (kế hoạch tiến độ, báo cáo hiện trường, nhật ký thi công,...), tạo nền tảng cho ứng dụng Machine Learning (ML) trong dự báo nguy cơ chậm tiến độ. Bài báo nghiên cứu ML dự báo chậm tiến độ, tập trung vào loại dữ liệu đầu vào, cách thiết kế bài toán theo chu kỳ tuần/tháng, các mô hình thường dùng và phương pháp đánh giá, cho thấy ML có thể khai thác dữ liệu đa biến, nhận diện quan hệ phi tuyến và cung cấp dự báo xác suất rủi ro để hỗ trợ cảnh báo sớm cho điều hành thi công. Dù vậy, việc triển khai thực tế vẫn gặp thách thức về tính nhất quán và chuẩn hóa dữ liệu, khả năng khái quát giữa dự án và mức độ diễn giải kết quả phục vụ ra quyết định, từ đó gợi mở các hướng nghiên cứu tiếp theo nhằm nâng cao tính khả thi và hiệu quả ứng dụng ML trong quản lý tiến độ dự án xây dựng.

Từ khóa: chậm tiến độ, dự án xây dựng, chi phí, học máy, rủi ro

ABSTRACT

Schedule delay is one of the most common risks in construction projects, increasing costs, extending project duration, and undermining investment efficiency as well as coordination among stakeholders. However, in practice, schedule control often relies on comparing planned versus actual progress and expert judgement, making it difficult to detect early signs of delay and providing limited quantitative basis for timely warnings. With the ongoing digital transformation of the construction industry, project data are increasingly available-such as baseline schedules, site reports, construction logs, etc. creating a solid foundation for applying ML to predict delay risk. This paper reviews ML-based approaches for forecasting schedule delays, focusing on input data types, problem formulation on weekly/monthly time cycles, commonly used ML models, and performance evaluation methods. The reviewed studies indicate that ML can leverage multivariate data, capture non-linear relationships, and generate probabilistic forecasts of schedule-delay risk to support early-warning mechanisms for construction operations management. Nevertheless, practical implementation remains challenging due to inconsistent data quality, difficulties in standardizing data across projects, limited model generalizability, and the interpretability of results for decision-making, thereby motivating future research directions to enhance the feasibility and effectiveness of ML applications in construction schedule management.

Keywords: schedule delay, construction projects, costs, Machine Learning, risks

1. GIỚI THIỆU

Trong quản lý dự án xây dựng, tiến độ là một trong ba mục tiêu cốt lõi bên cạnh chi phí

và chất lượng. Tuy nhiên, chậm tiến độ vẫn là hiện tượng phổ biến ở nhiều loại hình dự án, gây kéo dài thời gian hoàn thành, làm tăng chi

phí gián tiếp, ảnh hưởng đến kế hoạch khai thác công trình và gia tăng nguy cơ tranh chấp hợp đồng. Chậm tiến độ thường không xuất hiện do một nguyên nhân đơn lẻ mà là kết quả tích lũy của nhiều tín hiệu suy giảm hiệu suất trong quá trình thi công. Trong thực tiễn, kiểm soát tiến độ chủ yếu dựa vào các công cụ lập kế hoạch và theo dõi truyền thống như biểu đồ Gantt, CPM/PERT, cập nhật kế hoạch thực tế theo kỳ báo cáo và đánh giá kinh nghiệm của người quản lý. Cách tiếp cận này có vai trò nền tảng nhưng thường mang tính “phản ứng”, tức phát hiện sự cố sau khi sai lệch tiến độ đã hình thành rõ rệt. Khi đó, việc điều chỉnh thường tốn kém, khó điều chỉnh, và đôi khi không điều chỉnh được, bỏ qua các mốc quan trọng. Bên cạnh đó, các phương pháp truyền thống khó tận dụng đầy đủ dữ liệu phát sinh liên tục trong thi công như nhật ký hiện trường, năng suất tổ đội, tình trạng vật tư, hồ sơ RFI, shopdrawing, sự cố an toàn lao động và tiến độ thanh toán. Điều này đặt ra nhu cầu về các phương pháp dự báo sớm rủi ro chậm tiến độ dựa trên dữ liệu, giúp nhà quản lý chuyển từ điều hành bị động sang chủ động.

Những năm gần đây, chuyển đổi số trong ngành xây dựng tạo ra điều kiện thuận lợi cho việc thu thập và lưu trữ dữ liệu dự án theo thời gian. Dữ liệu tiến độ từ phần mềm quản lý (MS Project, Primavera), dữ liệu chất lượng, an toàn, hồ sơ vật tư, nhật ký thi công, báo cáo tuần/tháng và trao đổi giữa các bên ngày càng được số hóa. Đồng thời, việc áp dụng BIM, đặc biệt là BIM 4D, góp phần liên kết mô hình công trình với tiến độ thi công, hỗ trợ trực quan hóa và theo dõi mức độ hoàn thành theo thời gian. Sự sẵn có của dữ liệu đa nguồn và đa dạng mở ra cơ hội áp dụng các kỹ thuật phân tích hiện đại, trong đó học máy (Machine Learning) nổi bật nhờ khả năng học từ dữ liệu lịch sử, xử lý quan hệ phi tuyến, khai thác tương tác giữa nhiều biến đầu vào và đưa ra dự báo xác suất rủi ro.

Trong bối cảnh đó, nhiều nghiên cứu đã ứng dụng học máy để dự báo chậm tiến độ dưới các dạng bài toán khác nhau như phân loại: chậm, không chậm; mức rủi ro thấp, trung bình, cao hoặc hồi quy. Dữ liệu đầu vào được khai thác từ các nhóm yếu tố như kế hoạch, thực tế tiến độ, chỉ số hiệu suất, nguồn lực nhân công, thiết bị, vật tư và logistics, phê

duyet thiết kế, tài chính, thời tiết và điều kiện hiện trường. Tuy nhiên, kết quả nghiên cứu vẫn phân tán theo loại dữ liệu, phạm vi dự án, cách gắn nhãn chậm tiến độ, lựa chọn thuật toán và tiêu chí đánh giá. Ngoài ra, các thách thức về chất lượng dữ liệu, thiếu nhất quán giữa dự án, tính tổng quát của mô hình, cũng như khả năng diễn giải để hỗ trợ ra quyết định quản lý vẫn là rào cản lớn đối với triển khai thực tế.

Vì vậy, bài báo này thực hiện tổng quan có hệ thống về ứng dụng máy học trong dự báo chậm tiến độ dự án xây dựng, nhằm (i) khái quát các hướng tiếp cận và bài toán dự báo thường gặp; (ii) tổng hợp các loại dữ liệu và nhóm biến đầu vào được sử dụng; (iii) phân tích xu hướng lựa chọn mô hình, quy trình xây dựng dữ liệu và tiêu chí đánh giá hiệu suất; và (iv) chỉ ra các thách thức then chốt và định hướng nghiên cứu tương lai để nâng cao tính khả thi ứng dụng trong môi trường quản lý dự án. Về cấu trúc, bài báo lần lượt trình bày tổng quan nền tảng về chậm tiến độ và dữ liệu dự án; các hướng tiếp cận máy học và thiết kế bài toán dự báo theo chu kỳ; tổng hợp, so sánh xu hướng nghiên cứu; thảo luận thách thức triển khai; và đề xuất hướng phát triển, bao gồm tăng cường dữ liệu theo chuỗi thời gian, nâng cao khả năng diễn giải và tích hợp mô hình dự báo vào quy trình quản lý tiến độ.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Bài báo được thực hiện theo phương pháp tổng quan hệ thống nhằm tổng hợp có hệ thống các nghiên cứu về ứng dụng học máy trong dự báo chậm tiến độ dự án xây dựng, đồng thời nhận diện xu hướng phát triển và các thách thức triển khai.

Các tiêu chí lựa chọn được xác định vào: (1) là bài báo tạp chí hoặc hội nghị; (2) tập trung vào dự báo chậm tiến độ hoặc vượt thời gian trong dự án xây dựng; (3) sử dụng mô hình học máy/AI để dự báo (phân loại hoặc hồi quy); và (4) mô tả tối thiểu về dữ liệu đầu vào và cách đánh giá mô hình. Ngược lại, các tài liệu bị loại nếu: (i) không thuộc lĩnh vực xây dựng hoặc không liên quan đến tiến độ; (ii) chỉ phân tích nguyên nhân chậm mà không xây dựng mô hình dự báo; (iii) không có toàn văn hoặc thiếu thông tin phương pháp khiến không thể trích xuất dữ liệu; hoặc (iv) trùng lặp.

Quy trình sàng lọc tài liệu được thực hiện theo các bước PRISMA, gồm: (1) nhận diện tài liệu từ kết quả tìm kiếm và loại bỏ trùng lặp; (2) sàng lọc tiêu đề và tóm tắt theo tiêu chí lựa chọn/loại trừ; (3) đánh giá toàn văn để xác định mức độ phù hợp; và (4) chốt danh mục nghiên cứu cuối cùng đưa vào tổng hợp.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Tổng quan nghiên cứu

Các nghiên cứu nền tảng trong lĩnh vực quản lý xây dựng đã tiếp cận vấn đề chủ yếu theo hướng nhận diện nguyên nhân và phân nhóm yếu tố gây trễ. Assaf chỉ ra các nhóm nguyên nhân nổi bật trong dự án quy mô lớn, bao gồm thay đổi thiết kế, phạm vi, năng lực quản lý, điều phối, nguồn lực và phối hợp giữa các bên [1]. Al-Momani tiếp cận theo hướng định lượng và cho thấy sự liên hệ giữa trễ tiến độ với khâu ra quyết định, phê duyệt và hiệu quả kiểm soát trong giai đoạn thực hiện [2]. Kumaraswamy hệ thống hóa các tác nhân gây trễ theo nhóm chủ thể (chủ đầu tư, tư vấn, nhà thầu) và nhấn mạnh yếu tố bối cảnh dự án chi phối mức độ ảnh hưởng của từng nhóm nguyên nhân [3]. Ở cấp độ tổng hợp, nghiên cứu [4] cho thấy các nhóm nguyên nhân phổ biến có xu hướng lặp lại trên phạm vi toàn cầu, song mức độ ảnh hưởng thay đổi theo điều kiện thị trường, thể chế hợp đồng và năng lực quản trị. Các kết quả này tạo nền tảng quan trọng cho việc xây dựng khung yếu tố, biến đầu vào trong các nghiên cứu tiếp theo, nhưng phần lớn vẫn tập trung vào “giải thích” hơn là “dự báo”.

Trong những năm gần đây, chuyển đổi số trong ngành xây dựng đã tạo điều kiện để dịch chuyển từ cách tiếp cận dựa trên khảo sát sang hướng phân tích dữ liệu và dự báo rủi ro tiến độ. Dữ liệu dự án ngày càng được ghi nhận có hệ thống từ kế hoạch, thực tế tiến độ, nhật ký công trường, hồ sơ vật tư, thiết bị, quản lý chất lượng, an toàn và các thay đổi thiết kế, phê duyệt. Tổng quan về dữ liệu lớn trong xây dựng nhấn mạnh tiềm năng của phân tích dữ liệu và học máy trong hỗ trợ ra quyết định quản lý dự án, đồng thời chỉ ra các thách thức cốt lõi liên quan đến chuẩn hóa, tích hợp và chất lượng dữ liệu [5]. Là vấn đề có ý nghĩa đặc biệt đối với bài toán dự báo chậm tiến độ,

bởi dữ liệu thường mang tính phân tán, không đồng nhất theo dự án và chịu ảnh hưởng của cách thức ghi nhận tại hiện trường.

Song song với số hóa dữ liệu quản trị, các hướng tiếp cận dựa trên dữ liệu hiện trường và mô hình số cũng phát triển mạnh. Nghiên cứu tích hợp dữ liệu đám mây điểm với mô hình BIM theo chuẩn IFC cho thấy khả năng tự động hóa theo dõi tiến độ và phát hiện sai lệch kế hoạch, thực tế [6]. Những hệ thống giám sát tiến độ dạng này không chỉ phục vụ kiểm soát sau sự kiện, mà còn góp phần tạo ra dữ liệu có cấu trúc hơn cho mô hình dự báo theo thời gian. Ngoài ra, việc kết hợp các chỉ báo điều hành thi công với học máy đã được nghiên cứu nhằm dự báo hiệu suất tiến độ dựa trên chuỗi quan sát theo giai đoạn thực hiện, qua đó tăng khả năng cảnh báo sớm trong điều hành công trường [7]. Các kết quả này cho thấy xu hướng chung là tăng cường dữ liệu theo chu kỳ (tuần/tháng) và sử dụng các chỉ báo vận hành thay vì chỉ dựa trên các đặc trưng tĩnh của dự án.

Trên nền tảng dữ liệu ngày càng sẵn có, học máy được xem là một hướng tiếp cận phù hợp cho dự báo rủi ro chậm tiến độ nhờ khả năng xử lý dữ liệu đa biến, mô hình hóa quan hệ phi tuyến và tạo dự báo dạng xác suất. Các nghiên cứu hiện hữu thường mô hình hóa bài toán theo hai dạng chính: (i) phân loại nguy cơ chậm (chậm/không chậm hoặc theo mức rủi ro), và (ii) hồi quy mức độ trễ (số ngày trễ hoặc tỷ lệ vượt tiến độ). Egwim và cộng sự cho thấy các mô hình AI, ML có thể cải thiện hiệu năng dự báo so với tiếp cận truyền thống, đặc biệt khi sử dụng các cấu trúc tổ hợp để tăng tính ổn định trong điều kiện dữ liệu dự án không đồng nhất [8]. Trong nhóm mô hình hiện đại, các thuật toán boosting được quan tâm do hiệu năng tốt trên dữ liệu dạng bảng. Alsulamy thực hiện so sánh CatBoost, XGBoost và LightGBM trong dự báo rủi ro trễ và ghi nhận sự khác biệt hiệu năng theo cách thiết kế bài toán, chất lượng dữ liệu và tiêu chí đánh giá [9]. Nhìn chung, kết quả của các nghiên cứu này cho thấy việc lựa chọn thuật toán chỉ là một phần; hiệu quả dự báo còn phụ thuộc mạnh vào thiết kế biến, cách gán nhãn, chiến lược chia dữ liệu và cách đánh giá mô hình theo mục tiêu quản trị.

Về mặt nền tảng thuật toán, XGBoost được giới thiệu như một hệ thống cây quyết định có khả năng mở rộng và tối ưu hóa tốt, phù hợp với dữ liệu có nhiều tương tác phi tuyến [10]. Trong bối cảnh dự án xây dựng, lợi thế của nhóm mô hình này nằm ở khả năng xử lý tập biến hỗn hợp và khai thác tương tác giữa các nhóm yếu tố (nguồn lực, gián đoạn vật tư, thay đổi thiết kế, điều kiện thi công...). Tuy nhiên, một vấn đề then chốt khi triển khai mô hình dự báo trong quản lý dự án là khả năng diễn giải: nhà quản lý không chỉ cần biết dự án “có nguy cơ chậm”, mà còn cần biết yếu tố nào đang làm tăng rủi ro để ưu tiên can thiệp. Do đó, xu hướng gần đây dịch chuyển sang các mô hình dự báo có tính giải thích cao hơn. Nghiên cứu về phương pháp kết hợp trong bối cảnh xây dựng lắp ghép cho thấy việc kết hợp mô hình tổ hợp với cơ chế giải thích giúp làm rõ vai trò của các yếu tố đầu vào, tăng tính minh bạch và hỗ trợ quyết định quản trị rủi ro tiến độ [11]. Đây là hướng phát triển quan trọng để rút ngắn khoảng cách giữa kết quả mô hình và hành động quản lý tại hiện trường.

Đối với bài báo tổng quan hệ thống, việc trình bày minh bạch quy trình tìm kiếm, sàng lọc và tổng hợp bằng chứng là yêu cầu bắt buộc trong chuẩn mực quốc tế. Các hướng dẫn PRISMA 2020 được sử dụng rộng rãi nhằm chuẩn hóa cách báo cáo tổng quan hệ thống, giúp đảm bảo tính nhất quán, khả năng lặp lại và giảm sai lệch lựa chọn tài liệu [12], [13]. Trong khung tổng quan về ứng dụng học máy dự báo chậm tiến độ, các tiêu chí lựa chọn thường tập trung vào: bối cảnh dự án, loại dữ liệu đầu vào, phương pháp tiền xử lý, thuật toán/mô hình, chiến lược đánh giá và khả năng triển khai thực tế. Đồng thời, các nghiên cứu định lượng về nguyên nhân chậm tiến độ cũng cung cấp cơ sở tham chiếu quan trọng để đánh giá tính phù hợp theo bối cảnh và đề lý giải sự khác biệt kết quả giữa các thị trường [14].

Tại Việt Nam, nghiên cứu về chậm tiến độ vẫn chủ yếu dựa trên khảo sát, phân tích nhân tố và xếp hạng mức độ ảnh hưởng, qua đó cung cấp nền tảng quan trọng để xây dựng bộ biến đầu vào cho mô hình dự báo. Thỉnh thoảng khảo sát các dự án sử dụng vốn Nhà nước và chỉ ra các nguyên nhân tác động mạnh như năng lực tài chính nhà thầu, khó khăn nguồn

vốn của chủ đầu tư, thiếu vật tư, lao động và tổ chức thi công kém; đồng thời rút trích các nhóm nhân tố đại diện [15].

Ở góc độ lập tiến độ, Linh nhấn mạnh hạn chế thực hành khi bố trí thời gian dự phòng tại Việt Nam và đề xuất cách ước lượng dự phòng theo mô phỏng để giảm rủi ro chậm tiến độ do yếu tố ngẫu nhiên [16]. Các yếu tố vận hành tại công trường cũng được xem xét: Hưng và Khánh (2017) nghiên cứu hiện tượng công nhân “ngừng việc hoặc không có việc làm” và rút ra các nhóm nguyên nhân liên quan năng lực chủ đầu tư–tư vấn, năng lực tổ chức thi công của nhà thầu, quản lý tài nguyên và chính sách/tiến độ chưa phù hợp [17].

Về tối ưu hóa tiến độ, Học (2019) tiếp cận bài toán cân bằng thời gian–chi phí trong các dự án có công tác lắp lại và đề xuất cơ chế tối ưu đa mục tiêu, góp phần làm rõ quan hệ đánh đổi giữa rút ngắn thời gian và gia tăng chi phí [18]. Bên cạnh đó, thay đổi thiết kế là một tác nhân quan trọng làm phát sinh trễ tiến độ: Khánh (2020) đánh giá các yếu tố gây thay đổi thiết kế trong dự án nhà, cho thấy tác động đáng kể đến thời gian và chi phí và cung cấp nhóm biến nền cho các mô hình dự báo rủi ro [19].

Tổng hợp các kết quả cho thấy ba xu thế nổi bật gồm: dịch chuyển từ mô tả nguyên nhân sang dự báo dựa trên dữ liệu; gia tăng sử dụng dữ liệu theo chu kỳ và dữ liệu hiện trường, BIM để nâng chất lượng tín hiệu; và tăng cường khả năng diễn giải nhằm hỗ trợ ra quyết định. Tuy nhiên, vẫn tồn tại thách thức về không đồng nhất dữ liệu giữa dự án, thiếu chuẩn hóa định nghĩa “chậm” và quy tắc gắn nhãn theo kỳ, mất cân bằng lớp, cũng như dịch chuyển phân phối khi áp dụng mô hình giữa các bối cảnh khác nhau.

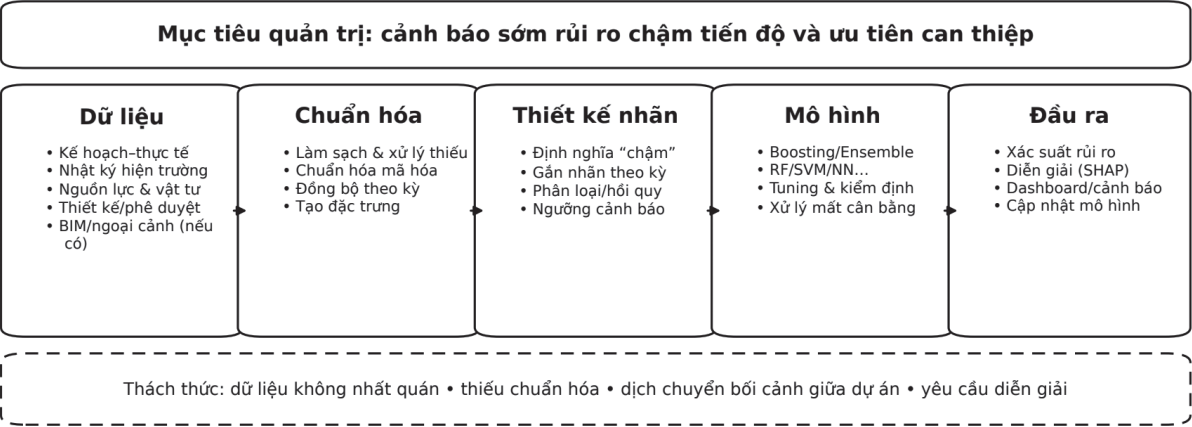
3.2. Thiết kế bài toán dự báo chậm tiến độ và quy tắc gắn nhãn theo chu kỳ

Kết quả tổng hợp cho thấy hiệu quả dự báo phụ thuộc đáng kể vào cách định nghĩa “chậm tiến độ” và cách gắn nhãn. Trong thực tiễn quản lý, “chậm” có thể được hiểu theo nhiều mức: chậm mốc trung gian, chậm hoàn thành, vượt thời gian hợp đồng, hoặc suy giảm hiệu suất ở một giai đoạn thi công. Sự khác biệt định nghĩa dẫn đến khác biệt trong thiết kế bài toán học máy và làm giảm tính so sánh

giữa các nghiên cứu. Do đó, một xu hướng quan trọng là thiết kế bài toán theo chu kỳ cập nhật (tuần/tháng) để phục vụ cảnh báo sớm, phù hợp với thực tế dữ liệu được cập nhật định kỳ từ phần mềm quản lý và báo cáo hiện trường [7]. Khi kết hợp thêm dữ liệu theo dõi tự động từ BIM và đo đạc hiện trường, việc gắn nhãn theo chu kỳ cũng giúp giảm độ trễ và phát hiện sai lệch kế hoạch so với thực tế [6].

Các nghiên cứu chủ yếu triển khai theo hai hướng. Hướng thứ nhất là phân loại, dự báo chậm/không chậm hoặc phân mức rủi ro, phù

hợp với mục tiêu kích hoạt cảnh báo và quy trình kiểm soát theo ngưỡng trong quản lý tiến độ [8], [9]. Hướng thứ hai là hồi quy, dự báo mức độ trễ (số ngày trễ hoặc tỷ lệ trễ so với tiến độ), hữu ích khi cần lượng hóa tác động và liên kết với bài toán tối ưu thời gian, chi phí [18]. Nhìn chung, cách tiếp cận dự báo theo xác suất rủi ro được đánh giá có tính thực hành cao vì phản ánh bất định, giúp nhà quản lý ưu tiên can thiệp theo mức độ rủi ro, nhưng đồng thời đặt yêu cầu cao về lựa chọn thước đo đánh giá và hiệu chỉnh mô hình theo bối cảnh.



Hình 1. Khung tổng hợp các thành phần chính trong dự án chậm tiến độ bằng học máy (tổng hợp từ các nghiên cứu hướng dẫn [5]–[11])

3.3. Dữ liệu đầu vào và nhóm biến: từ dữ liệu tĩnh sang dữ liệu động và dữ liệu hiện trường

Kết quả tổng quan cho thấy dữ liệu đầu vào thường được tổ chức theo hai lớp. Lớp thứ nhất là các biến tĩnh mô tả đặc trưng dự án như loại công trình, quy mô, mức độ phức tạp, cơ chế hợp đồng, cách tổ chức thi công và năng lực các bên. Các biến này thường xuất phát từ khung nguyên nhân gây trễ đã được ghi nhận trong các nghiên cứu nền tảng và tổng quan định lượng [1]–[4], [14], đồng thời được nội địa hóa qua các nghiên cứu bối cảnh Việt Nam [15], [19]. Lớp thứ hai là các biến động theo thời gian phản ánh trạng thái vận hành của dự án theo chu kỳ báo cáo, bao gồm sai lệch kế hoạch và thực tế, chỉ số hiệu suất, biến động nguồn lực nhân công, thiết bị, tình trạng cung ứng vật tư, tiến độ phê duyệt thiết kế, cũng như các yếu tố ngoại cảnh. Xu hướng khai thác dữ liệu động ngày càng rõ, bởi chậm tiến độ thường hình thành từ các tín hiệu suy

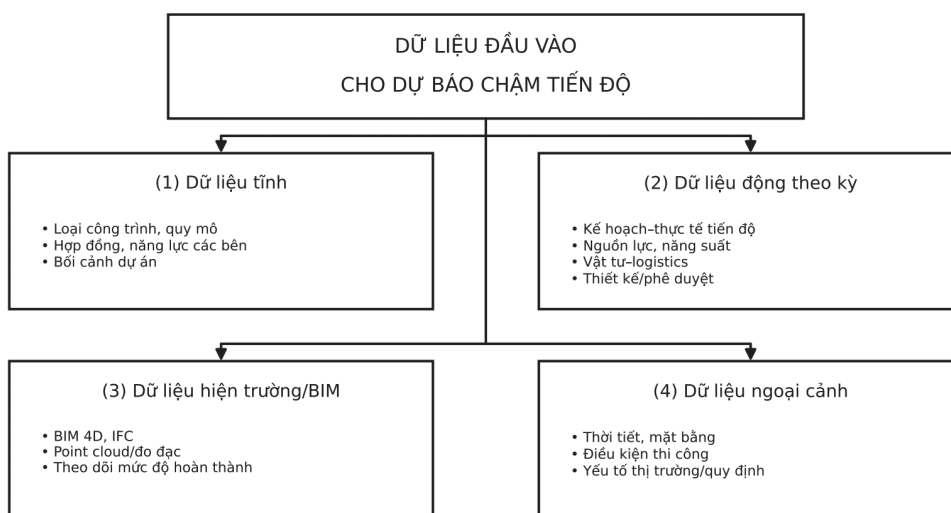
giảm hiệu suất lặp lại theo chu kỳ thay vì chỉ do đặc trưng tĩnh ban đầu.

Một nhóm dữ liệu đáng chú ý là các chỉ báo vận hành ở công trường. Nghiên cứu về hiện tượng công nhân nghỉ việc, cán bộ kỹ thuật không sắp xếp công việc phù hợp cho công nhân, gián đoạn vật tư hoặc vấn đề tổ chức thi công, qua đó gợi ý nhóm biến có ý nghĩa cho bài toán dự báo [17]. Ngoài ra, dữ liệu về dự phòng tiến độ và tính ngẫu nhiên trong lập kế hoạch cũng cần được xem xét, vì thiếu dự phòng hợp lý làm tăng xác suất trễ trước biến động hiện trường [16]. Ở chiều ngược lại, khi yêu cầu đánh đổi thời gian – chi phí xuất hiện, dữ liệu liên quan tăng cường nguồn lực hoặc biện pháp rút ngắn tiến độ cần được mô hình hóa để tránh “tối ưu cục bộ” theo thời gian nhưng gây tăng chi phí quá mức [18].

Bên cạnh dữ liệu quản trị, tổng quan cũng ghi nhận xu hướng gia tăng khai thác dữ liệu hiện trường và mô hình số. Dữ liệu đám mây điểm kết hợp BIM theo IFC cho phép mô hình

hóa mức độ hoàn thành và sai lệch tiến độ tương đối khách quan, hỗ trợ hình thành tập dữ liệu có cấu trúc và cập nhật tốt hơn [6]. Tuy nhiên, về dữ liệu lớn trong xây dựng cho thấy thách thức lớn nằm ở chuẩn hóa và tích hợp dữ liệu đa nguồn, vốn là điều kiện tiên quyết để mô hình học máy đạt hiệu quả ổn

định khi triển khai [5]. Vì vậy, trong nhiều bối cảnh (đặc biệt tại Việt Nam), lộ trình khả thi là ưu tiên chuẩn hóa dữ liệu dạng bảng sẵn có (tiền độ, nhật ký, vật tư, thay đổi thiết kế, nguồn lực) trước, sau đó mở rộng dần sang dữ liệu BIM và dự báo.



Hình 2. Phân loại nhóm dữ liệu, biến đầu vào cho mô hình dự báo chậm tiến độ

3.4. Xu hướng lựa chọn mô hình học máy

Trên dữ liệu dự án dạng bảng và có tương tác phi tuyến, nhóm mô hình cây quyết định và tổ hợp thường cho kết quả dự báo cạnh tranh. Các nghiên cứu so sánh cho thấy các mô hình AI/ML có giám sát có thể nâng cao độ chính xác dự báo so với phương pháp truyền thống, đặc biệt khi áp dụng “ensemble of ensembles” để tăng tính ổn định trước dữ liệu không đồng nhất [8]. Trong nhóm thuật toán hiện đại, boosting nhận được nhiều quan tâm nhờ hiệu năng tốt trên dữ liệu dạng bảng; nghiên cứu đối chiếu CatBoost, XGBoost và LightGBM chỉ ra rằng không có thuật toán “tối ưu tuyệt đối”, mà hiệu năng phụ thuộc vào quy tắc gắn nhãn, chất lượng dữ liệu và tiêu chí đánh giá [9]. XGBoost được xem là một nền tảng quan trọng của nhóm boosting nhờ cơ chế tối ưu hóa và khả năng mở rộng tốt [10].

Tuy nhiên, kết quả tổng hợp cho thấy hiệu năng dự báo cao chưa đủ để phục vụ quản lý dự án nếu mô hình thiếu khả năng diễn giải. Trong quản lý tiến độ, dự báo chỉ thực sự hữu ích khi có thể chỉ ra yếu tố nào đang làm tăng rủi ro để nhà quản lý ưu tiên nguồn lực và biện pháp can thiệp. Do đó, xu hướng gần đây nhấn

nhấn các tiếp cận “có thể giải thích”, trong đó mô hình tổ hợp đi kèm cơ chế giải thích giúp tăng minh bạch và hỗ trợ ra quyết định theo nhóm yếu tố [11]. Hướng này đặc biệt phù hợp với dự án xây dựng, nơi quyết định điều hành thường cần căn cứ vào nhóm nguyên nhân quen thuộc như vật tư, nguồn lực, thiết kế, điều phối đã được ghi nhận rộng rãi trong các nghiên cứu nguyên nhân [1]–[4], [14] và trong bối cảnh Việt Nam [15], [19].

3.5. Đánh giá mô hình, mất cân bằng lớp và tính tổng quát khi triển khai

Tổng hợp các nghiên cứu cho thấy đánh giá mô hình dự báo chậm tiến độ cần gắn với bối cảnh triển khai thực tế thay vì chỉ tối ưu một chỉ số duy nhất. Một thách thức phổ biến là mất cân bằng lớp trong bài toán phân loại, khi tỷ lệ “chậm” hoặc “không chậm” chênh lệch đáng kể tùy theo cách định nghĩa và gắn nhãn. Trong bối cảnh đó, việc chỉ dựa vào độ chính xác có thể gây sai lệch, và cần lựa chọn tiêu chí phản ánh đúng mục tiêu quản trị (ví dụ ưu tiên giảm bỏ sót các trường hợp có nguy cơ cao) [8], [9]. Ngoài ra, với dữ liệu theo chu kỳ, chiến lược chia dữ liệu và đánh giá cần tránh rò rỉ thông tin theo thời gian để đảm bảo

kết quả phản ánh đúng năng lực dự báo trong thực tiễn.

Một vấn đề khác là tính tổng quát khi áp dụng mô hình giữa các dự án khác nhau. Các phân tích định lượng về nguyên nhân chậm tiến độ cho thấy mức độ tác động thay đổi theo bối cảnh thị trường, cơ chế hợp đồng và năng lực quản trị [4], [14]. Điều này đồng nghĩa mô hình học máy có thể gặp dịch chuyển phân phối khi chuyển từ dự án này sang dự án khác, hoặc từ loại công trình này sang loại công trình khác. Vì vậy, các nghiên cứu nhấn mạnh tầm quan trọng của mô tả minh bạch quy trình tổng quan và lựa chọn nghiên cứu theo PRISMA 2020 nhằm bảo đảm tính nhất quán báo cáo và tạo điều kiện so sánh/đối chiếu kết quả [12], [13]. Về triển khai, thách thức không chỉ nằm ở thuật toán mà còn nằm ở quy trình dữ liệu, quy tắc gắn nhãn và cơ chế cập nhật mô hình theo bối cảnh vận hành.

3.6. Hàm ý cho triển khai trong quản lý tiến độ tại Việt Nam

Từ kết quả tổng hợp và đối chiếu với nghiên cứu trong nước, có thể thấy triển khai học máy dự báo chậm tiến độ tại Việt Nam cần tập trung vào ba nội dung. Thứ nhất là chuẩn hóa dữ liệu và thống nhất định nghĩa “chậm” phù hợp mục tiêu quản trị. Các nghiên cứu trong nước về nguyên nhân chậm tiến độ, dự phòng tiến độ và thay đổi thiết kế cung cấp bộ biên nền quan trọng, nhưng dễ dùng cho dự báo cần chuyển hóa thành dữ liệu có cấu trúc và cập nhật theo kỳ [15-16], [19]. Thứ hai là ưu tiên dữ liệu động theo chu kỳ, bởi các tín hiệu suy giảm hiệu suất thường thể hiện qua biến động nguồn lực, gián đoạn vật tư, chậm phê duyệt và các hiện tượng vận hành như ngừng việc, phân công không phù hợp [17]. Việc tổ chức dữ liệu theo tuần/tháng sẽ giúp chuyển từ kiểm soát “phản ứng” sang cảnh báo sớm và điều hành “chủ động”, phù hợp với xu hướng kết hợp LPS và học máy [7]. Thứ ba là cân bằng giữa hiệu năng và khả năng diễn giải, bởi mô hình dự báo chỉ có giá trị khi hỗ trợ ra quyết định can thiệp. Nhóm mô hình boosting và ensemble có tiềm năng về hiệu năng [9-10], nhưng cần gắn với cơ chế giải thích để làm rõ vai trò các yếu tố đầu vào và tăng tính chấp nhận của người dùng [11].

Tóm lại, kết quả và thảo luận cho thấy nghiên cứu đang dịch chuyển theo hướng dự báo rủi ro chậm tiến độ dựa trên dữ liệu, tăng cường khai thác dữ liệu theo chu kỳ và dữ liệu hiện trường, BIM, đồng thời nhấn mạnh nhu cầu diễn giải để phục vụ quyết định quản trị. Trong bối cảnh Việt Nam, trọng tâm nên đặt vào chuẩn hóa dữ liệu từ các nguồn sẵn có, xác lập quy tắc gắn nhãn nhất quán theo mục tiêu cảnh báo, và tích hợp mô hình dự báo vào quy trình điều hành thi công để tăng tính khả thi ứng dụng.

4. KẾT LUẬN

Bài báo đã tổng quan có hệ thống các nghiên cứu về ứng dụng học máy trong dự báo chậm tiến độ dự án xây dựng và làm rõ những xu hướng nổi bật cũng như các thách thức khi triển khai trong thực tiễn. Kết quả tổng hợp cho thấy nghiên cứu đang dịch chuyển từ các tiếp cận mang tính mô tả nguyên nhân sang các mô hình dự báo dựa trên dữ liệu, trong đó dữ liệu dạng bảng từ hệ thống quản lý dự án vẫn là nguồn chủ đạo, đồng thời xuất hiện xu hướng khai thác dữ liệu hiện trường và BIM nhằm nâng cao tính kịp thời và độ tin cậy của thông tin đầu vào.

Về phương pháp, các mô hình học máy có giám sát, đặc biệt là nhóm mô hình tổ hợp và thúc đẩy, được sử dụng phổ biến do khả năng mô hình hóa quan hệ phi tuyến và tương tác biến trong dữ liệu dự án; tuy nhiên hiệu quả dự báo phụ thuộc mạnh vào chất lượng dữ liệu, quy tắc gắn nhãn và chiến lược đánh giá.

Bên cạnh các kết quả tích cực, tổng quan cũng chỉ ra những hạn chế cốt lõi gồm: dữ liệu dự án phân tán và thiếu chuẩn hóa, định nghĩa “chậm tiến độ” và quy tắc gắn nhãn chưa thống nhất, hiện tượng mất cân bằng lớp, dịch chuyển phân phối khi áp dụng mô hình giữa các bối cảnh dự án, và khoảng cách giữa kết quả dự báo với quy trình ra quyết định tại công trường. Do đó, các hướng nghiên cứu tiếp theo cần tập trung vào chuẩn hóa dữ liệu theo chu kỳ (tuần/tháng), xây dựng quy tắc gắn nhãn nhất quán gắn với mục tiêu quản trị rủi ro, phát triển mô hình dự báo có khả năng thích nghi theo bối cảnh, tăng cường tích hợp yếu tố diễn giải để hỗ trợ xác định ưu tiên.

TÀI LIỆU THAM KHẢO

- [1] S. A. Assaf and S. Al-Hejji (2006). Causes of delay in large construction projects. *International Journal of Project Management*, 24(4), 349–357, DOI: 10.1016/j.ijproman.2005.11.010
- [2] A. H. Al-Momani (2000). Construction delay: A quantitative analysis. *International Journal of Project Management*, 18(1), 51–59. DOI: 10.1016/S0263-7863(98)00060-X
- [3] M. M. Kumaraswamy and D. W. M. Chan (1998). Contributors to construction delays,” *Construction Management and Economics*, 16(1), 17–29. DOI: 10.1080/014461998372556
- [4] M.O. Sanni-Anibire, R. Mohamad Zin, and S.O. Olatunji (2022). Causes of delay in the global construction industry: A meta-analytical review. *International Journal of Construction Management*, 22(8), 1395–1407. DOI: 10.1080/15623599.2020.1716132
- [5] M. Bilal *et al.*, (2016). Big data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, 30(3), 500–521. DOI: 10.1016/j.aei.2016.07.001
- [6] P. Kavaliauskas, J.B. Fernandez, K. McGuinness, and A. Jurelionis (2022). Automation of construction progress monitoring by integrating 3D point cloud data with an IFC-based BIM model. *Buildings*, 12(10), 1754, DOI: 10.3390/buildings12101754.
- [7] C.I. Lagos, R.F. Herrera, A.F. Mac Cawley, and L.F. Alarcón (2024). Predicting construction schedule performance with last planner system and machine learning. *Automation in Construction*, 167, 105716, DOI: 10.1016/j.autcon.2024.105716
- [8] C.N. Egwim, H. Alaka, L.O. Toriola-Coker, H. Balogun, and F. Sunmola (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications*, 6, 100166, DOI: 10.1016/j.mlwa.2021.100166
- [9] S. Alsulamy (2025). Predicting construction delay risks in Saudi Arabian projects: A comparative analysis of CatBoost, XGBoost, and LGBM. *Expert Systems with Applications*, 268, 126268, DOI: 10.1016/j.eswa.2024.126268
- [10] T. Chen and C. Guestrin (2016). XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794, DOI: 10.1145/2939672.2939785.
- [11] Y. Tian, Y. Chen, and Y. Ning (2026). Identifying and predicting delay risks in prefabricated construction: An explainable ensemble learning approach. *Advanced Engineering Informatics*, 71, 104219, DOI: 10.1016/j.aei.2025.104219
- [12] M.J. Page *et al.*, (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, 71, DOI: 10.1136/bmj.n71
- [13] M. J. Page *et al.*, (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, 134, 178–189, DOI: 10.1016/j.jclinepi.2021.03.001
- [14] E. Viles, N.C. Rudeli, and A. Santilli (2020). Causes of delay in construction projects: A quantitative analysis. *Engineering, Construction and Architectural Management*, 27(4), 917–935, DOI: 10.1108/ECAM-01-2019-0024
- [15] L.L. Thịnh (2017). Khảo sát và xây dựng mô hình đánh giá nguyên nhân chậm trễ tiến độ trong các dự án đầu tư xây dựng sử dụng vốn nhà nước. *Tạp Chí Khoa học Công nghệ xây dựng*, 11(1), 75–81.
- [16] L.Đ. Linh, N.Q. Toàn, N.H. Hải, and V.K. Cường (2016). Sử dụng phương pháp mô phỏng Monte Carlo để ước lượng thời gian dự phòng khi lập tiến độ thi công xây dựng. *Tạp Chí Khoa học công nghệ xây dựng*, 10(3), 64–67.
- [17] P.V. Hưng, H.D. Khánh (2017). Xác định nguyên nhân và giải pháp khắc phục sự ngừng việc của công nhân trên công trường xây dựng tại Vĩnh Long. *Tạp Chí Khoa học công nghệ xây dựng*, 11(2), 81–87.
- [18] T. Đức Học (2019). Tối ưu cân bằng thời gian chi phí trong tiến độ các dự án có công tác lặp lại. *Tạp Chí Khoa học công nghệ xây dựng*, 13(1V), 56–65, DOI: 10.31814/stce.nuce2019-13(1V)-06
- [19] H.D. Khanh (2020). Factors causing design changes in Vietnamese residential construction projects: An evaluation and comparison. *Journal of Science and Technology in Civil Engineering (JSTCE)*, NUCE, 14(3), 151–166, DOI: 10.31814/stce.nuce2020-14(3)-13