XÁC ĐỊNH CHI PHÍ ĐẦU TƯ XÂY DỰNG CỦA DỰ ÁN SỬ DỤNG VỐN ĐẦU TƯ CÔNG TẠI BƯỚC LẬP, THẨM ĐỊNH BÁO CÁO NGHIÊN CỨU KHẢ THI TRONG GIAI ĐOẠN CHUẨN BỊ DỰ ÁN BẰNG MÔ HÌNH HỌC MÁY TỐI ƯU

DETERMINING THE CONSTRUCTION INVESTMENT COSTS OF PUBLIC INVESTMENT PROJECTS AT THE FEASIBILITY STUDY PREPARATION AND APPRAISAL STAGE DURING THE PROJECT PREPARATION PHASE USING AN OPTIMAL MACHINE LEARNING MODEL.

ThS. Trần Quang Lâm[1], PGS.TS Trần Đức Học[2], NCS. Trần Nhật Quang[1,2*]
[1]Sở Xây dựng Thành Phố Hồ Chí Minh / [2]Trường Đại học Bách Khoa, Đại học Quốc gia Thành phố Hồ Chí Minh
** Tác giả liên hệ: tnquang.sdh251@hcmut.edu.vn*

Tóm tắt: Sự chính xác của việc xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công tại bước lập, thẩm định báo cáo nghiên cứu khả thi đóng vai trò quan trọng trong quá trình ra quyết định phê duyệt dự án, triển khai các bước tiếp theo và quản lý ngân sách; đặc biệt trong bối cảnh Luật Xây dựng 2025 (có hiệu lực thi hành từ ngày 01 tháng 7 năm 2026) đã bãi bỏ quy định về thẩm định thiết kế xây dựng triển khai sau thiết kế cơ sở của cơ quan chuyên môn về xây dựng (bao gồm nội dung thẩm định dự toán xây dựng công trình). Tuy nhiên, việc xác định chi phí ở giai đoạn đầu thường gặp nhiều khó khăn do sự hạn chế về thông tin và mức độ bất định cao. Nghiên cứu này đề xuất một mô hình học máy lai ghép kết hợp giữa hồi quy vectơ hỗ trợ bình phương tối thiểu và thuật toán tối ưu Cheetah sử dụng dữ liệu về chi phí của các dự án, công trình tương tự đã thực hiện nhằm nâng cao độ chính xác trong việc xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công. Trong đó, thuật toán Cheetah được sử dụng để tối ưu hóa các siêu tham số của mô hình LSSVR, qua đó cải thiện khả năng tổng quát hóa và độ ổn định của mô hình. Bộ dữ liệu nghiên cứu bao gồm 50 công trình đầu tư công được thu thập tại Thành phố Hồ Chí Minh, phản ánh đặc điểm thực tiễn của các dự án xây dựng trong điều kiện đô thị. Hiệu quả của mô hình đề xuất được đánh giá và so sánh với các phương pháp phổ biến gồm máy vectơ hỗ trợ (SVM) và mạng nơ-ron nhân tạo (ANNs). Kết quả nghiên cứu cho thấy mô hình CO-LSSVR vượt trội hơn các mô hình so sánh, đạt độ chính xác dự báo cao với trung bình 49,38, MAE trung bình 38,15, MAPE trung bình 11,92% và R^2 đạt 0,883. Các kết quả này khẳng định hiệu quả của việc kết hợp thuật toán tối ưu metaheuristic với mô hình học máy trong bài toán xác định chi phí đầu tư xây dựng tại bước lập, thẩm định báo cáo nghiên cứu khả thi của dự án. Mô hình đề xuất cung cấp cho người quyết định đầu tư, cơ quan quản lý nhà nước và các chủ đầu tư một công cụ hữu hiệu nhằm nâng cao độ tin cậy của việc xác định chi phí và hỗ trợ ra quyết định trong các dự án đầu tư công.

Từ khoá: Ước tính chi phí đầu tư xây dựng, Dự án đầu tư công, Báo cáo nghiên cứu khả thi, Hồi quy vectơ hỗ trợ bình phương tối thiểu, Thuật toán Cheetah, Học máy.

Abstract: The accuracy of determining the construction investment costs of public-funded projects at the feasibility study stage plays a crucial role in the decision-making process for project approval, implementation of subsequent steps, and budget management; especially in the context of the 2025 Construction Law (effective from July 1, 2026), which abolishes the regulation on the appraisal of construction design following the basic design stage by specialized construction agencies (including the appraisal of construction cost estimates). However, determining costs at the initial stage often faces many difficulties due to limited information and high uncertainty. This study proposes a hybrid machine learning model combining least-squares support vector regression (LSSVR) and the Cheetah optimization algorithm using cost data from similar projects and works already implemented to improve the accuracy in determining the construction investment costs of public-funded projects. In this study, the Cheetah optimizer was used to optimize the hyperparameters of the LSSVR model, thereby improving the model's generalization ability and stability. The research dataset includes 50 public investment projects collected in Ho Chi Minh City, reflecting the practical characteristics of construction projects in urban conditions. The effectiveness of the proposed model was evaluated and compared with common methods including support vector machines (SVM) and artificial neural networks (ANNs). The research results showed that the CO–LSSVR model outperformed the comparison models, achieving high prediction accuracy with an average of 49,38, average MAE of 38,15, average MAPE of 11,92%, and R^2 of 0,883. These results confirm the effectiveness of combining metaheuristic optimization algorithms with machine learning models in estimating construction investment costs during the feasibility study phase of a project. The proposed model provides investment decision-makers, government agencies, and investors with an effective tool to improve the reliability of cost estimates and support decision-making in public investment projects.

Key words: Construction investment cost estimation, Public investment projects, Feasibility study reports, Support Vector Regression (LSSVR), Cheetah Optimizer, Machine learning.

1. Mở đầu

Xác định chi phí đầu tư xây dựng ở giai đoạn chuẩn bị dự án là một trong những nhiệm vụ quan trọng nhất nhưng cũng tiềm ẩn nhiều rủi ro trong quá trình lập, thẩm định và phê duyệt dự án đầu tư xây dựng. Mức độ chính xác của việc xác định chi phí đầu tư xây dựng thuộc tổng mức đầu tư trong Báo cáo nghiên cứu khả thi đầu tư xây dựng ảnh hưởng trực tiếp đến việc ra quyết định phê duyệt dự án, lập và phân bổ kế hoạch vốn, cũng như hiệu quả quản lý chi phí trong suốt quá trình thực hiện dự án [1]. Đối với các dự án đầu tư công, vai trò này càng trở nên quan trọng do sự yêu cầu nghiêm ngặt của các quy định về quản lý ngân sách nhà nước và trách nhiệm trong quản lý và sử dụng nguồn lực công. Trong thực tiễn, việc áp dụng các quy định của pháp luật hiện hành để xác định chi phí đầu tư xây dựng của dự án tại bước lập báo cáo nghiên cứu khả thi thường gặp nhiều khó khăn do mức độ hoàn thiện, sự chi tiết của hồ sơ thiết kế cơ sở còn hạn chế, sự bất định cao của các yếu tố thị trường, sự thiếu tính đặc thù của công trình nếu sử dụng suất vốn đầu tư xây dựng do cơ quan nhà nước có thẩm quyền công bố và sự khó khăn trong việc xác định tính tương đương nếu sử dụng dữ liệu về chi phí các dự án, công trình đã thực hiện. Thời gian thực hiện dự án kéo dài trong nhiều năm có thể khiến giá vật liệu, nhân công, máy thi công và thiết bị có thể biến động đáng kể, đặc biệt trong bối cảnh thị trường xây dựng ngày càng chịu tác động mạnh từ các yếu tố kinh tế – xã hội và chuỗi cung ứng toàn cầu. Nhiều nghiên cứu đã chỉ ra rằng các yếu tố ảnh hưởng đến chất lượng ước tính chi phí bao gồm thông tin và thông số thiết kế, các yếu tố bên ngoài như giá và khả năng cung ứng vật liệu, cũng như năng lực và hiệu quả thực hiện của các nhà thầu. Những yếu tố này làm gia tăng sự sai lệch giữa chi phí xác định ban đầu và chi phí thực hiện thực tế, từ đó ảnh hưởng đến tiến độ, hiệu quả và tính bền vững của dự án [2, 3].

Đối với các dự án đầu tư công, sai lệch chi phí ở giai đoạn lập dự án không chỉ gây tác động tiêu cực đến từng dự án cụ thể mà còn ảnh hưởng đến công tác cân đối ngân sách và hiệu quả phân bổ vốn đầu tư công trong trung hạn và dài hạn. Trong nhiều trường hợp, xác định chi phí không sát với thực tế dẫn đến việc điều chỉnh tổng mức đầu tư nhiều lần ở các giai đoạn sau, kéo dài thời gian thực hiện dự án và làm giảm hiệu quả sử dụng vốn nhà nước, lãng phí nguồn lực công. Nguyên nhân chủ yếu xuất phát từ việc các phương pháp xác định chi phí đầu tư xây dựng truyền thống theo quy định hiện nay vẫn dựa nhiều vào suất vốn đầu tư, định mức và kinh nghiệm của người lập tổng mức, chưa phản ánh đầy đủ các đặc điểm riêng của từng dự án và biến động của thị trường xây dựng. Trước những hạn chế nêu trên, nhiều phương pháp và công cụ để ước tính, xác định chi phí đầu tư xây dựng đã được nghiên cứu và ứng dụng, từ các phương pháp định tính và định lượng truyền thống đến các kỹ thuật tiên tiến dựa trên mô hình học máy và trí tuệ nhân tạo. Các phương pháp định lượng có thể được phân thành các nhóm chính như phương pháp thống kê, mô hình tương tự và mô hình phân tích. Trong số đó, các mô hình học máy như mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN) và máy vectơ hỗ trợ (Support Vector Machine - SVM) đã được chứng minh có khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp giữa chi phí xây dựng và các biến đầu vào, từ đó cải thiện độ chính xác của dự báo chi phí ở giai đoạn sớm [4].

Tuy nhiên, hiệu quả của các mô hình học máy phụ thuộc lớn vào việc lựa chọn cấu trúc mô hình và các siêu tham số phù hợp. Việc xác định các siêu tham số này bằng phương pháp thử – sai truyền thống thường tốn nhiều thời gian và không đảm bảo đạt được nghiệm tối ưu toàn cục. Do đó, xu hướng kết hợp các mô hình học máy với các thuật toán tối ưu metaheuristic đang ngày càng được quan tâm nhằm tự động hóa quá trình tối ưu siêu tham số và nâng cao khả năng tổng quát hóa của mô hình. Trong bối cảnh đó, hồi quy vectơ hỗ trợ bình phương tối thiểu (Least Squares Support Vector Regression - LSSVR) nổi lên như một biến thể hiệu quả của SVM trong các bài toán hồi quy, nhờ khả năng xử lý dữ liệu phi tuyến và giảm chi phí tính toán [5]. Việc kết hợp LSSVR với các thuật toán tối ưu metaheuristic có tiềm năng cải thiện đáng kể hiệu quả dự báo, đặc biệt trong các bài toán xác định chi phí đầu tư xây dựng ở giai đoạn chuẩn bị dự án. Thuật toán tối ưu báo săn (Cheetah Optimizer - CO) [6], với khả năng cân bằng hiệu quả giữa khai thác và khám phá không gian tìm kiếm, là một công cụ phù hợp để tối ưu hóa các siêu tham số của mô hình LSSVR.

Căn cứ những phân tích trên và cơ sở pháp lý từ phương pháp xác định tổng mức đầu tư từ dữ liệu về chi phí của các dự án, công trình tương tự đã thực hiện [7], nghiên cứu này đề xuất một mô hình học máy lai ghép kết hợp giữa LSSVR và thuật toán Cheetah nhằm xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công ở giai đoạn chuẩn bị dự án. Mô hình được kiểm chứng thông qua bộ dữ liệu gồm 50 dự án dân dụng sử dụng vốn đầu tư công được thu thập tại Thành phố Hồ Chí Minh và được so sánh với các mô hình phổ biến như SVM và ANN. Mục tiêu của nghiên cứu là phát triển một công cụ xác định chi phí có độ tin cậy cao,

góp phần hỗ trợ cho người quyết định đầu tư, cơ quan quản lý nhà nước và các chủ đầu tư trong việc nâng cao độ tin cậy của việc ước tính chi phí đầu tư xây dựng trong quá trình lập, thẩm định báo cáo nghiên cứu khả thi cũng như phê duyệt dự án đầu tư xây dựng. Phần còn lại của bài báo được tổ chức như sau: Phần 2 trình bày tổng quan các nghiên cứu liên quan; Phần 3 trình bày phương pháp nghiên cứu; Phần 4 mô tả bộ dữ liệu, phân tích và thảo luận kết quả; và Phần 5 đưa ra kết luận cùng các định hướng nghiên cứu tiếp theo.

2. Tổng quan về các nghiên cứu liên quan

Trong những năm gần đây, học máy (machine learning), với vai trò là một nhánh quan trọng của trí tuệ nhân tạo, đã được ứng dụng rộng rãi trong bài toán ước tính chi phí đầu tư xây dựng, đặc biệt ở giai đoạn đầu của dự án khi thông tin còn hạn chế. Ưu điểm nổi bật của các mô hình học máy là khả năng khám phá và mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các biến đặc trưng của dự án và chi phí cuối cùng, điều mà các phương pháp truyền thống khó có thể thực hiện hiệu quả. Do đó, nhiều nghiên cứu đã tập trung phát triển và so sánh các chiến lược mô hình hóa khác nhau nhằm nâng cao độ chính xác của dự đoán chi phí giai đoạn đầu [8].

Một trong những nghiên cứu tiêu biểu trong lĩnh vực này là của Gwang-Hee Kim [9] trong đó tác giả so sánh hiệu quả của ba kỹ thuật gồm phân tích hồi quy, mạng nơ-ron nhân tạo và máy vectơ hỗ trợ trong ước lượng chi phí xây dựng trường học – một loại hình công trình công cộng. Nghiên cứu sử dụng bộ dữ liệu gồm 217 dự án xây dựng trường học được thực hiện tại tỉnh Kyeonggi (Hàn Quốc) trong giai đoạn 2004–2007. Dữ liệu được chia thành tập huấn luyện, tập kiểm định chéo và tập kiểm tra để đánh giá hiệu suất mô hình. Kết quả cho thấy mô hình ANN đạt độ chính xác cao nhất với sai số tuyệt đối trung bình thấp hơn đáng kể so với hồi quy truyền thống và SVM, qua đó khẳng định tiềm năng của ANN trong bài toán ước tính chi phí đầu tư xây dựng công trình công cộng.

Tiếp nối hướng nghiên cứu này, nhiều tác giả đã chuyển sang ứng dụng các kỹ thuật học máy trong xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công nhằm khắc phục các hạn chế của phương pháp truyền thống [10]. Alrasheed, et al. [4] đã tiến hành nghiên cứu các yếu tố chính ảnh hưởng đến ước tính chi phí đầu tư xây dựng và phát triển mô hình mạng nơ-ron nhân tạo để nâng cao độ chính xác dự báo chi phí trước khi xây dựng. Nghiên cứu sử dụng dữ liệu thu thập từ nhiều dự án xây dựng công cộng tại Kuwait và thử nghiệm 16 cấu hình ANN khác nhau để xác định cấu trúc tối ưu. Kết quả đánh giá bằng chỉ số sai số phần trăm tuyệt đối trung bình cho thấy mô hình ANN đạt sai số chỉ 0,72%, tương ứng với độ chính xác 99,28%. Mô hình tiếp tục được kiểm chứng trên các dự án độc lập, chứng minh tính mạnh mẽ và khả năng ứng dụng thực tiễn trong bối cảnh đầu tư công.

Bên cạnh ANN, các mô hình dựa trên máy vectơ hỗ trợ, đặc biệt là hồi quy vectơ hỗ trợ và hồi quy vectơ hỗ trợ bình phương tối thiểu, cũng được nhiều nhà nghiên cứu quan tâm do khả năng tổng quát hóa tốt và hiệu quả xử lý dữ liệu phi tuyến. Chen, et al. [11] đã nghiên cứu bài toán ước tính chi phí sơ bộ các dự án xây dựng tại Indonesia bằng mô hình SVR. Thông qua tổng quan tài liệu, nhóm tác giả xác định 14 yếu tố ảnh hưởng chính đến chi phí xây dựng và thu thập dữ liệu từ 104 dự án thực tế. Mô hình SVR sử dụng hàm nhân cơ sở xuyên tâm được huấn luyện và đánh giá bằng phương pháp kiểm chứng chéo năm lần, cho kết quả độ chính xác trung bình lên đến 96% và đạt hiệu quả cao trong việc giảm thời gian ước tính chi phí sơ bộ.

Ở hướng nghiên cứu nâng cao, Cheng and Hoang [12] đã đề xuất một mô hình lai ghép sử dụng hồi quy vectơ hỗ trợ bình phương tối thiểu ước lượng khoảng dựa trên học máy và thuật toán tiến hóa vi phân để dự đoán chi phí đầu tư xây dựng của dự án. Trong nghiên cứu này, thuật toán DE được sử dụng để tối ưu hóa các tham số của mô hình LS-SVM trong quá trình kiểm định chéo. Mô hình đề xuất không chỉ cung cấp giá trị dự đoán điểm mà còn xác định khoảng dự đoán trên và dưới nhằm phản ánh mức độ không chắc chắn của dự báo. Kết quả mô phỏng và so sánh cho thấy mô hình lai ghép đạt độ chính xác cao và độ tin cậy tốt hơn so với các phương pháp truyền thống.

Tại Việt Nam, các nghiên cứu ứng dụng học máy trong ước tính chi phí đầu tư xây dựng, đặc biệt cho các dự án đầu tư công, đã bước đầu được triển khai nhưng số lượng còn hạn chế. Khoa [13] đã ứng dụng mạng nơ-ron nhân tạo để ước lượng chi phí đầu tư xây dựng dự án chung cư, sử dụng bộ dữ liệu gồm 14 công trình với các biến đầu vào liên quan đến quy mô, cấp công trình và giá vật liệu. Nghiên cứu cho thấy ANN có khả năng tự động hóa quá trình ước lượng chi phí và cho kết quả khả quan trong điều kiện dữ liệu hạn chế. Quang [14] đã phát triển mô hình ANN để xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công, công trình Trường trung học phổ thông tại Thành phố Hồ Chí Minh. Ngoài ra, Phong, et al. [15] đã ứng dụng ANN trong ước lượng chi phí xây dựng nhà xưởng ở giai đoạn đấu thầu, sử dụng dữ liệu từ 35 công trình và 11 yếu tố đầu vào, cho thấy mô hình học máy có tiềm năng cải thiện độ chính xác

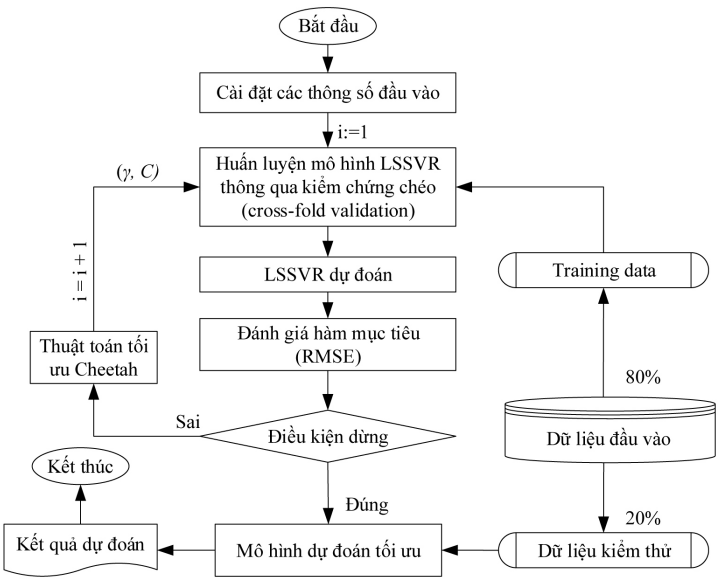
so với phương pháp truyền thống.

Từ tổng quan các nghiên cứu trong và ngoài nước có thể nhận thấy rằng, các mô hình học máy, đặc biệt là ANN và SVM/LSSVR, đã chứng minh được hiệu quả trong việc nâng cao độ chính xác xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công. Tuy nhiên, phần lớn các nghiên cứu hiện nay vẫn tập trung vào việc áp dụng các mô hình học máy đơn lẻ hoặc các thuật toán tối ưu truyền thống. Việc khai thác các thuật toán tối ưu metaheuristic mới để tối ưu hóa mô hình LSSVR cho bài toán xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công, đặc biệt dựa trên dữ liệu thực tế tại các đô thị lớn ở Việt Nam, vẫn còn là một khoảng trống nghiên cứu. Nghiên cứu này được thực hiện nhằm bổ sung vào khoảng trống đó thông qua việc đề xuất và kiểm chứng mô hình LSSVR được tối ưu hóa bằng thuật toán Cheetah.

3. Đề xuất mô hình CO-LSSVR

Mục này trình bày mô hình đề xuất CO-LSSVR, được xây dựng trên cơ sở tích hợp giữa hồi quy vectơ hỗ trợ bình phương tối thiểu (Least Squares Support Vector Regression – LSSVR) và thuật toán tối ưu Cheetah (Cheetah Optimizer – CO). Quy trình hoạt động tổng thể của mô hình lai ghép CO-LSSVR được minh họa trong Hình 1. Trong mô hình đề xuất, LSSVR được sử dụng như một thuật toán học có giám sát nhằm thiết lập mối quan hệ hàm phi tuyến tiềm ẩn giữa các biến đầu vào và chi phí xây dựng đầu ra. Thuật toán Cheetah đóng vai trò như một thủ tục tối ưu hóa, được tích hợp để tự động xác định các siêu tham số tối ưu của LSSVR, bao gồm hệ số phạt C và tham số của hàm nhân γ, nhằm nâng cao độ chính xác dự báo của mô hình.

Tập dữ liệu lịch sử được chia thành hai phần, trong đó 80% dữ liệu được sử dụng cho huấn luyện và 20% dữ liệu còn lại được dành cho kiểm tra (testing) nhằm đánh giá khả năng tổng quát hóa của mô hình. Trong quá trình huấn luyện, để tinh chỉnh các siêu tham số của LSSVR và tránh hiện tượng quá khớp, phương pháp kiểm định chéo 5-fold được áp dụng trực tiếp trên tập huấn luyện [16]. Cụ thể, tập dữ liệu huấn luyện được chia thành năm phần con có kích thước tương đương; trong mỗi vòng lặp, bốn phần được sử dụng để huấn luyện mô hình, trong khi phần còn lại được dùng để đánh giá hiệu suất thông qua hàm mục tiêu, điển hình là sai số căn phương trung bình (RMSE). Giá trị trung bình của các chỉ số đánh giá qua năm lần lặp được sử dụng làm tiêu chí dẫn hướng cho quá trình tối ưu của thuật toán Cheetah. Thông qua cơ chế kết hợp giữa học máy và tối ưu metaheuristic, mô hình CO-LSSVR không chỉ tận dụng được khả năng học phi tuyến mạnh mẽ của LSSVR mà còn khắc phục hạn chế trong việc lựa chọn siêu tham số thủ công, từ đó nâng cao độ tin cậy và độ chính xác đối với việc xác định chi phí đầu tư xây dựng của dự án trong giai đoạn chuẩn bị dự án, đặc biệt đối với các dự án sử dụng vốn đầu tư công.



Hình 1. Mô hình CO-LSSVR cho bài toán xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công

4. Áp dụng thực nghiệm

4.1. Dữ liệu

Bộ dữ liệu sử dụng trong nghiên cứu bao gồm 50 dự án đầu tư xây dựng công trình dân dụng sử dụng vốn đầu tư công, được thu thập tại Thành phố Hồ Chí Minh, phản ánh tương đối đầy đủ đặc điểm thực tiễn của các dự án xây dựng trong bối cảnh đô thị lớn. Các công trình trong tập dữ liệu được triển khai trong giai đoạn năm 2022-2025, có đầy đủ hồ sơ pháp lý và thông tin về chi phí đầu tư xây dựng rõ ràng, bảo đảm độ tin cậy cho mục tiêu xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công ở giai đoạn chuẩn bị dự án. Việc tập trung vào nhóm dự án đầu tư xây dựng công trình dân dụng giúp duy trì tính đồng nhất về công năng sử dụng, đồng thời vẫn thể hiện được sự đa dạng về quy mô, cấp công trình, điều kiện xây dựng và tổng mức đầu tư, qua đó nâng cao khả năng khái quát hóa của các mô hình dự báo.

Trên cơ sở tổng hợp từ hồ sơ dự án và tham khảo các nghiên cứu liên quan về dự báo chi phí xây dựng, bộ dữ liệu được cấu trúc với mười biến đầu vào (X_1–X_{10}) và một biến đầu ra (Y). Các biến đầu vào bao quát những đặc trưng chủ yếu của công trình, bao gồm quy mô xây dựng, đặc điểm kỹ thuật và điều kiện triển khai dự án. Cụ thể, các biến định lượng gồm tổng diện tích sàn xây dựng (X_1),

Bảng 1. Thống kê mô tả các biến đầu vào và đầu ra của bộ dữ liệu

Biến	Ký hiệu	Mô tả / Đơn vị	Min	Max	Mean	Std
Tổng diện tích sàn	X_1	m^2	1.759,5	50.113,4	10.073,64	9.406,37
Số tầng nổi	X_2	Tầng	2	15	5,8	3,64
Số tầng hầm	X_3	Tầng	0	3	0,7	0,81
Diện tích khu đất	X_4	m^2	879,75	20.484,7	3.523,26	3.316,6
Cấp công trình	X_5	Phân loại (I–III)	I	III	–	–
Loại kết cấu	X_6	Phân loại	BTCT	BTCT + vì kèo thép	–	–
Năm xây dựng	X_7	Năm	2022	2025	2023,32	0,99
Mức độ hoàn thiện & thiết bị	X_8	Thấp–TB–Cao	Thấp	Cao	–	–
Điều kiện xây dựng	X_9	Phân loại	Rộng rãi	Chật	–	–
Địa điểm xây dựng	X_{10}	Quận/Huyện TP.HCM	–	–	–	–
Chi phí xây dựng	Y	Tỷ đồng	42,09	761,66	172,38	186,53

số tầng nổi (X_2), số tầng hầm (X_3), diện tích khu đất (X_4) và năm xây dựng (X_7), trong khi các biến định tính phản ánh cấp công trình (X_5), loại kết cấu chính (X_6), mức độ hoàn thiện và sử dụng thiết bị (X_8), điều kiện xây dựng (X_9) và địa điểm xây dựng cụ thể tại các quận, huyện cũ của Thành phố Hồ Chí Minh (X10). Biến đầu ra Y là chi phí đầu tư xây dựng của dự án, được tính theo đơn vị tính là tỷ đồng, đóng vai trò là đại lượng mục tiêu trong bài toán dự báo.

Bảng 1 trình bày thống kê mô tả của toàn bộ các biến trong tập dữ liệu. Đối với các biến định lượng, các chỉ tiêu thống kê bao gồm giá trị nhỏ nhất (Min), giá trị lớn nhất (Max), giá trị trung bình (Mean) và độ lệch chuẩn (Std). Kết quả cho thấy tổng diện tích sàn dao động từ 1.759,5 m^2 đến 50.113,4 m^2, với giá trị trung bình 10.073,64 m^2, trong khi chi phí đầu tư xây dựng biến thiên trong khoảng từ 42,09 đến 761,66 tỷ đồng, phản ánh mức độ đa dạng đáng kể về quy mô và tổng mức đầu tư của các dự án trong tập dữ liệu. Các biến định tính được mô tả theo phạm vi giá trị và được mã hóa phù hợp trước khi đưa vào mô hình học máy. Sau khi kiểm tra tính nhất quán, chuẩn hóa dữ liệu và xử lý các giá trị bất thường, toàn bộ tập dữ liệu được sử dụng làm đầu vào cho các mô hình ANN, SVM, LSSVR và mô hình đề xuất CO-LSSVR trong các bước phân tích và so sánh tiếp theo.

4.2. Kết quả và so sánh

Hiệu quả xác định chi phí đầu tư xây dựng của mô hình đề xuất CO-LSSVR được đánh giá và so sánh với các mô hình học máy phổ biến, bao gồm LSSVR, mạng nơ-ron nhân tạo và máy vectơ hỗ trợ [17]. Toàn bộ quá trình đánh giá được thực hiện bằng phương pháp kiểm định chéo 5-fold nhằm đảm bảo tính khách quan và khả năng tổng quát hóa của các mô hình. Trong mỗi vòng lặp, bốn phần dữ liệu được sử dụng để huấn luyện, trong khi phần còn lại được dùng để kiểm tra; các chỉ số đánh giá cuối cùng được xác định dựa trên giá trị trung bình của năm lần lặp.

Các mô hình được đưa vào so sánh bao gồm mô hình đề xuất CO-LSSVR, LSSVR truyền thống, ANN và SVM. Các tham số của các mô hình đối sánh được lựa chọn dựa trên khuyến nghị từ các nghiên cứu trước kết hợp với phương pháp thử–sai, nhằm đảm bảo tính khách quan, khả năng tái lập và hiệu quả thực tiễn. Đối với các mô hình sử dụng hàm nhân RBF, các tham số quan trọng như hệ số phạt C và tham số nhân γ được lựa chọn trong các khoảng giá trị thường được sử dụng trong các bài toán dự báo chi phí đầu tư xây dựng, nhằm cân bằng giữa độ chính xác và khả năng tổng quát hóa. Để đánh giá toàn diện hiệu suất dự báo, nghiên cứu sử dụng đồng thời bốn chỉ số phổ biến gồm hệ số xác định (R^2), sai số căn phương trung bình (RMSE), sai số tuyệt đối trung bình (MAE) và sai số phần trăm tuyệt đối trung bình (MAPE); công thức tính các chỉ số này được trình bày trong các phương trình (1) đến (4).

$$R^2 = 1 - \sum_{i=1}^{n}\left(y_i - p_i\right)^2 / \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \quad (1)$$

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(y_i - p_i\right)^2} \quad (2)$$

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left|y_i - p_i\right| \quad (3)$$

$$MAPE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(\frac{\left|y_i - p_i\right|}{y_i}\right) * 100 \quad (4)$$

Trong đó y_i và p_i lần lượt là giá trị thực tế và giá trị dự báo; $\bar{y}$ là giá trị trung bình của các quan sát thực tế; và n là số lượng mẫu dữ liệu.

Bảng 2 trình bày kết quả so sánh hiệu suất dự

Bảng 2. So sánh kết quả giữa các mô hình

Mô hình	RMSE		MAE		MAPE (%)		R^2	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
ANN	58,74	2,91	45,63	2,38	14,92	0,84	0,842	0,018
SVM	55,18	2,64	42,87	2,11	13,96	0,77	0,858	0,016
LSSVR	52,06	2,32	40,21	1,94	12,84	0,69	0,871	0,014
CO-LSSVR (Nghiên cứu này)	49,38	2,08	38,15	1,73	11,92	0,62	0,883	0,012

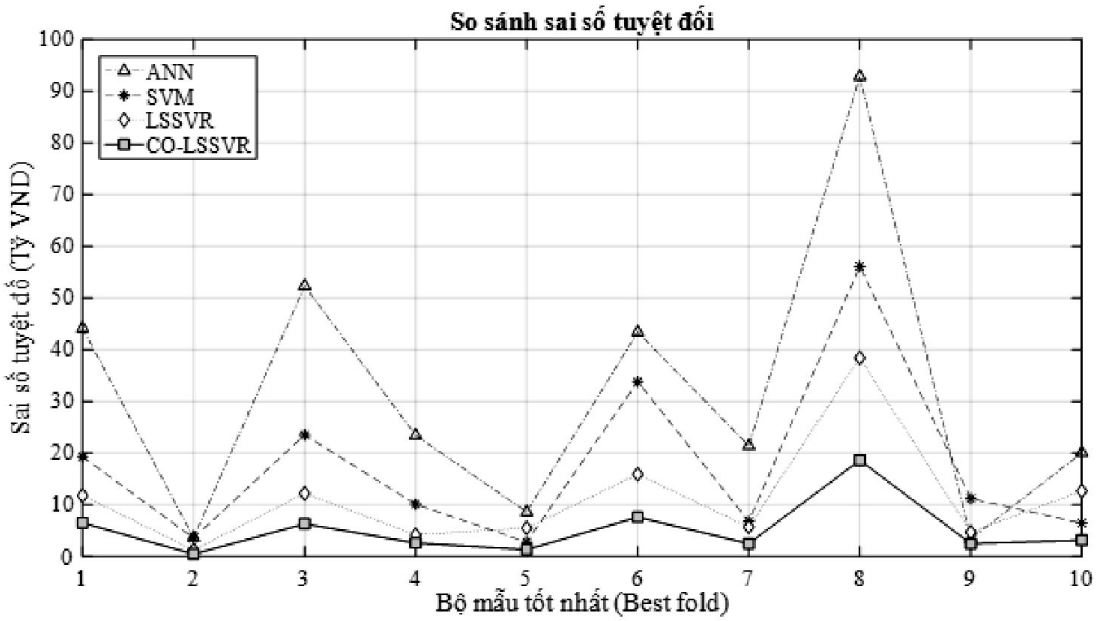
báo của các mô hình trên tập dữ liệu kiểm tra. Kết quả cho thấy trong nhóm các mô hình đơn lẻ, LSSVR cho độ chính xác cao hơn so với ANN và SVM, thể hiện qua các giá trị RMSE, MAE và MAPE thấp hơn, đồng thời R^2 cao hơn. Điều này khẳng định ưu thế của LSSVR trong việc mô hình hóa các mối quan hệ phi tuyến giữa các biến đầu vào và chi phí xây dựng.

Hình 2 thể hiện sai lệch tuyệt đối giữa chi phí dự báo và chi phí thực tế của các dự án sử dụng vốn đầu tư công ứng với fold kiểm tra tốt nhất của các mô hình được so sánh. Kết quả cho thấy mô hình CO-LSSVR đạt hiệu quả vượt trội khi duy trì sai lệch tuyệt đối thấp và ổn định trên toàn bộ tập dữ liệu kiểm tra. So với các mô hình ANN, SVM và LSSVR truyền thống, CO-LSSVR cho thấy khả năng kiểm soát sai số tốt hơn, đặc biệt đối với các dự án có quy mô vốn lớn và đặc điểm phi tuyến phức tạp. Tỷ lệ các trường hợp có sai lệch nằm trong ngưỡng chấp nhận được của mô hình đề xuất cao hơn rõ rệt, phản ánh hiệu quả của thuật toán Cheetah optimizer trong việc tối ưu các siêu tham số của LSSVR, qua đó nâng cao độ chính xác và tính ổn định của mô hình dự báo chi phí.

Bên cạnh độ chính xác, thời gian tính toán cũng được xem xét nhằm đánh giá khả năng ứng dụng thực tiễn. Các mô hình ANN và SVM có thời gian xử lý ngắn hơn do cấu trúc đơn giản và không yêu cầu quá trình tối ưu lặp phức tạp, trong khi mô hình CO-LSSVR cần nhiều thời gian tính toán hơn do tích hợp thuật toán tối ưu metaheuristic. Tuy nhiên, sự gia tăng chi phí tính toán này là hợp lý khi xét đến mức cải thiện đáng kể về các chỉ số RMSE, MAE, MAPE và R^2 đã đạt được. Trong bối cảnh lập và quản lý chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công, nơi độ chính xác có ý nghĩa quyết định đối với hiệu quả sử dụng vốn ngân sách, mô hình CO-LSSVR cho thấy tiềm năng ứng dụng cao và đáng tin cậy.

5. Kết luận

Nghiên cứu này đã đề xuất và kiểm chứng mô hình CO-LSSVR, một mô hình lai ghép giữa hồi quy véc-tơ hỗ trợ bình phương tối thiểu (LSSVR) và thuật toán tối ưu Cheetah, nhằm xác định chi phí đầu tư xây dựng của dự án sử dụng vốn đầu tư công trong giai đoạn chuẩn bị dự án. Trên cơ sở bộ dữ liệu thực tế thu thập từ các dự án sử dụng vốn đầu tư công, hiệu quả của mô hình đề xuất đã được đánh giá và so sánh với các mô hình học máy phổ biến như LSSVR truyền thống, mạng nơ-ron nhân tạo và máy véc-tơ hỗ trợ thông qua phương pháp kiểm định chéo 5-fold và bốn chỉ số đánh giá RMSE, MAE, MAPE và R^2. Kết quả thực nghiệm cho thấy mô hình CO-LSSVR đạt độ chính xác dự



Hình 2. Sai lệch tuyệt đối của các mô hình tại bộ mẫu kiểm tra tốt nhất

báo cao nhất và độ ổn định tốt nhất trong số các mô hình được so sánh; việc tích hợp thuật toán tối ưu Cheetah giúp xác định hiệu quả các siêu tham số của LSSVR, qua đó nâng cao khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các yếu tố đầu vào và chi phí đầu tư xây dựng. Mặc dù thời gian tính toán của mô hình đề xuất cao hơn so với các mô hình đơn lẻ, mức gia tăng này được xem là chấp nhận được trong bối cảnh lập chi phí đầu tư công, nơi độ chính xác và độ tin cậy được ưu tiên hàng đầu.

Mô hình CO-LSSVR có tiềm năng ứng dụng cao trong công tác lập, thẩm tra, thẩm định, phê duyệt, điều chỉnh và tổ chức quản lý chi phí đầu tư xây dựng của dự án đầu tư xây dựng sử dụng vốn đầu tư công, góp phần hỗ trợ ra quyết định và nâng cao hiệu quả sử dụng vốn ngân sách. Tuy nhiên, nghiên cứu vẫn tồn tại một số hạn chế liên quan đến phạm vi dữ liệu và tập biến đầu vào. Các hướng nghiên cứu tiếp theo có thể tập trung vào mở rộng dữ liệu cho các dự án với nhiều loại hình công trình và khu vực khác nhau, tích hợp thêm các yếu tố định tính, cũng như so sánh mô hình đề xuất với các thuật toán tối ưu metaheuristic khác nhằm tiếp tục nâng cao độ chính xác và khả năng ứng dụng trong thực tế.❑

Tài liệu tham khảo

[1]T. H. Tuấn, "Các nhân tố ảnh hưởng đến chi phí và thời gian hoàn thành dự án trong giai đoạn thi công trường hợp nghiên cứu trên địa bàn thành phố Cần Thơ," Tạp chí khoa học Trường Đại học Cần Thơ, no. 30, pp. 26-33, 2014.

[2]T. H. T. Nguyen, Q. T. Pham, K. V. T. Hoang, L. P. Vu, and T. H. Ha, "Identifying factors affecting cost management of investment projects in construction of technical infrastructure under the public-private partnership (PPP) approach," Journal of Construction, vol. 11, pp. 92-99, 2024.

[3]T. Q. Bằng and P. Đ. Thắng, "Nghiên cứu về tác động của chi phí lên vốn đầu tư trong xây dựng công trình theo tiêu chuẩn công trình xanh," Tạp chí Vật liệu và Xây dựng, vol. 14, no. 4, pp. 130-136, 2024.

[4]K. Alrasheed, M. A. Ballat, E. Soliman, and H. Albader, "Artificial Neural Network-based cost estimation for public construction projects in Kuwait," Journal of Engineering Research, 2025/09/08/ 2025.

[5]M.-Y. Cheng, N.-D. Hoang, and Y.-W. Wu, "Hybrid intelligence approach based on LS-SVM and Differential Evolution for construction cost index estimation: A Taiwan case study," Automation in Construction, vol. 35, pp. 306-313, 2013/11/01/ 2013.

[6]M. A. Akbari, M. Zare, R. Azizipanah-abarghooee, S. Mirjalili, and M. Deriche, "The cheetah optimizer: a nature-inspired metaheuristic algorithm for large-scale optimization problems," Scientific Reports, vol. 12, no. 1, p. 10953, 2022/06/29 2022.

[7](2021). Nghị định số 10/2021/NĐ-CP ngày 09 tháng 02 năm 2021 của Chính phủ về Quản lý chi phí đầu tư xây dựng (Điểm c khoản 1 Điều 6).

[8]H. H. Elmousalami, "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review," Journal of Construction Engineering and Management, vol. 146, no. 1, p. 03119008, 2020.

[9]J.-M. S. Gwang-Hee Kim, Sangyong Kim, Yoonseok Shin, "Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network, and Support Vector Machine," Journal of Building Construction and Planning Research, 2013.

[10] R. Wang, H. Salleh, J. Lyu, Z. Abdul-Samad, N. F. M. Radzuan, and K. C. Wen, "Application and prospect of machine learning techniques in cost estimation of building projects," Engineering, Construction and Architectural Management, vol. 32, no. 12, pp. 8445-8471, 2024/10/30/ 2024.

[11] J.-H. Chen, Y.-M. Su, D. W. Hayati, I. Wijatmiko, and R. Purnamasari, "Improving preliminary cost estimation in Indonesia using support vector regression," Proceedings of the Institution of Civil Engineers - Management, Procurement and Law, vol. 172, no. 1, pp. 25-33, 2019/01/24/ 2019.

[12] M.-Y. Cheng and N.-D. Hoang, "Interval estimation of construction cost at completion using least squares support vector machine," Journal of Civil Engineering and Management, vol. 20, no. 2, pp. 223-236, 2014/03/04 2014.

[13] P. V. Khoa, "Ước lượng chi phí đầu tư xây dựng dự án chung cư bằng Neural Networks," Khoa Kỹ thuật Xây dựng, Trường Đại học Bách Khoa - ĐHQG Tp. Hồ Chí Minh, 2006.

[14] N. M. Quang, "Ước lượng chi phí xây dựng công trình trường trung học phổ thông tại Tp. Hồ Chí Minh sử dụng mạng nơron nhân tạo," Khoa Kỹ thuật Xây dựng, Trường Đại học Bách Khoa - ĐHQG Tp. Hồ Chí Minh, 2017.

[15] L. H. Q. Phong, T. Đ. Học, N. N. Thụy, and P. Q. Trâm, "Ước lượng chi phí xây dựng nhà xưởng trong giai đoạn đấu thầu ứng dụng mạng neural nhân tạo (ANN)," Tạp chí Xây dựng, vol. 7, pp. 76-80, 2022.

[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," International Joint Conference on Artificial Intelligence, no. 14th Int, pp. 1137–1143, 1995.

[17] S. Tayefeh Hashemi, O. M. Ebadati, and H. Kaur, "Cost estimation and prediction in construction projects: a systematic review on machine learning techniques," SN Applied Sciences, vol. 2, no. 10, p. 1703, 2020/09/15 2020.