TRƯỜNG ĐẠI HỌC THỦY LỢI



NHÓM 03

BÁO CÁO BÀI TẬP LỚN MÔN HỌC

PHÂN TÍCH CHUỖI THỜI GIAN

Dự báo tải mạng viễn thông Quy mô quốc gia

|  |  |
| --- | --- |
| GIẢNG VIÊN HƯỚNG DẪN | TRẦN ANH ĐẠT |

Hà Nội, năm 2025

MỤC LỤC

**[MỞ ĐẦU 4](#_Toc18209)**

**[CHƯƠNG 1 GIỚI THIỆU BÀI TOÁN VÀ BỘ DỮ LIỆU 5](#_Toc22691)**

**[1. Giới thiệu bài toán 5](#_Toc18893)**

**[2. Giới thiệu tập dữ liệu 5](#_Toc29802)**

**[CHƯƠNG 2 CƠ SỞ LÝ THUYẾT 7](#_Toc27830)**

**[2.1. LSTM 7](#_Toc1051)**

**[2.1.1 Khái niệm 7](#_Toc17152)**

**[2.1.2 Cấu trúc 7](#_Toc9537)**

**[2.1.3 Các bước của mô hình 8](#_Toc22305)**

**[CHƯƠNG 3 THỰC NGHIỆM VÀ KẾT QUẢ 10](#_Toc22342)**

**[3.1 Tiền xử lý dữ liệu 10](#_Toc30331)**

**[3.2 Kết quả thực nghiệm 11](#_Toc22301)**

**[3.2.1 LSTM 11](#_Toc25321)**

**[3.2.2 Tạo giao diện 14](#_Toc15648)**

**[KẾT LUẬN 15](#_Toc24988)**

Thành viên thực hiện

|  |  |
| --- | --- |
| Họ tên | Mã sinh viên |
| Nguyễn Văn Quân | 2251262627 |

MỞ ĐẦU

Trong thời đại chuyển đổi số và bùng nổ dữ liệu hiện nay, hạ tầng viễn thông đóng vai trò then chốt trong việc duy trì sự kết nối liên tục của hàng triệu người dùng và thiết bị. Các nhà mạng viễn thông không chỉ đối mặt với áp lực về mặt kỹ thuật trong việc đảm bảo chất lượng dịch vụ (QoS), mà còn cần có khả năng **dự báo chính xác tải mạng** để tối ưu hóa việc phân bổ tài nguyên, nâng cao hiệu suất vận hành và giảm thiểu rủi ro gián đoạn mạng.

Một trong những hướng tiếp cận hiệu quả để giải quyết bài toán này là sử dụng các mô hình **dự báo chuỗi thời gian (time series forecasting)** nhằm ước lượng trước lưu lượng mạng trong tương lai tại từng khu vực cụ thể. Tuy nhiên, tải mạng viễn thông là một loại dữ liệu có tính **đa biến, phi tuyến và dao động mạnh** theo thời gian (ngày/đêm, ngày thường/cuối tuần, lễ Tết,...), nên đòi hỏi các mô hình dự báo có độ chính xác cao và khả năng khái quát tốt.

Thông qua đề tài này, em hướng tới mục tiêu không chỉ **ứng dụng các mô hình học máy hiện đại vào bài toán thực tiễn**, mà còn góp phần xây dựng nền tảng hỗ trợ ra quyết định trong việc quản lý và điều phối hệ thống mạng viễn thông quy mô lớn.

# GIỚI THIỆU BÀI TOÁN VÀ BỘ DỮ LIỆU

1. **Giới thiệu bài toán**

Trong bối cảnh nhu cầu kết nối viễn thông ngày càng gia tăng và hạ tầng mạng phải liên tục thích ứng với tải trọng không ổn định, việc **dự báo tải mạng** (bao gồm lưu lượng **Internet, cuộc gọi, và tin nhắn SMS**) đóng vai trò then chốt trong việc **quản lý tài nguyên mạng hiệu quả**, đảm bảo **chất lượng dịch vụ** và tối ưu hóa hoạt động vận hành của các nhà mạng.

Đặc biệt tại các quốc gia có **mật độ thuê bao cao**, tải mạng có thể thay đổi đáng kể theo **không gian**và **thời gian** (theo chu kỳ giờ, ngày, tuần,...), đòi hỏi những mô hình dự báo có khả năng xử lý **chuỗi thời gian đa biến và có tính phi tuyến cao**.

Đề tài này tập trung vào bài toán **dự báo tải mạng viễn thông** theo **quy mô toàn quốc**, dựa trên dữ liệu **CDRs (Call Detail Records)** là bản ghi chi tiết cuộc gọi, tin nhắn và dùng dữ liệu Internet do các nhà mạng thu thập mỗi khi người dùng thực hiện một hoạt động viễn thông. Dữ liệu được thu thập và tổ chức theo lưới không gian thời gian với độ phân giải cao (mỗi 10 phút), giúp phản ánh chi tiết hoạt động của người dùng trong từng vùng cụ thể.

1. **Giới thiệu tập dữ liệu**

Dữ liệu sử dụng trong đề tài được lấy từ một tập dữ liệu thực tế được công bố bởi công ty viễn thông Telecom Italia, bộ dữ liệu này sau đó đã được chuẩn hóa và lưu trữ trên nền tảng [Harvard Dataverse](https://dataverse.harvard.edu" \t "_new). Dữ liệu đã được thu thập trong hai tháng, từ ngày 1 tháng 11 năm 2013 đến ngày 1 tháng 1 năm 2014 ở hai thành phố là Trento và Milan. Với số lượng dữ liệu của thành phố Trento là gần 70 triệu dòng, của thành phố Milan là hơn 160 triệu dòng.

Bộ dữ liệu có 8 đặc trưng:

|  |  |  |
| --- | --- | --- |
| STT | Tên đặc trưng | Ý nghĩa |
| 1 | Square id | ID của ô vuông trong lưới không gian của thành phố |
| 2 | Time Interval | Khoảng thời gian ghi nhận mỗi 10 phút/lần. |
| 3 | Country code | Mã quốc gia của điện thoại. |
| 4 | SMS-in activity | Mức độ hoạt động của tin nhắn đến |
| 5 | SMS-out activity | Mức độ hoạt động của tin nhắn đi |
| 6 | Call-in activity | Mức độ hoạt động của cuộc gọi đến |
| 7 | Call-out activity | Mức độ hoạt động của cuộc gọi đi |
| 8 | Internet traffic activity | Mức độ sử dụng dữ liệu Internet. |

# CƠ SỞ LÝ THUYẾT

## **LSTM**

**2.1.1 Khái niệm**

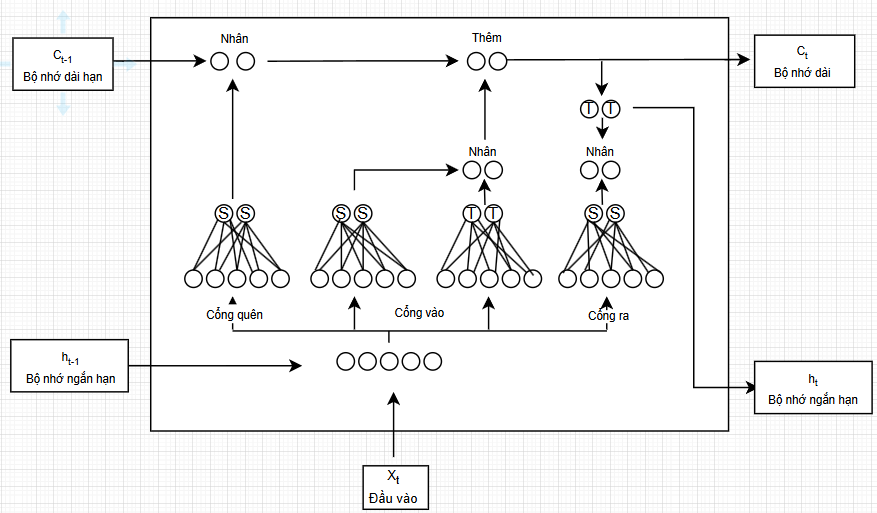
Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

Chìa khóa của LSTM là bộ nhớ dài hạn (cell state) một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các mắt xích (các nút mạng) và chỉ tương tác tuyến tính đôi chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay

đổi.

**2.1.2 Cấu trúc**



****Cell state (bộ nhớ dài hạn)****: Như một "băng chuyền" chuyên chở thông tin xuyên suốt chuỗi thời gian

**Ba cổng điều khuyển**:

**+ Forget Gate** : Quyết định phần nào của cell state cũ cần bị loại bỏ, phần nào sẽ được giữ lại

**+ Input Gate và **Candidate Cell State****  : Tạo thông tin mới bằng hàm tanh và dùng thàm sigmoid để quyết định phần nào nên lưu vào bộ nhớ.

****Candidate Cell State****  : Tạo ra giá trị đề xuất cập nhật cho cell state, được “nén” trong khoảng [−1,1]bởi hàm tanh.

**+ Output Gate** : Quyết định phần nào của cell state sẽ được “trích xuất” ra làm hidden state.

**2.1.3 Các bước của mô hình**

Bước đầu tiên của LSTM là quyết định xem thông tin nào cần bỏ đi từ bộ nhớ dài hạn . Quyết định này được đưa ra bởi tầng sigmoid - gọi là “cổng quên” . Nó sẽ lấy đầu vào là  và rồi đưa ra kết quả là một số trong khoảng [0,1]  cho mỗi số trong .Đầu ra là 1 thể hiện rằng nó giữ toàn bộ thông tin lại, còn 0 chỉ rằng toàn bộ thông tin sẽ bị bỏ đi.

Bước tiếp theo là quyết định xem thông tin mới nào ta sẽ lưu vào bộ nhớ dài hạn. Đầu tiên là sử dụng một tầng sigmoid được gọi là “cổng vào” để quyết định giá trị nào ta sẽ cập nhập. Tiếp theo là một tầng tanh tạo ra một véc-tơ cho giá trị mới nhằm thêm vào cho bộ nhớ dài hạn Trong bước tiếp theo, ta sẽ kết hợp 2 giá trị đó lại để tạo ra một cập nhập cho bộ nhớ.

Cuối cùng, ta cần quyết định xem ta muốn đầu ra là gì. Giá trị đầu ra sẽ dựa vào bộ nhớ dài hạn, nhưng sẽ được tiếp tục sàng lọc. Đầu tiên, ta chạy một tầng sigmoid để quyết định phần nào của bộ ta muốn xuất ra. Sau đó, ta đưa bộ nhớ qua một hàm tanh để co giá trị nó về khoảng [−1,1], và nhân nó với đầu ra của cổng sigmoid để được giá trị đầu ra ta mong muốn.

# CHƯƠNG 3 THỰC NGHIỆM VÀ KẾT QUẢ

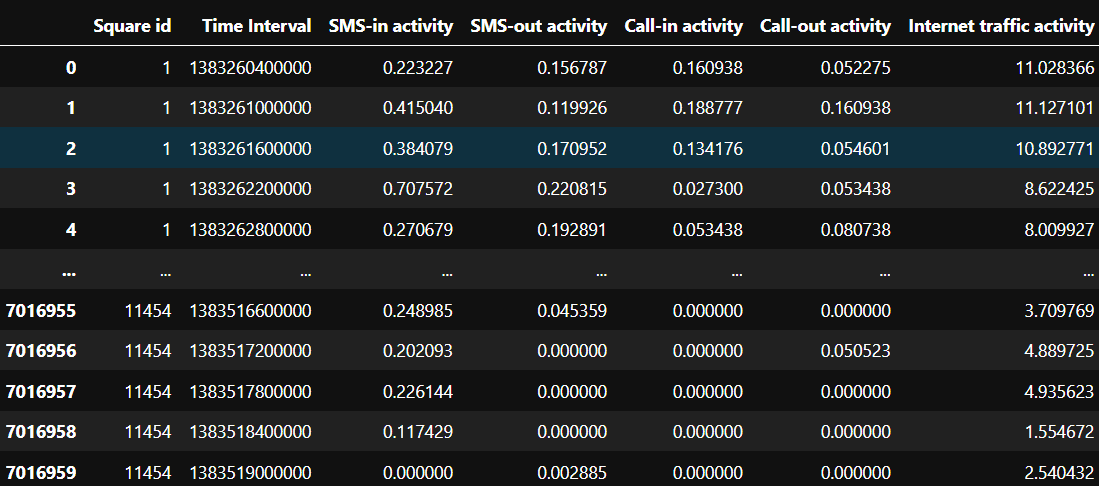
## 3.1 Tiền xử lý dữ liệu

Ban đầu dữ liệu gốc các Square id sẽ có nhiều giá trị Time Interval trùng nhau gây ra mất đi đặc trưng thời gian của dữ liệu nên cần phải gộp các giá trị có giá trị Time Interval trùng nhau bằng cách tính tổng các giá trị.

Nhóm nhận thấy rằng khi gộp các giá trị Country code không còn giống như ban đầu nữa và nó không có tác dụng khi thực hiện dự báo với đề tài của nhóm, nên sẽ xóa cột đó đi.

Dữ liệu gốc bao gồm 62 files text dữ liệu với mỗi file là 1 ngày dữ liệu. Và chúng em thấy rằng nếu gộp tất cả dữ liệu vào cùng một file sẽ rất lớn và không thể thực hiện gì được với mô hình (hơn 145 triệu dòng dữ liệu), nên chúng em chia dữ liệu thành các file batch dữ liệu (21 batch) để có thể dễ dàng thực hiện.

Dưới đây là 1 batch dữ liệu sau khi xử lý xong.



## 3.2 Kết quả thực nghiệm

**3.2.1 LSTM**

Đầu vào:

Input\_len : 432 số bước thời gian đầu vào (tức là 432 × 10 phút = 72 giờ dữ liệu trước).

Input\_dim : 5

Hidden\_dim : 64

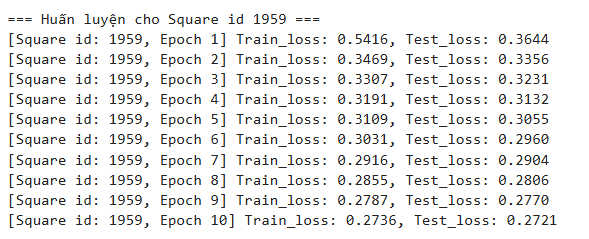
Learning\_rate : 0.01

Back\_size : 32

N\_epochs : 10 ( mỗi square id )

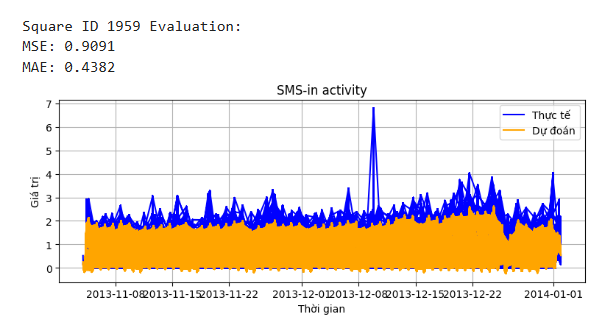
Num\_layer : 2

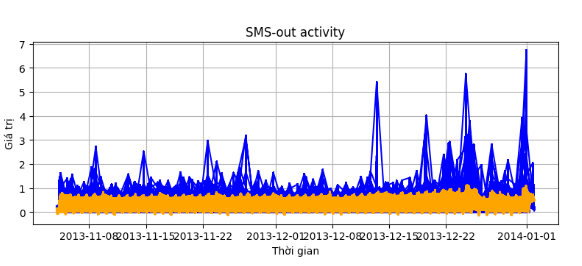
Chạy thử nghiệm trên square id 1959:

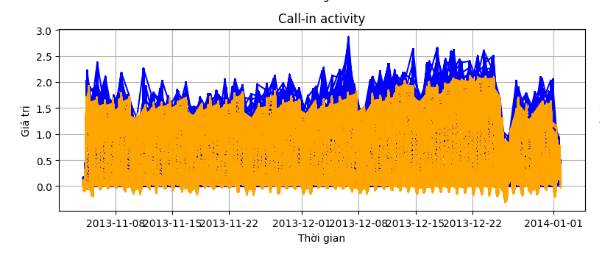


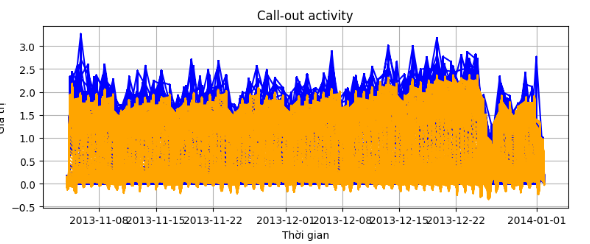
**Train loss và Test loss đều giảm đều qua các epoch** → cho thấy mô hình học được tốt, không overfit.

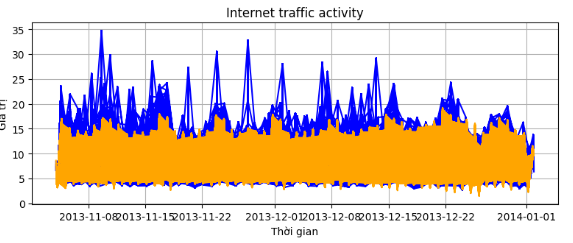
Không có hiện tượng tăng test loss ở cuối -> có thể tăng epoch để mô hình học tốt hơn



\







Mô hình **dự báo khá tốt các đặc trưng có xu hướng ổn định**, thể hiện rõ ở các biểu đồ như:, Call-out activity, Call-in activity

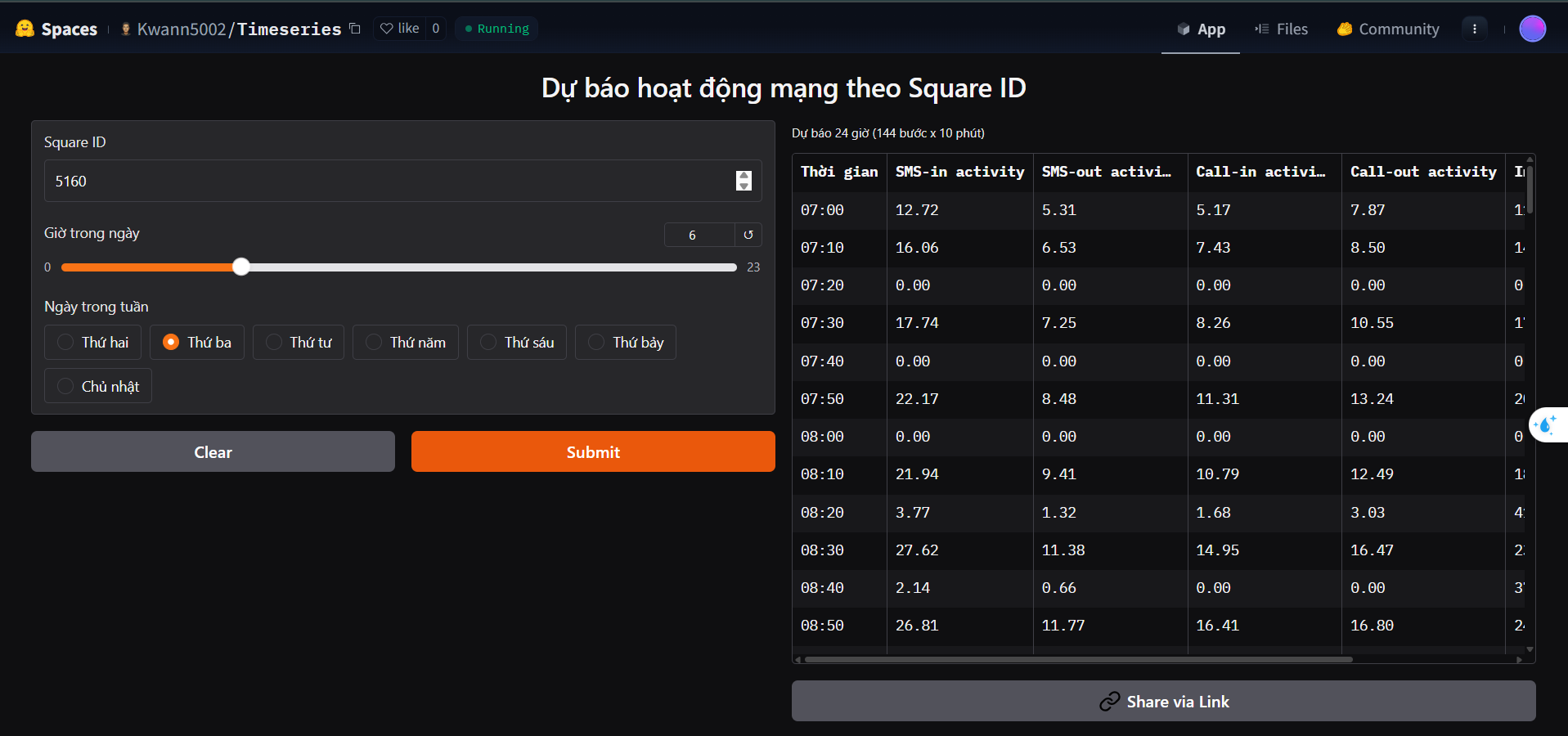
Mô hình **khó dự đoán các giá trị đột biến** như các đỉnh nhọn xuất hiện trong SMS-in, Call-in, và Internet traffic.

=> Mô hình LSTM hiện tại dự báo tốt với xu hướng dài hạn và các hoạt động ổn định, nhưng còn hạn chế với dữ liệu có độ dao động cao hoặc có nhiều điểm đột biến.

### **3.2.2 Tạo giao diện**

Trong quá trình xây dựng hệ thống dự báo tải mạng quy, em đã triển khai một giao diện người dùng trực quan nhằm giúp người sử dụng dễ dàng tương tác với mô hình mà không cần hiểu sâu về kỹ thuật lập trình. Giao diện này được phát triển trên nền tảng Hugging Face Spaces, sử dụng thư viện Gradio – một công cụ mã nguồn mở hỗ trợ xây dựng giao diện web nhanh chóng cho các mô hình học máy.

Người dùng có thể truy cập giao diện thông qua đường dẫn: <https://huggingface.co/spaces/Kwann5002/Timeseries>

Mục tiêu chính của giao diện là cho phép người dùng:

Nhập **mã trạm viễn thông (Square ID)** tương ứng với một ô vuông địa lý (cell tower),

Chọn **mốc thời gian cụ thể** (giờ và ngày trong tuần),

Nhận kết quả **dự báo hoạt động mạng trong vòng 24 giờ tiếp theo**, tương ứng với 144 bước thời gian (mỗi bước dài 10 phút),

Trực quan hóa kết quả dự báo dưới dạng bảng.

# KẾT LUẬN

Trong bối cảnh nhu cầu viễn thông ngày càng gia tăng, đặc biệt tại các khu vực đô thị và quốc gia có mật độ thuê bao cao, việc dự báo tải mạng viễn thông đóng vai trò thiết yếu trong quản lý và tối ưu hóa tài nguyên hạ tầng mạng. Đề tài này đã thực hiện xây dựng hệ thống dự báo tải mạng quy mô toàn quốc, dựa trên phân tích dữ liệu CDRs thực tế với độ phân giải thời gian 10 phút và độ phân giải không gian theo lưới (Square ID).

Mô hình **LSTM** cho kết quả tốtvới các đặc trưng có tính ổn định, nhưng gặp khó khăn với các tín hiệu dao động mạnh như lượng dữ liệu Internet hoặc các đỉnh hoạt động bất thường.

Bên cạnh việc xây dựng mô hình, nhóm cũng phát triển **giao diện người dùng trực quan**, cho phép người dùng nhập thông tin (Square ID, giờ, ngày trong tuần) và nhận kết quả dự báo tải mạng trong 24 giờ tiếp theo. Giao diện này góp phần làm tăng tính ứng dụng thực tiễn và khả năng mở rộng của hệ thống.