Transport and Communications Science Journal, Vol 77, Issue 5 (06/2026), 672-686

Transport and Communications Science Journal

A HYBRID IF–LOF MACHINE LEARNING FRAMEWORK FOR DETECTING GNSS-RTK OUTLIERS IN CABLE-STAYED BRIDGES DISPLACEMENT MONITORING

Le Khanh Giang*, Ho Thi Lan Huong, Nguyen Thuy Linh

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 07/03/2026
Revised: 10/06/2026
Accepted: 12/06/2026
Published online: 15/06/2026
https://doi.org/10.47869/tcsj.77.5.5

* *Corresponding author*
Email: gianglk@utc.edu.vn; Tel: +84983663031

Abstract. Global Navigation Satellite System Real-Time Kinematic (GNSS-RTK) technology is widely used in structural health monitoring (SHM) of long-span bridges because it enables continuous three-dimensional displacement measurements with millimetre-level accuracy. However, GNSS-RTK time-series data often contain abnormal observations caused by multipath effects, signal interruptions, and environmental disturbances. This study addresses the problem of detecting and removing outliers in GNSS-RTK displacement data used for cable-stayed bridge monitoring. A hybrid machine-learning framework combining Isolation Forest (IF) and Local Outlier Factor (LOF) is proposed. In the first stage, IF performs global screening to identify potentially anomalous observations. In the second stage, LOF evaluates the local density of these candidate samples to confirm true outliers through a dual-confirmation mechanism. The method was validated using a GNSS-RTK dataset collected from the Can Tho cable-stayed bridge in Vietnam, consisting of 12,615 displacement samples. The hybrid framework detected 177 outliers (1.40%), compared with 253 detected by IF and 379 detected by LOF. Furthermore, by restricting the LOF analysis to a subset of suspicious observations identified by IF, the proposed framework substantially reduced the computational burden of the local density evaluation stage while preserving most valid structural responses. The results demonstrate that the proposed IF–LOF framework provides an efficient preprocessing approach for improving the reliability of GNSS-based bridge monitoring systems.

Keywords: GNSS-RTK, outlier detection, Isolation Forest, Local Outlier Factor, structural health monitoring, cable-stayed bridges.

Tạp chí Khoa học Giao thông vận tải

KHUNG HỌC MÁY LAI GHÉP IF – LOF ĐỂ PHÁT HIỆN SỐ LIỆU GNSS-RTK NGOẠI LAI TRONG QUAN TRẮC CHUYỂN DỊCH CẦU DÂY VĂNG

Lê Khánh Giang*, Hồ Thị Lan Hương, Nguyễn Thuỳ Linh

Trường Đại học Giao thông vận tải, Số 3 Cầu Giấy, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

CHUYÊN MỤC: Công trình khoa học
Ngày nhận bài: 07/03/2026
Ngày nhận bài sửa: 10/06/2026
Ngày chấp nhận đăng: 12/06/2026
Ngày xuất bản Online: 15/06/2026
https://doi.org/10.47869/tcsj.77.5.5

* *Tác giả liên hệ*
Email: gianglk@utc.edu.vn; Tel: +84983663031

Tóm tắt. Công nghệ định vị vệ tinh GNSS-RTK được sử dụng rộng rãi trong các hệ thống quan trắc sức khỏe công trình (SHM) của cầu dây văng nhịp lớn nhờ khả năng đo chuyển dịch ba chiều liên tục với độ chính xác cao. Tuy nhiên, chuỗi dữ liệu GNSS-RTK thường chứa các giá trị bất thường do hiệu ứng đa đường, gián đoạn tín hiệu và các nhiễu môi trường. Nghiên cứu này tập trung vào bài toán phát hiện và loại bỏ các ngoại lai trong dữ liệu chuyển dịch GNSS-RTK của cầu dây văng. Một khung học máy lai ghép kết hợp hai thuật toán Isolation Forest (IF) và Local Outlier Factor (LOF) được đề xuất. Trong giai đoạn đầu, IF thực hiện sàng lọc toàn cục để xác định các điểm nghi ngờ. Sau đó, LOF đánh giá mật độ cục bộ của các điểm này nhằm xác nhận các ngoại lai thực sự thông qua cơ chế xác nhận kép. Phương pháp được kiểm chứng trên bộ dữ liệu GNSS-RTK của cầu Cần Thơ gồm 12.615 mẫu chuyển dịch. Kết quả cho thấy mô hình lai ghép phát hiện 177 ngoại lai (1,40%), thấp hơn so với IF (253 điểm) và LOF (379 điểm). Bên cạnh đó, việc chỉ áp dụng LOF trên tập các quan trắc nghi ngờ được xác định bởi IF giúp giảm đáng kể khối lượng tính toán của giai đoạn phân tích mật độ cục bộ trong khi vẫn bảo toàn phần lớn dữ liệu quan trắc hợp lệ. Kết quả nghiên cứu cho thấy phương pháp IF–LOF là một giải pháp tiền xử lý hiệu quả cho dữ liệu quan trắc cầu sử dụng GNSS.

Từ khóa: GNSS-RTK, phát hiện ngoại lai, Isolation Forest, Local Outlier Factor, quan trắc sức khỏe công trình, cầu dây văng.

1. ĐẶT VẤN ĐỀ

Cầu dây văng là một trong những dạng kết cấu phức tạp và nhạy cảm nhất trong hệ thống hạ tầng giao thông hiện đại. Với đặc trưng nhịp lớn, độ cứng tổng thể thấp và khả năng chịu ảnh hưởng mạnh của các tác động môi trường, loại cầu này thường chịu các tải trọng động phức tạp như gió, lưu lượng giao thông và biến thiên nhiệt độ. Những tác động này có thể gây ra các phản ứng chuyển dịch đáng kể của kết cấu, đặc biệt đối với các nhịp chính dài [1]. Do đó, các hệ thống quan trắc sức khỏe công trình (Structural Health Monitoring – SHM) ngày càng được triển khai rộng rãi nhằm theo dõi liên tục trạng thái làm việc của cầu và phát hiện sớm các dấu hiệu bất thường của kết cấu [2].

Trong các công nghệ quan trắc hiện nay, hệ thống định vị vệ tinh đo động thời gian thực GNSS-RTK (Global Navigation Satellite System – Real Time Kinematic) được xem là một giải pháp hiệu quả nhờ khả năng đo trực tiếp chuyển dịch ba chiều của kết cấu với độ chính xác ở mức milimet và khả năng cung cấp dữ liệu liên tục theo thời gian thực [3]. Khác với các cảm biến truyền thống như gia tốc kế, vốn chủ yếu ghi nhận dao động tần số cao và khó chuyển đổi trực tiếp sang chuyển vị tuyệt đối, GNSS-RTK cho phép thu thập chuỗi dữ liệu chuyển dịch dài hạn của công trình [4]. Nhờ đó, hệ thống GNSS-RTK có thể hỗ trợ hiệu quả cho việc phân tích cả biến dạng tĩnh và dao động động của kết cấu cầu nhịp lớn.

Tuy nhiên, dữ liệu GNSS-RTK thu thập trong môi trường cầu thường chịu ảnh hưởng của nhiều nguồn nhiễu khác nhau. Các yếu tố như hiệu ứng đa đường (multipath) do phản xạ tín hiệu từ mặt nước, tháp cầu và hệ thống dây văng, nhiễu khí quyển, hoặc gián đoạn tín hiệu vệ tinh có thể tạo ra các giá trị đo bất thường trong chuỗi dữ liệu. Những giá trị này thường được gọi là ngoại lai (outliers) và có thể xuất hiện dưới dạng các nhiễu xung cục bộ hoặc các sai lệch lớn so với xu hướng chuyển dịch chung của kết cấu. Nếu không được phát hiện và loại bỏ kịp thời, các ngoại lai này có thể gây sai lệch đáng kể trong quá trình phân tích phản ứng kết cấu, thậm chí dẫn đến các cảnh báo sai trong hệ thống quan trắc [5].

Trong các nghiên cứu trước đây, nhiều phương pháp truyền thống như các kỹ thuật thống kê, phương pháp cửa sổ trượt hoặc các thuật toán lọc đã được sử dụng để phát hiện ngoại lai trong chuỗi dữ liệu GNSS [6]. Tuy nhiên, các phương pháp này thường dựa trên các giả thiết đơn giản về phân bố dữ liệu và gặp khó khăn khi xử lý các tập dữ liệu phức tạp, có tính phi tuyến hoặc chứa nhiều loại nhiễu khác nhau. Trong những năm gần đây, các phương pháp học máy đã được nghiên cứu rộng rãi và cho thấy tiềm năng vượt trội trong phát hiện các giá trị đo bất thường. Chẳng hạn, nghiên cứu [7] đã đề xuất một khung học không giám sát kết hợp t-SNE và phân cụm K-means nhằm nhận dạng các mẫu chuyển vị tiềm ẩn và hỗ trợ phát hiện các hành vi bất thường từ dữ liệu GNSS-RTK phục vụ giám sát sức khỏe cầu dây văng. Ngoài ra, các thuật toán học máy khác như K-Nearest Neighbors (KNN), One-Class Support Vector Machine (OC-SVM), Gaussian Mixture Model (GMM), Density Peaks-based Fast Clustering (DPFC) và mạng nơ-ron tự mã hóa cũng đã được áp dụng trong nhiều nghiên cứu về phát hiện bất thường và phát hiện hư hỏng kết cấu [8].

Trong số các thuật toán học máy không giám sát, Isolation Forest (IF) và LOF được sử dụng phổ biến trong phát hiện ngoại lai [9, 10]. IF có ưu điểm trong việc phát hiện các ngoại lai toàn cục và có khả năng xử lý hiệu quả các tập dữ liệu lớn nhờ cơ chế phân tách ngẫu nhiên của cây quyết định [9]. Trong khi đó, LOF dựa trên phân tích mật độ lân cận và đặc biệt hiệu quả trong việc phát hiện các bất thường cục bộ trong dữ liệu [10]. Các nghiên cứu trước đây cũng chỉ ra rằng LOF thường có thời gian thực thi nhanh hơn IF trong các kịch bản quan trắc thời gian thực [11]. Ngoài ra, nghiên cứu [12] chỉ ra rằng IF được sử dụng rộng rãi trong phát

hiện ngoại lai của hệ thống quan trắc công trình nhờ khả năng xử lý hiệu quả các tập dữ liệu lớn và chi phí huấn luyện thấp hơn so với các mô hình phức tạp như mạng nơ-ron [12]. Tuy nhiên, việc áp dụng trực tiếp các thuật toán này cho dữ liệu quan trắc GNSS-RTK vẫn còn một số hạn chế. IF có thể đánh dấu nhầm các chuyển dịch thực của kết cấu là bất thường, trong khi LOF yêu cầu tính toán mật độ lân cận giữa các điểm dữ liệu, dẫn đến chi phí tính toán cao khi áp dụng cho các tập dữ liệu lớn.

Tại Việt Nam, một số nghiên cứu gần đây đã áp dụng các phương pháp học máy để phát hiện ngoại lai trong chuỗi tọa độ GNSS. Chẳng hạn, nghiên cứu [13] đã so sánh ba thuật toán IF, OC-SVM và LOF trong phát hiện ngoại lai của chuỗi tọa độ GNSS mô phỏng và cho thấy IF đạt hiệu suất tốt nhất, trong khi LOF có khả năng phát hiện ngoại lai kém hơn đáng kể. Tuy nhiên, các nghiên cứu hiện nay chủ yếu tập trung vào chuỗi dữ liệu GNSS địa động hoặc các hệ thống cảm biến khác, trong khi việc áp dụng các phương pháp học máy cho phát hiện ngoại lai trong dữ liệu chuyển dịch GNSS-RTK phục vụ quan trắc cầu dây văng vẫn còn tương đối hạn chế. Bên cạnh đó, phần lớn các nghiên cứu hiện nay thường sử dụng các thuật toán đơn lẻ, điều này có thể dẫn đến việc phát hiện nhiều ngoại lai giả hoặc làm tăng chi phí tính toán khi xử lý các tập dữ liệu lớn.

Để khắc phục những hạn chế nêu trên, nghiên cứu này đề xuất một khung học máy lai ghép kết hợp hai thuật toán IF và LOF nhằm phát hiện các giá trị ngoại lai trong chuỗi dữ liệu chuyển dịch GNSS-RTK. Trong khung phương pháp đề xuất, IF được sử dụng để sàng lọc nhanh các điểm dữ liệu nghi ngờ trên toàn bộ tập dữ liệu, sau đó LOF được áp dụng để phân tích mật độ cục bộ và xác nhận các ngoại lai thực sự. Cách tiếp cận này giúp giảm đáng kể khối lượng tính toán của LOF đồng thời hạn chế việc loại bỏ nhầm các giá trị chuyển dịch thực của kết cấu.

Hiệu quả của phương pháp đề xuất được kiểm chứng thông qua dữ liệu quan trắc chuyển dịch GNSS-RTK của cầu Cần Thơ. Kết quả nghiên cứu được đánh giá thông qua các chỉ tiêu về khả năng phát hiện ngoại lai, chi phí tính toán và mức độ cải thiện chất lượng dữ liệu chuyển dịch trong hệ thống quan trắc cầu dây văng.

2. PHƯƠNG PHÁP

2.1. Khung phương pháp

Nghiên cứu đề xuất một khung phát hiện ngoại lai hai giai đoạn cho chuỗi dữ liệu chuyển dịch GNSS-RTK, kết hợp giữa IF và LOF. Mục tiêu của khung phương pháp là đạt được sự cân bằng giữa khả năng phát hiện bất thường, chi phí tính toán và độ tin cậy của tín hiệu quan trắc. Quy trình xử lý gồm hai bước chính:

(1) Sàng lọc toàn cục bằng IF nhằm xác định các điểm dữ liệu nghi ngờ trong toàn bộ chuỗi.

(2) Xác nhận cục bộ bằng LOF để đánh giá mật độ lân cận và xác định các ngoại lai thực sự.

Cách tiếp cận hai giai đoạn này giúp giảm đáng kể khối lượng tính toán của LOF đồng thời hạn chế việc loại bỏ nhầm các chuyển dịch thực của kết cấu. Hình 1 minh họa chi tiết quy trình của phương pháp đề xuất, bao gồm các bước sau:

Hình 1. Khung phương pháp phát hiện ngoại lai GNSS-RTK dựa trên mô hình lai ghép IF–LOF.

- Tiền xử lý dữ liệu: Khôi phục dữ liệu thiếu, lọc trung vị và chuẩn hóa các thành phần chuyển dịch GNSS-RTK.
- Sàng lọc toàn cục (Giai đoạn 1 – IF): Áp dụng thuật toán Isolation Forest trên toàn bộ tập dữ liệu để tính toán điểm số bất thường (anomaly score) và xác định các điểm dữ liệu nghi ngờ dựa trên phân bố của điểm số.
- Tính toán mật độ không gian: Xây dựng cấu trúc mật độ lân cận bằng thuật toán LOF dựa trên toàn bộ tập dữ liệu nhằm giữ nguyên cấu trúc hình học của không gian chuyển dịch.
- Sàng lọc cục bộ (Giai đoạn 2 – LOF): Tính toán hệ số LOF cho tập con các điểm nghi ngờ được xác định từ giai đoạn IF để phát hiện các ngoại lai cục bộ.
- Cơ chế xác nhận kép: Một điểm dữ liệu chỉ được xác định là ngoại lai khi đồng thời được IF đánh dấu là nghi ngờ và có giá trị LOF vượt ngưỡng xác định.
- Đánh giá hiệu quả: Hiệu quả của phương pháp được đánh giá thông qua các chỉ tiêu thống kê nhằm đo lường mức độ cải thiện chất lượng chuỗi dữ liệu.

2.2. Dữ liệu và tiền xử lý

2.2.1. Dữ liệu đầu vào

Dữ liệu đầu vào là chuỗi tọa độ GNSS-RTK thu thập liên tục theo thời gian, được biểu diễn dưới dạng tập dữ liệu:

$$D = \{P_1, P_2, \dots, P_N\} \quad (1)$$

Trong đó mỗi điểm quan trắc $P_i = (x_i, y_i, z_i)$ biểu diễn chuyển dịch của điểm quan trắc theo ba phương không gian tại thời điểm t_i, và N là tổng số mẫu quan trắc trong chuỗi dữ liệu.

2.2.2. Tiền xử lý dữ liệu

(1) Làm sạch dữ liệu

Trước khi áp dụng các thuật toán học máy, dữ liệu được xử lý sơ bộ nhằm bảo đảm tính liên tục và độ tin cậy của chuỗi quan trắc GNSS-RTK. Trong tập dữ liệu tồn tại một số quan trắc không hợp lệ được mã hóa bằng giá trị −9999, phản ánh các thời điểm mất tín hiệu GNSS tạm thời hoặc không thu được nghiệm định vị đáng tin cậy. Các giá trị này được chuyển thành NaN và được khôi phục bằng quy trình nội suy gồm nội suy tuyến tính, Last Observation Carried Forward (LOCF) và Next Observation Carried Backward (NOCB). Sau quá trình nội suy, không còn giá trị thiếu nào trong tập dữ liệu.

(2) Lọc trung vị

Để giảm nhiễu xung cục bộ thường xuất hiện trong dữ liệu GNSS-RTK, một bộ lọc trung vị với kích thước cửa sổ bằng 3 được áp dụng cho từng thành phần tọa độ. Bộ lọc trung vị giúp loại bỏ các nhiễu đơn điểm trong chuỗi dữ liệu mà không làm suy giảm đáng kể biên độ dao động thực của kết cấu.

(3) Chuẩn hóa dữ liệu

Sau bước làm sạch và lọc trung vị, dữ liệu được chuẩn hóa bằng phương pháp Standardization nhằm đưa các biến về cùng thang đo. Mỗi thành phần tọa độ được chuẩn hóa theo công thức:

$$x_i^{norm} = \frac{x_i - \mu_x}{\sigma_x} \quad (2)$$

trong đó μ_x và σ_x lần lượt là giá trị trung bình và độ lệch chuẩn của chuỗi dữ liệu. Việc chuẩn hóa giúp các thuật toán dựa trên khoảng cách, đặc biệt là LOF, phản ánh chính xác cấu trúc hình học của dữ liệu trong không gian ba chiều.

2.3. Giai đoạn 1: Phát hiện ngoại lai toàn cục bằng Isolation Forest

IF là một thuật toán phát hiện ngoại lai không giám sát dựa trên nguyên lý cô lập dữ liệu thông qua các cây phân tách ngẫu nhiên. Các điểm ngoại lai thường dễ bị cô lập hơn so với các điểm bình thường do chúng hiếm và khác biệt về thuộc tính. Vì vậy, các điểm này có xu hướng xuất hiện ở các nút lá với độ sâu nhỏ trong cây phân tách [9].

Điểm số bất thường của một điểm dữ liệu x được tính toán như sau [9]:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(N)}} \quad (3)$$

Trong đó: $h(x)$ là độ sâu của điểm x trên một cây phân tách, $E(h(x))$ là giá trị kỳ vọng (trung bình) của độ sâu $h(x)$ trên toàn bộ "rừng" (Forest) cây, $c(N)$ là hệ số chuẩn hóa phụ thuộc vào kích thước mẫu:

$$c(N) = 2H(N-1) - \frac{2(N-1)}{N} \quad (4)$$

Trong đó: $H(N-1)$: là số điều hòa được tính theo:

$$H(N-1) = \sum_{k=1}^{N-1} \frac{1}{k} \quad (5)$$

Trong nghiên cứu này, mô hình IF được xây dựng với 500 cây phân tách và tham số contamination bằng 0.02 nhằm ổn định quá trình huấn luyện. Sau khi huấn luyện, điểm số bất thường được tính cho toàn bộ dữ liệu.

Để xác định các điểm nghi ngờ, các ngưỡng phân vị của phân bố anomaly score được sử dụng. Cụ thể, các quan trắc thuộc 2% thấp nhất của phân bố anomaly score được xem là ngoại lai tiềm năng (Outlier). Các quan trắc nằm trong khoảng 2%–7% được xem là các điểm nghi ngờ (Suspect), đại diện cho các quan trắc nằm gần ranh giới bất thường, trong khi các điểm còn lại được xem là dữ liệu bình thường (Normal). Ngưỡng 7% được lựa chọn sau nhiều thử nghiệm sơ bộ trên tập dữ liệu nghiên cứu nhằm cân bằng giữa khả năng loại bỏ nhiễu và bảo toàn dữ liệu. Hai nhóm Suspect và Outlier được gộp lại thành tập con D_{sub} để tiếp tục phân tích ở giai đoạn tiếp theo.

2.4. Giai đoạn 2: Phát hiện ngoại lai cục bộ bằng Local Outlier Factor

Trong giai đoạn thứ hai, thuật toán LOF được sử dụng để đánh giá mật độ cục bộ của các điểm dữ liệu. Khác với IF, LOF xác định ngoại lai dựa trên sự suy giảm mật độ của một điểm so với các điểm lân cận của nó [10]. Với mỗi điểm P, tập k lân cận gần nhất được ký hiệu là: $N_k(P)$.

Khoảng cách tiếp cận (Reachability Distance) từ điểm A đến điểm B được định nghĩa:

$$reach\_dist_k(A, B) = \max(k\text{-distance}(B), d(A, B)) \quad (6)$$

Trong đó $d(A, B)$ là khoảng cách Euclid, và k-distance(B) là khoảng cách từ B đến điểm lân cận thứ k của nó.

Mật độ tiếp cận cục bộ (Local Reachability Density - LRD) của điểm A là nghịch đảo của

khoảng cách tiếp cận trung bình tới k lân cận của nó:

$$LRD_k(A) = \left(\frac{1}{|N_k(A)|}\sum_{B\in N_k(A)} reach\_dist_k(A,B)\right)^{-1} \quad (7)$$

Cuối cùng, hệ số LOF của điểm A là trung bình tỷ số giữa mật độ của các lân cận và mật độ của chính nó:

$$LOF_k(A) = \frac{1}{|N_k(A)|}\sum_{B\in N_k(A)} \frac{LRD_k(B)}{LRD_k(A)} \quad (8)$$

Trong thực nghiệm, thuật toán LOF được thiết lập với tham số n_neighbors = 20. Khi áp dụng cho tập điểm nghi ngờ, mô hình sử dụng tham số contamination để xác định các quan trắc có mật độ suy giảm đáng kể so với lân cận và được xem là ngoại lai.

2.5. Mô hình lai ghép IF–LOF

Phương pháp đề xuất kết hợp hai thuật toán IF và LOF theo cơ chế phát hiện hai giai đoạn. Trong giai đoạn đầu, IF được sử dụng để sàng lọc nhanh toàn bộ tập dữ liệu và xác định một tập con các quan trắc nghi ngờ, ký hiệu là D_{sub}, bao gồm các điểm được gán nhãn Suspect và Outlier.

Trong giai đoạn thứ hai, thuật toán LOF được áp dụng trên tập con D_{sub} gồm các quan trắc được IF xác định là Suspect hoặc Outlier nhằm đánh giá mật độ cục bộ và xác nhận các ngoại lai thực sự.

Một điểm dữ liệu P_i được xác định là ngoại lai cuối cùng khi thỏa mãn hai điều kiện:

(1) Điểm đó được IF đánh dấu là Suspect hoặc Outlier trong giai đoạn sàng lọc toàn cục.

(2) Giá trị LOF của điểm đó cho thấy mật độ cục bộ suy giảm đáng kể so với các lân cận của nó.

Cơ chế xác nhận kép này giúp kết hợp ưu điểm của cả hai thuật toán: IF có khả năng phát hiện nhanh các ngoại lai toàn cục, trong khi LOF cung cấp đánh giá chi tiết dựa trên mật độ cục bộ. Nhờ đó, phương pháp đề xuất vừa giảm số lượng cảnh báo giả của IF vừa hạn chế chi phí tính toán khi áp dụng LOF trên các tập dữ liệu lớn.

2.6. Đánh giá hiệu quả mô hình

Do dữ liệu GNSS-RTK thực tế không có nhãn ngoại lai, hiệu quả của phương pháp được đánh giá gián tiếp thông qua sự cải thiện độ ổn định của chuỗi dữ liệu.

2.6.1. Độ lệch chuẩn

Đại lượng đánh giá sự phân tán xung quanh trục dịch chuyển trung bình của dao động. Với phương X (tương tự cho Y và Z) [14]:

$$\sigma_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_X)^2} \quad (9)$$

Khi các giá trị ngoại lai (mang sai số lớn) bị loại bỏ, σ sẽ giảm. Sự ổn định của chuỗi được định lượng qua Chỉ số cải thiện chất lượng (η).

2.6.2. Chỉ số cải thiện chất lượng dữ liệu

$$\eta = \frac{\sigma_{trước} - \sigma_{sau}}{\sigma_{trước}} \times 100\% \quad (10)$$

Giá trị η càng lớn cho thấy mức độ giảm nhiễu của chuỗi dữ liệu càng cao và độ tin cậy của dữ liệu quan trắc được cải thiện đáng kể.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

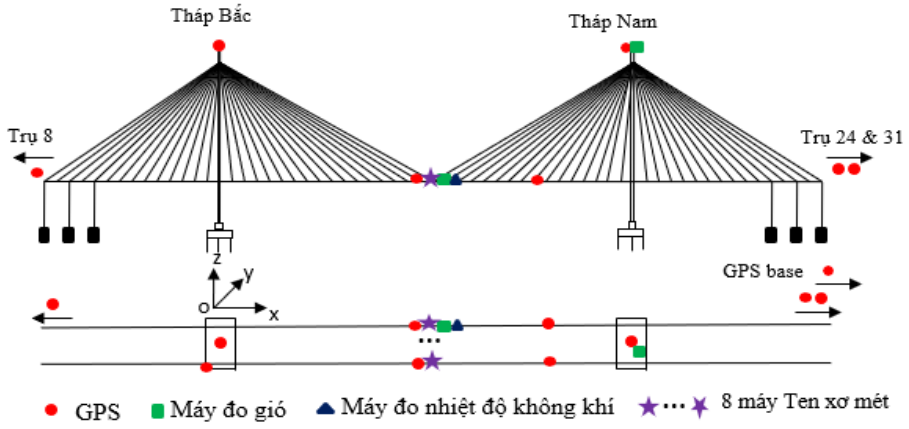
3.1. Dữ liệu nghiên cứu

Dữ liệu thực nghiệm được thu thập từ hệ thống quan trắc sức khỏe công trình của cầu Cần Thơ, một cầu dây văng hai mặt phẳng dây với nhịp chính dài 550 m và chiều cao trụ tháp 164,8 m. Cầu được thiết kế theo tiêu chuẩn AASHTO và có đặc trưng kết cấu mềm, do đó rất nhạy cảm với các tác động động lực học như gió, tải trọng giao thông và biến thiên nhiệt độ [15].

Hệ thống quan trắc GNSS-RTK được lắp đặt tại nhiều vị trí trên kết cấu cầu nhằm ghi nhận chuyển dịch theo thời gian thực. Sơ đồ bố trí các thiết bị quan trắc, bao gồm anten GNSS, cảm biến gió và cảm biến nhiệt độ, được minh họa trong Hình 2. Theo thông số kỹ thuật của hệ thống, độ chính xác danh định của phép đo đạt khoảng ±(10 mm + 1 ppm) theo phương ngang và ±(20 mm + 1 ppm) theo phương đứng [16].

Trong nghiên cứu này, một chuỗi dữ liệu chuyển dịch GNSS-RTK đã được trích xuất với khoảng thời gian lấy mẫu trung bình khoảng 10 phút (tức khoảng một quan trắc mỗi 10 phút) trong khoảng thời gian từ 01/10/2017 đến 30/12/2017. Tập dữ liệu ban đầu bao gồm 12.615 quan trắc liên tục theo ba phương không gian: X – chuyển dịch dọc cầu, Y – chuyển dịch ngang cầu, Z – chuyển dịch thẳng đứng.

Khảo sát sơ bộ cho thấy chuỗi dữ liệu chứa một số nhiễu xung cục bộ và các giá trị đo bất thường có thể phát sinh từ hiệu ứng đa đường (multipath), gián đoạn tín hiệu vệ tinh hoặc sai số đo đạc. Do đó, dữ liệu cần được tiền xử lý trước khi áp dụng các thuật toán học máy.



Hình 2. Sơ đồ hệ thống quan trắc GNSS-RTK của cầu Cần Thơ.

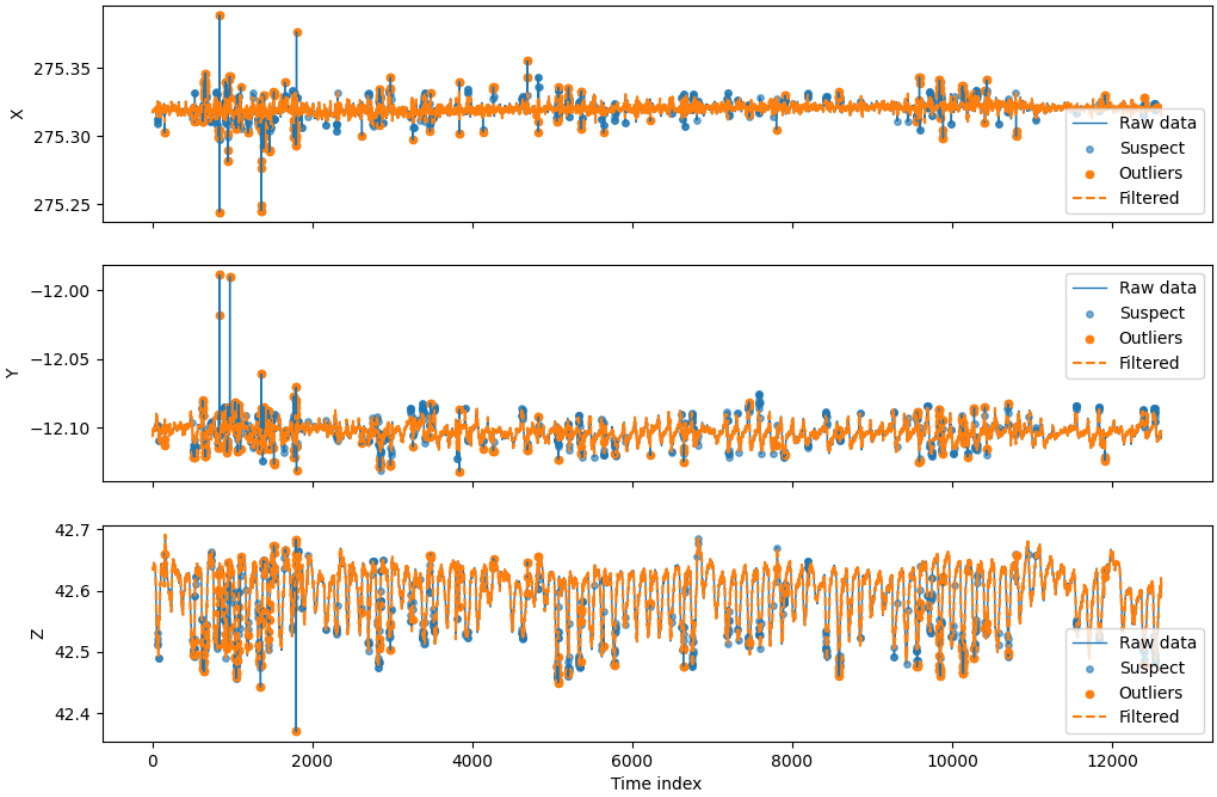
3.2. Kết quả tiền xử lý dữ liệu

Quy trình tiền xử lý dữ liệu được thực hiện theo phương pháp trình bày trong Mục 2.2.2. Trong bộ dữ liệu GNSS-RTK của cầu Cần Thơ, có 21 quan trắc không hợp lệ được ghi nhận. Sau quá trình tiền xử lý, toàn bộ các giá trị thiếu đã được khôi phục thành công và không còn giá trị thiếu nào trong tập dữ liệu.

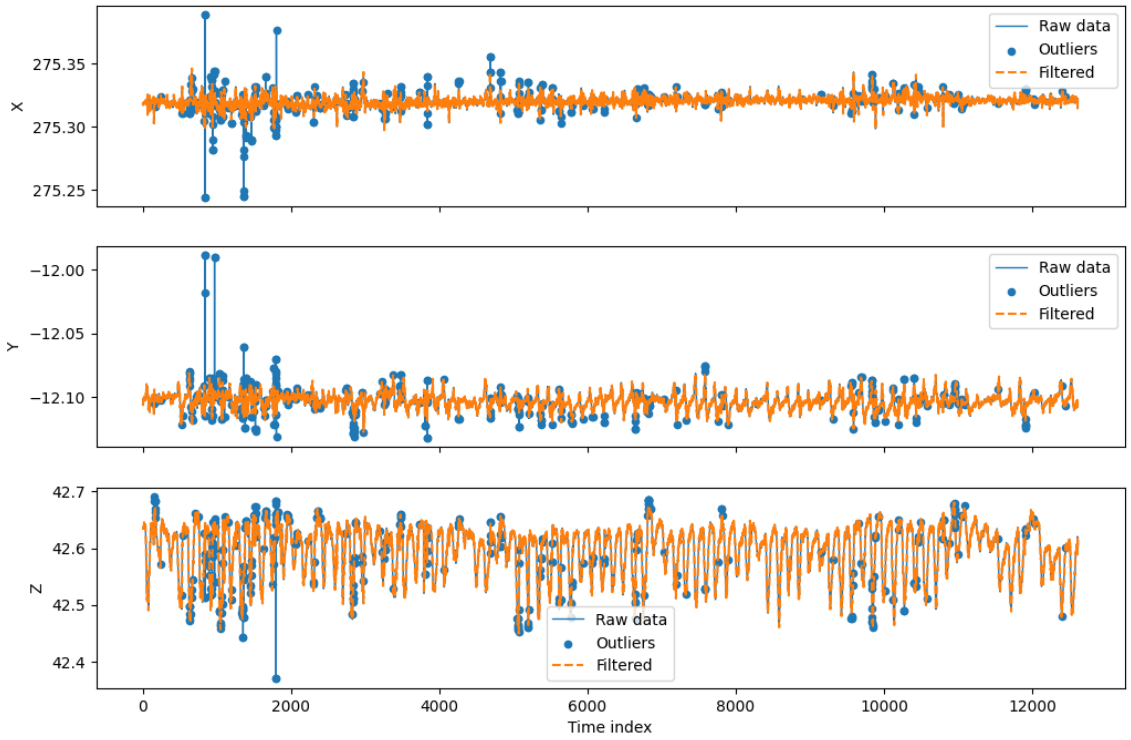
Tiếp theo, dữ liệu được lọc nhiễu và chuẩn hóa trước khi đưa vào các thuật toán phát hiện ngoại lai. Sau các bước tiền xử lý, toàn bộ 12.615 quan trắc được giữ lại để thực hiện các thí nghiệm được trình bày trong các mục tiếp theo.

3.3. Kết quả thực nghiệm

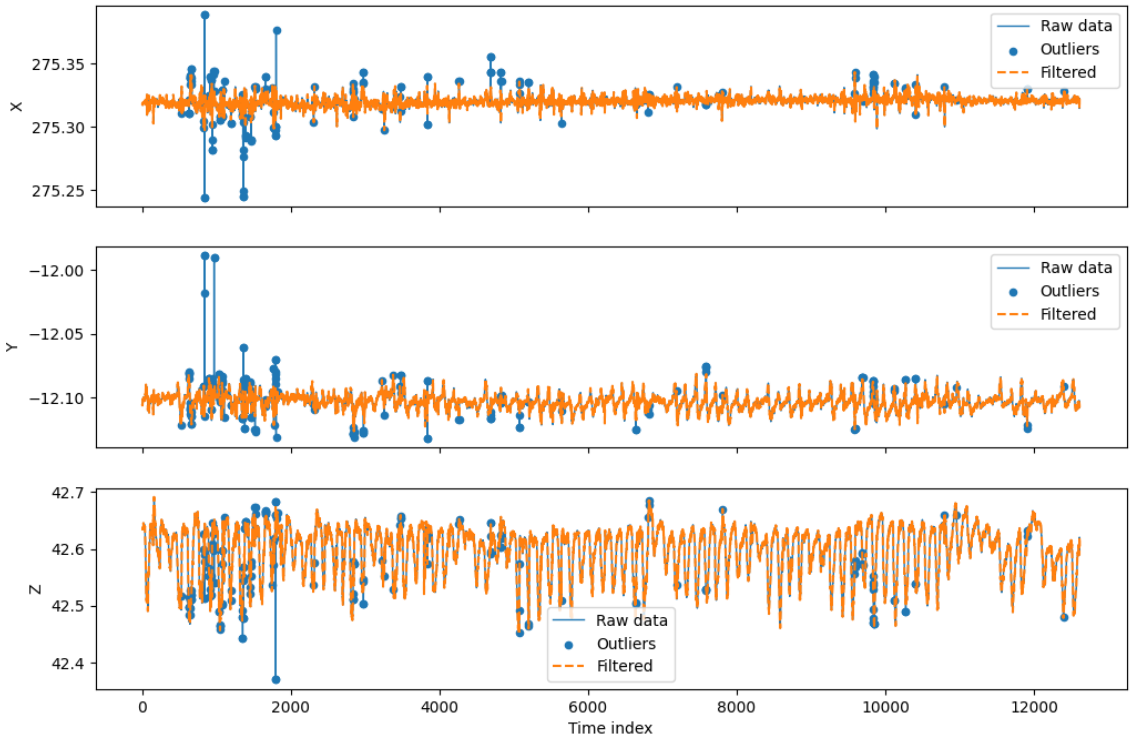
Kết quả phát hiện ngoại lai của ba phương pháp IF, LOF và mô hình lai ghép IF–LOF được trình bày trong Hình 3 dưới dạng chuỗi thời gian chuyển dịch theo ba phương không gian X, Y và Z. Hình 3 cho thấy phần lớn các điểm được đánh dấu ngoại lai tập trung tại các thời điểm xuất hiện các đột biến biên độ lớn hoặc các sai lệch cục bộ rõ rệt so với xu hướng dao động chung của kết cấu. Đường màu xanh biểu diễn dữ liệu gốc, các điểm được đánh dấu là ngoại lai (hoặc nghi ngờ trong trường hợp IF) được thể hiện bằng ký hiệu tròn, trong khi đường nét đứt màu cam biểu diễn chuỗi dữ liệu sau khi loại bỏ ngoại lai. Sau khi lọc, dạng dao động tổng thể của chuỗi chuyển dịch vẫn được duy trì, đặc biệt đối với thành phần chuyển dịch theo phương đứng (Z), cho thấy các phương pháp có xu hướng bảo toàn dạng dao động tổng thể của chuỗi chuyển dịch.



(a)



(b)



(c)

Hình 3. Kết quả phát hiện ngoại lai trên chuỗi chuyển dịch GNSS-RTK theo ba phương X, Y và Z bằng: (a) Isolation Forest (IF); (b) Local Outlier Factor (LOF); (c) mô hình lai ghép IF–LOF.

3.3.1. Phát hiện ngoại lai bằng Isolation Forest

Kết quả của thuật toán IF được thể hiện trong Hình 3a. Với cấu hình 500 cây phân tách, mô hình phát hiện:

- 11.732 điểm bình thường (93,0%)
- 630 điểm nghi ngờ (4,99%)
- 253 điểm ngoại lai (2,01%)

IF phát hiện hiệu quả các điểm nằm xa cụm dữ liệu chính. Tuy nhiên, do thuật toán dựa trên cơ chế phân tách toàn cục trong không gian đặc trưng, một số chuyển dịch có biên độ lớn của kết cấu cũng có thể bị đánh dấu nhầm là ngoại lai.

3.3.2. Phát hiện ngoại lai bằng Local Outlier Factor

Kết quả phát hiện ngoại lai của LOF được thể hiện trong Hình 3b. Với tham số n_neighbors = 20, thuật toán phát hiện 379 điểm ngoại lai, tương ứng khoảng 3% tổng số mẫu.

Do LOF đánh giá mật độ cục bộ của từng điểm so với các điểm lân cận, thuật toán có khả năng phát hiện các bất thường nhỏ nằm sâu trong cụm dữ liệu. Tuy nhiên, độ nhạy cao của phương pháp cũng có thể dẫn đến việc phát hiện số lượng ngoại lai lớn hơn cần thiết.

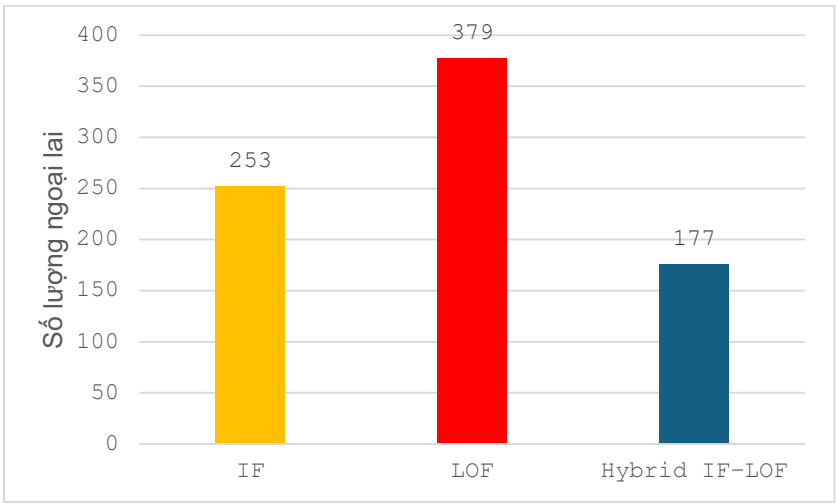
3.3.3. Kết quả của mô hình lai ghép IF–LOF

Kết quả của phương pháp lai ghép được thể hiện trong Hình 3c. Trong phương pháp đề xuất, Isolation Forest được sử dụng để sàng lọc các điểm nghi ngờ trên toàn bộ tập dữ liệu. Sau đó, thuật toán LOF chỉ được áp dụng cho tập con này để xác nhận các ngoại lai thực sự.

Kết quả cuối cùng cho thấy:

- 12.438 điểm bình thường (98,6%)
- 177 điểm ngoại lai (1,40%)

So với các thuật toán riêng lẻ, mô hình lai ghép giúp giảm đáng kể số lượng ngoại lai được phát hiện trong khi vẫn giữ lại phần lớn dữ liệu quan trắc hợp lệ. Kết quả so sánh được minh họa trong Hình 4.



Hình 4. So sánh số lượng ngoại lai được phát hiện của các phương pháp.

3.4. So sánh chi phí tính toán

So sánh chi phí tính toán của các thuật toán cho thấy thời gian thực thi của IF trên toàn bộ tập dữ liệu là 2,92 s, trong khi LOF cần 0,23 s khi được áp dụng trên toàn bộ dữ liệu. Đối với mô hình IF–LOF, thời gian thực hiện bước xác nhận cục bộ bằng LOF trên tập con các điểm nghi ngờ D_{sub} là khoảng 0,06 s.

Mặc dù IF thường được xem là thuật toán hiệu quả đối với các tập dữ liệu lớn, thời gian thực thi dài hơn trong thực nghiệm này chủ yếu do mô hình sử dụng số lượng cây phân tách lớn (n_estimators = 500). Quá trình xây dựng và đánh giá một tập hợp nhiều cây quyết định ngẫu nhiên có thể làm tăng chi phí tính toán khi kích thước dữ liệu ở mức trung bình.

Trong khi đó, thuật toán LOF trong thực nghiệm này chỉ yêu cầu tính toán lân cận gần nhất với số lượng lân cận nhỏ (n_neighbors = 20), do đó có thể thực thi nhanh hơn trên tập dữ liệu có kích thước vừa phải.

Đối với mô hình lai ghép IF–LOF, thuật toán LOF chỉ được áp dụng trên tập con các điểm nghi ngờ được xác định bởi IF thay vì toàn bộ tập dữ liệu. Điều này giúp giảm đáng kể số lượng phép tính khoảng cách cần thiết để ước lượng mật độ lân cận. Về mặt lý thuyết, chiến lược này giúp giảm đáng kể khối lượng tính toán của giai đoạn phân tích mật độ cục bộ, do số lượng điểm cần đánh giá giảm từ N xuống còn M (M ≪ N). Điều này đặc biệt có ý nghĩa khi xử lý các tập dữ liệu quan trắc GNSS có kích thước lớn.

3.5. Đánh giá chất lượng dữ liệu sau khi lọc

Hiệu quả làm sạch dữ liệu được đánh giá thông qua sự thay đổi của độ lệch chuẩn của chuỗi chuyển dịch trước và sau khi lọc. Kết quả được minh họa trong Hình 5 và tổng hợp trong Bảng 1.

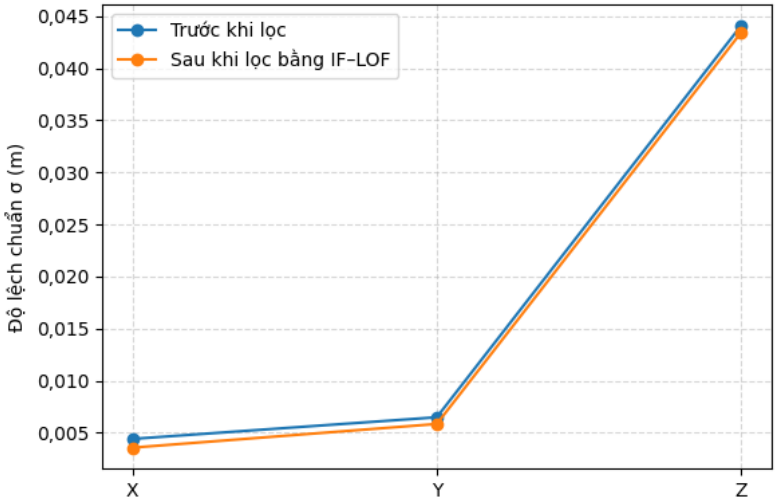
Phân tích kết quả trong Bảng 1 cho thấy thuật toán IF độc lập đạt giá trị cải thiện độ lệch chuẩn lớn nhất theo phương X (ηx = 33,50%). Tuy nhiên, kết quả này cần được diễn giải thận trọng trong bối cảnh quan trắc chuyển dịch kết cấu cầu.

Trong các hệ thống quan trắc sức khỏe công trình, chuỗi dữ liệu chuyển dịch thường chứa cả nhiễu đo đạc và các phản ứng dao động thực của kết cấu. Việc giảm độ lệch chuẩn quá mạnh có thể cho thấy một số giá trị dao động biên độ lớn của kết cấu đã bị loại bỏ cùng với nhiễu đo. Điều này có thể xảy ra khi thuật toán phát hiện ngoại lai toàn cục như Isolation Forest đánh dấu các quan trắc nằm xa trung tâm phân bố là ngoại lai.

Ngược lại, mô hình lai ghép IF–LOF giữ lại số lượng mẫu lớn nhất (12.438 quan trắc) trong khi vẫn giảm đáng kể độ phân tán của dữ liệu (ηx = 19,05%). Kết quả này cho thấy phương pháp đề xuất có khả năng loại bỏ hiệu quả các nhiễu đo rõ rệt trong dữ liệu GNSS-RTK, đồng thời hạn chế việc loại bỏ nhầm các dao động hợp lệ của kết cấu.

Do đó, trong bối cảnh quan trắc cầu dây văng, việc đạt được mức cải thiện độ lệch chuẩn vừa phải trong khi bảo toàn phần lớn dữ liệu quan trắc được xem là một kết quả hợp lý hơn so với việc giảm mạnh độ lệch chuẩn bằng cách loại bỏ nhiều quan trắc.

Do đó, mục tiêu của nghiên cứu không phải là tối đa hóa số lượng ngoại lai được phát hiện, mà là đạt được sự cân bằng giữa khả năng giảm nhiễu và việc bảo toàn các phản ứng thực của kết cấu. Theo quan điểm này, một phương pháp phát hiện nhiều ngoại lai hơn chưa chắc đã mang lại kết quả tốt hơn nếu đồng thời làm mất đi các thông tin dao động có ý nghĩa của công trình.



Hình 5. So sánh độ lệch chuẩn của chuỗi chuyển dịch theo ba phương không gian trước và sau khi áp dụng các phương pháp lọc ngoại lai.

Bảng 1. So sánh hiệu quả của các phương pháp phát hiện ngoại lai trên dữ liệu GNSS-RTK.

Phương pháp	Số mẫu giữ lại	ηx (%)	ηy (%)	ηz (%)
IF	11.732	33,50	17,89	9,83
LOF	12.236	18,83	10,69	2,51
Hybrid IF–LOF	12.438	19,05	9,92	1,51

3.6. Thảo luận

Các kết quả thực nghiệm cho thấy mô hình lai ghép IF–LOF mang lại một số lợi ích đáng kể cho việc xử lý dữ liệu GNSS trong các hệ thống quan trắc sức khỏe công trình. Trước hết, cơ chế xác nhận kép của phương pháp giúp giảm nguy cơ loại bỏ nhầm các chuyển dịch hợp lệ của kết cấu. Trong các hệ thống quan trắc cầu, các quan trắc có biên độ lớn không phải lúc nào cũng là nhiễu đo mà có thể phản ánh phản ứng động lực học thực của công trình. Việc chỉ loại bỏ các điểm dữ liệu được cả hai thuật toán xác nhận là ngoại lai giúp hạn chế hiện tượng phát hiện sai (false positives) khi sử dụng các thuật toán phát hiện ngoại lai riêng lẻ.

Bên cạnh đó, phương pháp đề xuất giúp giảm khối lượng tính toán của giai đoạn phân tích mật độ cục bộ. Như đã trình bày trong Mục 3.4, mô hình LOF được huấn luyện trên toàn bộ tập dữ liệu nhằm bảo toàn cấu trúc mật độ của không gian dữ liệu. Tuy nhiên, việc đánh giá điểm số LOF chỉ được thực hiện đối với tập con các quan trắc nghi ngờ được xác định bởi IF. Chiến lược này giúp giảm đáng kể số lượng phép tính khoảng cách cần thiết để ước lượng mật độ lân cận, từ đó cải thiện hiệu suất xử lý của hệ thống.

Ngoài ra, khung phương pháp đề xuất có tính linh hoạt cao và có thể dễ dàng tích hợp vào các quy trình xử lý dữ liệu quan trắc hiện có. Phương pháp này có thể được sử dụng như một bước tiền xử lý trước khi thực hiện các phân tích tiếp theo, chẳng hạn như nhận dạng dao động riêng, phân tích xu hướng chuyển dịch dài hạn hoặc phát hiện hư hỏng kết cấu.

Mặc dù đạt được những kết quả khả quan, nghiên cứu vẫn tồn tại một số hạn chế cần được thừa nhận. Do bộ dữ liệu GNSS-RTK được sử dụng trong nghiên cứu này được thu thập từ một hệ thống quan trắc cầu thực tế, số lượng và vị trí chính xác của các giá trị ngoại lai thực sự không được biết trước. Vì vậy, hiệu năng của khung phương pháp được đề xuất không thể được đánh giá bằng các chỉ số phân loại truyền thống như độ chính xác, độ bao phủ hoặc F1-score. Thay vào đó, việc so sánh được thực hiện dựa trên các chỉ số gián tiếp, bao gồm mức độ giảm độ lệch chuẩn, mức độ giảm RMSE, tỷ lệ mẫu được giữ lại và mức độ giảm khối lượng tính toán của giai đoạn LOF.

Ngoài ra, mặc dù khung phương pháp đề xuất đã cho thấy hiệu năng đầy hứa hẹn trên bộ dữ liệu của cầu dây văng Cần Thơ, việc kiểm định bổ sung bằng các bộ dữ liệu tổng hợp có chèn trước các giá trị ngoại lai đã biết, cũng như các bộ dữ liệu quan trắc từ nhiều loại cầu khác nhau, sẽ giúp đánh giá nghiêm ngặt hơn khả năng khái quát hóa của phương pháp. Những nội dung này sẽ được tiếp tục nghiên cứu và làm rõ trong các công trình tiếp theo.

4. KẾT LUẬN

Nghiên cứu này đã đề xuất một khung học máy lai ghép kết hợp hai thuật toán IF và LOF nhằm phát hiện và loại bỏ các giá trị ngoại lai trong chuỗi dữ liệu chuyển dịch GNSS-RTK phục vụ quan trắc cầu dây văng. Phương pháp được kiểm chứng thông qua bộ dữ liệu quan trắc thực tế của cầu Cần Thơ với 12.615 mẫu chuyển dịch theo ba phương không gian.

Kết quả thực nghiệm cho thấy mô hình lai ghép IF–LOF có khả năng phát hiện ngoại lai hiệu quả trong khi vẫn bảo toàn các phản ứng chuyển dịch thực của kết cấu. So với các thuật toán đơn lẻ, phương pháp đề xuất chỉ xác nhận 177 ngoại lai (1,40%), thấp hơn so với IF (253 điểm) và LOF (379 điểm). Điều này cho thấy cơ chế xác nhận kép của mô hình giúp giảm đáng kể hiện tượng phát hiện sai (false positives) trong dữ liệu quan trắc GNSS.

Bên cạnh đó, việc chỉ áp dụng LOF trên tập các quan trắc nghi ngờ được xác định bởi IF giúp giảm đáng kể khối lượng tính toán của giai đoạn phân tích mật độ cục bộ so với việc áp dụng trực tiếp LOF trên toàn bộ tập dữ liệu. Điều này cho thấy khung IF–LOF có tiềm năng triển khai cho các bài toán xử lý dữ liệu quan trắc GNSS quy mô lớn.

Nhìn chung, nghiên cứu đã chứng minh rằng khung học máy lai ghép IF–LOF là một phương pháp tiền xử lý hiệu quả nhằm nâng cao độ tin cậy của dữ liệu GNSS-RTK trong các hệ thống quan trắc sức khỏe công trình. Trong các nghiên cứu tiếp theo, phương pháp cần được kiểm chứng trên nhiều bộ dữ liệu quan trắc khác nhau để đánh giá tính tổng quát, đồng thời có thể được mở rộng để xử lý các chuỗi dữ liệu dài hạn hoặc tích hợp với các phương pháp học sâu nhằm nâng cao hơn nữa khả năng phát hiện bất thường trong các hệ thống SHM quy mô lớn.

TÀI LIỆU THAM KHẢO

[1]. H. Wang, J.X. Mao, Z.D. Xu, Investigation of dynamic properties of a long-span cable-stayed bridge during typhoon events based on structural health monitoring, J. Wind Eng. Ind. Aerodyn., 201 (2020) 104172. https://doi.org/10.1016/j.jweia.2020.104172

[2]. H.T. Al-Khateeb, H. W. Shenton, M. J. Chajes, C. Aloupis, Structural health monitoring of a cable-stayed bridge using regularly conducted diagnostic load tests. Front. Built Environ., 5 (2019) 41. https://doi.org/10.3389/fbuil.2019.00041

[3]. N. Shen, L. Chen, J. Liu, L. Wang, T. Tao, D. Wu, R. Chen, A review of global navigation satellite system (GNSS)-based dynamic monitoring technologies for structural health monitoring. Remote Sens., 11 (2019) 1001. https://doi.org/10.3390/rs11091001

[4]. T.H. Yi, H.N. Li, M. Gu, Recent research and applications of GPS-based monitoring technology for high-rise structures, Struct. Control Health Monit., 20 (2013) 649–670. https://doi.org/10.1002/stc.1501

[5]. Y. Deng, Z. Yingjie, J. Hanwen, Y. Ting-Hua, L. Aiqun, Abnormal data detection for structural health monitoring: State-of-the-art review, Dev. Built Environ., 17 (2024) 100337. https://doi.org/10.1016/j.dibe.2023.100337

[6]. L.V. Hien, L.M. Ngoc, T.D. Cong, Nghiên cứu phương pháp tiền xử lý dữ liệu quan trắc liên tục GNSS của cầu dây văng nhiều trụ tháp, Tạp chí Khoa học Giao thông Vận tải, 75 (2024) 2345–2355. https://doi.org/10.47869/tcsj.75.9.9

[7]. H.T.L. Huong, T.D. Cong, L.V. Vu, L.K. Giang, Latent pattern recognition in GNSS-based SHM using t-SNE and adaptive time-series modeling, J. Civ. Struct. Health Monit., 15 (2025) 3743–3766. https://doi.org/10.1007/s13349-025-01014-9

[8]. Z. Wang, Y.J. Cha, Unsupervised machine and deep learning methods for structural damage detection: a comparative study, Eng. Rep., 7 (2025) e12551. https://doi.org/10.1002/eng2.12551

[9]. F.T. Liu, K.M. Ting, Z.H. Zhou, Isolation Forest, in: Proceedings of the 8th IEEE International Conference on Data Mining, IEEE, Pisa, Italy, 2008, pp. 413–422. https://doi.org/10.1109/ICDM.2008.17

[10]. M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: identifying density-based local outliers, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Dallas, USA, 2000, pp. 93–104. https://doi.org/10.1145/342009.335388

[11]. I. Bayane, J. Leander, R. Karoumi, An unsupervised machine learning approach for real-time damage detection in bridges, Eng. Struct., 308 (2024) 117971. https://doi.org/10.1016/j.engstruct.2024.117971

[12]. Z. Sun, D. Siringoringo, C. Shi-Zhi, Cumulative displacement-based detection of damper malfunction in bridges using data-driven isolation forest algorithm, Eng. Fail. Anal., 143 (2023) 106849. https://doi.org/10.1016/j.engfailanal.2022.106849

[13]. H.D. Nguyen, T.D. Tran, Detecting outliers in GNSS position time series using machine learning techniques, J. Min. Earth Sci., 64 (2023) 22–30.

[14]. D.C. Montgomery, G.C. Runger, Applied Statistics and Probability for Engineers, 5th ed., John Wiley & Sons, New York, 2010.

[15]. AASHTO, AASHTO LRFD Bridge Design Specifications, 9th ed., American Association of State Highway and Transportation Officials, Washington DC, 2020.

[16]. Nippon Koei Co. Ltd., Chodai Co. Ltd., TEDI South, Tóm tắt Báo cáo kỹ thuật thiết kế hệ thống quan trắc kết cấu cầu Cần Thơ, 2010.