

# Nghiên cứu ứng dụng phương pháp LightGBM để dự báo lưu lượng giao thông

A study on applying the LightGBM method for traffic flow prediction

> TS PHẠM THỊ LÝ<sup>1</sup>, TS VÕ ĐỨC NHÂN<sup>2</sup>, THS NGUYỄN THỊ HỒNG HOA<sup>1,\*</sup>

<sup>1</sup>Trường Đại học Giao thông vận tải

<sup>2</sup>Cục Đổi mới sáng tạo, Bộ Khoa học và Công nghệ

\*Email: hoanthe@utc.edu.vn

## TÓM TẮT

Nội dung bài báo nghiên cứu ứng dụng thuật toán LightGBM cho mô hình dự báo lưu lượng giao thông sử dụng bộ dữ liệu gồm 2.976 bản ghi và 9 đặc trưng đầu vào: Time, Date, Day of Week, Carcount, Bikecount, buscount, truckcount, total, situation. Nhóm tác giả đã tiến hành thử nghiệm và đánh giá kết quả qua hiệu suất của mô hình, ma trận nhầm lẫn và mối tương quan giữa giá trị thực và giá trị dự báo. Các kết quả thu được cho thấy thuật toán LightGBM là một lựa chọn rất phù hợp các hệ thống phân tích và dự báo lưu lượng giao thông hiện đại.

**Từ khóa:** Dự báo giao thông; LightGBM; Hệ thống giao thông thông minh; LightGBM Regressor; LightGBM Classifier.

## ABSTRACT

This study explores the application of the LightGBM algorithm in developing a traffic-flow forecasting model using a dataset comprising 2,976 records with nine input features: Time, Date, Day of Week, Carcount, Bikecount, Buscount, Truckcount, Total and Situation. The authors conducted a series of experiments and evaluated the model using performance metrics, the confusion matrix and correlation analysis between actual and predicted values. The experimental results demonstrate that LightGBM is a highly effective and robust approach for modern traffic analysis and forecasting systems.

**Keywords:** Traffic forecasting; Intelligent transport systems; LightGBM; LightGBM Classifier; LightGBM Regressor.

## 1. ĐẶT VẤN ĐỀ

Dự báo lưu lượng giao thông là một vấn đề nghiên cứu quan trọng trong nhiều ứng dụng của Hệ thống giao thông thông minh (ITS). Sự mở rộng của Internet vạn vật (IoT) đã dẫn đến các giải pháp sáng tạo mới, chẳng hạn như các thành phố thông minh, giúp cuộc sống của chúng ta trở nên hiệu quả, tiện lợi và thông minh hơn. Cốt lõi của các thành phố thông minh là Hệ thống giao thông thông minh (ITS) đã được tích hợp vào nhiều ứng dụng thành phố thông minh nhằm cải thiện giao thông và khả năng di chuyển. ITS nhằm giải quyết nhiều vấn đề giao thông, chẳng hạn như vấn đề tắc nghẽn giao thông. Gần đây, các mô hình và khung dự đoán lưu lượng giao thông mới đã được phát triển nhanh chóng cùng với việc áp dụng các phương pháp trí tuệ nhân tạo nhằm cải thiện độ chính xác của dự đoán lưu lượng giao thông. Dự báo giao thông là một nhiệm vụ quan trọng trong ngành Giao thông vận tải. Nó có

thể ảnh hưởng đáng kể đến thiết kế các công trình và dự án đường bộ, bên cạnh tầm quan trọng của nó đối với việc lập kế hoạch tuyến đường và quy định giao thông. Hơn nữa, tình trạng tắc nghẽn giao thông là một vấn đề nghiêm trọng ở các khu vực đô thị và các thành phố đông đúc. Do đó, nó cần được đánh giá và dự báo một cách chính xác. Vì vậy, một phương pháp đáng tin cậy và hiệu quả để dự đoán lưu lượng giao thông là điều cần thiết. Phương pháp dự báo dùng LightGBM được rất nhiều tác giả trên thế giới sử dụng [1-6]. Có nhiều cách khác nhau sử dụng LightGBM để dự báo lưu lượng giao thông, trong đó [1] sử dụng LightGBM kết hợp với GRU để dự báo tắc nghẽn với dữ liệu được thu thập từ dịch vụ gọi xe tại Thành Đô, Trung Quốc. Kết quả của [1] cho thấy mô hình được đề xuất có những cải thiện đáng kể về độ chính xác và hiệu suất từng phần so với các nghiên cứu trước đó nhờ việc kết hợp LightGBM và GRU để sinh đặc trưng và dự báo chỉ số tắc nghẽn; báo cáo cải thiện so

với mô hình đơn lẻ nhờ tận dụng cả cây boosting và mạng hồi tiếp. [2] lại ứng dụng LightGBM để dự báo lưu lượng hành khách tuyến metro (short-term), cho thấy LightGBM nhanh và chính xác hơn một số mô hình truyền thống.

Mục tiêu chính của nghiên cứu này là: Trình bày một tổng quan toàn diện về thuật toán LightGBM được áp dụng trong dự đoán giao thông và từ đó thử nghiệm đánh giá cho mô hình để xác định tính phù hợp của mô hình với yêu cầu dự báo lưu lượng giao thông.

**2. MÔ HÌNH TOÁN HỌC CỦA THUẬT TOÁN LIGHTGBM**

**2.1. Đặc điểm của bộ dữ liệu sử dụng**

Tập dữ liệu được sử dụng trong nghiên cứu này được thu thập từ camera giám sát giao thông và lưu trữ theo định dạng CSV được đưa ra trong Bảng 1. Tập dữ liệu bao gồm các cột như: Thời gian tính theo giờ, ngày, ngày trong tuần và số lượng cho từng loại phương tiện (CarCount, BikeCount, BusCount, TruckCount). Cột "Tổng" biểu thị tổng số lượng của tất cả các loại phương tiện được phát hiện trong khoảng thời gian 15 phút. Bộ dữ liệu được cập nhật 15 phút một điểm. Ngoài ra, bộ dữ liệu bao gồm một cột chỉ ra tình hướng giao thông được phân loại thành 4 loại: 1-Heavy, 2-High, 3-Normal và 4-Low. Dữ liệu tình hướng giao thông này giúp đánh giá mức độ nghiêm trọng của tình trạng tắc nghẽn và theo dõi tình trạng giao thông tại các thời điểm và ngày khác nhau trong tuần.

**2.2. Thuật toán LightGBM để dự báo lưu lượng giao thông**

Mỗi bản ghi trong bộ dữ liệu CSV được biểu diễn bằng vector đặc trưng:

$$x_i = [t_i, d_i, w_i, c_i^{car}, c_i^{bike}, c_i^{bus}, c_i^{truck}] \in \mathbb{R}^7$$

Trong đó:

- $t_i$  - Thời gian trong ngày;
- $d_i$  - Ngày tháng;
- $w_i$  - Ngày trong tuần;
- $c_i^{(.)}$  - Số lượng xe theo từng loại trong 15 phút.

Biến mục tiêu là tổng lưu lượng:

$$y_i = Total_i = C_i^{car} + C_i^{bike} + C_i^{bus} + C_i^{truck}$$

Tổng các cây hồi quy của LightGBM:

$$F_M(x) = \sum_{m=1}^M f_m(x)$$

Trong đó, mỗi  $f_m$  là một cây quyết định mô tả các ngưỡng chia liên quan trực tiếp đến dữ liệu giao thông. Hàm mục tiêu tối ưu tại bước boosting thứ  $m$  như sau:

$$\mathcal{L}^{(m)} = \sum_{i=1}^N l(y_i, \hat{y}_i^{(m-1)} + f_m(x_i)) + \Omega(f_m)$$

Hàm mất mát:

$$l(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2$$

Khai triển Taylor ta có:

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^N [g_i f_m(x_i) + \frac{1}{2} h_i f_m(x_i)^2] + \Omega(f_m)$$

$g_i = \hat{y}_i^{(m-1)} - y_i$  (gradient - thể hiện mức sai lệch lưu lượng dự báo so với thực tế)  
 $h_i = 1$  (Hessian cho L2-loss)

Những mẫu có sai lệch lớn (cao điểm, tắc đường, dữ liệu biến động mạnh) sẽ có  $|g_i|$  lớn và được ưu tiên trong quá trình học.

Một cây có  $T$ , mỗi lá tương trưng cho một nhóm điều kiện giao thông giống nhau (ví dụ: "giờ cao điểm sáng với nhiều xe máy"). Với tập mẫu thuộc lá  $f$ :

$$I_j = \{i: x_i \text{ rơi vào lá } f\}$$

Giá trị tốt nhất của lá là:

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

Ý nghĩa:

- Nếu tại lá đó dự báo luôn thấp hơn thực tế  $\rightarrow$  gradient dương  $\rightarrow$  âm  $\rightarrow$  mô hình tăng dự báo.

- Nếu dự báo cao hơn thực tế  $\rightarrow$  gradient âm  $\rightarrow$  mô hình giảm dự báo.

Như vậy, mỗi lá được học để hiệu chỉnh sai số lưu lượng theo đúng dữ liệu thực tế giao thông.

Chọn Split dựa trên biến giao thông có lợi nhất ta cho điểm tách lá:

$$Gain = \frac{1}{2} \left( \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma$$

Nếu tách theo "carcount > 22" cho Gain cao nhất  $\rightarrow$  mô hình nhận ra số xe con ở mức này dẫn đến thay đổi đáng kể trong tổng lưu lượng.

Trong dữ liệu giao thông, biến động thường xảy ra:

- Giờ cao điểm;
- Xe lớn (truck/bus) xuất hiện đột ngột;
- Ngày lễ/không lễ;
- Sự kiện bất thường.

Các mẫu này có gradient cao. GOSS giữ toàn bộ các mẫu gradient lớn (ví dụ top 20%) và lấy mẫu ngẫu nhiên ở phần còn lại.

Gradient của mẫu được giữ lại hoặc lấy mẫu được điều chỉnh:

$$g'_i = \begin{cases} g_i & \text{nếu } i \in A \text{ (gradient lớn)} \\ \frac{1-a}{b} g_i & \text{nếu } i \in B \text{ (gradient nhỏ)} \end{cases}$$

Từ đó, mô hình học tập trung vào các đoạn giao thông quan trọng nhất.

Các đặc trưng xe thường không cùng khác 0 bất kỳ lúc nào (ví dụ: Xe buýt rất ít xuất hiện cùng lúc với số lượng xe truck lớn trong một giai đoạn nhỏ). Do đó, LightGBM gộp các đặc trưng để giảm số chiều mà không làm mất thông tin để tăng tốc độ huấn luyện:

$$x^{bundle} = x_{car} + x_{bus} + x_{truck}$$

Cập nhật cuối mỗi vòng Boosting:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta f_m(x_i)$$

Trong đó:  $\eta$  - Tốc độ học của mô hình.

Ý nghĩa trong dữ liệu giao thông:

- Mỗi cây hiệu chỉnh mô hình dựa trên sai số còn lại của dự báo lưu lượng theo từng khoảng thời gian;

- Mô hình ngày càng khớp hơn với quy luật di chuyển của xe trên từng loại.

**3. THỬ NGHIỆM VÀ ĐÁNH GIÁ**

Tiến hành thử nghiệm dự báo với mô hình LightGBM sử dụng bộ dữ liệu gồm 2.976 bản ghi dưới định dạng tệp CSV và 9 đặc trưng đầu vào: Time, Date, Day of Week, Carcount, Bikecount, buscount, truckcount, total, situation. Kết quả đánh giá cho mô hình như sau:

**3.1. Đánh giá hiệu suất của mô hình**

Trong nghiên cứu này, hai mô hình LightGBM được xây dựng nhằm giải quyết hai mục tiêu đánh giá khác nhau: (i) Mô hình hồi quy (LightGBM Regressor) dự báo tổng lưu lượng phương tiện trong mỗi khoảng thời gian và (ii) mô hình phân loại (LightGBM Classifier)

nhận diện mức độ ùn tắc giao thông (“low”, “normal”, “high”, “heavy”). Các kết quả thực nghiệm cho thấy cả hai mô hình đều đạt hiệu suất rất cao và ổn định, phản ánh khả năng mô hình hóa tốt cấu trúc dữ liệu và mối quan hệ giữa các đặc trưng đầu vào như trong Bảng 1 dưới đây:

Bảng 1. Kết quả đánh giá hiệu suất của hai mô hình LightGBM phân loại và hồi quy

Mô hình hồi quy LightGBM Regressor		Mô hình phân loại LightGBM Classifier	
MAE	1,63	Accuracy	0,9798
RMSE	2,31	Precision	0,9801
R <sup>2</sup>	0,9985	Recall	0,9798
		F1-score	0,9799

Từ Bảng 1 cho thấy, mô hình hồi quy cho kết quả MAE = 1,63, RMSE = 2,31 và R<sup>2</sup> = 0,9985, chứng tỏ mức độ dự báo chính xác rất cao trên toàn bộ tập dữ liệu. Giá trị MAE thấp cho thấy sai lệch trung bình giữa dự đoán và thực tế chỉ khoảng 1,63 phương tiện, điều này đặc biệt ấn tượng khi tổng số phương tiện trong các phiên đo có thể dao động lên tới hơn 200. Chỉ số RMSE nhỏ phản ánh rằng sai số lớn hầu như không xuất hiện và mô hình không gặp phải tình trạng dự đoán lệch có hệ thống. Đặc biệt là chỉ số xác định R<sup>2</sup> = 0,9985, chứng minh rằng mô hình giải thích được 99,85% phương sai của dữ liệu thực tế. Mức R<sup>2</sup> này cho thấy sự phù hợp gần như hoàn hảo giữa giá trị dự báo và giá trị thực, vượt xa mức thông thường trong các mô hình dự báo lưu lượng vốn thường chịu tác động mạnh bởi nhiễu thời gian thực, thay đổi hành vi giao thông hoặc sai lệch cảm biến. Kết quả này đồng thời cho thấy LightGBM tận dụng hiệu quả các đặc trưng đầu vào như thời gian, ngày trong tuần và số lượng từng loại phương tiện (car, bike, bus, truck) để học và suy luận quan hệ phi tuyến có trong dữ liệu.

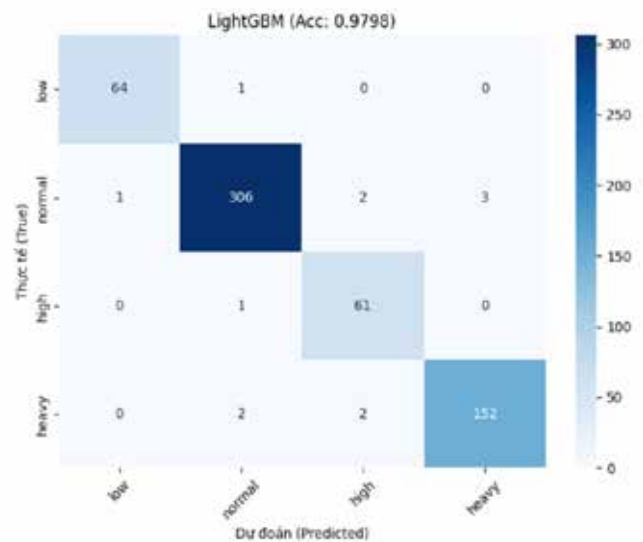
Đối với nhiệm vụ phân loại mức độ lưu thông, mô hình LightGBM Classifier đạt Accuracy = 0,9798, Precision = 0,9801, Recall = 0,9798 và F1-score = 0,9799. Kết quả này chứng minh mô hình có khả năng phân biệt rất tốt giữa các mức độ lưu lượng khác nhau và duy trì sự cân bằng giữa dự báo đúng và hạn chế lỗi bỏ sót. Giá trị Accuracy gần 98% cho thấy mô hình phân loại đúng phần lớn các trường hợp. Precision cao chứng tỏ mô hình dự đoán từng lớp rất chính xác, gần như không xảy ra dự báo nhầm sang lớp khác một cách không cần thiết. Recall đạt gần tương đương precision, cho thấy mô hình vẫn duy trì khả năng phát hiện đầy đủ các trường hợp thuộc từng lớp, bao gồm cả các lớp có tần suất xuất hiện thấp hơn như “heavy” hoặc “high”. Chỉ số F1-score xấp xỉ 0,98 phản ánh sự cân bằng rất tốt giữa precision và recall, đặc biệt quan trọng trong ngữ cảnh giao thông, nơi bỏ sót trạng thái ùn tắc (false negative) hoặc báo động giả (false positive) đều có thể gây ảnh hưởng tiêu cực đến hiệu quả vận hành hệ thống. Kết quả này chứng minh rằng LightGBM không chỉ phân loại chính xác các trạng thái giao thông mà còn duy trì độ tin cậy cao trong nhiều điều kiện vận hành khác nhau.

Như vậy, từ hai nhóm kết quả trên cho thấy, LightGBM là mô hình cực kỳ phù hợp cho cả bài toán hồi quy và phân loại liên quan đến dữ liệu lưu lượng giao thông. Mô hình đạt sai số thấp, mức độ giải thích cao và độ ổn định tốt. Sự nhất quán giữa các chỉ số hiệu suất xác nhận rằng mô hình không bị overfitting và có khả năng tổng quát hóa mạnh mẽ trên tập dữ liệu đa dạng theo thời gian. Với

độ chính xác cao và khả năng mô hình hóa quan hệ phi tuyến phức tạp, LightGBM có tiềm năng ứng dụng thực tế trong các hệ thống dự báo lưu lượng giao thông theo thời gian thực.

### 3.2. Đánh giá theo ma trận nhầm lẫn (Confusion Matrix)

Để đánh giá toàn diện hiệu quả mô hình phân loại lưu lượng giao thông, nghiên cứu sử dụng ma trận nhầm lẫn (confusion matrix) như một công cụ phân tích chi tiết bên cạnh các chỉ số tổng hợp như Accuracy, Precision, Recall hay F1-score. Ma trận nhầm lẫn cho phép quan sát trực quan mức độ dự đoán đúng/sai của mô hình cho từng lớp cụ thể, từ đó cung cấp các thông tin quan trọng về hành vi của mô hình đối với từng trạng thái giao thông (“low”, “normal”, “high”, “heavy”). Kết quả thu được như trên Hình 1:



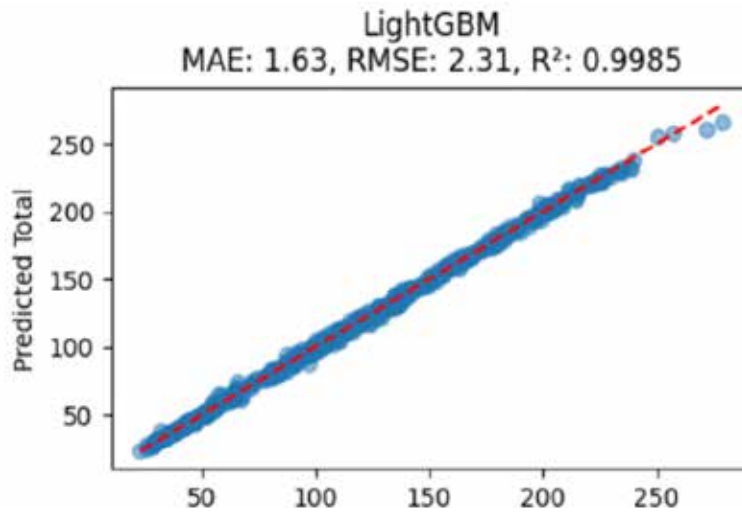
Hình 1. Ma trận nhầm lẫn khi dùng thuật toán LightGBM

Từ Hình 1 ta thấy mô hình LightGBM đạt độ chính xác tổng thể Accuracy = 0,9798, đồng thời duy trì mức lệch phân loại cực thấp giữa các lớp. Cụ thể, lớp *normal* vốn chiếm tỷ lệ lớn nhất trong tập dữ liệu, đạt 306 mẫu được dự đoán chính xác, chỉ có tổng cộng 6 mẫu bị nhầm sang các lớp khác. Điều này minh chứng rằng mô hình học tốt các quy luật phổ biến của trạng thái giao thông trung bình và hoạt động ổn định ngay cả khi dữ liệu không hoàn toàn cân bằng.

Đối với lớp *low*, mô hình dự đoán chính xác 64 trường hợp, chỉ có 1 trường hợp bị nhầm sang lớp *normal*. Đây là một kết quả tích cực, cho thấy mô hình có khả năng phân biệt tốt giữa mức lưu lượng thấp và mức trung bình - hai lớp thường dễ bị nhầm lẫn do biên độ dữ liệu gần nhau. Tương tự, lớp *high* cũng đạt hiệu suất rất tốt với 61 dự đoán chính xác, chỉ 1 mẫu bị nhầm sang *normal*, chứng minh rằng mô hình vẫn duy trì độ nhạy (recall) cao cho những mức lưu lượng ít xuất hiện hơn.

Lớp *heavy*, vốn có ảnh hưởng quan trọng trong các hệ thống điều khiển giao thông vì liên quan trực tiếp đến tình trạng ùn tắc, cho kết quả 152 mẫu được phân loại đúng và chỉ có 2 mẫu nhầm sang *high* và 2 mẫu nhầm sang *normal*. Tỷ lệ nhầm lẫn này là rất nhỏ so với tổng số mẫu, cho thấy mô hình có thể nhận diện trạng thái lưu lượng nặng một cách đáng tin cậy, từ đó giúp hỗ trợ tốt cho các hệ thống điều khiển đèn giao thông thích ứng hoặc cảnh báo ùn tắc.

### 3.3. Đánh giá bằng tương quan giữa giá trị thực và giá trị dự đoán



Hình 2. Kết quả mô hình LightGBM dựa trên quan hệ giữa giá trị thực và giá trị dự đoán

Hình 2 dưới đây minh họa mối tương quan giữa giá trị lưu lượng giao thông thực tế (Actual Total) và giá trị dự đoán (Predicted Total) của mô hình LightGBM Regressor trên tập dữ liệu kiểm thử. Đường chuẩn lý tưởng  $y = x$ .

Kết quả thu được cho thấy các điểm dữ liệu phân bố tập trung và bám sát đường tham chiếu, chứng tỏ mô hình học tốt quan hệ giữa các đặc trưng đầu vào và biến mục tiêu. Không thấy có các điểm “ngoại lai mạnh” (outliers) vượt quá đường lý tưởng một cách rõ rệt. Các điểm ở khoảng 200 - 260 (lưu lượng cao) cũng bám sát đường lý tưởng. Sai số ở vùng giá trị thấp (<50) cũng rất nhỏ. Điều này chứng minh LightGBM không bị sai lệch theo phân bố (no bias shift) - một ưu điểm rất quan trọng trong ứng dụng thực tế.

#### 4. KẾT LUẬN

Nội dung bài báo đã nghiên cứu ứng dụng thuật toán LightGBM để dự báo lưu lượng giao thông sử dụng bộ dữ liệu gồm 2.976 bản ghi và 9 trường dữ liệu đầu vào: Time, Date, Day of Week, carcount, bikecount, buscount, truckcount, total, situation. Kết quả thử nghiệm được đánh giá qua hiệu suất của mô hình, ma trận nhầm lẫn và tính tương quan giữa giá trị thực và giá trị dự đoán. Tất cả các kết quả thu được đều cho thấy mô hình dự báo sử dụng LightGBM là có độ tin cậy cao và phù hợp cho việc ứng dụng vào các hệ thống phân loại mức độ lưu lượng thời gian thực. Điều đó chứng minh rằng LightGBM là lựa chọn mạnh mẽ và hiệu quả cho các hệ thống phân tích và dự báo lưu lượng giao thông hiện đại.

#### TÀI LIỆU THAM KHẢO

- [1]. W. Cheng, J.-I. Li, H.-C. Xiao, và L.-n. Ji (2022), Combination predicting model of traffic congestion index in weekdays based on LightGBM-GRU, Scientific Reports, vol.12, art.2912. Doi: 10.1038/s41598-022-06975-1.
- [2]. Y. Zhang, C. Zhu and Q. Wang, LightGBM-based model for metro passenger volume forecasting, IET Intelligent Transport Systems, vol.14, no.13, pp.1815-1823. Doi: 10.1049/iet-its.2020.0396.
- [3]. I. Ahmed, I. Kumara, V. Reshadat, A. S. M. Kayes, W.-J. van den Heuvel and D. A. Tamburri (2022), Travel Time Prediction and Explanation with Spatio-Temporal Features: A Comparative Study, Electronics, vol.11, no.1, art.106. Doi: 10.3390/electronics11010106.

- [4]. Z. Chu, J. Yu and A. Hamdulla (2020), LPG-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis and deep gated recurrent unit network, Information Sciences, vol.535, pp.107-129. Doi: 10.1016/j.ins.2020.05.042.

- [5]. X. Zeng, Y. Wang, J. Wang, et al. (2021), Short-Term Traffic Flow Prediction Based on Ensemble Machine Learning Strategies, in Proceedings of IEEE 10th International Conference on Big Data (BigData Congress/Data Science).

- [6]. M. Patel (2024), Unleashing the Potential of Boosting Techniques to Station-Pair Short-Term Passenger Flow Forecast, Procedia Computer Science/Conference Proceedings.