BỘ TÀI NGUYÊN VÀ MÔI TRƯỜNG

TỔNG CỤC KHÍ TƯỢNG THỦY VĂN

TRUNG TÂM THÔNG TIN VÀ DỮ LIỆU KHÍ TƯỢNG THỦY VĂN

CHƯƠNG TRÌNH KH&CN CẤP QUỐC GIA "KHOA HỌC VÀ CÔNG NGHỆ ỨNG PHÓ VỚI BIẾN ĐỔI KHÍ HẬU QUẢN LÝ VỀ TÀI NGUYÊN VÀ MÔI TRƯỜNG", MÃ SỐ BĐKH/16-20

BÁO CÁO TỔNG HỢP

ĐỀ TÀI: NGHIÊN CỨU CƠ SỞ KHOA HỌC VÀ GIẢI PHÁP ỨNG DỤNG TRÍ TUỆ NHÂN TẠO ĐỂ NHẬN DẠNG, HỖ TRỢ DỰ BÁO VÀ CẢNH BÁO MỘT SỐ HIỆN TƯỢNG KHÍ TƯỢNG THỦY VĂN NGUY HIỂM TRONG BỐI CẢNH BIẾN ĐỔI KHÍ HẬU TẠI VIỆT NAM

Mã số: BĐKH.34/16-20

Tổ chức chủ trì: Trung tâm Thông tin và Dữ liệu khí tượng thủy văn

Chủ nhiệm đề tài: ThS. Ngô Văn Mạnh

Thời gian thực hiện: 2018 - 2020

HÀ NỘI - 2020

BỘ TÀI NGUYÊN VÀ MÔI TRƯỜNG
TỔNG CỤC KHÍ TƯỢNG THỦY VĂN
TRUNG TÂM THÔNG TIN VÀ DỮ LIỆU KHÍ TƯỢNG THỦY VĂN

CHƯƠNG TRÌNH KH&CN CẤP QUỐC GIA "KHOA HỌC VÀ CÔNG NGHỆ ỨNG PHÓ VỚI BIẾN ĐỔI KHÍ HẬU QUẢN LÝ VỀ TÀI NGUYÊN VÀ MÔI TRƯỜNG", MÃ SỐ BĐKH/16-20

BÁO CÁO TỔNG HỢP

ĐỀ TÀI: NGHIÊN CỨU CƠ SỞ KHOA HỌC VÀ GIẢI PHÁP ỨNG DỤNG TRÍ TUỆ NHÂN TẠO ĐỂ NHẬN DẠNG, HỖ TRỢ DỰ BÁO VÀ CẢNH BÁO MỘT SỐ HIỆN TƯỢNG KHÍ TƯỢNG THỦY VĂN NGUY HIỂM TRONG BỐI CẢNH BIẾN ĐỔI KHÍ HẬU TẠI VIỆT NAM

Mã số: BĐKH.34/16-20

CHỦ NHIỆM ĐỀ TÀI	TỔ CHỨC CHỦ TRÌ
(Ký ghi rõ họ tên)	(Ký ghi rõ họ tên và đóng dấu)
ThS. Ngô Văn Mạnh	Trần Văn Nghĩa

HÀ NỘI - 2020

DANH SÁCH CÁN BỘ THAM GIA THỰC HIỆN ĐỀ TÀI

TT	Họ và tên, học hàm học vị	Chức danh thực hiện đề tài	Tổ chức công tác
1	ThS. Ngô Văn Mạnh	Chủ nhiệm	Trung tâm Thông tin và Dữ liệu khí tượng thủy văn
2	PGS.TS. Nguyễn Xuân Hoài	Thư ký khoa học	Công ty cổ phần công nghệ VDSpaces
3	TS. Ban Hà Bằng	Thành viên chính	Trường Đại học Bách khoa Hà Nội
4	TS. Nguyễn Đăng Quang	Thành viên chính	Trung tâm Dự báo khí tượng thủy văn quốc gia
5	KS. Vũ Trọng Thành	Thành viên chính	Trung tâm Ứng dụng công nghệ khí tượng thủy văn
6	PGS.TS. Nguyễn Bá Thủy	Thành viên chính	Trung tâm Dự báo khí tượng thủy văn quốc gia
7	TS. Võ Văn Hòa	Thành viên chính	Đài Khí tượng thủy văn khu vực Đồng bằng Bắc Bộ
8	ThS.Bùi Đình Lập	Thành viên chính	Trung tâm Dự báo khí tượng thủy văn quốc gia
9	ThS. Lê Đại Thắng	Thành viên chính	Trung tâm Thông tin và Dữ liệu khí tượng thủy văn
10	KS. Vũ Duy Tiến	Thành viên chính	Trung tâm Thông tin và Dữ liệu khí tượng thủy văn

CÁC TỔ CHỨC PHỐI HỢP

(Ghi các tổ chức phối hợp chính thực hiện đề tài)

STT	Các tổ chức phối hợp
1	Trung tâm Dự báo khí tượng thủy văn quốc gia
2	Đài Khí tượng Thủy văn khu vực Đồng bằng Bắc Bộ
3	Đài Khí tượng Thủy văn khu vực Đông Bắc
4	Đài Khí tượng Thủy văn khu vực Bắc Trung Bộ
5	Tạp chí Khí tượng Thủy văn

THÔNG TIN KẾT QUẢ NGHIÊN CỨU

1. Thông tin chung

- Tên đề tài: Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam;
- Mã số: BĐKH.34/16-20.
- Chủ nhiệm đề tài: Ngô Văn Mạnh.
- Tổ chức chủ trì: Trung tâm Thông tin và Dữ liệu khí tượng thủy văn.
- Thời gian thực hiện: 30 tháng (Từ tháng 07/2018 đến tháng 12/2020).

2. Mục tiêu

- Xác định và đưa ra được cơ sở khoa học và giải pháp ứng dụng của AI để nhận dạng và dự báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam.
- Đề xuất và ứng dụng được một số mô hình AI để nhận dạng và dự báo một số hiện tượng khí tượng thủy văn nguy hiểm gồm bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão.
- Xây dựng được hệ thống nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm dựa trên mô hình AI phù hợp và bước đầu triển khai thử nghiệm trong dự báo nghiệp vụ.

3. Kết quả nghiên cứu

Hệ thống nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm dựa trên mô hình AI phù hợp.

4. Sản phẩm đã đạt được

- Cơ sở dữ liệu lớn Big Data lưu trữ toàn bộ dữ liệu về bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão.
- Báo cáo cơ sở khoa học và thực tiễn của việc ứng dụng AI để nhận dạng và dự báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam.
- Mô hình, thuật toán, cơ chế học máy của một số mô hình AI để nhận dạng và dự báo một số hiện tượng khí tượng thủy văn nguy hiểm gồm bão, mưa lớn diện rộng, không khí lạnh, lũ (trên hệ thống sông Hồng), nước biển dâng do bão (ven biển Bắc Bộ và Bắc Trung Bộ).

- Báo cáo đánh giá chất lượng nhận dạng và dự báo một số hiện tượng khí tượng thủy văn nguy hiểm dựa trên mô hình AI.
- Hệ thống nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm dựa trên mô hình AI phù hợp.
- Báo cáo quy trình, chuyển giao và đào tạo sản phẩm. Thử nghiệm đánh giá khả năng ứng dụng của công nghệ trong thực tiễn.
- Báo cáo tổng thuật, báo cáo tóm tắt, các báo cáo chuyên đề và 8 bài báo khoa học.

5. Phương thức chuyển giao, địa chỉ ứng dụng, tác động và lợi ích mang lại của kết quả nghiên cứu

- P hương thức chuyển giao: Sản phẩm của đề tài sẽ được chuyển giao cho các đơn vị sử dụng bằng hình thức chuyển giao có tập huấn cho các cán bộ ở Trung tâm Dự báo KTTV quốc gia và các Đài KTTV khu vực Bắc Trung Bộ, Đông Bắc, Đồng bằng Bắc Bộ, Trung tâm Thông tin và Dữ liệu KTTV. Một số sản phẩm có thể được chuyển giao ngay trong quá trình thực hiện đề tài.
- Địa chỉ ứng dụng: Trung tâm Thông tin và Dữ liệu KTTV, Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc Bộ và Bắc Trung Bộ.
- Tác động và lợi ích mang lại: Kết quả của đề tài sẽ góp phần nâng cao năng lực dự báo, cảnh báo thời tiết nguy hiểm như bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão, góp phần đáp ứng được yêu cầu của Luật Phòng, Chống thiên tai và Luật KTTV.

INFORMATION ON RESEARCH RESULTS

1. General information

- Project title: Research scientific bases and solutions to apply artificial intelligence to identify, support forecasts and warnings about some dangerous hydro-meteorological phenomena in the context of climate change in Vietnam.
- Code number: BĐKH.34/16-20.
- Coordinator: Ngo Van Manh.
- Implementing institution: Information Center and Hydro-Meteorological Data.
- Duration: 30 months (From July 2018 to December 2020)

2. Objectives

- Identify and provide scientific foundations and solutions to apply artificial intelligence t1000o identify and forecast some dangerous hydro-meteorological phenomena in the context of climate change in Vietnam.
- Proposing and applying a few artificial intelligence models to identify and forecast a number of dangerous hydro-meteorological phenomena including storms, heavy rains, cold air, floods and rising sea levels.
- Building a system of identification, support for forecasting and warning of some dangerous hydro-meteorological phenomena based on an appropriate artificial intelligence model and initially deploying experiments in professional forecasting.

3. Research results

Identification system, support for forecasting and warning of some dangerous hydro-meteorological phenomena based on appropriate artificial intelligence model.

4. Products

- Big Data Database Big Data stores all data about storms, large rainfall, cold air, floods, and storm surge.
- Report on scientific and practical basis of the application of artificial intelligence to identify and forecast some dangerous hydro-meteorological phenomena in the context of climate change in Vietnam.
- Models, algorithms, machine learning mechanisms of some artificial intelligence models to identify and forecast a number of dangerous hydro-meteorological

phenomena including storms, large heavy rains, cold air, floods (on the Red River system), storm surge (coastal of North and North Central).

- Report assessing the quality of identification and forecast of some dangerous hydro-meteorological phenomena based on artificial intelligence model.
- The system of identification, support for forecasting and warning of some dangerous hydro-meteorological phenomena based on an appropriate artificial intelligence model.
- Report process, transfer and product training. Testing and evaluating applicability of technology in practice.
- General reports, summary reports, topical reports and 8 scientific papers.

5. Transfer alternatives of research results, application address, impact and benefits of research results

- Transfer alternatives: The products of the project will be transferred to the units for use in the form of transfer with training for officials at the National Center for Economic Forecasting and the Northern Central Vietnam Television Channels., Northeast, Northern Delta, Center for Information and Data on KTTV. Some products may be delivered immediately during the course of the project.
- Application address: Information and data center on KTTV, National Center for Hydrometeorology, Hydrological and Hydrometeorological Station in the Northeast region, the Northern Delta and the North Central Coast.
- Benefits: The results of the project will contribute to improving the capacity for forecasting and warning of dangerous weather such as storms, heavy rains, cold air, floods and storm surges, contributing to meeting the requirements of the Law on Prevention and Combat. and Resist. natural disasters and meteorology and hydrometeorology and hydrological law.

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

STT	Chữ viết tắt	Ý nghĩa
1.	AI	Trí tuệ nhân tạo - Artificial Intelligence
2.	ANN	Mạng thần kinh nhân tạo - Artificial Neural Network
3.	ATNĐ	Áp thấp nhiệt đới
4.	CA	Thuật toán gom cụm - Clustering Algorithms
5.	CSDL	Cơ sở dữ liệu
6.	CNN	Mạng nơ-ron tích chập - Convolutional Neural Networks
7.	DL	Học sâu - Deep Learning
8.	DM	Phương pháp bản đồ phổ biến - Diffusion Maps
9.	ĐTĐM	Điện toán đám mây
10.	GMĐB	Gió mùa Đông Bắc
11.	HDFS	Hệ thống tập tin phân tán - Hadoop Distributed File System
12.	HPC	Tính toán hiệu năng cao - High Performance Computing
13.	JMA	Cơ quan khí tượng Nhật Bản
14.	KKL	Không khí lạnh
15.	k-NN	Thuật toán K láng giềng gần nhất - k-Nearest Neighbor
16.	KTTV	Khí tượng thủy văn
17.	LDA	Phương pháp phân tích khác biệt tuyến tính - Linear discriminant analysis
18.	LIME	Phép diễn giải cục bộ cho mô hình bất khả tri - Local Interpretable Model-Agnostic Explanations
19.	LR	Phương pháp hồi quy tuyến tính - Linear Regression
20.	LSTM	Mạng bộ nhớ dài - ngắn - Long Short Term Memory Networks
21.	MNM	Mã nguồn mở
22.	MCDM	Mô hình ra quyết định đa tiêu chuẩn - Multiple Criteria Decision Making
23.	MAE	Sai số tuyệt đối trung bình - Mean Absolute Error
24.	MDA	Phân tích phân biệt đa nhóm - Multiple Discriminant Analysis
25.	ME	Sai số trung bình - Mean Error
26.	ML	Học máy - Machine Learning

STT	Chữ viết tắt	Ý nghĩa
27.	MLR	Hồi quy tuyến tính đa biến - Multiple Linear Regression
28.	MLP	Thuật toán dựa trên Perceptron nhiều lớp - Multi-layer Perceptron
29.	MPI	Giao diện truyền thông điệp - Message Passing Interface
30.	RMSE	Sai số quân phương - Root Mean Square Error
31.	PCA	Phương pháp phân tích thành phần chính - Principal Component Analysis
32.	SL	Mô hình học có giám sát - Supervised learning
33.	SVD	Thuật toán phân rã giá trị riêng - Singular Value Decomposition
34.	UL	Mô hình học không có giám sát - Unsupervised Learning

MỤC LỤC

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

1. Bối cảnh và tính cấp thiết của đề tài

Trên thế giới và Việt Nam, trí tuệ nhân tạo (Artificial Intelligence - AI) được xem là một trong những công nghệ cốt lõi của cuộc cách mạng công nghiệp 4.0. Nhiều quốc gia bắt đầu ghi nhận xu thế phát triển tất yếu và tác động chuyển đổi to lớn của AI trong mọi mặt đời sống xã hội, từ phát triển kinh tế xã hội, thay đổi cán cân quyền lực kinh tế, đến quân sự, chính trị để đem lại sự ổn định và thịnh vượng cho các quốc gia.

Trong những năm gần đây các kỹ thuật trong trí tuệ nhân tạo (AI) và học máy (ML - machine learning) đã được áp dụng rộng rãi trong lĩnh vực khí tượng thủy văn (KTTV) và biến đổi khí hậu. So với các phương pháp phân tích truyền thống dựa trên các hệ phương trình vi-tích phân, AI và ML có các lợi thế lớn sau: (i) các mô hình dựa trên AI/ML được xây dựng dựa trên dữ liệu (data driven), do đó có khả năng phản ánh chính xác hơn các quan hệ nhân quả của các đặc tính (attribute/feature) KTTV liên quan đến khí hậu và các hình thái thời tiết; (ii) AI/ML cho phép xử lý lượng dữ liệu lớn (big data), dữ liệu phức tạp, dữ liệu có nhiễu, phi cấu trúc, vố thường có trong các bài toán của KTTV. Điều này giúp cho phép xây dựng các mô hình hỗ trợ dự đoán, dự báo các hiện tượng KTTV bất thường, nguy hiểm.

Thực tế, trong thời gian gian gần đây, AI và ML đã được áp dụng vào việc dự đoán, dự báo các hiện tượng KTTV bất thường, cụ thể, về dự báo mưa lớn trên diện rộng, Cramer et al. (2017) đã đánh giá bảy phương pháp phổ biến trong ML để dự báo lượng mưa của 42 thành phố dựa vào dữ liệu trong quá khứ. Các thực nghiệm cho thấy các phương pháp ML cho kết quả vượt trội so với phương pháp truyền thống. Những kỹ thuật ML hiện đại ứng dụng trong xử lý hình ảnh dựa trên học sâu (Deep Learning) (Shi X. et al., 2015) đã được áp dụng và cho kết quả vượt trội so với giải thuật dự báo tốt nhất được ứng dụng trong thực tế trong dự báo lượng mưa. Giải thuật này cũng đề xuất một cấu trúc hiệu quả cho mạng thần kinh nhân tạo (ANN) để tận dụng mối liên hệ không gian-thời gian trong việc dự báo. Các phương pháp ML được ứng dụng chủ yếu là phương pháp học truyền thống như mạng Nơ-ron, máy véc tơ hỗ trợ Support Vector Machines, và học sâu (DL). Trong đó mạng Nơ-ron nhân tạo là mô hình AI/ML được áp dụng nhiều nhất. Các kỹ thuật của ML/AI được áp dụng nhiều nhất trong dự báo hiện tượng mưa lớn diện rộng dựa trên dự báo lượng mưa. Trong đó, mô hình chủ yếu được sử dụng là các mạng Nơ-ron nhân tạo (đa lớp, hướng tiến, truy hồi) kết hợp với một số mô hình thống kê (như ARIMA) cho các tập dữ liệu của các vùng địa lý khác nhau, khoảng chia dữ liệu khác nhau (dữ

liệu theo ngày, tuần, tháng, năm), với các hàm lỗi khác nhau. Việc sử dụng mạng Nơ-ron nhân tạo cho kết quả dự đoán tương đối chính xác, nhưng chưa đưa ra được sai số (và xác suất cho sai số) hay độ tin cậy của dự báo. Hơn nữa mạng Nơ-ron nhân tạo có tính chất hộp đen (black-box) do đó khả năng giải thích cho mô hình dự báo không cao. Chính vì vậy, một hướng nghiên cứu khác của bài toán dự đoán lượng mưa và hiện tượng mưa lớn diện rộng là áp dụng các phương pháp học không giám sát (unsupervised) cho dữ liệu không nhãn nhằm giúp phân cụm (clustering) và giảm chiều dữ liệu (dimensionality reduction) để giúp hiểu được các đặc trưng quan trọng của dữ liệu có ảnh hưởng chính tới việc dự báo chính xác lượng mưa.

Đối với hiện tượng thời tiết bất thường rét đậm, rét hại, hiện chưa có nhiều nghiên cứu liên quan đến việc dùng ML/AI cho việc dự báo (dù là hạn ngắn). Các nghiên cứu ứng dụng AI/ML mới chỉ dừng ở việc dự báo nhiệt độ và khoảng thay đổi nhiệt độ nhưng độ chính xác chưa cao.

Trên thế giới đã có những nghiên cứu áp dụng AI/ML vào công tác dự đoán, dự báo các hiện tượng thời tiết nguy hiểm như bão, nước biển dâng do bão, lũ (storm surge, storm tide). Các cách tiếp cận ban đầu thường dựa trên kết hợp các phương pháp học máy thông dụng (mạng Nơ-ron, cây quyết định mờ, học dựa vào cá thể intance-based learning) kết hợp với các mô hình hệ thống động động lực hay mô hình thống kê cho chuỗi thời gian (như lọc Kalman). Đặc biệt gần đây, mạng Nơ-ron được dùng như công cụ học máy phổ biến để xử lý dữ liệu và dự báo, dự đoán các hiện tượng bất thường bão, lũ, nước biển dâng cho bão.

Như vậy, sự phát triển của công nghệ tính toán và các giải thuật AI, học máy hiện đại đang tạo ra những cơ hội lớn để phát triển các giải pháp mới để giải quyết các vấn đề trong lĩnh vực KTTV, đặc biệt trong bài toán hỗ trợ dự đoán, dự báo các hiện tượng khí tượng bất thường dựa trên dữ liệu KTTV. Các mô hình AI/ML được áp dụng rất đa dạng từ các mô hình học máy cơ bản đến các mô hình hiện đại như học học sâu (Deep Learning). Trong đó, mô hình mạng Nơ-ron được dùng phổ biến nhất (kể cả trong các mô hình học sâu). Các mô hình học máy được đưa ra để giải quyết nhiều bài toán và xử lý dữ liệu như cải thiện chất lượng dữ liệu, lựa chọn đặc trưng, giảm chiều, phân cụm, …. Tuy nhiên, vẫn còn có một số vấn đề còn tồn tại với các mô hình AI/ML khi áp dụng cho xử lý dữ liệu KTTV để dự báo các hiện tượng thời tiết bất thường bao gồm: (i) Chất lượng, độ chính xác của tiền xử lý dữ liệu vẫn cần được cải thiện; (ii) Việc cấu hình cho các máy học (như mạng Nơ-ron) còn mang tính định tính và heuristics, cần có những phương pháp hỗ trợ xác định cấu hình các hệ thống AI/ML một cách tự động để hỗ trợ người thiết kế mô hình; (iii) Thời gian huấn luyện của các mô hình học máy còn lớn; (iv) Chất lượng, độ chính

xác của dự đoán, dự báo vẫn cần được cải thiện; (v) Khả năng giải thích của mô hình AI/ML chưa cao do phần lớn dựa trên mô hình hộp đen (black-box); (vi) Hầu như chưa có các đánh giá độ tin cậy cho các dự đoán, dự báo được đưa ra bởi các mô hình AI/ML; (vii) Việc kết hợp giữa các mô hình AI/ML với các mô hình truyền thống cần được tiếp tục nghiên cứu và cải thiện.

Vì vậy, một trong các nhiệm vụ của đề tài này sẽ hướng tới giải quyết các vấn đề còn tồn tại nêu trên khi áp dụng mô hình AI/ML để hỗ trợ dự báo các hiện tượng KTTV nguy hiểm như bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão dựa trên việc phân tích, xử lý dữ liệu KTTV.

Tại Việt Nam, kể từ năm 2000, khi mô hình dự báo thời tiết số đầu tiên HRM bắt đầu được đưa vào chạy dự báo tại Việt Nam, cho đến nay có nhiều mô hình khác nhau đang được chạy nghiên cứu hay dự báo thử nghiệm tại Việt Nam như HRM, Eta, RAMS, HRM, MM5, WRF, …. Tuy nhiên, tại các đơn vị sử dụng mô hình dự báo, sản phẩm dự báo cuối cùng vẫn là sản phẩm dự báo trực tiếp từ mô hình, chưa có bất kỳ hiệu chỉnh nào.

Cho đến nay, tập số liệu bề mặt (địa hình, thảm phủ thực vật, sử dụng đất) sử dụng trong tất cả các mô hình hiện có tại Việt Nam đều được lấy từ tập số liệu toàn cầu của Mỹ. Tập số liệu tương ứng, chi tiết hơn trên lãnh thổ Việt Nam do Việt Nam thực hiện, có cập nhật theo từng tháng hay từng năm chưa được tích hợp vào trong bất cứ mô hình nào. Tập số liệu của Mỹ chủ yếu dựa trên các quan trắc vệ tinh trong thập kỷ 90 và ít được cập nhật. Với tốc độ phát triển kinh tế xã hội của Việt Nam, tập số liệu này đã không còn mô tả đúng một số khu vực trên lãnh thổ Việt Nam khi có những biến động lớn trong khu dân cư tại các đô thị hay vấn đề sử dụng đất. Do đó, sai số hệ thống tồn tại trong sản phẩm dự báo của mô hình là không thể tránh khỏi. Bởi vậy, một dự báo thống kê như MOS trở nên rất cần thiết nhằm tăng cường chất lượng dự báo, loại bỏ các sai số hệ thống từ mô hình cũng như điều kiện ban đầu. Bên cạnh đó, đòi hỏi ngày càng cao của xã hội trong chất lượng dự báo hạn ngắn cũng thúc đẩy xây dựng dự báo sau mô hình.

Đã có nhiều phương pháp được triển khai thực hiện đề tăng cường chất lượng dự báo sau mô hình như: Võ Văn Hòa và cộng sự (2007) đã thử nghiệm lọc Kalman vào hiệu chỉnh dự báo từ mô hình; Bùi Minh Tăng và cộng sự (2009) đã thử nghiệm dự báo định lượng mưa cho hạn dựa báo 24 giờ dựa trên một số phương pháp MOS truyền thống như hồi quy tuyến tính đa biến (MLR), phân tích riêng biệt (MDA), mạng thần kinh nhân tạo (ANN) và hồi quy logistic (LR) từ sản phẩm của mô hình HRM và GSM. Tuy nhiên, các phương pháp được sử dụng trong nghiên cứu này vẫn bộc lộ những hạn chế của phương pháp MOS truyền thống; Đỗ Lệ Thủy và cộng sự

(2009) đã nghiên cứu ứng dụng các phương pháp UMOS và KF để thử nghiệm dự báo cho một số yếu tố khí tượng bề mặt như nhiệt độ, gió, độ ẩm cho đến hạn dự báo 72 giờ cho từng 6 giờ một. Chất lượng dự báo các yếu tố này đã được tăng lên đáng kể sau khi thực hiện hiệu chỉnh thống kê sau mô hình UMOS và KF trên các sản phẩm mô hình. Tuy nhiên, lượng mưa lại chưa được đề cập đến trong nghiên cứu theo hướng thống kê sau mô hình có cập nhật này. Mặc dù đã có rất nhiều nghiên cứu trong nước áp dụng các phương pháp thống kê cổ điển và hiện đại để nâng cao chất lượng dự báo các yếu tố khí tượng. Tuy nhiên, cho đến nay chưa có nghiên cứu nào về ứng dụng AI cho bài toán dự báo khí tượng ở Việt Nam.

Về dự báo nước dâng do bão, có 3 hướng chủ yếu được thực hiện, đó là tự xây dựng mô hình riêng; nghiên cứu phát triển mô hình mã nguồn mở và sử dụng mô hình thương mại từ nước ngoài. Tổng quan về nghiên cứu nước dâng bão tại Việt Nam cho thấy, hướng nghiên cứu chính tập chung vào mô hình số trị, nghiên cứu sử dụng AI trong dự báo nước biển dâng do bão chưa được thực hiện.

Hiện nay, các tiếp cận theo hướng dữ liệu dựa trên thống kê, học máy, trí tuệ nhân tạo đang ngày càng trở nên phổ biến và chứng mình tính hiệu quả trên thế giới. Tuy nhiên về mặt phương pháp luận vẫn còn một số vấn đề cần phải được tiếp tục nghiên cứu, cải thiện (như chất lượng dự đoán, dự báo, tính giải thích, …). Thực tiễn, tại Việt nam cũng chưa có nhiều nghiên cứu theo hướng AI/ML, khoa học dữ liệu cho các bài toán dự đoán, dự báo các hiện tượng thời tiết bất thường dựa trên dữ liệu KTTV. Trên cơ sở đó, nhóm thực hiện đề ra mục tiêu của đề tài gồm các pha nghiên cứu cơ sở khoa học và các giải pháp kỹ thuật trong xử lý dữ liệu để tiến hành dự đoán, dự báo các hiện tượng KTTV nguy hiểm.

Các phân tích trên cho thấy việc nghiên cứu phát triển các mô hình thống kê, học máy, trí tuệ nhân tạo cụ thể cho việc xử lý dữ liệu khí tượng, thủy văn và hải văn trong các hiện tượng khí tượng bất thường nguy hiểm hay xảy ra có tại Việt nam như bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão là nhiệm vụ ***cần thiết và cấp bách***. Sản phẩm của đề tài sẽ được xây dựng thành các công cụ phần mềm để hỗ trợ cho các chuyên gia trong công tác dự đoán, dự báo các hiện tượng khí hậu bất thường nói trên, và được tiến hành triển khai thử nghiệm để đánh giá hiệu quả trong thực tế.

2.Mục tiêu nghiên cứu

- Xác định và đưa ra được cơ sở khoa học và giải pháp ứng dụng của trí tuệ nhân tạo để nhận dạng và dự báo một số hiện tượng KTTV nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam

- Đề xuất và ứng dụng được một số mô hình trí tuệ nhân tạo để nhận dạng và dự báo một số hiện tượng KTTV nguy hiểm gồm bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão.

- Xây dựng được hệ thống nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng KTTV nguy hiểm dựa trên mô hình trí tuệ nhân tạo phù hợp và bước đầu triển khai thử nghiệm trong dự báo nghiệp vụ.

3.Nội dung nghiên cứu

- Nội dung 1: Nghiên cứu tổng quan và đánh giá hiện trạng.

- Nội dung 2: Điều tra, khảo sát, thu thập, số hóa, biên tập, tính toán, thống kê, phân tích, đánh giá, tổng hợp dữ liệu KTTV và tài liệu liên quan về bão, mưa lớn diện rộng, không khí lạnh, lũ (hệ thống sông Hồng), nước biển dâng do bão (ven biển Bắc Bộ, Bắc Trung Bộ).

- Nội dung 3: Nghiên cứu giải pháp quản lý, lưu trữ và phân tích dữ liệu theo mô hình cơ sở dữ liệu lớn Big Data.

- Nội dung 4: Nghiên cứu xác định cơ sở khoa học và thực tiễn của việc sử dụng trí tuệ nhân tạo trong việc nhận dạng, hỗ trợ dự báo, cảnh báo một số hiện tượng khí tượng, hải văn nguy hiểm tại Việt Nam.

- Nội dung 5: Nghiên cứu, phát triển thử nghiệm và đánh giá các phương pháp học máy, nhận dạng, hỗ trợ dự báo, cảnh báo một số hiện tượng KTTV nguy hiểm (bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão).

- Nội dung 6: Ứng dụng trí tuệ nhân tạo trong việc nhận dạng hình thế và hỗ trợ dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh (miền Bắc).

- Nội dung 7: Ứng dụng trí tuệ nhân tạo trong việc nhận dạng hình thế và hỗ trợ dự báo, cảnh báo lũ (trên hệ thống sông Hồng).

- Nội dung 8: Ứng dụng trí tuệ nhân tạo trong việc nhận dạng và hỗ trợ cảnh báo, dự báo nước biển dâng do bão (ven biển Bắc Bộ và Bắc Trung Bộ).

- Nội dung 9: Xây dựng một framework có tính mở để tích hợp và kết nối được với các mô-đun nhận dạng, hỗ trợ dự báo, cảnh báo các hiện tượng bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão thuộc khu vực nghiên cứu.

- Nội dung 10: Nghiên cứu xây dựng quy trình, chuyển giao và đào tạo sản phẩm. Thử nghiệm, đánh giá khả năng ứng dụng của công nghệ trong thực tiễn.

- Nội dung 11: Viết báo cáo tổng kết và báo cáo tóm tắt đề tài.

4.Thời gian và kinh phí thực hiện

- Thời gian thực hiện: 30 tháng (Từ tháng 07/2018 đến tháng 12/2020);
- Kinh phí thực hiện: 7.800 triệu đồng (Bảy tỷ tám trăm triệu đồng).

5.Ý nghĩa khoa học và thực tiễn của kết quả nghiên cứu

a) Ý nghĩa khoa học:

- Cung cấp cơ sở khoa học và thực tiễn của việc ứng dụng trí tuệ nhân tạo để nhận dạng và dự báo một số hiện tượng KTTV nguy hiểm gồm bão, mưa lớn diện rộng, không khí lạnh, lũ (trên hệ thống sông Hồng), nước biển dâng do bão (ven biển Bắc Bộ và Bắc Trung Bộ);
- Cung cấp cơ sở khoa học vững chắc để tiếp tục mở rộng các hướng nghiên cứu mới trong việc ứng dụng trí tuệ nhân tạo trong dự báo, cảnh báo KTTV;
- Cung cấp cơ sở khoa học vững chắc để đề xuất được các giải pháp giám sát và ứng phó với bão, không khí lạnh, mưa lớn diện rộng, lũ và nước biển dâng do bão trong một số lĩnh vực kinh tế - xã hội trong bối cảnh biến đổi khí hậu.

b) Ý nghĩa thực tiễn:

- Nâng cao trình độ khoa học công nghệ cho đội ngũ cán bộ tham gia đề tài;
- Phát triển công nghệ dự báo hiện đại cho ngành KTTV, nhất là trong dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh, lũ và nước biển dâng do bão góp phần đáp ứng được yêu cầu của Luật Phòng, Chống thiên tai và Luật KTTV;
- Phát triển công cụ hỗ trợ hiện đại để sử dụng trong đào tạo và nghiên cứu trong trường đại học và sau đại học;
- Tăng cường năng lực phòng, chống thiên tai nhất là đối với các hiện tượng KTTV nguy hiểm như bão, mưa lớn diện rộng, không khí lạnh, lũ và nước biển dâng do bão;
- Đẩy mạnh ứng dụng trí tuệ nhân tạo trong công tác dự báo, cảnh báo nghiệp vụ KTTV;
- Hỗ trợ sự phát triển của nhiều ngành/ lĩnh vực và các thành phần kinh tế xã hội trên khu vực phía Bắc do giảm thiểu rủi ro gây ra bởi bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão trong bối cảnh biến đổi khí hậu.

6.Kết cấu báo cáo

Ngoài phần Mở đầu, Kết luận, tài liệu tham khảo, kết cấu Báo cáo tổng kết của đề tài bao gồm 5 chương, cụ thể: Chương 1: Tổng quan về vấn đề nghiên cứu;

Chương 2: Phạm vi, đối tượng, phương pháp và số liệu nghiên cứu; Chương 3: Phát triển công cụ, phương pháp học máy, AI và xây dựng Big data KTTV; Chương 4: Xây dựng và triển khai hệ thống AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm; Chương 5: Chuyển giao, đào tạo, thử nghiệm và đánh giá.

Các kết quả đề tài nhận được có sự góp sức tích cực của tập thể cán bộ tham gia đề tài trong suốt quá trình hơn 2 năm thực hiện. Tham gia thực hiện đề tài, ngoài các thành viên chính còn có đội ngũ các nhà khoa học thuộc Trung tâm Thông tin dữ liệu KTTV, Trung tâm Dự báo KTTV quốc gia, Trung tâm Ứng dụng công nghệ KTTV, Đại học Bách khoa Hà Nội, Công ty cổ phần công nghệ VDSpaces và các Đài KTTV khu vực: Đồng bằng Bắc Bộ, Đông Bắc, Bắc Trung Bộ là các đơn vị tiếp nhận, thử nghiệm sản phẩm của đề tài. Đóng góp hết sức quan trọng vào sự thành công của đề tài là sự hỗ trợ, giúp đỡ tận tình, tạo mọi điều kiện thuận lợi trong quá trình thực hiện đề tài của Bộ Tài nguyên và Môi trường, Văn phòng Chương trình Khoa học công nghệ phục vụ mục tiêu quốc gia ứng phó với BĐKH, Tổng cục KTTV, Trung tâm Thông tin và Dữ liệu KTTV và các đơn vị trực thuộc Tổng cục KTTV.

Tập thể tác giả xin chân thành cám ơn Bộ Tài nguyên và Môi trường, Ban chủ nhiệm và Văn phòng Chương trình Khoa học công nghệ phục vụ mục tiêu Quốc gia ứng phó với BĐKH, Tổng cục KTTV và các nhà khoa học đã chỉ đạo và giúp đỡ trong quá trình thực hiện đề tài.

1. CHƯƠNG 1: TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU

1.1. Tổng quan về trí tuệ nhân tạo (AI) và dữ liệu lớn (Big data)

1.1.1. Định nghĩa và các lĩnh vực liên quan đến AI

1.1.1.1. Định nghĩa AI

Trong khoa học máy tính, trí tuệ nhân tạo hay AI (Artificial Intelligence), đôi khi được gọi là trí thông minh nhân tạo, là trí thông minh được thể hiện bằng máy móc, trái ngược với trí thông minh tự nhiên của con người. Thông thường, thuật ngữ "trí tuệ nhân tạo" thường được sử dụng để mô tả các máy móc (hoặc máy tính) có khả năng bắt chước các chức năng "nhận thức" mà con người thường phải liên kết với tâm trí, như "học tập" và "giải quyết vấn đề"[1].

1.1.1.2. Các lĩnh vực liên quan đến AI

a) Các bài toán điển hình áp dụng các phương pháp AI

Bao gồm: (i) Nhận dạng mẫu: Nhận dạng chữ cái quang học; nhận dạng chữ viết tay; nhận dạng tiếng nói; nhận dạng khuôn mặt; (ii) Xử lý ngôn ngữ tự nhiên, dịch tự động (dịch máy) và Chatterbot; (iii) Điều khiển phi tuyến và Robotics; (iv) Computer vision, Thực tại ảo và Xử lý ảnh; (v) Lý thuyết trò chơi và Lập kế hoạch (Strategic planning); (vi) Trò chơi trí tuệ nhân tạo và computer game bot[1].

b) Các lĩnh vực khác áp dụng các phương pháp AI

Bao gồm: Tự động hóa; Bio-inspired computing; Điều khiển học; Hệ thống thông minh lai; Agent thông minh; Điều khiển thông minh; Suy diễn tự động; Khai phá dữ liệu; rô-bốt nhận thức (Cognitive robotics); rô-bốt phát triển (developmental robotics); rô-bốt tiến hóa (evolutionary robotics); Chatbot[1].

1.1.2. Định nghĩa, đặc trưng và các vấn đề liên quan đến Big data

1.1.2.1. Định nghĩa Big data

Big data là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Big data bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan, truy vấn và tính riêng tư. [2].

1.1.2.2. Các đặc trưng của Big data

Khối lượng dữ liệu (Volume): Đây là đặc điểm tiêu biểu nhất của dữ liệu lớn, khối lượng dữ liệu rất lớn. Công nghệ "đám mây" sử dụng để lưu trữ Big data[2].

Tốc độ (Velocity): Tốc độ có thể hiểu theo 2 khía cạnh: (a) Khối lượng dữ liệu gia tăng rất nhanh; (b) Xử lý dữ liệu nhanh ở mức thời gian thực (real-time), có nghĩa dữ liệu được xử lý ngay tức thời ngay sau khi chúng phát sinh (mili giây)[2].

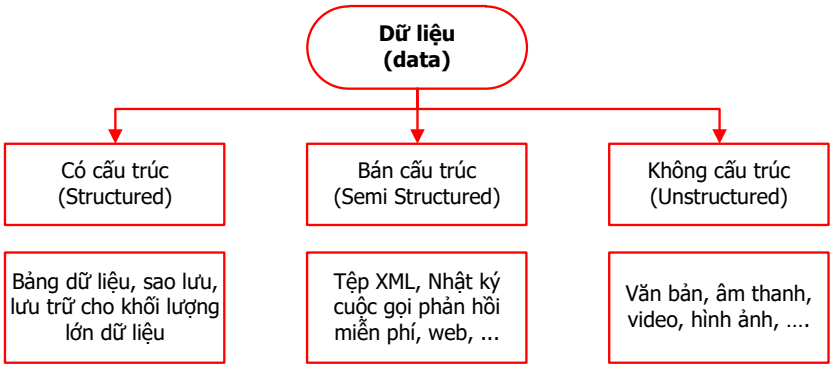
Đa dạng (Variety): Big data cho phép liên kết, phân tích nhiều dạng dữ liệu có cấu trúc và phi cấu trúc (tài liệu, blog, hình ảnh, …)[2].

Độ tin cậy/chính xác (Veracity): Một trong những tính chất phức tạp nhất của dữ liệu lớn là độ tin cậy/chính xác của dữ liệu. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của Big data[2].

Giá trị (Value): Giá trị là đặc điểm quan trọng nhất của dữ liệu lớn [2].

1.1.2.3. Kiểu dữ liệu của Big data

Các loại dữ liệu trong big data gồm: (i) Dữ liệu có cấu trúc (Data Structured); (ii) Dữ liệu bán cấu trúc (Data Semi Structured); (iii) Dữ liệu không cấu trúc (Data Unstructured)[20].



Hình 1.1: Các loại dữ liệu trong Big data

1.1.2.4. Sự khác biệt giữa Big data và dữ liệu truyền thống

Dữ liệu lớn khác với dữ liệu truyền thống (ví dụ, kho dữ liệu - Data Warehouse) ở một số điểm cơ bản: Dữ liệu đa dạng hơn; lưu trữ dữ liệu lớn hơn; truy vấn dữ liệu nhanh hơn; độ chính xác cao hơn [20].

1.2. Đánh giá tổng quan tình hình nghiên cứu ở ngoài nước

1.2.1. Ứng dụng AI trong dự báo thời tiết tại Mỹ

- **Sản phẩm mô hình dự báo Deep Thunder của IBM sử dụng công nghệ AI và Big data**: Tập đoàn công nghệ máy tính IBM đã mua Công ty Thời tiết The Werther Company và kết hợp số liệu có sẵn của IBM vào hệ thống AI của The Werther Company. Với sự kết hợp này, IBM đã cho ra đời sản phẩm Deep Thunder, một mô hình dự báo thời tiết tiên tiến sẽ sử dụng công nghệ máy học để dự báo chính xác tác động của thời tiết đối với các doanh nghiệp. Mô hình này cũng sẽ sử dụng dữ liệu thời tiết lịch sử để cải thiện dự báo thời tiết. Weather Company phân tích khoảng

100 terabyte dữ liệu hàng ngày đến từ các nguồn khác nhau bao gồm nguồn của bên thứ ba, dữ liệu vị trí và thông qua mạng lưới hơn 195.000 trạm thời tiết cá nhân của Weather Underground[110]. Tích hợp sử dụng dữ liệu của Earth Networks, công ty có trụ sở tại Maryland, vận hành 12.000 trạm thời tiết tại hơn 90 quốc gia, cùng với 1.800 cảm biến ở 50 quốc gia cho mạng lưới định vị sét; Total Lightning Network có thể phát hiện sét trong đám mây để tạo ra các cảnh báo bão liên quan đến thời tiết khắc nghiệt như lốc xoáy và mưa đá. Công ty Thời tiết của IBM có quyền truy cập vào hệ thống cảm biến quan trắc và mô hình, một lượng lớn dữ liệu thời tiết đưa vào nền tảng AI của IBM Watson để cải thiện dự đoán. Tính đến năm 2016, hàng ngày, các mô hình AI của IBM đã sử dụng hơn 100 terabyte dữ liệu. TempoQuest có trụ sở tại Colorado tập trung vào việc thu thập dữ liệu thời tiết được thu thập bởi các vệ tinh, máy bay không người lái và radar. Sau đó, nó sử dụng phần mềm chạy trên các đơn vị xử lý đồ họa NVIDIA (GPU) cực mạnh có thể tạo ra dự báo thời tiết có độ phân giải cao trong vài phút, kết quả là một dự báo chính xác và nhanh chóng được sử dụng cho các doanh nghiệp vận tải và các hãng hàng không. Deep Thunder có khả năng cung cấp bản tin dự báo thời tiết với thời hạn dự báo từ 24 đến 48 giờ với độ phân giải cực cao (500 m đến 1 km). Dữ liệu thời tiết có thể được kết hợp với phân tích và hình ảnh hóa phù hợp với nhu cầu kinh doanh cá nhân.

- **Cơ quan Khí quyển và Đại dương Quốc gia Hoa Kỳ (NOAA)** gần đây đã sử dụng máy học nhiều hơn để cải thiện dự báo của họ. Một nhóm các nhà nghiên cứu từ NOAA phát hiện ra rằng "việc áp dụng các kỹ thuật AI cùng với hiểu biết vật lý về môi trường có thể cải thiện đáng kể kỹ năng dự đoán đối với nhiều loại thời tiết có tác động mạnh". Thời tiết có tác động mạnh bao gồm các sự kiện như giông bão nghiêm trọng, lốc xoáy và bão. Bài báo của họ kết luận rằng những cải tiến này có các ứng dụng thương mại rõ ràng nêu rõ, "Các phương pháp AI mở rộng dễ dàng dự đoán trực tiếp các tác động của thời tiết có tác động mạnh, chẳng hạn như năng lượng được tạo ra bởi các nguồn biến đổi như năng lượng mặt trời hoặc gió, tiêu thụ năng lượng trong một khu vực hoặc công suất đến sân bay" [111].

Monsanto - Climate Corporation: Năm 2013, Monsanto đã mua lại Climate Corporation với giá chỉ dưới 1 tỷ USD. Trong số các dịch vụ mà Climate Corporation cung cấp, một dịch vụ nếu trọng tâm chính của nó là thông tin dự báo thời tiết siêu địa phương cho nông dân. Nó sử dụng nhiều nguồn khác nhau và máy học để tối ưu hóa các dự đoán thời tiết dành riêng cho nông nghiệp [111]. Để giúp tăng cường khả năng dự báo và phân tích, Monsanto đã mua lại HydroBio , một công ty phần mềm nông nghiệp có trụ sở tại Denver, CO. HydroBio sử dụng kết hợp hình ảnh vệ tinh, dữ liệu đất và dữ liệu thời tiết địa phương để cung cấp thông tin chi tiết về tưới tiêu cho nông dân. Nó giúp người nông dân biết rõ hơn khi nào họ nên hay

không nên tưới. Điều này ngăn các trang trại lãng phí nguồn nước hạn chế và có thể cải thiện năng suất. Những công cụ này cuối cùng sẽ được tích hợp vào Climate FieldView của Climate Corporation [111].

- **Hội Khí tượng học Mỹ (AMS): AI giúp cải thiện chất lượng dự báo thời tiết:** AI có thể giúp đưa dự báo chính xác hơn thành hiện thực. Dự báo thời tiết là một trường hợp sử dụng lý tưởng của AI, bởi vì có một kho tàng dữ liệu thời tiết lịch sử và hiện tại, có thể cung cấp dữ liệu vào thuật toán theo nghĩa đen, yêu cầu cả chất lượng và số lượng khi giao phối các lần xuất hiện trong quá khứ với các dự đoán trong tương lai. Trong khi không ai có thể dự đoán đầy đủ tương lai, các kỹ thuật AI áp dụng việc học tăng cường cho các dự đoán trước đây và các kết quả thực tế. Bằng cách so sánh các dự đoán với kết quả quan trắc, mô hình có thể tìm hiểu và cải thiện khả năng mô phỏng, dự báo nhiều hơn trong tương lai và với độ chính xác cao hơn. Để hỗ trợ dự báo thời tiết, dữ liệu được đưa vào một thuật toán sử dụng các kỹ thuật học tập sâu để tìm hiểu và đưa ra các dự đoán dựa trên dữ liệu trong quá khứ. Các mô hình thời tiết được tạo thành từ một số điểm dữ liệu phức tạp, khiến dự báo thời tiết trở thành một bài tập chuyên sâu về dữ liệu và tính toán cao. Điều này cho thấy việc sử dụng sức mạnh tính toán đáng kể, cũng như rất nhiều dung lượng lưu trữ để giữ lại dữ liệu. Tuy nhiên, sức mạnh của AI và hệ thống học tập sâu có thể được sử dụng để nhập và phân tích các bộ dữ liệu đa yếu tố này nhanh hơn và chính xác hơn bao giờ hết. Hội Khí tượng học Mỹ (AMS) đã chỉ ra rằng việc sử dụng kỹ thuật AI hiện đại đang cải thiện khả năng sàng lọc qua lượng dữ liệu thời tiết để trích xuất những hiểu biết chính xác và hướng dẫn kịp thời và những người ra quyết định. Lợi ích chính là AI có thể cung cấp các mô hình linh hoạt và mạnh mẽ hơn có khả năng xác định các mối quan hệ phức tạp giữa một số lượng lớn các tính năng thời tiết được mô hình hoá và quan sát được[14].

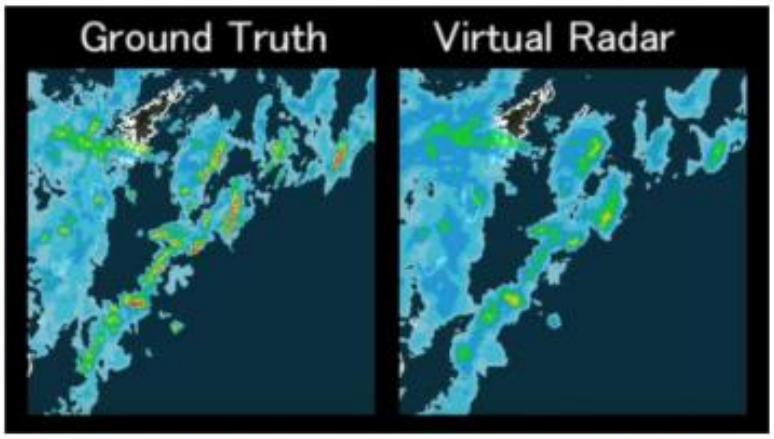
- ***Tractica: Mở rộng nguồn dữ liệu trong dự báo thời tiết:*** Lý do khiến AI cần ứng dụng nhiều trong dự báo thời tiết là dữ liệu KTTV. Hơn 1.000 vệ tinh tập trung vào thời tiết hiện đang bay trên quỹ đạo, theo dõi và gửi dữ liệu về các mẫu mây, gió, nhiệt độ và hệ thống thời tiết. Hàng trăm ngàn trạm mặt đất đang liên tục thu thập dữ liệu thời gian thực. Công nghệ Internet of Things (IoT), các nguồn dữ liệu mới đang trực tuyến mỗi ngày (đèn giao thông có internet, tấm pin mặt trời, nhiệt kế và điều hòa không khí thông minh, …) có thể cấp dữ liệu chi tiết hơn vào thuật toán dự báo thời tiết. Tractica dự báo công nghệ AI được sử dụng trong dự báo thời tiết sẽ đạt 553,4 triệu USD mỗi năm cho chi tiêu cho phần cứng, phần mềm và dịch vụ vào năm 2025, tăng từ 31,7 triệu USD vào năm 2018. Thúc đẩy thị trường là các công ty lớn đã đầu tư vào các hệ thống dự đoán thời tiết công cộng và dữ liệu thời tiết tư nhân[14].

1.2.2. Ứng dụng mạng ANN để dự báo nước biển dâng do bão tại Nhật Bản

Tại Nhật Bản, GS. Sooyoul Kim các công sự đã sử dụng mạng ANN để dự báo nước biển dâng do bão. Dữ liệu thử nghiệm là dữ liệu quan trắc tại các trạm Hamada, Matsue, Yonago, Ama và Saigo thuộc bờ biển Tottori (Nhật Bản) áp dụng cho các cơn bão Maemi 2003, Songda 2004, Megi 2004. Mô hình mạng ANN đã đưa ra được kết quả dự báo nước biển dâng. Hạn chế của phương pháp ANN của Kim, Sooyoul là đề xuất theo kinh nghiệm 14 tập con từ 10 thuộc tính để chọn ra tập con thuộc tính tốt nhất để dự báo nước biển dâng. Tuy vậy với 10 thuộc tính chúng ta có 2^{10} tập con. Chưa có phương pháp tự động chọn tập con các thuộc tính tối ưu để dự báo nước biển dâng. *Theo Kim, Sooyoul, et al. "A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori Coast, Japan"*[109].

1.2.3. Ứng dụng AI để giám sát mưa của Weathernews Nhật Bản

Weathernews cộng tác với NVIDIA trong dự án ứng dụng AI để phát triển khả năng giám sát thời gian thực độ chính xác cao và dự báo phân phối lượng mưa toàn cầu. Dự án sử dụng công nghệ Deep Learning để giảm thiểu thiệt hại tại các khu vực có thiên tai mưa lớn thường xuyên xảy ra như Đông Nam Á[15]. Ở những vùng thường xuyên xảy ra thiên tai mưa lũ, khả năng nắm bắt điều kiện thời tiết và dự báo chi tiết là rất quan trọng. Trong khi đó, công nghệ dự báo KTTV tiên tiến và cơ sở hạ tầng, nhân lực để vận hành vẫn đang ở giai đoạn phát triển và gặp khó khăn về tài chính và tiến độ. Hơn nữa, công nghệ dự báo dựa trên các mô hình vật lý hiện tại đang tiến gần tới các giới hạn của nó. Do đó, Weathernews đang thực hiện một phương pháp tiếp cận mới là dựa trên AI, sử dụng công nghệ Deep Learning tiên tiến từ NVIDIA để phân tích và dự đoán phân phối lượng mưa. Bằng cách này, sẽ giúp giảm thiểu thiệt hại do mưa lớn bằng cách cung cấp thông tin dự báo kết hợp với số liệu quan trắc của các trạm thời tiết địa phương. Hình 1.2 cho thấy, ảnh bên trái là những đám mây mưa được quan sát bởi radar thời tiết hiện tại, và hình ảnh bên phải là một hình ảnh radar đám mây mưa ảo được tạo ra từ một ảnh vệ tinh sử dụng công nghệ Deep Learning. Điều này cho phép khả năng dự báo chính xác hơn.



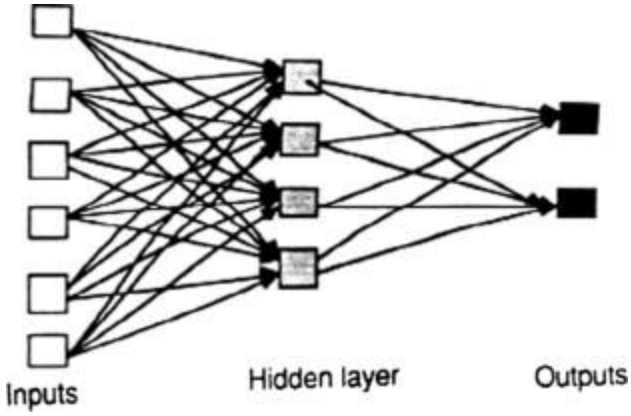
Hình 1.2: Ảnh đám mây mưa quan trắc radar và đám mây mưa ảo tạo bằng DL

Weathernews sử dụng các siêu máy tính AI hàng đầu thế giới của NVIDIA để vận hành công nghệ Deep Learning. Trong dự án này, các ảnh vệ tinh có độ chính xác cao tập trung xung quanh Nhật Bản cùng với các ảnh phản chiếu rađa, ảnh radar sẽ được sử dụng làm dữ liệu huấn luyện để tạo ra các hình ảnh giống mô phỏng radar dựa trên ảnh vệ tinh để hiển thị và dự báo điều kiện lượng mưa. Nói cách khác, ngay cả đối với những khu vực mà cơ sở hạ tầng quan sát như radar thời tiết không đầy đủ, hoặc trên đại dương, việc thiết lập hạ tầng quan sát ở đó sẽ không cần thiết. Weathernews dự định bắt đầu bằng cách sử dụng khu vực Đông Nam Á làm khu vực mục tiêu để phân tích, sau đó mở rộng sang các khu vực khác[15]. Một hình ảnh radar ảo được tạo ra từ một hình ảnh vệ tinh (mô phỏng). Các đường màu vàng biểu thị phạm vi nơi có thể bổ sung các ảnh radar thời tiết hiện tại. Dự án này sẽ giúp có thể hiểu được điều kiện lượng mưa ở các khu vực trong khu vực Đông Nam Á nơi trước đây không thể thực hiện được.

Hình 1.3: Ảnh ra đa ảo được tạo ra từ một hình ảnh vệ tinh (mô phỏng)

1.2.4. Ứng dụng mạng ANN để dự báo lũ lụt trên sông Nile, Sudan

Nghiên cứu này sử dụng mô hình ANN để dự báo lũ lụt dọc theo sông Nile. Loại mạng nơ-ron nhân tạo phổ biến nhất bao gồm ba nhóm hoặc lớp đơn vị: (1) một lớp các đơn vị " đầu vào " được kết nối với (2) một lớp các đơn vị " ẩn ", được kết nối với (3) một lớp đơn vị "đầu ra" [112].



Hình 1.4: Mạng ANN phổ biến

Để quan sát hiệu suất mô hình hóa của ANN, các trạm khác nhau trên sông Nile đã được sử dụng trong nghiên cứu này. Tất cả các dữ liệu được Bộ Thủy lợi Sudan thu thập dưới dạng đọc hàng ngày. Do tính liên tục và sẵn có của dữ liệu, 1970-1985 được chọn làm giai đoạn huấn luyện (hiệu chuẩn), trong khi 1986 - 1987 được chọn làm giai đoạn xác minh. Bốn kịch bản đã được kiểm tra (Bảng 1.1). Các trạm khác nhau được sử dụng để cung cấp dữ liệu đầu vào, trong khi Trạm Dongola được sử dụng làm đầu ra. Một mô hình kết nối tiêu chuẩn ba lớp đã được áp dụng cho bốn tình huống. Để đánh giá hiệu suất của các mô hình trong các luồng dự báo, sai số bình phương trung bình gốc (RMSE) đã được sử dụng[112].

Bảng 1.1: Bốn kịch bản đã chọn và kết quả hiệu quả của mô hình

Tình huống	**Các trạm được sử dụng làm đầu vào**	**Sức mạnh của trạm**	**Cấu trúc tấm và chức năng quy mô**	**Hiệu quả mô hình**	
				Thời gian hiệu chuẩn	Thời gian xác minh
Đầu tiên	Tamaniat	0,34672	2,58,1	R 2 = 94,32%	R 2 = 91,59%
	Atbara	0,65328	Tuyến tính (−1, 1), Logistic, Logistic	RMSE = 0,338	RMS = 0,374
Thứ hai	Eddeim	0,23007	3,58,1	R 2 = 94,51%	R 2 = 91,62%
	Tamaniat	0,31869	Tuyến tính (−1,1), Logistic, Logistic	RMS = 0,333	RMS = 0,374

Đối với thảm họa lũ lụt năm 1998 tại Dongola, hiệu suất của mô hình ANN được so sánh với kết quả thực tế từ Sông Nile. Mô hình ANN đã được chạy để dự đoán mức lũ tại Trạm Dongola. Mực nước mặt bằng giá trị dự báo tại trạm cộng với mức 0 của trạm này. Do đó, mực nước mặt tại Dongola bằng với giai đoạn dự đoán sử dụng ANNs + 212,03 (m) - Bảng 1.2.

Bảng 1.2: Giai đoạn dự đoán cho Dongola trong tháng 8 và tháng 9 năm 1998

Ngày	Tamaniat	Atbara	Dongola quan sát	Dự đoán	lỗi	Mức nước
	Máy đo (m)	Máy đo (m)	Máy đo (m)	(m)	(m)	(m)
08/01/1998	14.30	14.02	12,7	1.315211	−0,45211	225,18
08/02/1998	14.42	14,22	12,98	1.333011	−0.35011	225,36
08/03/1998	14,58	14.365	13,32	1.346949	−0.14949	225,5
08/04/1998	14,6	14,51	13,36	1.358692	−0.22692	225,62
08/05/1998	14,69	14,6	13,56	1.367016	−0.11016	225,7

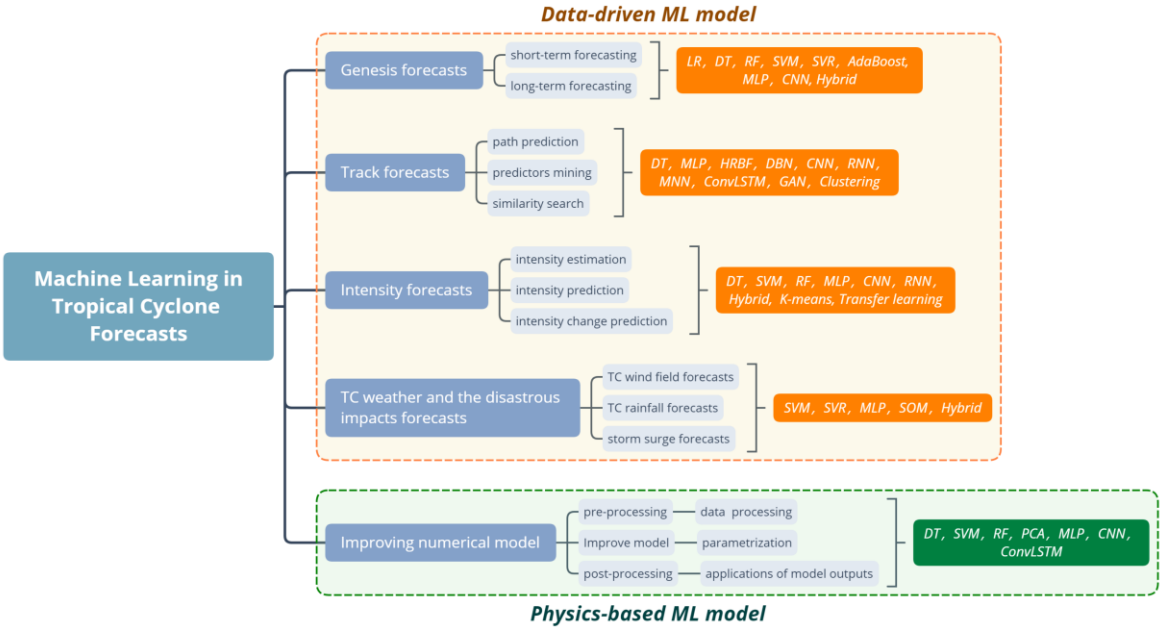
Nghiên cứu này dự đoán mức độ lũ lụt dọc theo Sông Nile bằng cách sử dụng ANN. Phương pháp này có lợi vì chỉ cần một biến, trong khi các mô hình khác yêu cầu nhiều biến để đưa ra dự đoán chính xác. ANN có lợi thế về sự đơn giản khi so sánh với các mô hình phức tạp hơn khác. Do đó, trong các tình huống thiếu hoặc khó thu thập thông tin, phương pháp ANN cung cấp phương án khả thi nhất để dự báo lũ. Mạng nơron (ANN) cung cấp một phương tiện giảm chi phí phân tích thông tin địa hình và thủy văn bằng cách giảm lượng thời gian dành cho việc phân tích dữ liệu.

1.2.5. Ứng dụng mô hình học máy (ML) để dự báo bão ở Trung Quốc

Nghiên cứu này do Rui Chen và các cộng sự thực hiện theo tài trợ bởi Chương trình Nghiên cứu và Phát triển Trọng điểm Quốc gia của Trung Quốc và Quỹ Khoa học Tự nhiên Quốc gia Trung Quốc. Kết quả nghiên cứu ứng dụng mô hình ML để dự báo bão cụ thể như sau:

a) Tổng quan về mô hình ML dự báo bão

Trong nghiên cứu, mô hình ML trong dự báo bão được chia thành năm khía cạnh. Về dự báo nguồn gốc bão, mục tiêu cuối cùng là tạo ra dự báo xác suất về một khu vực cố định trong thời gian thực và dự báo định lượng về thời gian và địa điểm hình thành bão. Tuy nhiên, ở giai đoạn này, ML chỉ có khả năng dự đoán liệu tiền thân có thể phát triển thành bão hay không và tần suất theo mùa của nguồn gốc bão trong mỗi khu vực, tương ứng với nhiệm vụ phân loại và tác vụ hồi quy trong ML. Do đó, các nghiên cứu chủ yếu sử dụng một số thuật toán điển hình, bao gồm DT, hồi quy logistic (LR), SVM và các thuật toán tổng hợp, như AdaBoost và rừng ngẫu nhiên (RF), để dự đoán nguồn gốc bão. Các thuật toán tổng hợp này về mặt lý thuyết tốt hơn một thuật toán đơn lẻ; tuy nhiên, chúng vẫn cần được đánh giá theo từng trường hợp cụ thể. Ngoài ra, các thuật toán học sâu (DL) như perceptron nhiều lớp (MLP) và CNN với các chức năng phức tạp và xử lý dữ liệu hình ảnh, cũng đóng một vai trò quan trọng trong việc cải thiện kỹ thuật dự báo nguồn gốc bão [114].



Hình 1.5: Sơ đồ mô hình ML dự báo bão

Hình 1.5 là sơ đồ tổ chức mô hình học máy dự báo bão. Các thuật toán sử dụng trong mô hình ML gồm: hồi quy logistic (LR), cây quyết định (DT), rừng ngẫu nhiên (RF), máy vectơ hỗ trợ (SVM), hồi quy vectơ hỗ trợ (SVR), perceptron nhiều lớp (MLP), mạng nơron tích hợp (CNN), mạng đối nghịch chung (GAN), mạng nơron tái phát (RNN), mạng chức năng cơ sở xuyên tâm kết hợp (HRBF), bản đồ tự tổ chức (SOM), phân tích thành phần chính (PCA), tích lũy dài hạn mạng bộ nhớ (ConvLSTM) [114].

b) Mô hình ML dự báo nguồn gốc bão

Các kỹ thuật dự báo truyền thống chủ yếu bao gồm các mô hình số và thống kê. Do đó, học máy đã được sử dụng để dự báo nguồn gốc và tất cả các thuật toán ML được thể hiện trong Bảng 1.3.

Bảng 1.3: Máy học trong dự báo nguồn gốc bão (xoáy thuận nhiệt đới)

Nhiệm vụ	Thuật toán	Ý chính
Dự báo ngắn hạn	LR	Chọn các yếu tố dự báo tối ưu và mô hình hóa cho các dự báo nguồn gốc
	DT	Dự đoán bão trong tương lai dựa trên các nhiễu động nhiệt đới
		Phát hiện nguyên nhân của bão bằng cách sử dụng các công cụ dự đoán từ dữ liệu vệ tinh
	RF	Dự đoán sự phát triển của ATNĐ
	AdaBoost	Xác định liệu ATNĐ sẽ phát triển thành bão
	SVM	Dự đoán sự hình thành bão từ dữ liệu ảnh vệ tinh

Nhiệm vụ	Thuật toán	Ý chính
	CNN	Phát hiện bão và tiền thân của chúng dựa trên mô phỏng các mô hình số
Dự báo dài hạn	SVR	Tạo dự báo về hoạt động bão cho mùa sắp tới
		Giảm lỗi dự đoán theo mùa của bão
		Cải thiện độ chính xác của các dự đoán bão theo mùa
	MLP	Cung cấp dự đoán theo mùa về hoạt động của bão
	SOM, FNN	Xác định GPI cho một tập hợp các mô hình khí hậu toàn cầu

Dự báo hạn ngắn: Các tác giả của đã xác định nguồn gốc hình thành bão bằng cách sử dụng các bộ phân loại phi tuyến (cây quyết định (DT), K-láng giềng gần nhất (KNN), MLP, phân tích dữ liệu định tính (QDA), SVM), hoặc bộ phân loại tập hợp phi tuyến (AdaBoost và RF)). Kết quả cho thấy AdaBoost là thuật toán hiệu quả nhất với độ chính xác dự báo là 97,2% cho nguồn gốc bão so với các mô hình thống kê tuyến tính thông thường sử dụng các yếu tố dự báo môi trường kết hợp với MCS / TCs trong thời gian dự đoán là 6 giờ. Độ tin cậy cũng được đảm bảo khi thời gian kéo dài đến 12, 24 hoặc thậm chí 48 giờ. DT, RF và SVM cũng được so sánh về kỹ năng dự đoán của chúng. Qua đó, SVM là thuật toán học máy hiệu quả nhất, hoạt động tốt hơn về kỹ năng dự đoán so với các thuật toán khác, với tỷ lệ truy cập nằm trong khoảng từ 94% đến 96%, cao hơn đáng kể so với hiệu suất LDA (77%). Học sâu có thể được sử dụng để giải quyết vấn đề này [114].

Dự báo hạn dài: Dự báo sự hình thành bão dài hạn nhằm dự đoán số lượng ở các khu vực dễ bị tổn thương cho mùa bão sắp tới. Để có được các dự đoán theo mùa tốt hơn về tần suất bão, các thuật toán học máy cũng phù hợp. Richman và cộng sự đã sử dụng hồi quy vectơ hỗ trợ (SVR) để dự đoán số lượng bão theo mùa. Trong thử nghiệm của họ, SVR có các yếu tố dự báo giống như hồi quy tuyến tính (MLR), nhưng với kết quả cải thiện 40,1% so với MLR. Để cải thiện hơn nữa kết quả, các tác giả cũng kết hợp dao động bán niên hai năm một lần (QBO) vào SVR, dẫn đến cải thiện đáng kể 121% kết quả dự đoán của SVR, trong khi sai số tuyệt đối trung bình (MAE) (0,97) và trung bình gốc sai số bình phương (RMSE) (1,19) được cải thiện lần lượt là 31,2% và 29,2%[114].

c) Mô hình ML theo dõi dự báo

Mô hình thống kê - động, là một trong những kỹ thuật dự báo hiện đại cho các đường đi của bão. Trong những năm gần đây, để tạo ra kết quả nhanh chóng và hiệu quả hơn cho các dự báo đường đi của bão, nhiều nhà nghiên cứu đã cố gắng áp dụng

học máy để xây dựng các mô hình dự đoán mới (Bảng 1.4), và họ chủ yếu tập trung vào việc cải tiến các kỹ thuật dự đoán và lựa chọn các yếu tố dự báo.

Bảng 1.4: Máy học trong dự báo theo dõi bão

Nhiệm vụ	Thuật toán	Ý chính
Dự đoán đường đi của bão	HRBF	Phát triển hệ thống khai thác theo dõi và nhận dạng bão tự động và tích hợp
	MLP	Dự đoán theo dõi lốc xoáy dựa trên MLP với thuật toán BP
	RNN	Đề xuất RNN thưa thớt với cấu trúc liên kết linh hoạt để dự đoán quỹ đạo
		Đề xuất RNN được kết nối đầy đủ để dự đoán quỹ đạo của các XTNĐ
	MNN	Phát triển một mô hình dự đoán lưu giữ thông tin không gian từ các rãnh xoáy
	GAN	Dự đoán theo dõi bão bằng GAN với hình ảnh vệ tinh và dữ liệu khí tượng
	ConvLSTM	Đề xuất mô hình không gian-thời gian dựa trên ConvLSTM để theo dõi và dự đoán quỹ đạo bão
	CNN	Thiết kế mô hình kết hợp dữ liệu đa nguồn để dự đoán các đường dẫn của XTNĐ
Khai thác dự báo	DT	Khám phá các yếu tố dự đoán và các quy tắc chi phối sự đổ bộ và định kỳ của XTNĐ
Tìm kiếm sự giống nhau	DBN	Tìm các XTNĐ tương tự trong lịch sử và tham khảo dữ liệu này để cải thiện dự báo XTNĐ
	Phân cụm	Áp dụng K-mean, c-mean mờ và SOM để phân nhóm các rãnh bão
		Nghiên cứu tính chất bão và các yếu tố quy mô lớn

Dự báo đường đi của bão: Các tác giả sử dụng MLP để dự đoán kinh độ và vĩ độ trong 24 giờ tiếp theo của lốc xoáy bằng cách nhập các số liệu quan trắc trong 12 giờ qua. Kết quả cho thấy RMSE giữa giá trị dự đoán và vĩ độ thực (kinh độ) là 1,01 (1,16) độ và hệ số tương quan của giá trị dự đoán là 0,98 (0,99). So với mô hình Quasi Lagrangian (QLM), MLP đạt được hiệu quả chính xác hơn. Tuy nhiên, độ chính xác của dự đoán giảm sau 24 giờ, chứng tỏ rằng MLP vẫn còn những hạn chế đáng kể trong dự báo. Chuyển động của bão xảy ra trên cả quy mô không gian và thời gian, và việc bỏ qua một trong hai quy mô này có thể làm giảm độ chính xác của hiệu ứng dự đoán. Do đó, các nhà nghiên cứu đã cố gắng sử dụng các sản phẩm của mô hình số làm đầu vào để xây dựng các mô hình không gian-thời gian dựa trên ConvLSTM, chẳng hạn như Deep-Hurricane-Tracker cho các đường đi của bão. Các thử nghiệm cho thấy mô hình hoạt động tốt hơn đáng kể so với đường cơ sở hiện có.

d) Mô hình ML dự báo cường độ bão

Các nhà nghiên cứu và dự báo đang bắt đầu đưa ra dự báo cường độ trực tiếp hoặc gián tiếp thông qua học máy. Các mục tiêu cho tất cả các trường hợp được tập trung vào ba loại chính: dự đoán cường độ có thay đổi nhanh chóng hay không, dự đoán cường độ trong tương lai. Trong phần mô tả các trường hợp thành công tiếp theo (Bảng 1.5), không khó để nhận thấy rằng các thuật toán học máy thực hiện tốt các bài toán dự đoán cường độ.

Bảng 1.5: Máy học trong dự báo cường độ bão

Nhiệm vụ	Thuật toán	Ý chính
Ước tính cường độ bão	DT	Ước tính cường độ sử dụng dữ liệu hình ảnh vi sóng
	SVM	Đề xuất khung học máy trong việc gắn nhãn các mức cường độ bão
	CNN	Thiết kế kiến trúc mạng để phân loại bão dựa trên cường độ
		Khám phá các khả năng ước tính cường độ bão từ các hình ảnh vệ tinh
		Ước tính cường độ bão như một nhiệm vụ hồi quy
		Sử dụng 2D-CNN và 3D-CNN để phân tích mối quan hệ giữa hình ảnh vệ tinh và cường độ bão
Dự báo cường độ bão	MLP	Dự đoán trực tiếp các giá trị cường độ bão
		So sánh các mô hình dựa trên mạng khác nhau để xác định mô hình dự báo cường độ tốt nhất
	RNN	Thiết kế mô hình dự đoán cường độ theo hướng dữ liệu thuần túy
	CNN-LSTM	Thiết kế mô hình không gian-thời gian dựa trên mạng kết hợp 2D-CNN, 3D-CNN và LSTM
	Chuyển giao học tập	Phát triển một mô hình dự đoán mạnh mẽ với học chuyển giao và xếp chồng
Dự đoán thay đổi cường độ bão	EA, PSO	Áp dụng EA hoặc PSO để dự đoán liệu các cơn bão sẽ tăng cường hay suy yếu
	DT	Xác thực các yếu tố dự báo liên quan đến RI và dự đoán sự thay đổi cường độ
	RNN	Sử dụng RNN với hệ số hợp tác để dự đoán RI
	SVM	Áp dụng kỹ thuật ML để phân loại bão là RI hoặc không RI

Nhiệm vụ	Thuật toán	Ý chính
	SVM, ANN, RF	Định lượng khả năng dự đoán RI bằng cách sử dụng một nhóm các phương pháp AI
	K-NN	Khám phá các cấu hình bão - đáy thích hợp cho RI

e) Mô hình ML dự báo thời tiết nguy hiểm và tác động thảm khốc do bão

Có nhiều ứng dụng thành công của ML trong dự báo gió, mưa và thủy triều do bão. Các nhà khoa học cũng tin rằng gió lớn, mưa lớn và nước dâng trong bão trong quá trình chuyển động của bão có thể được dự đoán hiệu quả bằng máy học và ý tưởng này đã được chứng minh là không chỉ cải thiện các mô hình dự báo hiện có mà còn đưa ra các chỉ dẫn đáng tin cậy hơn cho các cảnh báo nguy hiểm (Bảng 1.6).

Bảng 1.6: Máy học trong thời tiết bão và các dự báo tác động thảm khốc

Nhiệm vụ	Thuật toán	Ý chính
Dự báo trường gió trong bão	SVR	Phát triển kỹ thuật dự báo tốc độ gió bề mặt có độ tin cậy cao
	LSSVM	Ước tính cấu trúc trường gió bề mặt TC innercore 2D
	MLP	Mô phỏng trường gió bên trong lớp biên TC
		Xây dựng mô hình mô phỏng vận tốc gió
		Tối ưu hóa gió bão từ dữ liệu vệ tinh
Dự báo lượng mưa của bão	SVM	Thiết kế các mô hình dựa trên SVM để dự báo lượng mưa giờ do bão
		Thiết kế mô hình dự báo lũ do bão hai giai đoạn
	MLP	Xây dựng MLP nông để dự báo lượng mưa do bão
	SOM, MLP	Phát triển mô hình mạng nơ-ron kết hợp để dự báo lượng mưa do bão
	ANN-MRA	Xây dựng mô hình dự báo tổng lượng mưa và mực nước
	FNN-LLE	Thiết kế sơ đồ dự báo lượng mưa do bão
Dự báo triều cường	MLP	Dự đoán mức tăng đột biến và độ lệch tăng ngắn hạn
	SVR	
	BPN-ANFIS	Định lượng khả năng dự đoán RI bằng cách sử dụng một nhóm các phương pháp AI
	MLP	Phát triển mô hình triều cường phụ thuộc vào thời gian để dự đoán nhanh chóng

Nhiệm vụ	Thuật toán	Ý chính
		Dự đoán giá trị đỉnh của triều cường bằng cách sử dụng các thông số bão nhiệt đới
	ANN-SFM	Describe một mô hình dự báo triều cường và một quy trình lựa chọn khách quan

Sử dụng học máy đã đạt được tiến bộ trong các khía cạnh khác nhau trong dự báo bão, bao gồm dự báo nguồn gốc dự báo nguồn hình thành bão, theo dõi bão, dự báo cường độ, thời tiết nguy hiểm và dự báo tác động do bão gây ra; tuy nhiên, nhiều vấn đề nan giải và phức tạp vẫn còn tồn tại trong dự báo sử dụng học máy. Cách tư duy mới và phương pháp mới để đối phó với các vấn đề chính vừa là cơ hội vừa là thách thức đối với các nhà dự báo và nghiên cứu về các mô hình dự báo bão.

1.2.6. Ứng dụng mô hình học sâu (DL) để dự báo lượng mưa ở Đài Loan

Meng-Hua Yen và các cộng sự sử dụng thuật toán ESN (Echo State Network) và DeepESN (Deep Echo State Network) để dự đoán lượng mưa ở miền nam Đài Loan. Kết quả cho thấy hệ số tương quan bằng cách sử dụng DeepESN tốt hơn ESN và các thuật toán mạng nơ-ron thương mại (Mạng lan truyền ngược (BPN) và hồi quy vectơ hỗ trợ (SVR), MATLAB, The MathWorks co.), và độ chính xác của lượng mưa dự đoán bằng cách sử dụng DeepESN được cải thiện đáng kể. Tóm lại, DeepESN là một phương pháp đáng tin cậy và tốt để dự đoán lượng mưa[115].

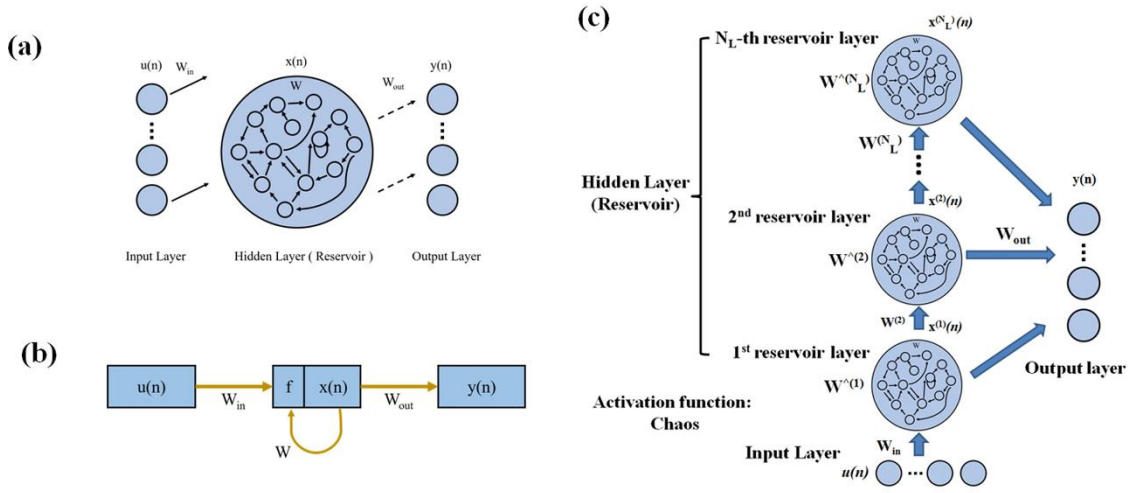
a) Về dữ liệu quan trắc và chuẩn hóa dữ liệu

Trong nghiên cứu này, dữ liệu được lấy từ Cục Thời tiết Trung ương Đài Loan (https://www.cwb.gov.tw/V7/index.htm) và Trung tâm Mực nước biển, Đại học Hawaii (https:// uhslc). soest.hawaii.edu/). Tổng cộng có 7 thông số, bao gồm áp suất không khí, nhiệt độ, độ ẩm, tốc độ gió, hướng gió, lượng mưa và mực nước biển. Do các giá trị của các thông số khí tượng sử dụng trong nghiên cứu này tập trung, nên phương pháp chuẩn hóa được sử dụng là chuẩn hóa mapminmax.

b) Quy trình thiết lập và dự đoán mô hình

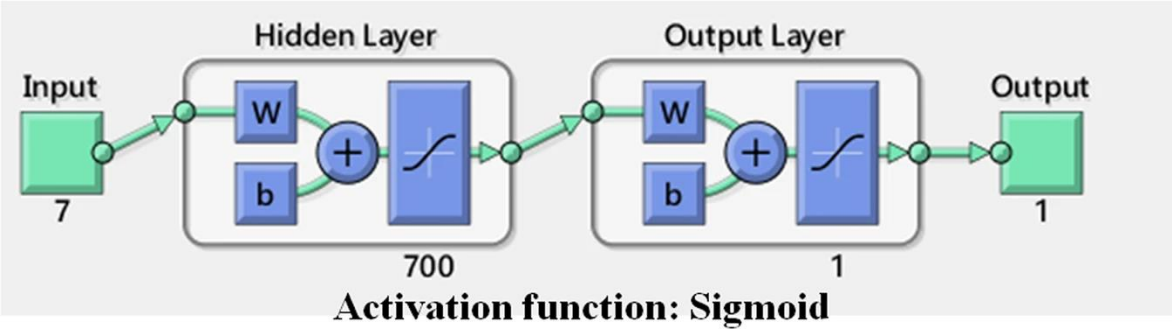
ESN chủ yếu cung cấp kiến trúc và các nguyên tắc học tập có giám sát cho các RNN 23. Cấu trúc của ESN có thể được phân tách thành ba phần: lớp đầu vào, lớp ẩn (hồ chứa) và lớp đầu ra (Hình 1.6a). Kiến trúc mạng được thể hiện trong Hình 1.8b, trong đó u (n) là dữ liệu đầu vào của mạng với n là thời gian rời rạc, W in là các lớp trọng số chưa được đào tạo, f là các hàm phi tuyến tính, x (n) là vectơ trạng thái của hồ chứa tại thời điểm bước n, W ra là trọng số thu được sau khi huấn luyện, W là ma trận trọng số bên trong bể chứa và y (n) là đầu ra của mạng. Hàm kích hoạt (f) được

sử dụng trong mô hình ESN là hỗn loạn. Số nơ-ron của các lớp ẩn dưới dạng Ổ chứa của mô hình ESN trong nghiên cứu này là 100. Hơn nữa, DeepESN dựa trên kiến trúc ESN và được cải tiến. Kiến trúc DeepESN được thể hiện trong Hình 1.6c. Để thuận tiện cho việc giải thích, chức năng chuyển đổi trạng thái của lớp đầu tiên của DeepESN được định nghĩa như sau:



Hình 1.6: Cấu trúc của ESN và DeepESN

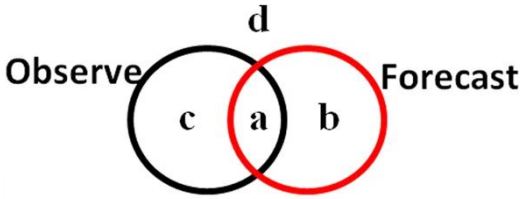
Để xác minh thêm tính khả thi của mô hình ESN / DeepESN, nhóm tác giả sử dụng mô hình BPN và mô hình SVR để thực hiện dự báo lượng mưa (Hình 1.7).



Hình 1.7: Kiến trúc BPN

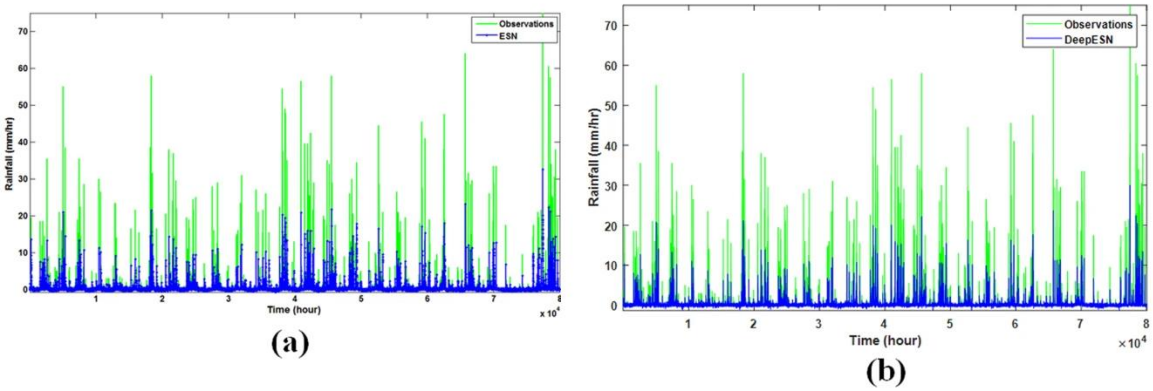
Ngoài ra, nếu coi dự báo lượng mưa định lượng (QPF) là dự báo về lượng mưa lớn hơn một ngưỡng nhất định, thì bài toán QPF có thể được chia thành một loạt các bảng dự phòng 2 × 2, mỗi bảng cho một giá trị ngưỡng khác nhau (Hình 1.8).

Forecast	Observed		
	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d



Hình 1.8: Kiểm tra kết quả dự báo và quan trắc

So sánh lượng mưa quan trắc (đường cong màu xanh lá cây) tại thông tin Đài quan sát Zengwen ở thành phố Đài Nam với lượng mưa dự báo (đường cong màu xanh lam) bằng cách sử dụng mô hình ESN (a) và mô hình DeepESN (b).



Hình 1.9: So sánh kết quả dự báo sử dụng mô hình ESN (a) và DeepESN (b)

Trong nghiên cứu này, nhóm tác giả chứng minh rằng dự đoán lượng mưa có thể đạt được nhờ các mạng nơron (mô hình ESN/ DeepESN). Qua kiểm tra và xác minh thực tế đã chứng minh rằng có thể thực hiện dự báo lượng mưa bằng mô hình ESN / DeepESN. Hiệu suất của mô hình DeepESN tốt hơn so với mô hình ESN và các thuật toán mạng nơ-ron thương mại (Mạng lan truyền ngược và hồi quy vectơ hỗ trợ, MATLAB, The MathWorks co.). Do đó, DeepESN là một mô hình tốt hơn để dự đoán lượng mưa so với các mô hình khác. Cuối cùng, nhóm tác giả kiểm tra tác động của từng tham số đầu vào bằng cách tắt một tham số đầu vào khác dựa trên mô hình DeepESN. Nó chỉ ra rằng lượng mưa, áp suất và độ ẩm là những thông số quan trọng nhất và ảnh hưởng lớn đến hiệu suất dự báo lượng mưa trong mô hình DeepESN.

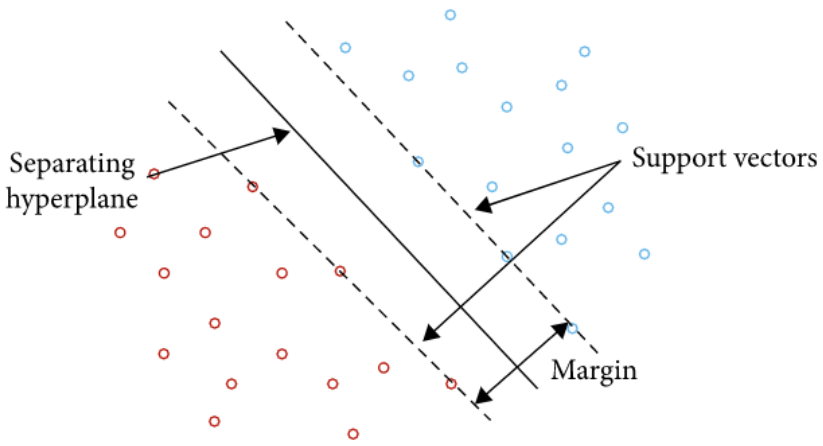
1.2.7. Ứng dụng mô hình ML để dự báo chất lượng không khí ở California

Nghiên cứu này được hỗ trợ bởi quỹ quốc gia FCT thông qua các dự án GADgET (2018), BINDER (2017) và AICE (2019) và LASIGE (2020). Trong nghiên cứu này, nhóm tác giả sử dụng một phương pháp học máy hỗ trợ hồi quy vectơ (SVR) để dự báo mức độ ô nhiễm bụi hạt và dự đoán chỉ số chất lượng không

khí (AQI). Các kết quả được trình bày chứng minh rằng SVR với hạt nhân RBF cho phép dự đoán chính xác nồng độ chất ô nhiễm hàng giờ, như carbon monoxide, sulfur dioxide, nitrogen dioxide, ozone tầng mặt đất và vật chất dạng hạt 2,5, cũng như AQI hàng giờ cho bang California. Theo tiêu chuẩn AQI do Cơ quan Bảo vệ Môi trường Hoa Kỳ xác định kết quả dự báo đã được thực hiện với độ chính xác trên 94,1%.

a) Sử dụng Máy vectơ hỗ trợ (SVM)

Máy vectơ hỗ trợ (SVM) áp dụng cho các bài toán phân loại. Mục tiêu là tìm kiếm siêu phẳng phân tách tối ưu giữa các lớp. Các điểm nằm trên ranh giới của các lớp được gọi là vectơ hỗ trợ và không gian ở giữa được gọi là siêu phẳng. Hình 1.10 trình bày một ví dụ về bài toán phân loại có thể phân tách tuyến tính được giải quyết bằng cách sử dụng SVM. SVM nhằm mục đích tối đa hóa lợi nhuận giữa các vectơ hỗ trợ và siêu phẳng.



Hình 1.10: Bài toán phân tách tuyến tính

b) Mô tả dữ liệu và tiền xử lý dữ liệu

Bộ dữ liệu được sử dụng trong nghiên cứu này được trích xuất từ Chất lượng không khí của EPA. Tất cả các tệp chứa dữ liệu hàng giờ, được phân tách bằng chất ô nhiễm hoặc thông số đang được đo CO, SO_2, NO_2, ozone, $PM_{2.5}$, nhiệt độ, độ ẩm và gió, với các quan sát từ bang California. Các sự kiện hàng giờ được thu thập từ 01/01/2016 đến 01/5/2018. Tổng cộng 102090 bản ghi đã được sử dụng. Bước tiền xử lý dữ liệu thường ảnh hưởng đến khả năng tổng quát hóa của thuật toán học máy. Tiền xử lý dữ liệu thường bao gồm việc nhập dữ liệu bị thiếu, loại bỏ hoặc sửa đổi các quan sát ngoại lệ, biến đổi dữ liệu (thường là chuẩn hóa).

c) Cài đặt thử nghiệm

SVM có ba siêu tham số cần được người dùng xác định: chức năng loại hạt nhân, hằng số chính quy C và độ lệch tối đa cho phép ε. Tham số C từ 10 - 100; ε là

từ 0,001 - 0,1. Số lần lặp được chọn để chạy tìm kiếm ngẫu nhiên cần trung bình 60 lần lặp để đạt được kết quả tốt (Bảng 1.7).

Bảng 1.7: Tìm kiếm ngẫu nhiên tham số tối ưu trên mỗi tập dữ liệu chất ô nhiễm

Chất ô nhiễm	Bộ dữ liệu PCA				Tập dữ liệu chuẩn hóa			
	Kernel	Chọn C	Chọn ε	R^2	Kernel	Chọn C	Chọn ε	R^2
CO	RBF	3	0,08	0,783	RBF	2	0,033	0,916
$SÔ_2$	RBF	3	0,025	0,882	RBF	1	0,067	0,948
SO_2	RBF	1	0,062	0,712	RBF	2	0,086	0,718
Khí quyển	RBF	2	0,076	0,903	RBF	2	0,02	0,979
$PM_{2.5}$	RBF	1	0,055	0,765	RBF	3	0,032	0,767

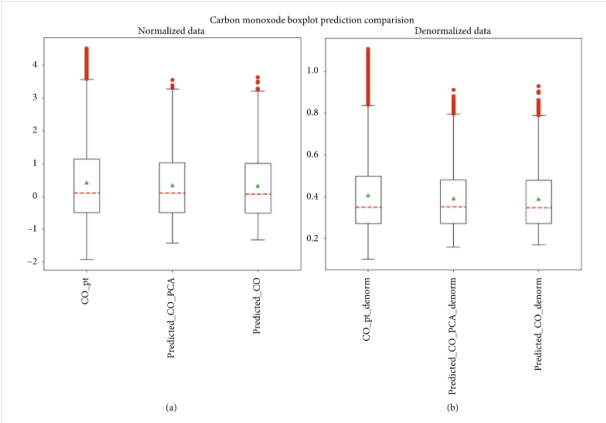
Trong Bảng 1.5, các giá trị khác nhau của *C* và ε thu được đối với các chất ô nhiễm khác nhau. Nhóm tác giả đã sử dụng mối tương quan Pearson, sai số MAE, sai số RMSE và nRMSE để so sánh hai mô hình này với nhau. Các mô hình được đào tạo sử dụng 70% tập dữ liệu có sẵn.

d) Kết quả thực nghiệm

Các kết quả thí nghiệm mà SVR-RBF và PCA SVR-RBF đạt được trong việc dự báo năm chất ô nhiễm khác nhau (CO, NO_2, SO_2, ozone, và $PM_{2.5}$, và dự báo chỉ số chất lượng không khí (AQI). Phần mềm được sử dụng để thực hiện giai đoạn thử nghiệm này được phát triển bằng Python (phiên bản 3.6), chủ yếu sử dụng gói Pandas và Scikit-learning. Kết quả như sau:

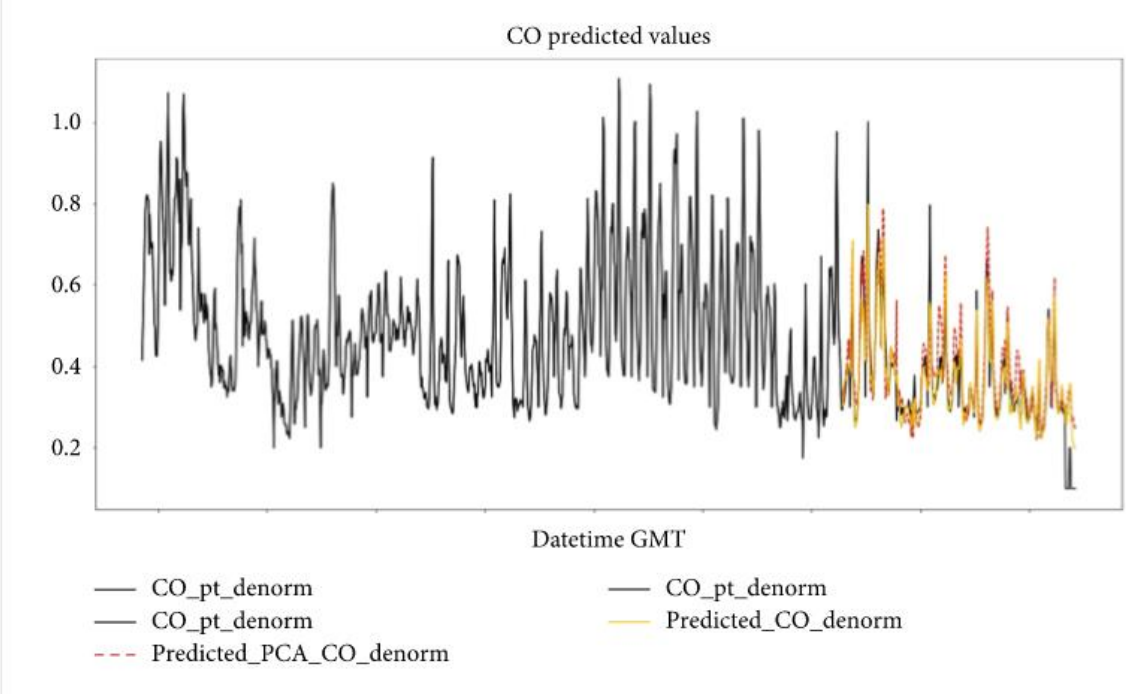
Dự báo Carbon Monoxide (CO)

Các kết quả thu được để dự báo carbon monoxide được thể hiện trong các ô vuông của hình 1.11.



Hình 1.11: Biểu đồ giá trị carbon monoxide dự đoán và quan trắc

Dự báo các phép đo carbon monoxide ở California bằng cách sử dụng các mô hình dự báo SVR. Đường màu đen hiển thị các giá trị quan trắc, đường màu đỏ và màu vàng hiển thị các dự báo của mô hình (hình 1.12).



Hình 1.12: Dự báo các phép đo carbon monoxide

Thống kê về sai số dự báo carbon monoxide được trình bày trong Bảng 1.8.

Bảng 1.8: Thống kê sai số mô hình dự báo với tập dữ liệu huấn luyện và xác nhận

Số liệu lỗi	PCA SVR-RBF		SVR-RBF	
	Tập huấn luyện	Bộ xác thực	Tập huấn luyện	Bộ xác thực
MAE	0,119	0,311	0,080	0,211
R^2	0,948	0,769	0,976	0,868
RMSE	0,184	0,492	0,128	0,367
nRMSE	0,032	0,076	0,022	0,057

Nhìn chung, cả hai mô hình dự báo được nghiên cứu đều đạt được hiệu suất tốt trong việc dự đoán các giá trị carbon monoxide quan sát được ở California.

Tương tự như kết quả dự báo CO, các kết quả dự báo nhìn chung, các mô hình dự báo được nghiên cứ đều đạt được hiệu suất tốt trong việc dự đoán các giá trị về Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Ground-Level Ozone, Particulate Matter 2.5 (PM2.5) và chất lượng không khí Air Quality Index (AQI) quan sát được ở California.

e) Kết luận

Dự báo chất lượng không khí là một công việc phức tạp do tính chất động lực học, tính dễ bay hơi và sự biến đổi lớn theo không gian và thời gian của các chất ô nhiễm và các hạt. Công trình này đã trình bày một nghiên cứu về hồi quy vectơ hỗ trợ (SVR) để dự báo mức độ ô nhiễm và hạt và để xác định chính xác AQI. Phương pháp được nghiên cứu đã tạo ra một mô hình phù hợp về ô nhiễm khí quyển hàng giờ, cho phép nhóm tác giả có được độ chính xác tốt trong việc mô hình hóa nồng độ chất ô nhiễm như O_3, CO và SO_2, cũng như AQI hàng giờ cho bang California. Trong tương lai, nhóm nghiên cứu dự định cải thiện và điều tra việc sử dụng SVR để dự báo chất lượng không khí thông qua các chủ đề sau: Tập dữ liệu và lựa chọn biến đổi dữ liệu với một tập dữ liệu lớn với nhiều thông số và phép đo hơn, có thể hỗ trợ các mô hình dự báo chính xác hơn cho các chất ô nhiễm không khí và các hạt, cụ thể là NO_2 và $PM_{2.5}$. Tối ưu hóa tham số SVR, vì hiệu suất của mô hình SVR bị ảnh hưởng rất nhiều bởi việc lựa chọn chức năng hạt nhân và tham số phạt *C*, sẽ thử nghiệm các phương pháp khác với tìm kiếm ngẫu nhiên, để tối ưu hóa siêu tham số như thuật toán di truyền hoặc tối ưu hóa bầy hạt. Cuối cùng, nhóm nghiên cứu dự định so sánh kết quả thu được từ SVR với kết quả đạt được bằng các thuật toán học máy khác như mạng NN, mạng Bayes, cây quyết định, rừng ngẫu nhiên và lập trình di truyền.

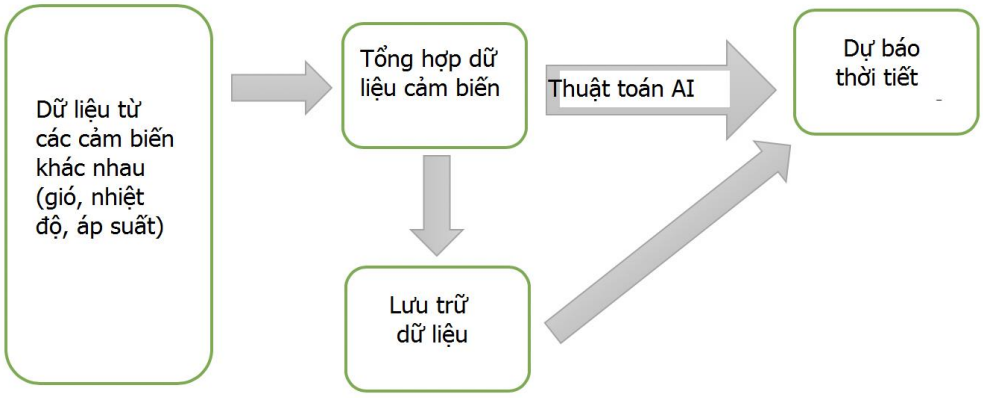
1.2.8. Cung cấp giải pháp công nghệ phần mềm AI trong dự báo thời tiết

Công ty Allerin là nhà cung cấp giải pháp phần mềm quốc tế, đã có nhiều nghiên cứu và cung cấp các giải pháp phần mềm về ứng dụng AI, Big data trong dự báo thời tiết [16].

- **Thay đổi dự báo thời tiết bằng AI**: Dự báo thời tiết là một khoa học phức tạp. Nó đề cập đến các tập dữ liệu khổng lồ được thu thập từ hàng ngàn vệ tinh và trạm quan trắc thời tiết mỗi ngày. Thu thập dữ liệu, xác định các mẫu trong các quan sát được thực hiện, và sau đó suy ra kết quả để có được các dự báo thời tiết chính xác có thể khá vất vả. Điều quan trọng là dữ liệu thời tiết được thu thập và phân tích trong thời gian thực. AI sử dụng các chương trình toán học do máy tính tạo ra và các phương pháp giải quyết vấn đề tính toán trên các tập Big data để xác định các mẫu và đưa ra một giả thuyết có liên quan, khái quát hóa dữ liệu. Xem xét các phức tạp khác nhau liên quan đến dự báo thời tiết, các nhà khoa học hiện đang sử dụng AI để dự báo thời tiết để có được kết quả chính xác, nhanh chóng.

- **Mô hình AI, ML cho dự báo thời tiết**: AI sử dụng các mô hình toán học sâu có thể học hỏi từ các dữ liệu thời tiết trong quá khứ. Một trong những mô hình phổ biến là dự báo thời tiết số (NWP). Mô hình nghiên cứu và điều khiển các tập Big

data được chuyển tiếp từ các vệ tinh thời tiết, các trạm chuyển tiếp để cung cấp các dự báo thời tiết ngắn hạn hoặc các dự báo khí hậu. Rất nhiều công ty đã đầu tư vào việc ứng dụng AI để dự đoán thời tiết.



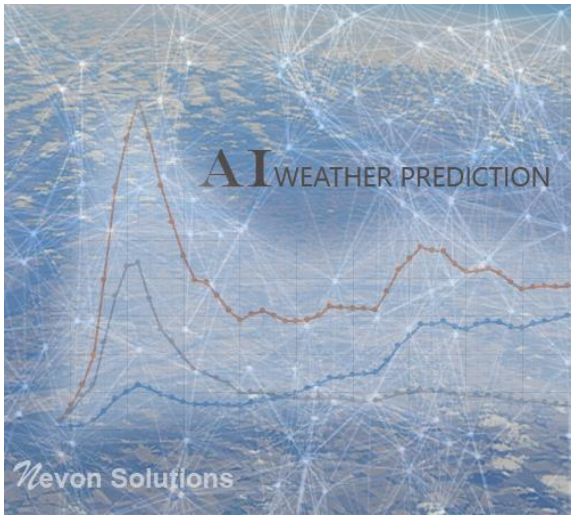
Hình 1.13: Mô hình công nghệ AI trong dự báo thời tiết

Ví dụ, IBM đã mua các thông tin dữ liệu thời tiết từ weather.com, Weather Underground, Weather Company Brand và WSI. Do đó, IBM đã có được quyền truy cập vào tập dữ liệu thống kê thời tiết từ khắp nơi trên thế giới. Cùng với nền tảng AI của riêng mình, IBM đã đạt được tiềm năng dự báo thời tiết có độ chính xác được cải thiện. Kết quả của việc mua lại này là Deep Thunder của IBM.

1.2.9. Sản phẩm công nghệ AI trong dự báo thời tiết tại Ấn Độ

Sản phẩm do Công ty giải pháp công nghệ phần mềm Nevon Solutions Pvt. Ltd. - 810, Corporate Annexe Sonawala Lane, Goregaon East, Mumbai, India, Planet Earth giới thiệu.

Nevon Solutions đưa ra một kỹ thuật bằng cách kết hợp dự báo với AI. Có nhiều kỹ thuật AI để dự đoán thời tiết với độ chính xác nhất định. Các kỹ thuật AI được áp dụng trong dự báo thời tiết như mạng nơ ron nhân tạo, mạng Ensemble Nơ-ron, mạng back-propagation, mạng cơ sở xuyên tâm, mạng Nơ-ron tổng hợp hồi quy, thuật toán di truyền, Perceptron đa lớp, phân cụm mờ, thuật toán backpropagation với lỗi tối thiểu.



Hình 1.14: Sản phẩm công nghệ AI trong dự báo thời tiết của Nevon Solutions

Nevon Solutions sử dụng nhiều thông số đầu vào để dự báo thời tiết dựa trên các yếu tố như nhiệt độ, lượng mưa, độ ẩm, mây, thời tiết trong ngày, …. [17].

1.2.10. Vai trò và ứng dụng của Big data trong lĩnh vực KTTV

1.2.10.1. Vai trò và sự cần thiết của của Big data trong lĩnh vực KTTV

Dữ liệu lớn Big data là một thuật ngữ được sử dụng để xác định một lượng lớn dữ liệu trên quy mô lớn, có thể là có cấu trúc, bán cấu trúc và phi cấu trúc, từ một số tài nguyên như phương tiện truyền thông và dữ liệu công khai, dữ liệu cảm biến, dữ liệu kho, v.v. . khác nhau về các định dạng như tệp .txt và .csv, tệp hình ảnh, tệp html, v.v. Dữ liệu được thu thập và chuẩn bị với tốc độ rất nhanh với sự trợ giúp của siêu máy tính và được xử lý tốc độ cao cho các ứng dụng thời gian thực và phạm vi rộng. Để đưa dữ liệu này vào hành động và thông tin, cần phải phân tích dữ liệu và do đó các bước phân tích dữ liệu lớn được đưa ra. Ba đặc trưng chính là Khối lượng, Vận tốc và Biến thể để phân tích các bộ dữ liệu lớn nhằm đảm bảo thông tin chính xác [113]. Việc gia tăng những thay đổi đáng chú ý về thời tiết trở thành một vấn đề nghiêm trọng cần quan tâm, những biến động từng ngày của thời tiết thu hút sự chú ý của không chỉ các nhà khí tượng học mà còn cả các nhà phân tích, đặc biệt là các dữ liệu dự báo KTTV. Nó đặt ra thách thức cho các nhà nghiên cứu là đằng sau sự kiện thời tiết, thiên tai là sự kiện gì sẽ xảy ra vào ngày mai và trong thời gian tới. Nghiên cứu những thay đổi của thời tiết là cần thiết để có được nhiều lợi ích như cứu người, vượt qua rủi ro, tăng cường lợi nhuận và chất lượng cuộc sống dựa vào thời tiết, v.v. Để dự báo thời tiết, chúng ta cần phân tích lượng dữ liệu khổng lồ, và do đó là dữ liệu lớn được sử dụng như một con át chủ bài cung cấp nhiều hướng dẫn cho các thảm họa thiên nhiên sắp tới như bão, lũ mưa lớn, sấm sét, lốc xoáy, sóng thần, ... [113]. Dự báo thời tiết như một chức năng của công nghệ Dữ liệu lớn và ứng dụng của nó, vai trò của phân tích dữ liệu lớn trong việc dự đoán thời tiết thay đổi cơ bản liên quan đến thiên tai và nhiều sự kiện khác. Cuộc sống hàng ngày của chúng ta phụ thuộc trực tiếp hoặc gián tiếp vào thời tiết về mặt kinh tế và môi trường, nó ảnh hưởng đến chúng ta với nhiều yếu tố khác nhau như sự kiện, thời gian, địa điểm, v.v ... Trong các yếu tố này, dự báo thời tiết hoạt động với các thông số nhiệt độ , độ ẩm và tốc độ gió. Dự báo thời tiết và các hiện tượng KTTV nguy hiểm là hoạt động phức tạp và đầy thách thức, sự tương tác giữa các yếu tố và thông số là điều cần thiết để xác thực dự báo. Big data rõ ràng là một công cụ thiết yếu để triển khai dự báo thời tiết, trong thời đại dữ liệu, nhiều cách đã được phát triển trong lĩnh vực phân tích để có kết quả chính xác và nhanh hơn, nhờ vào các thuật toán phân tích áp dụng trong Big data trong các mô hình AI dự báo KTTV.

1.2.10.2. Ứng dụng của Big data trong lĩnh vực KTTV

Dự báo thời tiết luôn cực kỳ thách thức, với số lượng biến liên quan và các tương tác phức tạp giữa các biến đó. Sự gia tăng về khả năng thu thập, xử lý dữ liệu đã giúp tăng cường khả năng báo thời tiết để xác định thời gian và mức độ nghiêm trọng của thiên tai bão, lũ lụt, bão tuyết và các sự kiện thời tiết khác.

- **Big data cho dự báo thời tiết điểm**: Một ví dụ về ứng dụng Big data trong dự báo thời tiết là Deep Thunder của IBM. Không giống như nhiều hệ thống dự báo thời tiết cung cấp thông tin chung về một khu vực địa lý rộng lớn, Deep Thunder cung cấp dự báo cho các địa điểm cực kỳ cụ thể, chẳng hạn như một sân bay để chính quyền địa phương có thể có được thông tin quan trọng trong thời gian thực. Dưới đây là một số ví dụ về thông tin mà Deep Thunder có thể cung cấp [18]: (Ước tính các khu vực có khả năng xảy ra lũ lụt nghiêm trọng; Sức mạnh và hướng của bão nhiệt đới; Lượng tuyết hoặc mưa nhiều khả năng sẽ rơi ở một khu vực cụ thể; Các vị trí đường dây điện có khả năng bị ảnh hưởng nhất; Ước tính các khu vực có tốc độ gió có khả năng cao nhất; Những vị trí mà cầu và đường có khả năng bị thiệt hại do bão nhất; Khả năng các chuyến bay bị hủy tại các sân bay cụ thể. Thông tin này rất cần thiết cho kế hoạch khẩn cấp. Sử dụng Big data, chính quyền địa phương có thể dự đoán tốt hơn các vấn đề gây ra bởi thời tiết trước khi chúng xảy ra. Ví dụ, các nhà quy hoạch có thể chuẩn bị để sơ tán các khu vực trũng thấp có khả năng bị ngập lụt. Cũng có thể lập kế hoạch nâng cấp các cơ sở hiện có. (Ví dụ: các đường dây điện dễ bị vô hiệu hóa bởi gió lớn có thể được nâng cấp).

- **Big data cho dự báo thời tiết sự kiện**: Thành phố Rio de Janeiro, Brazil, sử dụng Big data của Deep Thunder để lên kế hoạch cho Thế vận hội 2016. Sử dụng công nghệ Big data, thành phố sẽ sử dụng các dự báo được cải thiện về bão, lũ lụt và các thảm họa tự nhiên khác để đảm bảo Thế vận hội sẽ không bị gián đoạn bởi các sự kiện như vậy.

- **Big data cho xử lý dữ liệu dự báo**: IBM cũng đang cung cấp sức mạnh tính toán khổng lồ cho Cơ quan Khí tượng Hàn Quốc (KMA) để nắm bắt hoàn toàn công nghệ Big data. KMA thu thập hơn 1,5 terabyte dữ liệu khí tượng mỗi ngày, đòi hỏi một lượng lưu trữ và sức mạnh xử lý đáng kinh ngạc để phân tích. Bằng cách sử dụng Big data, KMA sẽ có thể cải thiện dự báo về cường độ và vị trí của bão nhiệt đới và các loại thời tiết khác. Một terabyte tương đương với một nghìn tỷ byte. Đó là 1.000.000.000.000 byte thông tin, theo ký hiệu khoa học là 1.0 x 10 12 và để lưu trữ sẽ cần khoảng 1.500 đĩa CD để lưu trữ một terabyte. Bao gồm cả vỏ nhựa của chúng, sẽ xếp chồng lên nhau như một tháp CD cao 40 feet (12 mét).

- **Big data cho dự báo, xác định thiên tai:** sử dụng Big data trong dự báo bão Sandy năm 2012 (bão thế kỷ). Trung tâm Bão quốc gia đã sử dụng công nghệ Big data để dự đoán đổ bộ của cơn bão để trong vòng 30 dặm một cho năm ngày trước. Đó là một sự gia tăng đáng kể về độ chính xác từ những sự kiện có thể xảy ra từ 20 năm trước. Do đó, Fema và các tổ chức quản lý thảm họa khác đã chuẩn bị tốt hơn nhiều để đối phó với mớ hỗn độn hơn những gì họ có thể đã xảy ra vào những năm 1990 hoặc trước đó. Các tập đoàn bán bảo hiểm thu thập và xử lý dữ liệu thời tiết để phòng ngừa rủi ro cho khách hàng bởi các thiệt hại thời tiết. Ví dụ Climate Corporation, thành lập năm 2006 bán các dịch vụ dự báo thời tiết và bảo hiểm chuyên ngành cho nông dân để phòng ngừa rủi ro từ thiệt hại mùa màng. Công ty sử dụng Big data để xác định các loại rủi ro liên quan đến một khu vực cụ thể, dựa trên lượng dữ liệu khổng lồ về độ ẩm, loại đất, năng suất cây trồng trong quá khứ, v.v.

1.3. Đánh giá tổng quan tình hình nghiên cứu ở trong nước

1.3.1. Hiện trạng, xu thế phát triển AI và Big data tại Việt Nam

1.3.2. Ứng dụng ANN trong dự báo định lượng mưa

Bùi Minh Tăng và cộng sự (2009) đã thử nghiệm dự báo định lượng mưa cho hạn dự báo 24 giờ dựa trên một số phương pháp MOS truyền thống như hồi quy tuyến tính đa biến (MLR), phân tích riêng biệt (MDA), mạng thần kinh nhân tạo (ANN) và hồi quy logistic (LR) từ sản phẩm của mô hình HRM và GSM. Các phương trình MOS được phát triển cho cả mục đích dự báo định lượng và dự báo xác suất. Các kết qủa đánh giá đã cho thấy kỹ năng dự báo mưa đã được cải thiện so với dự báo trực tiếp từ mô hình trong đó phương pháp MLR có kỹ năng tốt nhất và hiệu quả nhất về mặt tính toán. Ngoài ra, với cùng một phương pháp thống kê thì áp dụng cho mô hình GSM sẽ đem lại nhiều hiệu quả hơn so với mô hình HRM. Tuy nhiên, các phương pháp được sử dụng trong nghiên cứu này vẫn bộc lộ những hạn chế của phương pháp MOS truyền thống [43].

1.3.3. Ứng dụng AI để khôi phục các dữ liệu thủy văn

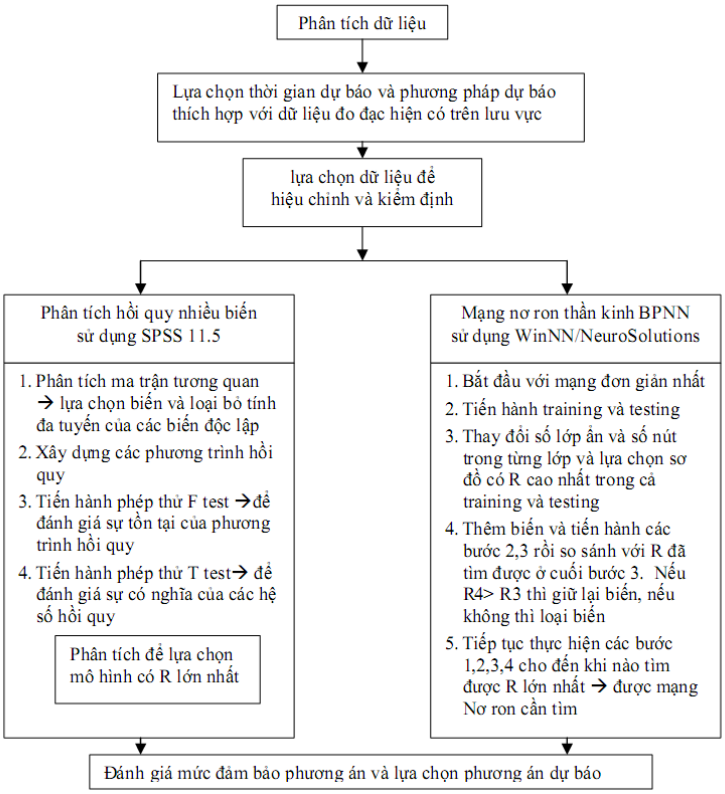
Mới đây, Tiến sĩ Trịnh Quang Toàn, Trưởng phòng nghiên cứu KTTV, Đại học UC Davis (Mỹ) đã giới thiệu công nghệ tính toán lượng mưa, KTTV và nước ngầm để dự báo thiên tai theo thời gian thực ở Việt Nam sử dụng AI. Việc kết hợp AI để xử lý dữ liệu ảnh vệ tinh, ảnh radar với mô hình tính toán thủy văn để khôi phục các dữ liệu về dòng chảy, mực nước... có thể giúp Việt Nam tự chủ số liệu trên lưu vực sông Hồng, Thái Bình và Mekong, vốn trước đây phụ thuộc hoàn toàn vào Trung Quốc. Sau đó sẽ đưa ra đánh giá ảnh hưởng từ các hồ chứa từ phía Trung Quốc tới Việt Nam. Công nghệ này cũng cho phép quan sát toàn bộ mực nước hồ của Trung

Quốc phía thượng nguồn thông qua giải đoán ảnh vệ tinh khí tượng theo ngày và dần tiến tới theo giờ. Nguồn số liệu trên là cơ sở để các đơn vị chức năng sẽ dự báo, tính toán để đưa ra cảnh báo mở cửa xả hồ chứa, đảm bảo an toàn hồ đập, di dân để tránh thiệt hại về người và của tài sản trên các lưu vực sông. Việc sử dụng ảnh vệ tinh để dự báo mưa, từ đó lấy số liệu thủy văn thì nhiều nước đã sử dụng từ lâu. Việt Nam là một trong những nước đang sử dụng phương pháp này. Độ chính xác, độ tin cậy về mô hình dự báo KTTV đã được xác lập là tương đối ổn định. Tuy nhiên, cái khó trong bài toán dự đoán lũ hiện nay nằm ở việc không có số liệu thiết kế, địa hình và vận hành của hàng trăm hồ đập thủy điện, thủy lợi từ phía Trung Quốc, đây là bài toán khó trong việc tính toán lưu lượng dòng chảy trên các hệ thống sông [11].

1.3.4. Ứng dụng AI dự báo lũ

1.3.4.1. Phương pháp nghiên cứu

PGS. TS Lê Văn Nghinh và các cộng sự đã triển khai nghiên cứu ứng dụng mạng nơ ron thần kinh vào sự báo lũ các sông ở tỉnh Bình Định và Quảng Trị[12]. Qua nghiên cứu, phân tích hệ thống sông, thời gian truyền lũ, số liệu đo đạc của các trạm quan trắc, ngoài việc lựa chọn phương pháp phân tích hồi quy nhiều biến, nhóm nghiên cứu đã ứng dụng phương pháp mạng AI (ANN) để dự báo báo lũ cho các sông ở hai tỉnh Bình Định và Quảng Trị. Cả hai phương pháp trên đều dựa triển các quan hệ giữa mực nước dự báo và các yếu tố ảnh hưởng như mực nước tại thời điểm dự báo, mực nước trạm quan trắc trên tại thời điểm dự báo, lượng mưa đo được cho đến thời điểm dự báo của các trạm trong lưu vực, …vv. Tuy nhiên, cách giải hay thuật toán của 2 phương pháp trên là khác nhau, một dựa trên thuật toán tối ưu hàm tuyến tính, còn một dựa trên thuật toán tối ưu hàm phi tuyến (Hình 1.15).

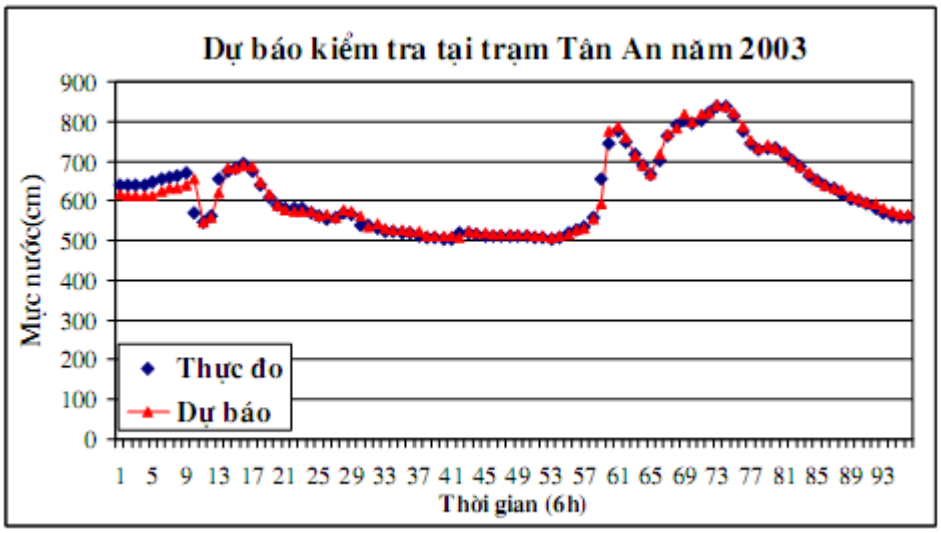


Hình 1.15: Các bước xây dựng phương án dự báo lũ sử dụng AI

1.3.4.2. Kết quả thu được

Thực hiện các bước tính toán như đã trình bày và sử dụng phần mềm thống kế SPSS version 11.5 để phân tích hồi quy nhiều biến, WinNN, Neuro Solution phiên bản 4.2 để xây dựng mạng thần kinh nhân tạo tối ưu dùng cho dự báo, nhóm nghiên cứu đã xây dựng được một số phương án dự báo mực nước trước 6 h với mức đảm bảo phương án là khá tốt (>80%) cho các sông như sông Thạch Hãn tại trạm Thạch Hãn, sông Bến Hải tại trạm Gia Vòng, sông Hiếu tại trạm Đông Hà (tỉnh Quảng Trị), sông Kon tại trạm Tân An, sông Hà Thành tại trạm Diêu Trì (Bình Định). Một số kết quả dự báo kiểm tra được minh hoạ ở hình 1.16. Nhìn chung các phương án dự báo mực nước lũ trước 6 h cho các sông lớn ở miền trung bằng phân tích hồi quy nhiều biến (MVR) và bằng phương pháp mạng thần kinh nhân tạo (BPNN) đều có mức đảm bảo trên 80%. Thông thường thì mô hình mạng nơ ron thần kinh cho kết quả tốt hơn MVR. Tuy nhiên, việc sử dụng mô hình BPNN để dự báo lại khó hơn việc sử dụng phương trình dự báo đơn giản của phương pháp MVR, do vậy nhóm nghiên cứu khuyến nghị nên dùng mô hình MVR. Nếu người sử dụng đã quen thuộc với phần mềm WinNN 32 hoặc Neuro Solutions rồi thì nên dùng BPNN. Trong trường hợp dự báo thấy có giá trị khác thường thì nên tham khảo thêm kết quả dự báo của cả hai mô hình.

Ngoài ra, việc theo dõi dự báo các hình thế thời tiết cũng cho chúng ta những tiên đoán về lượng mưa có thể xảy ra trên các lưu vực sông nhờ vào các kết quả tổng hợp và phân tích thống kê mưa lũ do các hình thế thời tiết gây ra, và vì vậy sẽ tạo điều kiện rất lớn cho công tác dự báo mực nước lũ của các sông miền Trung và qua đó có thể tăng cường thời gian cảnh báo lũ.



Hình 1.16: Kết quả dự báo kiểm tra mực nước trên sông Kone tại trạm Tân An bằng mô hình hồi quy nhiều biến

1.3.4.3. Kết luận

Kết quả đạt được của nghiên cứu cho thấy khả năng ứng dụng rất tốt của mô hình mạng trí tuệ nhân tạo ANN - mạng Nơ ron thần kinh với thuật toán lan truyền ngược BPNN vào dự báo thủy văn. Khả năng ứng dụng này không chỉ dừng lại ở hướng nghiên cứu BPNN mà còn ở các hướng khác như mạng Fuzzy và sự kết hợp với các thuật toán giải đoán gen GA để tìm mạng tốt nhất nâng cao hiệu quả và giảm thời gian chạy mô hình. Cần phải có những nghiên cứu tiếp theo về ANN cho những khu vực khác, đặc biệt là những khu vực mà tài liệu đo đạc tốt hơn để tổng kết đánh giá và đưa mô hình vào công tác dự báo tác nghiệp.

1.3.5. Áp dụng AI trong dự báo lưu lượng đến hồ lưu vực sông Ba

Cao Hoàng Hải và các cộng sự đã nghiên cứu sử dụng mô hình AI vào dự báo lưu lượng đến hồ lưu vực sông Ba [44]. Trong nghiên cứu này, hai mô hình AI là Random Forest (RF) và Support Vector Machine (SVM/SVR) đã được áp dụng thử nghiệm đối với một hồ chứa lớn - hồ Sông Hinh trên lưu vực Sông Ba, Việt Nam. Ba trường hợp tính toán là dự báo lưu lượng trung bình 3 ngày, 7 ngày và 1 tháng (hạn-ngắn, trung và hạn dài) đến hồ sử dụng số liệu khí tượng, thủy văn trong khu vực đã được xây dựng để kiểm nghiệm khả năng dự báo của hai mô hình RF và SVR. Kết quả cho thấy, cả hai mô hình đều đưa ra kết quả dự báo với độ chính xác cao thể hiện qua chỉ số NSE trung bình đạt trên 0,8, đặc biệt trong một số trường hợp tính toán như dự báo lưu lượng trung hạn và dài hạn, chỉ số NSE trung bình trên 0,9. Trong 2

mô hình được thử nghiệm thì mô hình SVR nhìn chung cho kết quả tốt nhất đối với dự báo ngắn và dài hạn, trong khi đó mô hình RF lại cho thấy sự vượt trội ở dự báo trung hạn. Các mô hình AI thử nghiệm đều không dự báo chính xác một cách đồng nhất dòng chảy lũ do các mô hình không được huấn luyện tập trung vào dự báo dòng chảy lũ mà ưu tiên vào quá trình dòng chảy. Bên cạnh đó, việc lựa chọn số liệu đầu vào có độ tương quan cao với lưu lượng đến hồ đóng vai trò quan trọng trong việc nâng cao hiệu quả dự báo của mô hình. Đây hoàn toàn có thể là một phương án bổ sung cho công tác dự báo lưu lượng tới hồ bên cạnh các cách tiếp cận đang được sử dụng hiện nay[44].

Các biến đầu vào được chọn tương ứng với các trường hợp tính toán như sau:

TH1: Sử dụng số liệu mưa và bốc hơi của kỳ tính toán (P(t) và E(t)) tại các trạm Tuy Hòa, Sơn Hòa, Mdrak, Củng Sơn, số liệu lưu lượng trung bình của 2 kỳ trước đó tại trạm Sông Hinh ($Q(t-_2)$ và $Q(t-_1)$), và số liệu lưu lương lớn nhất và nhỏ nhất của kỳ trước đó ($Qmax(t-_1)$, $Qmin(t-_1)$).

TH2: Sử dụng số liệu mưa và bốc hơi của kỳ tính toán (P(t) và E(t)) tại các trạm Tuy Hòa, Sơn Hòa, Mdrak, Củng Sơn, số liệu lưu lượng trung bình, lớn nhất và nhỏ nhất của kỳ trước đó tại trạm Sông Hinh (Q(t-1), Qmax(t-1), Qmin(t-1)).

TH3: Sử dụng số liệu mưa và bốc hơi của kỳ tính toán (P(t) và E(t)) tại các trạm Tuy Hòa, Sơn Hòa, Mdrak, Củng Sơn, số liệu lưu lượng trung bình của kỳ trước đó tại trạm Sông Hinh (Q(t-1)).

Nhằm đánh giá hiệu quả của các mô lình, các thông số chính của hai mô hình sẽ được tối ưu bằng công cụ GridSearchCV sẵn có trong thư viện scikit-learn. Sau khi được hiệu chỉnh bằng GridSearchCV, các thông số tối ưu của mô hình được trình bày trong Bảng dưới đây.

Bảng 1.9: Các thông số tối ưu của các mô hình trong 3 trường hợp tính toán

Thông số	Mô hình SVR			Thông số	Mô hình RF		
	TH1	TH2	TH3		TH1	TH2	TH3
kernel	rbf	rbf	rbf	*n_estimators*	50	50	50
gamma	0,01	0,01	0,01	*max_depth*	8	8	15
C	5	5	10				
epsilon	0,1	0,1	0,1				

Sau khi có được bộ thông số tối ưu, các mô hình được áp dụng cho bộ dữ liệu kiểm tra từ tháng 01/2014 đến tháng 12/2017. Đây là chuỗi dữ liệu mà mô hình chưa "nhìn thấy" (unseen data), do đó kết quả dự báo của mô hình trên chuỗi dữ liệu này sẽ được dùng để đánh giá hai mô hình thử nghiệm trong nghiên cứu. Các nội dung

đánh giá bao gồm: (i) đánh giá kết quả dự báo quá trình dòng chảy; (ii) đánh giá kết quả dự báo theo mùa; (iii) đánh giá kết quả dự báo đỉnh lũ tiêu biểu. Kết quả tính toán cho thấy diễn biến dòng chảy trong giai đoạn kiểm tra được cả hai mô hình dự báo với độ chính xác cao. Các chỉ số thống kê đều đạt mức tốt với NSE dao động từ 0,84 - 0,93 và RMSE dao động từ 31,98 đến 60,24 (Bảng 1.10). Có thể thấy rằng các mô hình cho kết quả dự báo chính xác hơn ở TH2 và TH3.

Bảng 1.10: Tổng hợp kết quả đánh giá khả năng dự báo dòng chảy của hai mô hình

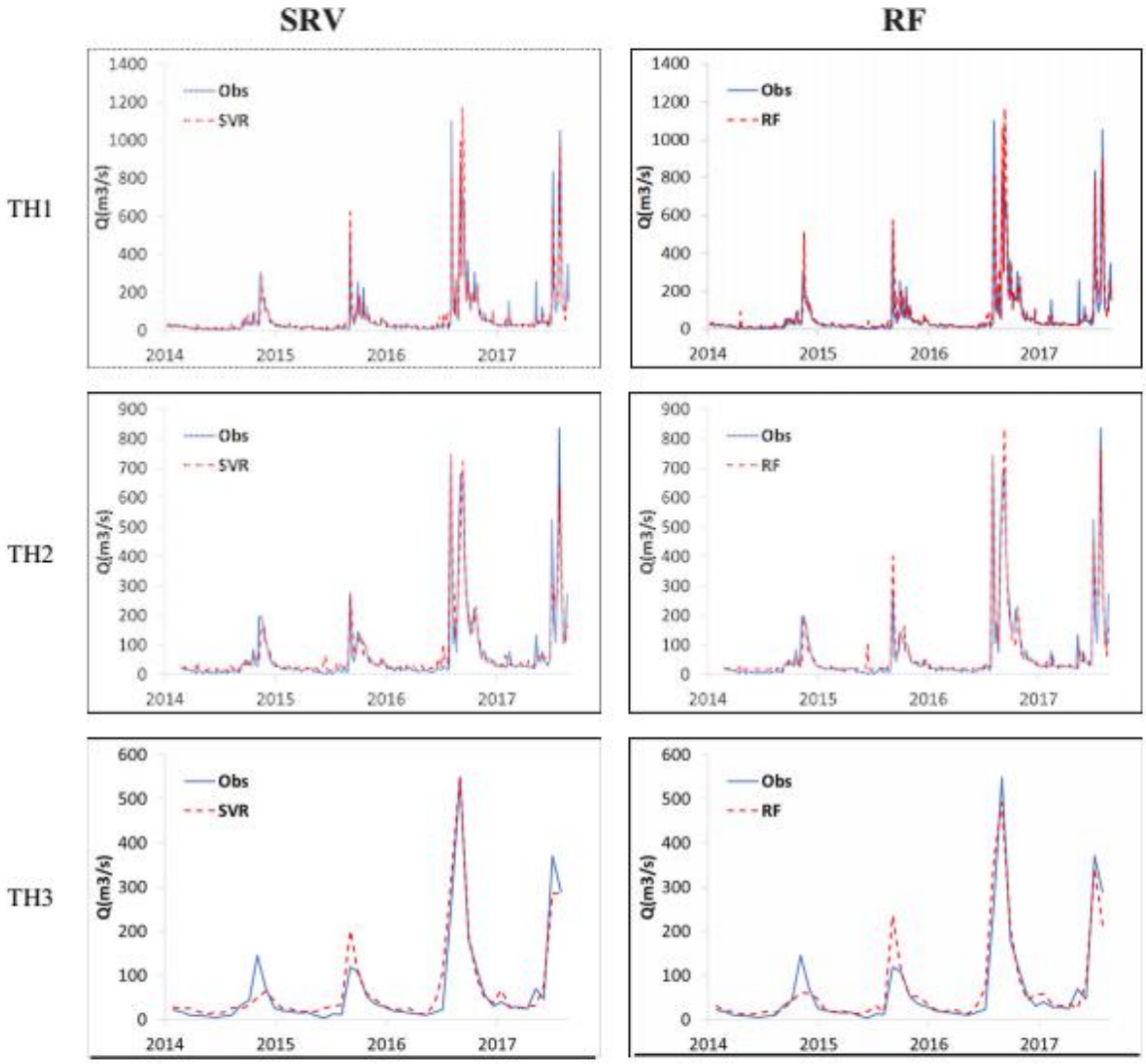
	NSE			RMSE		
	TH1	TH2	TH3	TH1	TH2	TH3
SVR	0,85	0,89	0,93	53,37	45,65	30,88
RF	0,84	0,92	0,91	60,24	40,91	31,98

Trên thực tế, việc đánh giá kết quả dự báo theo mùa được thực hiện cho hai giai đoạn: tháng 1 - 5, giai đoạn khô hạn nhất và tháng 9 - 12, giai đoạn xảy ra nhiều trận lũ nhất, của giai đoạn kiểm định (2104 - 2017).

Bảng 1.11: Tổng hợp kết quả đánh giá khả năng dự báo theo mùa của hai mô hình

Mùa		NSE			RMSE		
	Mô hình	TH1	TH2	TH3	TH1	TH2	TH3
Mùa khô	SVR	0,89	0,90	0,93	15,24	14,68	11,01
	RF	0,85	0,86	0,90	18,08	18,17	12,79
Mùa mưa	SVR	0,81	0,87	0,90	89,82	67,29	48,45
	RF	0,79	0,90	0,88	95,77	57,10	53,78

Kết quả kiểm nghiệm cho thấy, trong cả ba trường hợp, mô hình SVR chiếm ưu thế khi cho kết quả dự báo tốt hơn, đặc biệt là trong mùa khô, chỉ duy nhất ở TH2 mô hình RF có kết quả dự báo tốt hơn trong mùa mưa.



Hình 1.17: Kết quả dự báo lưu lượng vào hồ của hai mô hình SVR và RF theo 3 trường hợp tính toán trong giai đoạn kiểm tra từ 01/2014 - 12/2017

Nghiên cứu đã bước đầu thử nghiệm thành công hai mô hình AI là SVR và RF trong dự báo lưu lượng đến hồ, áp dụng cho hồ Sông Hinh thuộc lưu vực sông Ba. Ba trường hợp tính toán là dự báo báo dòng chảy trung bình 3 ngày, trung bình 7 ngày và trung bình 1 tháng tương ứng với dự báo ngắn hạn, trung hạn và dài hạn, đã được thử nghiệm. Kết quả cho thấy, cả hai mô hình ở cả ba trường hợp đều cho kết quả có độ chính xác khá cao đặc biệt là đối với trường hợp dự báo lưu lượng trung bình 7 ngày và 1 tháng. Trong 2 mô hình được thử nghiệm thì mô hình SVR nhìn chung cho kết quả tốt nhất đối với dự báo ngắn và dài hạn, trong khi đó mô hình RF lại cho thấy sự vượt trội ở dự báo trung hạn. Đối với dự báo theo mùa, các mô hình cho kết quả dự báo tốt trong cả mùa khô (tháng 1-5) và mùa mưa (tháng 9-12) với kết quả nhỉnh hơn trong mùa khô một điểm đáng chú ý là, các mô hình AI đều không dự báo chính xác một cách đồng nhất dòng chảy lũ. Lý do của hiện tượng này là các mô hình không được huấn luyện tập trung vào dự báo dòng chảy lũ mà ưu tiên vào quá trình dòng chảy. Kết quả tính toán ở trường hợp dự báo dòng chảy trung bình 3

ngày có độ chính xác thấp hơn đáng kể so với hai trường hợp còn lại, điều này là do ở bước thời gian này sự dao động trong dữ liệu cao hơn các trường hợp dữ liệu trung bình tuần hay tháng. Kết quả dự báo có thể được cải thiện nếu có dữ liệu có độ dài và chất lượng tốt hơn. Ngoài ra, việc lựa chọn dữ liệu đầu vào phù hợp là yếu tố rất quan trọng quyết định nên hiệu quả dự báo của mô hình. Trong đó, dòng chảy trong quá khứ là một trong những biến đầu vào quan trọng. Bên cạnh đó, số lượng dữ liệu đầu vào (số lượng features) cần phải đủ lớn để hỗ trợ cho mô hình AI trong việc khái quát hóa được mối quan hệ giữa biến đầu vào và dòng chảy đầu ra. Dựa trên những phân tích và đánh giá đã thực hiện, nhóm nghiên cứu đề xuất sử dụng mô hình SVR cho dự báo lưu lượng trung bình 3 ngày và 1 tháng, và RF cho dự báo lưu lượng trung bình 1 tuần. Tuy nhiên, đối với từng trường hợp dự báo, hay yêu cầu dự báo (đỉnh lũ, dòng chảy mùa khô, dòng chảy mùa mưa) có thể sử dụng các mô hình khác thay thế cho các mô hình được đề xuất do hiệu quả dự báo là khá tương đồng như đã phân tích ở các phần trên. Như vậy, bên cạnh các phương pháp truyền thống, các mô hình AI như SVR và RF sẽ cung cấp một công cụ mới, hiệu quả để bổ trợ cho công tác quản lý và vận hành hồ chứa nói chung và hồ sông Hinh nói riêng. Tuy nhiên, việc ứng dụng trong tương lai phụ thuộc rất lớn vào điều kiện và chất lượng số liệu ở địa điểm áp dụng [44].

1.3.6. Phát triển mô hình AI để đo ô nhiễm không khí

Phát triển mô hình AI để đo ô nhiễm không khí giao thông từ cảm biến đa cảm biến và dữ liệu thời tiết là kết quả nghiên cứu của Hai - Bang Ly thuộc Trường Đại học Công nghệ Giao thông vận tải, Hà Nội và các cộng sự thuộc Đại học Nông nghiệp, Đại học Đà Nẵng và Đại học Illinois tại Chicago, Chicago, IL 60607, Hoa Kỳ. Mục tiêu chính của nghiên cứu này là khám phá tác động của một số biến đầu vào trong việc đào tạo các chỉ số chất lượng không khí khác nhau bằng cách sử dụng logic mờ kết hợp với hai tối ưu hóa metaheuristic: ủ mô phỏng (SA) và tối ưu hóa bầy hạt (PSO). Trong công trình này, nồng độ của NO_2 và CO được dự đoán bằng cách sử dụng năm điện trở suất từ các thiết bị cảm ứng đa điểm và ba biến thời tiết (nhiệt độ, độ ẩm tương đối và độ ẩm tuyệt đối). Để xác nhận kết quả, một số biện pháp đã được tính toán, bao gồm hệ số tương quan và sai số tuyệt đối trung bình. Nhìn chung, PSO được cho là hoạt động tốt nhất. Cuối cùng, điện trở suất đầu vào của NO 2 và hydrocacbon phi kim (NMHC) được cho là nhạy nhất để dự đoán nồng độ của NO 2 và CO [117].

a) Phương pháp được sử dụng

Các phương pháp học máy được sử dụng gồm:

- Hệ thống suy luận mờ dựa trên mạng thích ứng: Thuật toán ANFIS kết hợp hệ thống mờ với mạng nơ-ron.

- Phương pháp tối ưu bầy đàn: PSO là phương pháp tiến hóa được sử dụng phổ biến nhất để tối ưu hóa tham số.

- Ủ mô phỏng: Quá trình ủ mô phỏng được phát triển sau PSO và nó đã trở thành một công cụ mạnh mẽ để tối ưu hóa toàn cầu.

Mẫu chính thức: Hiệu suất của mô hình chủ yếu được đánh giá bằng cách sử dụng ba phép đo thống kê: sai số tuyệt đối trung bình (MAE), sai số bình phương trung bình gốc (*RMSE*) và hệ số tương quan (*R*). Giá trị của *R* nằm trong khoảng từ 0 đến 1; giá trị *R* cao hơn (tức là gần 1) cho thấy hiệu suất tốt hơn. Ngược lại, giá trị *RMSE* và *MAE* thấp hơn cho thấy hiệu suất tốt hơn [117].

b) Tập dữ liệu

Dữ liệu về chất lượng không khí rất dồi dào, nhưng bộ dữ liệu đa biến lớn để phát triển các mô hình thì hạn chế. Trong nghiên cứu, nhóm tác giả sử dụng dữ liệu thu thập được từ tháng 3/2004 đến tháng 2/2005 tại một thành phố ô nhiễm do giao thông của Ý; dữ liệu có sẵn trong truy cập mở từ kho lưu trữ máy học của Đại học California, Irvine (UCI). Tập dữ liệu ban đầu chứa 9357 bản ghi, máy phân tích đã không hoạt động và dữ liệu tương ứng phải được xóa. Một thiết bị đa cảm ứng đã được sử dụng để cung cấp giá trị trung bình hàng giờ của điện trở suất biểu thị bằng CO-, NO_x-, O_3- và NO_2- chất MOX cụ thể, một cảm biến MOX có mục tiêu là hydro-cacbon (NMHC) phi kim loại. Thiết bị đa cảm biến cũng gắn các cảm biến để đo nhiệt độ và độ ẩm tương đối và tuyệt đối. Cuối cùng, các tham số đầu vào chứa 6941 phản hồi từ 8 đầu vào. Song song, năm trạm cố định thường xuyên cung cấp ước tính nồng độ tham chiếu cho CO (mg/m^3), NMHC (g/m^3), C_6H_6 (g/m^3), NO_x (ppb) và NO_2 (g/m^3). Những kết quả này được coi là đầu ra của hệ thống, được ghi lại hàng giờ bằng cách lấy giá trị trung bình của nồng độ. Trong khi tập dữ liệu ban đầu có năm đầu ra, nhóm tác giả chỉ tập trung vào ước tính nồng độ của NO_2 và CO. Bảng 1.12 thống kê tóm tắt tất cả các biến được sử dụng trong nghiên cứu [117].

Bảng 1.12: Tham số tập dữ liệu và phân tích thống kê

Thông số	**Cảm biến CO**	**Cảm biến NMHC**	**Cảm biến NO_x**	**Cảm biến SO_2**	**Cảm biến O_3**	**Nhiệt độ**	**Độ ẩm tương đối**	**Độ ẩm tuyệt đối**	**C (NO_2) ***	**C (CO) ****
Vai trò	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu ra	Đầu ra
Ký hiệu	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Min (α)	647	390	322	551	221	−1,9	9.2	0,18	2	0,1
Trung	1120	959	817	1453	1058	17,8	48,9	0,99	114	2,18

Thông số	Cảm biến CO	Cảm biến NMHC	Cảm biến NO_x	Cảm biến SO_2	Cảm biến O_3	Nhiệt độ	Độ ẩm tương đối	Độ ẩm tuyệt đối	C (NO_2) *	C (CO) **
bình cộng										
Trung bình	1085	931	786	1457	1006	16,8	49,2	0,95	110	1,90
Tối đa (β)	2040	2214	2683	2775	2523	44,6	88,7	2,2	333	11,9
Std	219	264	252	353	407	8,84	17.4	0,40	47	1,44
CV (%)	20	28	31	24	38	50	36	41	42	66

c) Kết quả và thảo luận

Quy trình tối ưu hóa

Có 250 tham số ANFIS được tối ưu hóa, tương ứng với không gian đầu vào tám chiều. Các tham số của ANFIS được tạo bằng cách sử dụng phân cụm C-mean. Bảng 1.13 và 1.14 hiển thị các tham số về mô phỏng ủ và tối ưu hóa cuối cùng được chọn cho SA và PSO tương ứng.

Bảng 1.13: Các thông số của ủ mô phỏng (SA)

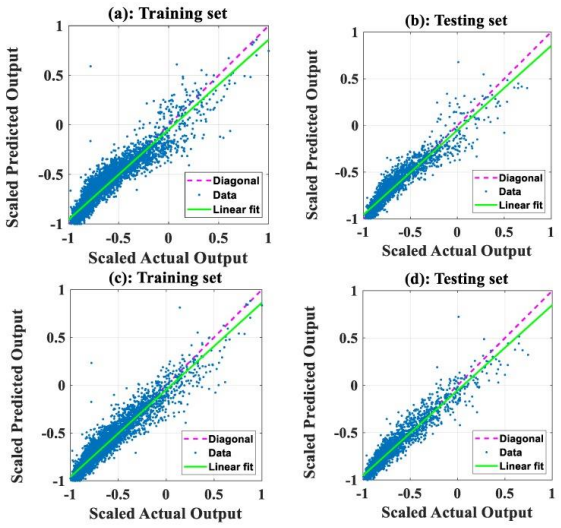
Tham số	Nồng độ NO_2	Nồng độ CO
Quy mô dân số	40	60
Số lần lặp lại tối đa	1000	2000
Nhiệt độ ban đầu	0,1	0,1
Tỷ lệ giảm nhiệt độ	0,99	0,99
Số hàng xóm trên mỗi cá nhân	5	5
Tỷ lệ đột biến	0,5	0,5
Độ lệch chuẩn đột biến	10%	10%

Bảng 1.14: Các thông số về tối ưu hóa bầy hạt (PSO) được sử dụng

Tham số	Nồng độ NO_2	Nồng độ CO
Kích thước bầy đàn	30	50
Số lần lặp lại tối đa	1000	2000
Quán tính trọng lượng	0,4	0,4
Hệ số học tập cá nhân	1	1
Hệ số học tập toàn cầu	2	2
Vận tốc tối đa	5	5
Vận tốc tối thiểu	−5	−5

Hiệu suất mô hình

Hiệu suất của hai mô hình đã phát triển được tóm tắt trong Bảng 1.15. Hình 1.18 là khả năng dự đoán nồng độ NO 2 ở dạng hồi quy. Các số liệu sử dụng ANFIS-SA cho (a) tập dữ liệu đào tạo và (b) tập dữ liệu thử nghiệm. Các số liệu sử dụng ANFIS-PSO cho (c) tập dữ liệu đào tạo và (d) tập dữ liệu kiểm tra. Kết luận, cả hai mô hình đều hoạt động tốt và có ý nghĩa thống kê, ANFIS-PSO được chứng minh là vượt trội hơn so với ANFIS-SA về nồng độ CO và NO2 trong mô hình.



Hình 1.18: Khả năng dự đoán nồng độ NO_2 ở dạng hồi quy

Bảng 1.15: Thông tin tóm tắt về khả năng dự đoán của dữ liệu được chia tỷ lệ

Đầu ra	Dataset	Mô hình	*R*	*RMSE*	*MAE*	Lỗi Std	Dốc
Nồng độ NO_2	Đào tạo	ANFIS-SA	0,934	0,103	0,075	0,102	42,23
		ANFIS-PSO	0,950	0,090	0,063	0,089	42.34
	Thử nghiệm	ANFIS-SA	0,935	0,101	0,075	0,100	42,11
		ANFIS-PSO	0,951	0,088	0,064	0,087	42.03
Nồng độ CO	Đào tạo	ANFIS-SA	0,885	0,134	0,100	0,134	37,73
		ANFIS-PSO	0,910	0,119	0,088	0,119	39,51
	Thử nghiệm	ANFIS-SA	0,883	0,135	0,102	0,135	37,65
		ANFIS-PSO	0,907	0,121	0,090	0,121	39,16

Phân tích độ nhạy

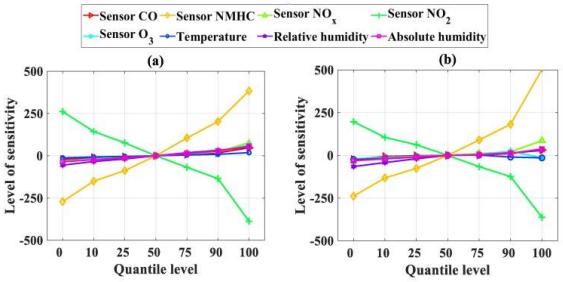
Bảng 1.16 tóm tắt các giá trị của mỗi đầu vào theo tỷ lệ phần trăm, Bảng 1.17 tóm tắt giải pháp đầu ra của các mô hình AI đã phát triển tương ứng với từng phân vị. Độ nhạy như một hàm của phân vị, được vẽ trong Hình 1.19 cho NO_2 và trong Hình 1.20 đối với CO ((a) sử dụng ANFIS-SA và (b) sử dụng ANFIS-PSO). Các thông số đầu vào khác có tác động thấp đến kết quả dự đoán so với cảm biến NMHC và NO_2.

Bảng 1.16: Các giá trị đầu vào mô hình

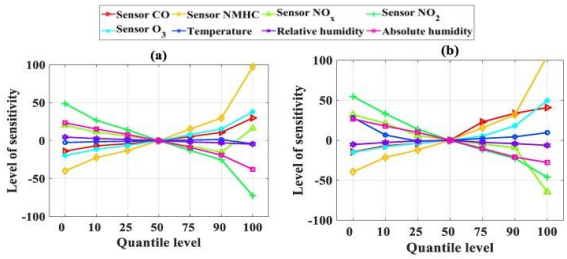
Biến / Phần trăm	P 0	P 10	P 25	P 50	P 75	P 90	P 100
Cảm biến CO	−1,00	−0,68	−0,55	−0,38	−0,13	0,13	1,00
Cảm biến NMHC	−1,00	−0,74	−0,61	−0,42	−0,19	0,02	1,00
Cảm biến NO_x	−1,00	−0,83	−0,73	−0,61	−0,47	−0,31	1,00
Cảm biến SO_2	−1,00	−0,64	−0,43	−0,19	0,02	0,22	1,00
Cảm biến O_3	−1,00	−0,72	−0,54	−0,32	−0,05	0,22	1,00
Nhiệt độ	−1,00	−0,64	−0,44	−0,20	0,10	0,37	1,00
Độ ẩm tương đối	−1,00	−0,60	−0,32	0,04	0,37	0,64	1,00
Độ ẩm tuyệt đối	−1,00	−0,73	−0,50	−0,23	0,06	0,39	1,00

Bảng 1.17: Các giá trị đầu ra mô hình

Đầu ra	Mô hình được sử dụng	Biến đổi	Q 0	Q 10	Q 25	Q 75	Q 90	Q 100
Nồng độ NO	ANFIS-SA	Cảm biến CO	−21,37	−10,57	−6,11	8,44	17,36	47.12
		Cảm biến NMHC	−270,95	−150,42	−87,62	105,13	203.01	383,83
		Cảm biến NO_x	−31,77	−17,89	−9,63	10,89	23,76	77,76
		Cảm biến SO_2	261,32	143,70	76,46	−68,35	−134,97	−387,09
		Cảm biến O_3	−32,66	−18,99	−10,31	13,29	26.12	64,10
		Nhiệt độ	−12,33	−6,84	−3,70	4,62	8,77	18.41
		Độ ẩm tương đối	−55,45	−33,83	−18,95	17,94	32,28	51,37
		Độ ẩm tuyệt đối	−35,46	−23,21	−12,45	13,31	28.47	56,79



Hình 1.19: Phân tích độ nhạy đối với nồng độ NO_2



Hình 1.20: Phân tích độ nhạy đối với nồng độ CO

Kết luận: Nghiên cứu đã phát triển một mô hình AI để dự đoán đáng tin cậy nồng độ NO_2 và CO hàng giờ từ cảm biến đa cảm biến khí và dữ liệu thời tiết. 8 biến đầu vào đã được sử dụng, bao gồm 5 biến cảm biến khí và 3 biến thời tiết. Hai mô hình AI đã được đào tạo và thử nghiệm là ANFIS-PSO và ANFIS-SA. Kết quả cho thấy cả hai mô hình đều hoạt động tốt và có ý nghĩa thống kê nhưng ANFIS-PSO hoạt động tốt hơn một chút. Để khảo sát thêm vai trò của từng biến đầu vào riêng lẻ trong các mô hình AI, một phân tích độ nhạy chi tiết đã được thực hiện, cảm biến NMHC và NO_2 đặc biệt ảnh hưởng đến độ nhạy của cả hai mô hình nồng độ NO_2 và CO. Mô hình nồng độ CO được chứng minh là nhạy cảm hơn với tất cả các biến. Tuy nhiên, ba biến thời tiết không ảnh hưởng quá nhiều đến độ chính xác của mô hình.

1.4. Kết chương 1

Các phân tích, đánh giá tổng quan ở trên khẳng định tầm quan trọng của công nghệ AI, Big data và ngày càng được phát triển mạnh mẽ trên thế giới và trong nước. Các ứng dụng của AI và Big data phong phú, đa dạng cho nhiều lĩnh vực của đời sống kinh tế xã hội. Hiện tại và xu hướng trong tương lai, công nghệ AI và Big data sẽ ứng dụng nhiều trong hoạt động KTTV bao gồm cả công tác đo đạc, quan trắc, xử lý thông tin dữ liệu và dự báo, cảnh báo KTTV và các thiên tai nguy hiểm. Kết quả của phần nghiên cứu tổng quan về công nghệ AI, Big data rất quan trọng, cung cấp thông tin nền tảng cơ sở để nghiên cứu các nội dung tiếp theo của đề tài.

Phần tiếp theo sẽ trình bày chi tiết về phạm vi và đối tượng nghiên cứu, mô tả tập số liệu và phương pháp nghiên cứu. Trong đó, sẽ đi sâu vào phân tích chuỗi số liệu KTTV quá khứ, hiện tại phục vụ nghiên cứu; phân tích các thành phần kỹ thuật, giải pháp của hệ thống Big data KTTV; phân tích các thành phần, yêu cầu kỹ thuật của công nghệ học máy (ML) và AI để nhận dạng, hỗ trợ dự báo, cảnh báo KTTV.

2. CHƯƠNG 2: PHẠM VI, ĐỐI TƯỢNG, SỐ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Trong chương này, chúng tôi sẽ tập trung trình bày về đối tượng, phạm vi nghiên cứu, các nguồn số liệu và phương pháp, công nghệ sử dụng trong nghiên cứu, trong đó tập trung vào phân tích phương pháp thu thập, xử lý số liệu phục vụ nghiên cứu. Kế tiếp, là phân tích về thành phần, giải pháp, yêu cầu kỹ thuật của công nghệ Big data KTTV. Cuối cùng, là luận giải về nghiên cứu cho công nghệ học máy, AI để nhận dạng, hỗ trợ dự báo KTTV sẽ được đi sâu phân tích.

2.1. Đối tượng, phạm vi và sơ đồ nghiên cứu

a) Đối tượng nghiên cứu: các hiện tượng KTTV nguy hiểm gồm bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão.

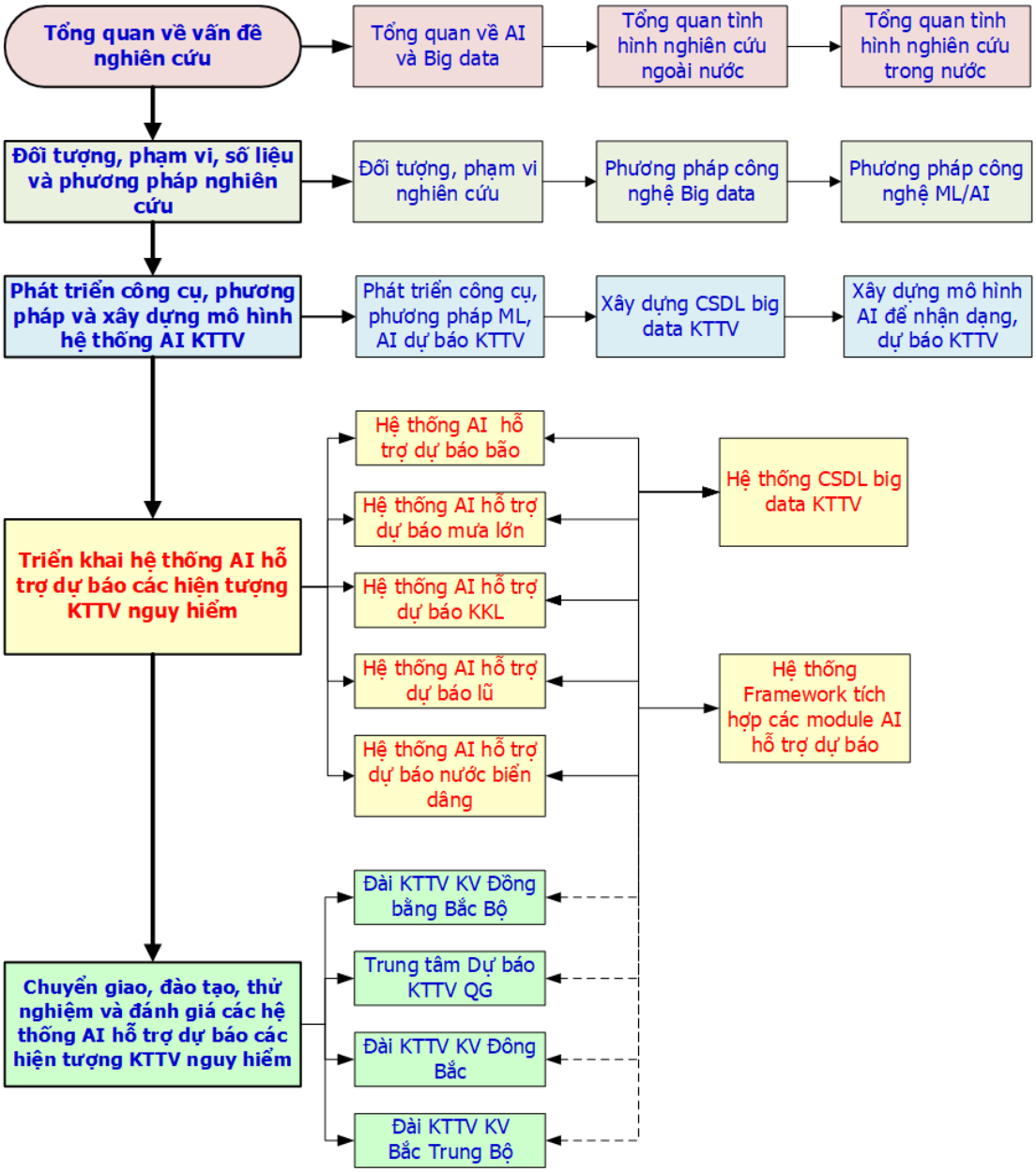
b) Phạm vi nghiên cứu:

- Về ứng dụng AI trong việc nhận dạng hình thế và hỗ trợ dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh khu vực miền Bắc trong giai đoạn 10 năm, từ năm 2008 - 2017.

- Về ứng dụng AI việc nhận dạng hình thế và hỗ trợ dự báo, cảnh báo lũ trên hệ thống sông Hồng trong giai đoạn 10 năm, từ năm 2008 - 2017.

- Về ứng dụng AI trong việc nhận dạng và hỗ trợ cảnh báo, dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ trong giai đoạn 10 năm, từ năm 2008 - 2017.

c) Sơ đồ nghiên cứu: Để đạt được các mục tiêu của đề tài, các nội dung nghiên cứu thực hiện theo hình 2.1.



Hình 2.1: Sơ đồ nghiên cứu thực hiện đề tài

2.2. Số liệu phục vụ nghiên cứu

Trong ứng dụng công nghệ AI để nhận dạng hình thế và hỗ trợ dự báo, cảnh báo hiện tượng KTTV nguy hiểm, số liệu KTTV đóng một vai trò hết sức quan trọng; trong phạm vi nghiên cứu của đề tài, nguồn số liệu được thu thập gồm:

- Số liệu khí tượng bề mặt;
- Số liệu thủy văn và hải văn;
- Dữ liệu tái phân tích của các yếu tố khí tượng hải văn và dữ liệu vệ tinh;
- Dữ liệu về bão ảnh hưởng đổ bộ vào Việt Nam;
- Dữ liệu về lũ và các dữ liệu liên quan;

- Dữ liệu về mưa lớn diện rộng;
- Dữ liệu về không khí lạnh.

Chi tiết về số liệu phục vụ nghiên cứu như sau:

2.2.1. Số liệu khí tượng bề mặt

- Các loại số liệu khí tượng bề mặt thu thập gồm: Nhiệt độ trung bình ngày, nhiệt độ tối thấp ngày, nhiệt độ tối cao ngày, lượng mưa ngày, độ ẩm, gió (hướng và tốc độ), khí áp.

- Khối lượng, phạm vi thu thập số liệu: 187 trạm quan trắc khí tượng bề mặt trên phạm vi toàn quốc.

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu khí tượng bề mặt tại Bảng 2.1

Bảng 2.1: Minh họa thu thập số liệu nhiệt độ trung bình ngày của trạm Sơn La

Trạm: Sơn La Tỉnh : Sơn La	KẾT QUẢ TÍNH TOÁN, CHỈNH LÝ, BIÊN TẬP TÀI LIỆU NHIỆT ĐỘ TRUNG BÌNH									Kinh độ: 103.54 Vĩ độ: 21.20 Đơn vị: oC		
Năm/ Ngày	***Tháng***											
2008	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
1	11.8	15.5	13.4	22.2	25.7	23.6	24.9	25	24	20.8	23.3	11.2
2	10.1	16.3	16.3	19.7	26.2	23.1	24.9	25.8	24.6	23.5	23.2	11.8
3	11.2	17.1	16	21.6	25.8	23.7	23.8	24.8	24.4	24.2	18.7	13.4
4	12.6	18.9	17.8	23.7	24.4	24.7	24.5	23.6	22.7	22.6	19.7	14.3
5	13	18.6	17	22.9	23.5	24.7	25.6	25.3	23.9	21.9	21.7	14.2
6	15.3	18.3	16.1	24.2	23.6	23.5	24.1	26.2	24.2	22.2	22.9	14.7
7	16.1	17.6	17.5	25.8	25.1	24.2	25.6	25.1	23.7	21.7	22.2	15
8	17	18.3	18.9	27.5	24.5	24.5	25.8	22.9	24.5	21.8	18.8	14.3
9	18.4	19.7	16.4	28.4	23.9	23.8	25.1	23.2	24.9	23	16.4	12

Kết quả thu thập, xử lý số liệu khí tượng bề mặt phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 2: Thu thập, biên tập, phân tích, đánh giá, tổng hợp số liệu khí tượng bề mặt: Nhiệt độ trung bình ngày, tối thấp ngày, tối cao ngày; lượng mưa ngày; độ ẩm, gió; khí áp.

2.2.2. Số liệu thủy văn và hải văn

- Các loại số liệu thủy văn và hải văn thu thập gồm:

+ Số liệu thủy văn: mực nước, lưu lượng, lũ của các trạm thủy văn trên lưu vực của hệ thống sông Hồng;

+ Số liệu hải văn: độ cao sóng, chu kỳ sóng, nước biển dâng do bão, mực nước thủy triều tại các trạm quan trắc.

+ Số liệu thiên tai thủy văn, hải văn: Số lần xảy ra, cường độ, phạm vi ảnh hưởng thiệt hại do thiên tai tại từng khu vực địa lý khí hậu, lưu vực sông chính, độ dài bờ biển, vùng biển.

- Khối lượng, phạm vi thu thập số liệu:

+ Dữ liệu thủy văn: 260 trạm trạm thủy văn trên hệ thống sông Hồng - Thái Bình và các trạm thủy văn trên phạm vi cả nước.

+ Dữ liệu hải văn: 06 trạm trên vùng biển Quảng Ninh đến Thanh Hóa.

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu thủy văn và hải văn tại Bảng 2.2.

Bảng 2.2: Minh họa thu thập số liệu mực nước ngày của trạm Mù Cang Chải

Trạm: Mù Cang Chải Sông : Nậm Kim				KẾT QUẢ TÍNH TOÁN, CHỈNH LÝ, BIÊN TẬP TÀI LIỆU MỰC NƯỚC						Kinh độ: 104.85 Vĩ độ: 21.85 Đơn vị: mm		
Năm/ Ngày	***Tháng***											
2011	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
1	0	0	0	23422	31230	31233	23440	23435	23427	23427	31233	23424
2	0	0	0	23422	31230	31233	23437	31239	31236	23426	23425	23424
3	0	0	0	23422	18744	31233	31239	31239	31236	23426	23425	23424
4	0	0	0	23422	13391	31233	31242	31247	31238	23426	23425	23424
5	0	0	0	23422	15619	31233	31238	31239	31237	23426	31233	23424
6	0	0	0	23422	31233	31238	31238	31238	31237	23426	23425	23424
7	0	0	0	23424	18741	31237	31237	31239	31236	23426	23425	23424
8	0	0	0	23423	31235	31233	18748	31238	31236	23426	31233	23424
9	0	0	0	23422	31235	31234	31242	31238	31236	23426	23424	23424

Kết quả thu thập, xử lý số liệu thủy văn, hải văn phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 3: Thu thập, biên tập, phân tích, đánh giá, tổng hợp số liệu thủy văn và hải văn.

2.2.3. Số liệu tái phân tích

- Các loại số liệu tái phân tích và vệ tinh thu thập gồm:

+ Dữ liệu tái phân tích: Gió, độ ẩm, áp suất, nhiệt độ không khí, nhiệt độ mặt nước biển, hơi nước cột tổng.

+ Dữ liệu vệ tinh: định dạng SATAID, có các kênh phổ IR1, IR2, IR3, VIS với tần suất 10 phút/ 1 kênh phổ dữ liệu.

- Phạm vi thu thập số liệu:

+ Dữ liệu tái phân tích: dữ liệu tái phân tích của Nhật Bản (JRA-55)

+ Dữ liệu vệ tinh: Các vệ tinh địa tĩnh Himawari được điều hành bởi Cơ quan Khí tượng Nhật Bản (JMA).

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu tái phân tích tại Bảng 2.3 và thống kê dung lượng thu thập dữ liệu vệ tinh tại Bảng 2.4.

Bảng 2.3: Minh họa thu thập các thông số của các trường phân tích đẳng áp

Mã	Thông số trường	Đơn vị	Tên file
1	Độ cao thế năng	gpm	anl_p125_hgt
2	Nhiệt độ	K	anl_p125_tmp
3	Độ lệch điểm sương	K	anl_p125_depr
4	Thành phần u của gió	$m\ s^{-1}$	anl_p125_ugrd
5	Thành phần v của gió	$m\ s^{-1}$	anl_p125_vgrd
6	Hàm dòng	$m^2\ s^{-1}$	anl_p125_strm

Bảng 2.4: Thống kê dung lượng thu thập dữ liệu ảnh vệ tinh Nhật Bản

STT	**Năm**	**Tổng số file**	**Dung lượng**	**STT**	**Năm**	**Tổng số file**	**Dung lượng**
1	2005	37,564	19 GB	8	2012	52,425	39,6 GB
2	2006	51,774	30,6 GB	9	2013	52,284	39,6 GB
3	2007	52,44	31,2 GB	10	2014	52,26	40,3 GB
4	2008	52,572	31,7 GB	11	2015	102,12	239 GB
5	2009	52,44	31,8 GB	12	2016	155,6	453 GB
6	2010	52,23	35 GB	13	2017	155,2	458 GB
7	2011	52,179	39,9 GB	14	2018	89,949	264 Gb

Kết quả thu thập, xử lý số liệu tái phân tích và dữ liệu vệ tinh phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 4: Thu thập, biên tập, phân tích, đánh giá, tổng hợp dữ liệu tái phân tích của các yếu tố khí tượng và dữ liệu vệ tinh.

2.2.4. Số liệu bão

- Các loại số liệu bão thu thập gồm: thống kê và thông tin chi tiết về các cơn bão, áp thấp nhiệt đới; số liệu về nước biển dâng do bão.

- Khối lượng, phạm vi thu thập số liệu: 103 cơn bão đổ bộ hoặc ảnh hưởng ở miền Bắc, miền Trung và miền Nam;

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

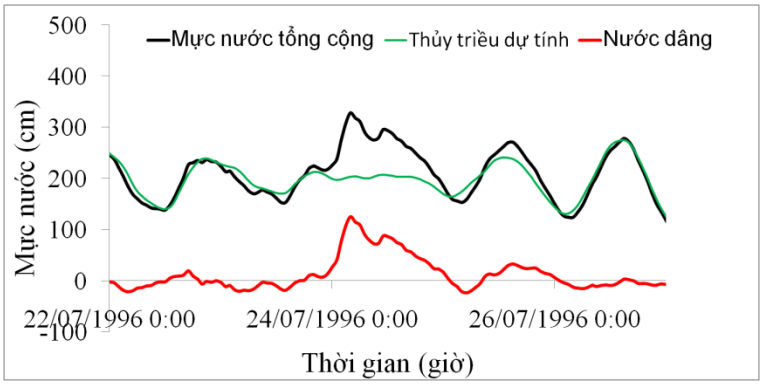
- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu bão, ATNĐ tại Bảng 2.5.

Bảng 2.5: Minh họa thông tin chi tiết của 1 cơn bão, ATNĐ

STT	QT Name	VN Name	Long	Lat	Pmin	Vmax	R1	R2	R3	Thời gian
1	MANGKHUT	SIEU BAO	139,8	14	915	105	30	110	200	9/11/2018 0:00
2	MANGKHUT	SIEU BAO	138,6	13,7	915	105	30	110	200	9/11/2018 0:00
3	MANGKHUT	SIEU BAO	137,5	13,9	915	105	30	110	200	9/11/2018 0:00
4	MANGKHUT	SIEU BAO	136,2	13,9	905	110	50	120	240	9/12/2018 0:00
5	MANGKHUT	SIEU BAO	135,2	14	905	110	50	150	250	9/12/2018 0:00

Minh họa kết quả thu thập, xử lý số liệu nước biển dâng do bão tại Hình 2.2.



Hình 2.2: Dao động của mực nước, thủy triều và nước dâng sau bão tại trạm Hòn Dấu trước và sau khi bão Frankie-1996 đổ bộ vào bờ

Kết quả thu thập, xử lý số liệu bão phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 5: Thu thập, biên tập, phân tích, đánh giá, tổng hợp số liệu bão, nước biển dâng do bão.

2.2.5. Số liệu lũ

- Các loại số liệu lũ và hồ thủy điện, thủy lợi thu thập gồm: Đặc điểm thủy văn mùa lũ, số trận lũ, mực nước cao nhất năm, thời gian xuất hiện đỉnh lũ, một số trận lũ điển hình trên các sông chính, lũ quét.

- Phạm vi thu thập số liệu: Trên các sông chính ở Bắc Bộ.

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu lũ tại các Bảng 2.6, 2.7 và 2.8.

Bảng 2.6: Phân bố các trận lũ (biên độ trên 1m) năm 2008

TT	**Sông**	**Tháng**								**Cộng**
		V	VI	VII	VIII	IX	X	XI	XII	
1	Hồng	1	2	3	2	1	1	-	-	10
2	Đà	1	3	4	1	1	1	1	-	12
3	Thao	2	3	2	2	1	1	1		12
4	Lô	2	3	4	2	1	-		-	12
5	Thái Bình	-	-	2	1	2	-	1	-	6
6	Câu	-	-	2	1	2	-	1	-	6
7	Thương	-	-	2	1	2	-	1	-	6
8	Lục Nam	-	-	2	1	2	-	1	-	6
9	Hoàng Long	-	-	-	1	1	1		-	3
Cộng		**6**	**11**	**21**	**12**	**13**	**4**	**6**	**0**	**73**

Bảng 2.7: Mực nước cao nhất năm 2008 trên các sông chính tại Bắc Bộ

TT	Sông	Trạm	Hmax 2008		Mức BĐ		So với TBNN		So với 2007	
			Ngày tháng	Mực nước (cm)	Cấp báo động	Cao hơn BĐ (cm)	TBNN (cm)	Chênh lệch (cm)	Cao nhất 2007 (cm)	Chênh lệch (cm)
1	Đà	Hoà Bình	11/VIII	11800	II	1800	11000	800	14500	-2700
2	Thao	Yên Bái	10/VIII	3426	m	226	3075	351	3138	288
3	Lô	T. Quang	9/VIII	2584	II	184	2533	51	2217	367
4	Hông	Hà Nội	11/VIII	1211	III	61	1077	134	993	218
5	Câu	Đáp Câu	4/XI	607	III	27	623	-16	394	213
6	Thương	P.LThương	28/IX	689	III	109	610	79	393	296
7	Lục Nam	Lục Nam	27/IX	788	III	208	607	181	476	312
8	Thái Bình	Phả Lại	28/IX	539	II	89	535	4	399	140
9	H. Long	Bến Đế	1/XI	469	III	69	348	121	517	-48

Bảng 2.8: Minh họa số liệu thu thập hồ thủy điện, thủy lợi trên hệ thống sông Hồng

STT	**Tên tỉnh**	**Tên hồ**	**Thiết kế**			
			MNDBT(m)	W(triệu m^3)	H max(m)	Flv(km2)
1	Lai Châu	Bản Chát	475,00	2137,70	130,00	
2		Huội Quảng				
3		Lai Châu				
4	Điện Biên	Pa Khoang	922,40	37,20	26,00	77,20
5		Nậm Ngám	1140,66	5,94	29,00	14,80

STT	Tên tỉnh	Tên hồ	Thiết kế			
			MNDBT(m)	W(triệu m^3)	H max(m)	Flv(km2)
6		Pe Luông	525,22	3,28	26,36	23,50
7		Hồng Sạt	480,20	2,79	23,00	8,60
8		Hồng Khếnh	515,30	2,58	23,00	5,40
9		Bản Ban	638,50	1,77	30,00	22,70

Kết quả thu thập, xử lý số liệu lũ phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 6: Số liệu về lũ, hồ thủy điện và các dữ liệu liên quan trên hệ thống sông Hồng.

2.2.6. Số liệu về mưa lớn diện rộng

- Các loại số liệu mưa lớn diện rộng thu thập gồm: Đặc trưng về số đợt mưa lớn diện rộng, đặc trưng phân bố mưa lớn theo tháng, đặc trưng về tổng lượng mưa, đặc trưng hình thế Synop chính gây mưa lớn, đặc trưng về số ngày kéo dài của các đợt mưa lớn diện rộng.

- Phạm vi thu thập số liệu: Các trạm đo mưa trên phạm vi cả nước.

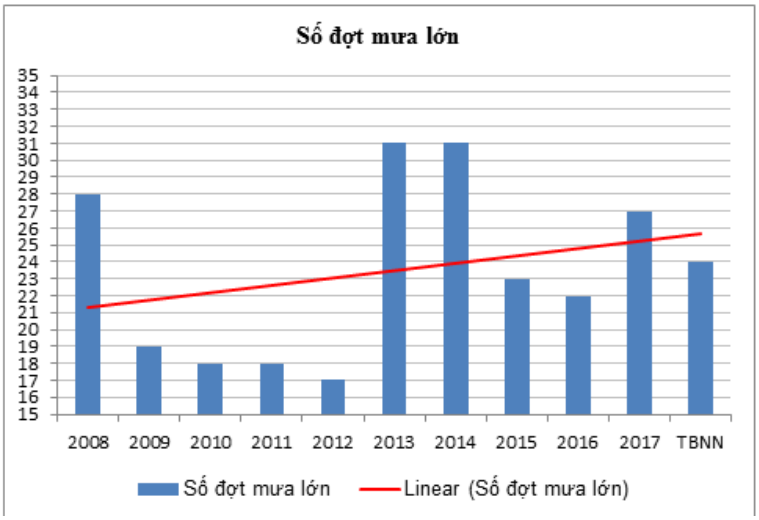
- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu lũ tại Bảng 2.9 và Hình 2.3.

Bảng 2.9: Các đợt mưa lớn diện rộng theo tháng của các khu vực từ năm 2008-2017

Khu vực	Tháng									
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Bắc Bộ	2	6	12	12	31	19	22	9	4	2
Bắc TB	2	2	2	7	15	9	23	17	6	0
Trung TB	2	1	2	0	3	2	17	28	24	7
Nam TB	0	2	2	0	2	2	6	16	20	6
Tây Nguyên	0	2	5	8	9	3	12	10	7	1
Nam Bộ	0	1	2	9	7	2	11	6	2	0
Tổng	**6**	**14**	**25**	**36**	**67**	**37**	**91**	**86**	**63**	**16**



Hình 2.3: Biểu đồ số đợt mưa lớn từ năm 2008 -2017 và TBNN

Kết quả thu thập, xử lý số liệu mưa lớn diện rộng phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 7: Thu thập, biên tập, phân tích, đánh giá, tổng hợp số liệu mưa lớn diện rộng.

2.2.7. Số liệu về không khí lạnh

- Các loại số liệu không khí lạnh thu thập gồm: các đợt gió mùa đông bắc và không khí lạnh tăng cường.

- Phạm vi thu thập số liệu: Khu vực Bắc Bộ.

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Phương pháp, kỹ thuật thu thập dữ liệu: Phương pháp chỉnh biên, chỉnh lý, biên tập thông tin dữ liệu theo quy định của ngành KTTV.

Minh họa kết quả thu thập, xử lý số liệu KKL tại Bảng 2.10 và Hình 2.4.

Bảng 2.10: Tổng hợp thống kê số đợt KKL trong 10 năm (2008-2017)

Năm	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	Cộng
2008	4	6	2	1	4					1	5	5	28
2009	6	4	2	1	2					1	5	5	28
2010	5	2	4	5	0	1				3	4	5	29
2011	9	2	8	2	4				1	4	3	5	38
2012	5	5	3	2	1				2	2	5	6	31
2013	5	3	2	3		1			1	1	2	5	23
2014	4	4	3	1	1					1	4	7	25
2015	2	2	4	3	1					2	2	5	21
2016	4	4	3	1	1					2	4	5	24
2017	2	2	6	3	3					2	4	6	28
TB	4.6	3.4	3.7	2.2	1.9				1.3	1.9	3.8	5.4	27.5

Hình 2.4: Số đợt KKL trong các tháng theo trung bình nhiều năm và tần suất xuất hiện đợt không khí lạnh trong các tháng

Kết quả thu thập, xử lý số liệu không khí lạnh diện rộng phục vụ nghiên cứu được trình bày tại sản phẩm Công việc 8: Thu thập, biên tập, phân tích, đánh giá, tổng hợp số liệu không khí lạnh.

2.3. Kỹ thuật Big data

2.3.1. Các giải pháp tìm kiếm, phân tích, thống kê trong Big data

Phân tích dữ liệu lớn là một bước quan trọng để trích rút các thông tin cần thiết. Một số phương pháp phân tích dữ liệu lớn thông dụng gồm:

Phương pháp giá trị trung bình: Giá trị trung bình là bằng tổng tất cả giá trị có trong tập dữ liệu chia cho số lượng các giá trị đó. Phương pháp giá trị trung bình được áp dụng trong đánh giá chất lượng dự báo của mô hình AI dự báo KTTV.

Phương pháp hồi quy tuyến tính - LR (Linear Regression): LR thuộc nhóm Supervised Learning (học có giám sát). Hồi quy chính là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập. LR là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác LR là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Phương pháp LR được áp dụng trong quá trình huấn luyện học máy của mô hình AI hỗ trợ dự báo KTTV.

Phương pháp lấy mẫu: Khi muốn thống kê để tìm quan hệ giữa các yếu tố ngẫu nhiên ta thường xuyên phải làm việc với các tập dữ liệu rất lớn. Vì vậy, phải chọn một tập mẫu nhỏ trong tập dữ liệu lớn để mô phỏng là rất cần thiết. Quá trình lấy mẫu đòi hỏi nhiều kĩ thuật để mẫu được lấy đại diện được cho toàn bộ tập dữ liệu.

Phương pháp lấy mẫu được áp dụng trong quá trình xử lý dữ liệu đầu vào cho các mô hình AI hỗ trợ dự báo KTTV.

2.3.2. Kỹ thuật xử lý tính toán trên CSDL đồ thị

2.3.2.1. Tổng quan về CSDL đồ thị (Graph Database)

Trong điện toán, CSDL đồ thị (GDB) là CSDL sử dụng các cấu trúc biểu đồ cho các truy vấn ngữ nghĩa với các nút , cạnh và thuộc tính để biểu diễn và lưu trữ dữ liệu [22]. CSDL đồ thị hiệu quả hơn CSDL quan hệ khi làm việc với dữ liệu kết nối. Điểm khác biệt cơ bản giữa CSDL quan hệ và CSDL đồ thị là ngôn ngữ xử lý dữ liệu. Với CSDL đồ thị, ngôn ngữ là các phương pháp xử lý đồ thị. Theo lý thuyết đồ thị, các loại giải thuật chung nhất về đồ thị gồm: vẽ đồ thị, vẽ lại đồ thị, mở rộng cây, các luồng mạng lưới, các bài toán tìm đường đi, …. Điều này cho thấy sử dụng CSDL đồ thị với dữ liệu được mô hình hướng đồ thị sẽ có nhiều sự hỗ trợ hơn xử lý dữ liệu. Mô hình hóa dữ liệu trong CSDL đồ thị gồm [45]: Ngôn ngữ mô hình hóa dữ liệu trong CSDL đồ thị; Mô hình hóa quan hệ trong miền quản lý hệ thống; Mô hình hóa đồ thị trong miền quản lý hệ thống.

2.3.2.2. Các phương pháp mô hình hóa dữ liệu

- Mô tả mô hình bằng các thuật ngữ của ứng dụng: Sử dụng các câu hỏi đối với dữ liệu để xác định các thực thể cũng như các mối quan hệ. Mã hóa trực tiếp các câu hỏi của người dùng để mô hình hóa dữ liệu.

- Mô tả mô hình bằng các nút (node) và các mối quan hệ (relationship): Sử dụng các nút (node) đại diện cho các thuộc tính thực thể (vật); các mối quan hệ (relationship) đại diện cho sự kết nối giữa các thực thể (làm rõ cấu trúc).

- Mô tả mô hình bằng mối quan hệ: Sử dụng các mối quan hệ chi tiết khi ta có một tập kín các kiểu quan hệ; sử dụng các quan hệ chung với các thuộc tính trong các đồ thị đánh trọng số và có những yêu cầu như tìm đường đi ngắn nhất và không cần thiết phải là một tập kín các quan hệ.

- Mô hình hóa các sự kiện: Khi hai hoặc nhiều thực thể khác nhau tương tác trong cùng một khoảng thời gian thì một sự kiện xuất hiện. Có thể biểu diễn những sự kiện này bằng các node riêng, và kết nối với các thực thể tham gia sự kiện đó.

2.3.2.3. Ngôn ngữ xử lý dữ liệu trong CSDL đồ thị

- Ngôn ngữ truy vấn đồ thị Cypher: Cypher là ngôn ngữ truy vấn cho CSDL đồ thị, có đặc điểm dễ đọc và dễ hiểu đối với cả các nhà phát triển, các chuyên gia CSDL. Cypher cho phép người dùng tìm kiếm thông tin trên CSDL theo một mô

hình cụ thể nào đó. Cypher cũng có các câu lệnh riêng. Câu truy vấn đơn giản nhất gồm một mệnh đề START, MATCH và RETURN.

- Hệ quản trị CSDL đồ thị Neo4j: Neo4j là hệ quản trị CSDL đồ thị. Trong Neo4j, đối tượng được mô tả thành các đỉnh của đồ thị, đặc điểm của đối tượng được mô tả qua thuộc tính của đỉnh và mối quan hệ giữa các đối tượng được mô tả bằng liên kết có hướng giữa các đỉnh.

2.3.3. Kỹ thuật làm sạch và tiền xử lý dữ liệu

Giai đoạn tiền xử lý dữ liệu là quá trình xử lý dữ liệu thô/gốc nhằm cải thiện chất lượng dữ liệu.

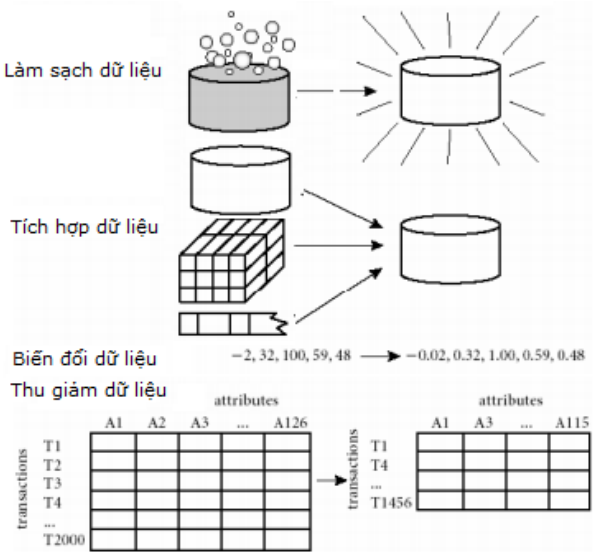
2.3.3.1. Các bước tiền xử lý dữ liệu

Bước 1: Làm sạch dữ liệu: loại bỏ nhiễu, hiệu chỉnh những phần dữ liệu không nhất quán;

Bước 2: Tích hợp dữ liệu: trộn dữ liệutừ nhiều nguồn khác nhau;

Bước 3: Biến đổi dữ liệu: chuẩn hoá dữ liệu (data normalization);

Bước 4: Thu giảm dữ liệu: thu giảm kích thước dữ liệu (giảm số phần tử) bằng kết hợp dữ liệu, loại bỏ các đặc điểm dư thừa (giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu.



Hình 2.5: Các bước tiền xử lý dữ liệu

2.3.3.2. Các kỹ thuật tiền xử lý dữ liệu

- Làm sạch dữ liệu: Tóm tắt hoá dữ liệu, nhận diện đặc điểm chung của dữ liệu, xử lý dữ liệu bị thiếu và dữ liệu bị nhiễu.

- Tích hợp dữ liệu: Tích hợp lược đồ và xử lý vấn đề dư thừa.

- Phát hiện và xử lý mâu thuẫn giá trị dữ liệu.

- Biến đổi dữ liệu: Làm trơn dữ liệu, kết hợp dữ liệu, tổng quát hóa dữ liệu, chuẩn hóa dữ liệu.

- Thu giảm dữ liệu: Kết hợp khối dữ liệu, chọn tập con các thuộc tính, thu giảm chiều, thu giảm lượng, tạo phân cấp ý niệm và rời rạc hóa dữ liệu.

2.3.3.3. Kỹ thuật làm sạch dữ liệu

- Xử lý dữ liệu bị thiếu: Các giải pháp xử lý dữ liệu bị thiếu gồm: (i) Bỏ qua; (ii) Xử lý tay (không tự động, bán tự động); (iii) Dùng giá trị thay thế (tự động): hằng số toàn cục, trị phổ biến nhất, trung bình toàn cục, trung bình cục bộ, trị dự đoán, … ; (iv) Ngăn chặn dữ liệu bị thiếu: thiết kế tốt CSDL và các thủ tục nhập liệu (các ràng buộc dữ liệu).

- Giải pháp nhận diện phần tử biên gồm: (i) Dựa trên phân bố thống kê; (ii) Dựa trên khoảng cách; (iii) Dựa trên mật độ; (iv) Dựa trên độ lệch.

2.3.4. Kỹ thuật học máy trong phân tích Big data

2.3.4.1. Các mô hình học máy trong phân tích Big data

Machine Learning - ML (Học máy hoặc Máy học) là một tập con của AI. Các mô hình thuật toán ML thường được ứng dụng trong phân tích Big data chia làm 4 nhóm: Supervise Learning, Unsupervised Learning, Semi-supervised Learning và Reinforcement Learning [21].

2.3.4.2. Kỹ thuật đánh giá, kiểm định kết quả của mô hình phân tích

Đánh giá và lựa chọn các mô hình ML là tập trung đánh giá, kiểm định khả năng dự đoán của mô hình. Các phương pháp đánh giá mô hình gồm:

a) Phương pháp ma trận nhầm lẫn (Confusion Matrix)

Trong đó:

a: TP (true positive) - mẫu mang nhãn dương được phân lớp **đúng** vào lớp **dương**.

b:FN (falsenegative) - mẫu mang nhãn dương bị phân lớp **sai** vào lớp **âm**.

c:FP (false positive) - mẫu mang nhãn âm bị phân lớp **sai** vào lớp **dương**.

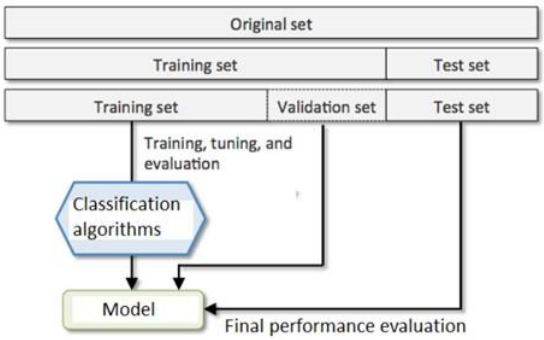
d:TN (true negative) - mẫu mang nhãn âm được phân lớp **đúng** vào lớp **âm**.

Bảng 2.11: Ma trận nhầm lẫn

Lớp thực tế	Lớp dự đoán		
		Yes	No
	Yes	a	b
	No	c	d

b) Phương pháp chia đôi (Hold Out)

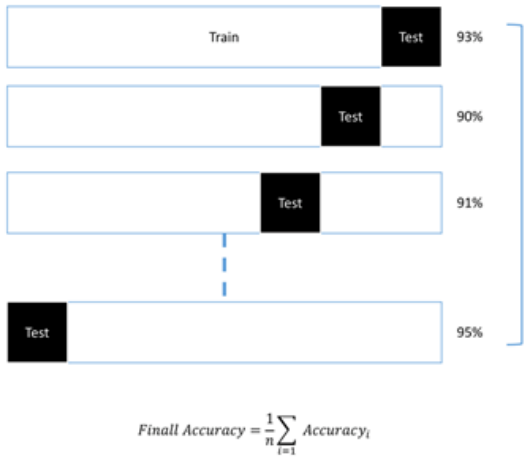
Phương pháp Hold Out là phương pháp chia đôi ngẫu nhiên tập dữ liệu thành 2 tập độc lập là tập dữ liệu huấn luyện (training set) và tập kiểm định mô hình (testing set) [23]. Thực hiện Hold Out k lần và độ chính xác của mô hình acc (M) = trung bình cộng k giá trị chính xác.



Hình 2.6: Phương pháp Hold Out

c) Phương pháp Cross Validation

Phương pháp Cross Validation là kỹ thuật phân chia tập dữ liệu ban đầu thành k tập dữ liệu con có cùng kích thước. Tại mỗi vòng lặp sử dụng tập con dữ liệu huấn luyện (training set) để huấn luyện mô hình và một tập con dữ liệu độc lập để đánh giá mô hình. Giá trị k thường là = 10 [23].



Hình 2.7: Phương pháp Cross Validation

2.4. Kỹ thuật học máy, AI để nhận dạng, hỗ trợ dự báo KTTV

2.4.1. Các phương pháp lưu trữ và tiền xử lý dữ liệu về hiện tượng KTTV

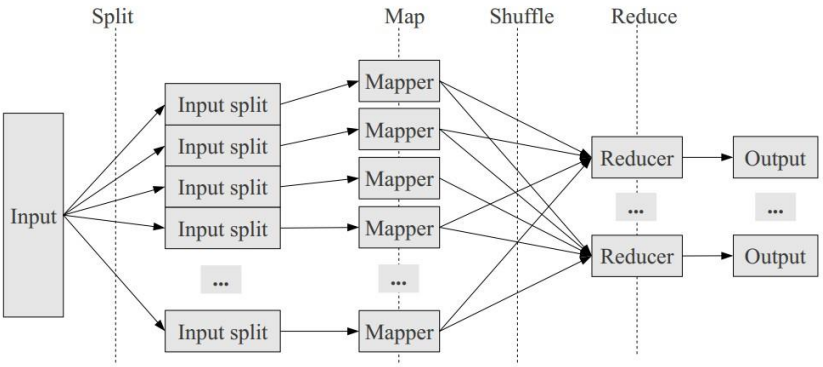
2.4.1.1. Phương pháp lưu trữ và tích hợp dữ liệu lớn

a) Mô hình xử lý dữ liệu phân tán MapReduce

MapReduce là một mô hình/ giải thuật lập trình, chuyên dùng để giải quyết vấn đề về xử lý dữ liệu lớn Big data. Mô hình cơ bản gồm các thao tác [46]:

- Split: Phân chia dữ liệu đầu vào;
- Map: Phân chia dữ liệu cho nhiều máy tính cùng thực hiện;
- Shuffle: Gom nhóm dữ liệu đầu ra của quá trình Map, tạo tiền đề cho Reduce;
- Reduce: Tổng hợp kết quả dữ liệu.

MapReduce định nghĩa dữ liệu dưới dạng các cặp <key, value> - <khóa, giá trị>. Dữ liệu được định nghĩa theo dạng này linh hoạt hơn các bảng dữ liệu quan hệ hai chiều truyền thống (quan hệ cha - con hay còn gọi là khóa chính - khóa phụ).



Hình 2.8: Mô hình tổng quát của Mapreduce

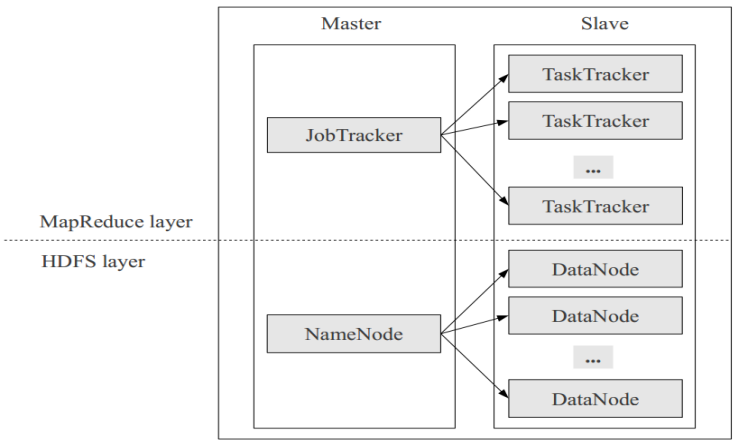
b) Nền tảng lập trình Hadoop

Hadoop là một nền tảng nguồn mở, viết bằng Java, dùng để hỗ trợ xây dựng, thực thi các ứng dụng tính toán phân tán theo mô hình MapReduce. Hadoop cluster là hệ thống máy tính đã được triển khai nền tảng Hadoop [49].

Kiến trúc của Hadoop cluster: Một Hadoop cluster bao gồm hai thành phần cơ bản là kiến trúc MapReduce và hệ thống tập tin phân tán HDFS, trong đó:

- Kiến trúc MapReducer gồm hai phần: TaskTracker - trực tiếp thực thi các tác vụ xử lý dữ liệu, JobTracker - quản lý và phân chia công việc cho các TaskTracker.
- Hệ thống HDFS gồm hai phần: DataNode - nơi trực tiếp lưu trữ dữ liệu, mỗi DataNode chịu trách nhiệm lưu trữ một phần dữ liệu của hệ thống; NameNode - quản lý các DataNode, dẫn đường cho các yêu cầu truy xuất dữ liệu.

Kiến trúc của Hadoop cluster là kiến trúc Master-Slave, và cả hai thành phần MapReduce và HDFS đều tuân theo kiến trúc này.



Hình 2.9: Các thành phần của Hadoop Cluster

c) Sử dụng HDFS để quản lý lưu trữ Big data

HDFS (Hadoop Distributed File System): là một hệ thống tập tin phân tán, được thiết kế để chạy trên hệ thống nhiều máy tính được nối mạng với nhau. HDFS có khả năng hỗ trợ tốt cho các ứng dụng xử lý dữ liệu lớn Big data[49].

Đặc điểm của HDFS:

- Lưu trữ dữ liệu rất lớn: HDFS được thiết kế để lưu trữ những tập tin với kích thước hàng trăm megabyte, gigabyte hay terabyte và lên đến petabyte dữ liệu.

- Xử lý lỗi phần cứng: HDFS không đòi hỏi phần cứng cấu hình cao đối với hệ thống máy tính. Một hệ thống HDFS có thể bao gồm hàng nghìn máy xử lý, mỗi máy (node) lưu trữ một phần của dữ liệu.

- Dữ liệu chặt chẽ: HDFS hoạt động theo cơ chế ghi một lần - đọc nhiều lần. Mỗi tập tin sẽ được tạo, ghi dữ liệu và đóng lại hoàn toàn. Việc cập nhật ghi thêm dữ liệu vào tập tin là không thể thực hiện trên HDFS.

- Di chuyển tính toán thay vì di chuyển dữ liệu: Các yêu cầu tính toán của ứng dụng sẽ được thực hiện tại node chứa dữ liệu gần nhất. HDFS cung cấp giao diện cho ứng dụng tìm kiếm và di chuyển chính nó đến vị trí dữ liệu gần nhất.

- Chạy trên nhiều nền tảng và thiết bị: HDFS được thiết kế để dễ dàng di chuyển từ nền tảng này sang nền tảng khác, thiết bị này sang thiết bị khác. Điều này tạo điều kiện thuận lợi cho việc ứng dụng HDFS một cách rộng rãi.

d) Xây dựng chương trình quản lý Big data trên nền Hadoop

Để xây dựng chương trình quản lý Big data, Hadoop xây dựng gói org.apache.hadoop.io hỗ trợ các kiểu dữ liệu phù hợp với Hadoop cho Java, gồm có:

- NullWritale: tương ứng với kiểu dữ liệu Null trong Java.

- Text: tương ứng với kiểu dữ liệu String trong Java.

- BytesWritable: tương ứng với kiểu Byte trong Java.
- BooleanWritable: tương ứng với kiểu Boolean trong Java.
- IntWritable: tương ứng với kiểu Integer trong Java.
- LongWritable: tương ứng với kiểu Long trong Java.
- FloatWritable: tương ứng với kiểu Float trong Java.
- DoubleWritable: tương ứng với kiểu Double trong Java.

e) Tích hợp dữ liệu Big data

- Tích hợp dữ liệu: là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu có sẵn cho quá trình khai phá dữ liệu.

- Yêu cầu về tích hợp dữ liệu: (i) Giảm thiểu các dư thừa và mâu thuẫn của dữ liệu, giúp cải thiện tốc độ quá trình khai phá dữ liệu; (ii) Đảm bảo không ảnh hưởng đến hoạt động và cấu trúc dữ liệu hiện có.

- Các công cụ tích hợp dữ liệu:
  + Centerprise Data Integrator: Công cụ cho phép tích hợp dữ liệu tự động [64].
  + Oracle Data Service Integrator: Nền tảng tích hợp dữ liệu toàn diện bao gồm tất cả các yêu cầu tích hợp dữ liệu: từ số lượng lớn, hiệu suất cao, …
  + Talend Data Integration là công cụ khác cho phép tích hợp dữ liệu lớn.

2.4.1.2. Phương pháp xử lý dữ liệu mất mát/ khuyết thiếu

a) Khái niệm và các dạng dữ liệu mất mát/ khuyết thiếu (Missing data)

Dữ liệu mất mát/ thiếu: Trong thống kê, **thiếu dữ liệu** hoặc **thiếu giá trị** xảy ra khi không có giá trị dữ liệu nào được lưu trữ cho biến trong một quan sát. Dữ liệu bị thiếu là một sự xuất hiện phổ biến và có thể có ảnh hưởng đáng kể đến các kết luận có thể được rút ra từ dữ liệu [104].

Các dạng dữ liệu bị mất mát/ thiếu sau đây:

- Thiếu ngẫu nhiên - MAR (Missing at Random): Là xu hướng giá trị của một đặc trưng bị khuyết không liên quan đến tính chất của đặc trưng đó nhưng liên quan đến một vài đặc trưng không bị khuyết khác.

- Thiếu hoàn toàn ngẫu nhiên - MCAR (Missing Completely at Random): Là xu hướng bị khuyết của một đặc trưng là hoàn toàn ngẫu nhiên. Không có mối quan hệ nào giữa đặc trưng bị khuyết với các giá trị giả định hoặc các ràng buộc trên các đặc trưng khác.

- Thiếu không ngẫu nhiên - MNAR (Missing not at Random): Xảy ra khi một điểm dữ liệu bị khuyết phụ thuộc vào các giá trị giả định và các giá trị của các đặc trưng khác.

b) Phương pháp xử lý dữ liệu mất mát/ khuyết thiếu

Các phương pháp chính để xử lý dữ liệu bị thiếu gồm [104]:

- Phương pháp áp đặt: điền giá trị áp đặt vào vị trí của dữ liệu bị thiếu

- Phương pháp xóa một phần: Phương pháp liên quan đến việc giảm dữ liệu có sẵn cho một tập dữ liệu không có giá trị bị thiếu bao gồm: Xóa theo danh sách / xóa theo dòng và xóa theo cặp.

- Phương pháp nội suy (nội suy song tuyến): là một phương pháp xây dựng các điểm dữ liệu mới trong phạm vi của một tập hợp rời rạc các điểm dữ liệu đã biết.

- Phương pháp phân tích đầy đủ: Các thuật toán cực đại hóa kỳ vọng; Phương pháp tiếp cận phân biệt đối xử; Phương pháp nhận dạng từng phần.

- Các kỹ thuật dựa trên mô hình, thường sử dụng đồ thị, cung cấp các công cụ bổ sung để kiểm tra các kiểu dữ liệu bị thiếu (MCAR, MAR, MNAR) và để ước tính các tham số trong điều kiện dữ liệu bị thiếu.

2.4.1.3. Phương pháp xử lý dữ liệu không chắc chắn

a) Khái niệm dữ liệu không chắc chắn

Dữ liệu không chắc chắn là dữ liệu có chứa **nhiễu** khiến nó đi chệch khỏi các giá trị chính xác, dự định. Trong thời đại của dữ liệu lớn, tính không chắc chắn hoặc tính xác thực của dữ liệu là một trong những đặc điểm xác định của dữ liệu. Dữ liệu liên tục phát triển về khối lượng, sự đa dạng, tốc độ và độ không chắc chắn. Các dạng dữ liệu không chắc chắn gồm: Tính không chắc chắn thuộc tính; Độ không đảm bảo đo tương quan; Sự không chắc chắn tuple [105].

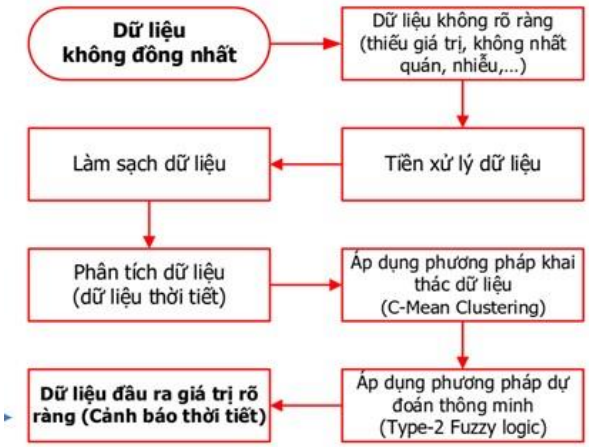
b) Phương pháp xử lý dữ liệu không chắc chắn

Hệ thống logic mờ loại 2 có thể quản lý dữ liệu bất định trong dự báo thời tiết. Nhưng hệ thống này dễ bị ảnh hưởng bởi dữ liệu nhiễu và dữ liệu dị thường. Điều khó khăn nữa, để xử lý một lượng lớn dữ liệu cần hệ thống tính toán hiệu năng cao và thời gian thực hiện lớn.

Fuzzy C-Mean clustering là một trong các kỹ thuật khai thác dữ liệu được sử dụng với Type -2 fuzzy logic system để phát hiện dữ liệu dị thường và nhóm dữ liệu mong muốn để cải thiện độ chính xác và hiệu quả. Do đó, chúng ta có thể xử dụng phương pháp lai bằng cách sử dụng “Fuzzy C-mean clustering và Type -2 fuzzy

logic system" để khắc phục những khó khăn này [105]. Phương pháp này gồm 4 bước để tăng tính chính xác của dữ liệu bất định và cải thiện hiệu quả (hình 2.10):

- Bước 1: Nhận dạng các thiết bị (cảm biến, vệ tinh, vv) thu dữ liệu thô.
- Bước 2: Làm sạch các dữ liệu, tiền xử lý dữ liệu (giá trị khuyết, dữ liệu và tiếng ồn không phù hợp).
- Bước 3: Áp dụng thuật toán C-mean Clustering phát hiện dữ liệu dị thường để tăng độ chính xác dự báo và nâng cao hiệu năng.
- Bước 4: Sử dụng hệ thống logic mờ loại 2 để làm sạch số liệu và có được quyết định chính xác và có được kết quả tốt nhất cho dự báo.



Hình 2.10: Mô hình xử lý dữ liệu mất mát, không chắc chắn

2.4.1.4. Phương pháp phát hiện và xử lý dữ liệu bất thường (ngoại lai)

a) Khái niệm và giới thiệu về dữ liệu ngoại lai

Theo Charu C. Aggarwal, điểm ngoại lai là điểm dữ liệu khác biệt đáng kể so với phần còn lại của tập dữ liệu [52]. Các giá trị ngoại lai là các mẫu dữ liệu đặc biệt, cách xa khỏi phần lớn dữ liệu khác trong tập dữ liệu. Có rất nhiều nguyên nhân chủ quan và khách quan dẫn tới sự xuất hiện của các điểm ngoại lai trong tập dữ liệu như: các lỗi nhập dữ liệu do con người gây ra; các lỗi đo lường do thiết bị, dụng cụ lấy mẫu, thí nghiệm gây ra, ….. [53]. Các giá trị ngoại lai thường chứa đựng những thông tin hữu ích về những đặc điểm bất thường của hệ thống và thực thể ảnh hưởng tới quá trình sinh dữ liệu. Một số phương pháp phát hiện các điểm dữ liệu ngoại lai gồm [52]: Phân tích giá trị cực trị (Extreme Value Analysis; Các mô hình xác suất và thống kê (Probabilistic and Statistical Models); Các mô hình tuyến tính (Linear Models); Các mô hình dựa trên lân cận (Proximity-based Models); Các mô hình dựa trên lý thuyết thông tin (Information Theoretic Models).

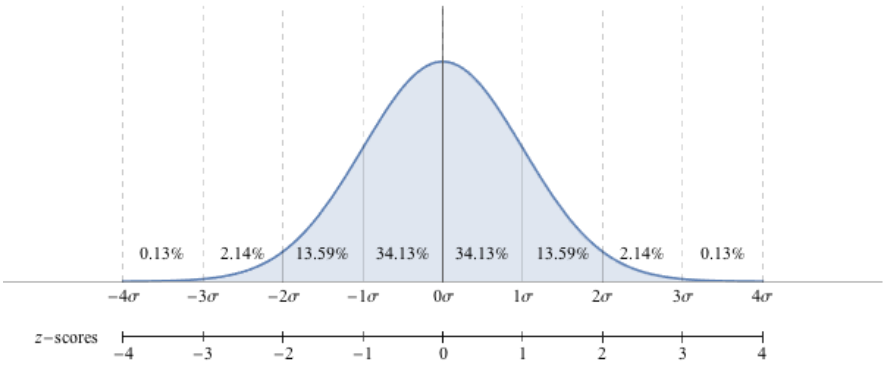
b) Phát hiện ngoại lai cho dữ liệu một chiều bằng Z-Score (Độ lệch chuẩn)

Điểm tiêu chuẩn hay Z-Score chỉ ra một thành phần chênh lệch so với trung bình là bao nhiêu độ lệch chuẩn. Z-Score của bất kỳ một điểm dữ liệu nào được tính theo công thức [53]:

$$z = \frac{(x-\mu)}{\sigma} \quad (2.4.1)$$

Trong đó: x là giá trị của điểm dữ liệu cần tính Z-Score; μ là giá trị trung bình của tập dữ liệu; σ là độ lệch chuẩn của tập dữ liệu. Nếu z < 0 thể hiện điểm dữ liệu đó nhỏ hơn giá trị trung bình; nếu z > 0 thể hiện điểm dữ liệu đó lớn hơn giá trị trung bình; nếu z = 0 thể hiện điểm dữ liệu đó bằng với giá trị trung bình.

Sau khi tính toán Z-Score cho từng điểm trong tập dữ liệu, một ngưỡng (threshold) sẽ được thiết lập để lọc các điểm này so với giá trị trung bình. Nếu tập dữ liệu theo phân phối chuẩn như chỉ ra trong hình 2.11, cho thấy: Với ngưỡng 2.5 (-2.5 < Z-Score < +2.5) có 99% điểm dữ liệu nằm trong phạm vi 2.5 lần độ lệch chuẩn; với ngưỡng 3.0 (-3.0 < Z-Score < +3.0) có 99.8% điểm dữ liệu nằm trong phạm vi 3.0 lần độ lệch chuẩn; với ngưỡng 5.0 (-5.0 < Z-Score < +5.0) có 99.9999426% điểm dữ liệu nằm trong phạm vi 5.0 lần độ lệch chuẩn.

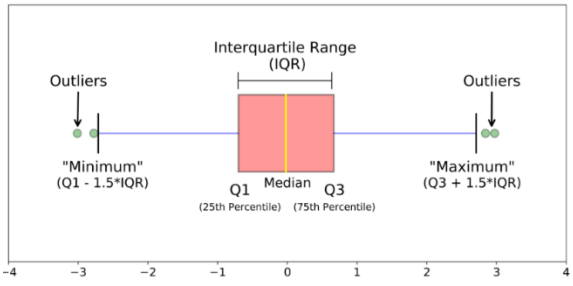


Hình 2.11: Tỷ lệ điểm dữ liệu theo ngưỡng Z-Score với phân phối chuẩn

c) Phát hiện ngoại lai cho dữ liệu một chiều bằng biểu đồ Box-plot

Biểu đồ Box-plot được sử dụng để đo khuynh hướng phân tán và xác định các giá trị ngoại lai của tập dữ liệu. Biểu đồ Box-plot chia tập dữ liệu thành các khoảng phần tư, phần thân của biểu đồ bao gồm một chiếc hộp, biểu đồ thể hiện 5 giá trị của tập dữ liệu bao gồm: (i) Giá trị nhỏ nhất của tập dữ liệu được xác định bằng Q1-1.5 * IQR; (ii) Tứ phân vị thứ nhất (Q1) của tập dữ liệu; (iii) Tứ phân vị thứ hai (Q2) chính là giá trị trung vị (Median) của tập dữ liệu; (iv) Tứ phân vị thứ ba (Q3) của tập dữ liệu; (v) Giá trị lớn nhất (Maximum) của tập dữ liệu có giá trị bằng Q3 + 1.5* IQR. Nếu tập dữ liệu có chứa các giá trị ngoại lai thì chiều dài tối đa của 2 râu tính từ mỗi cạnh hộp sẽ được xác định bằng 1.5 lần độ trải giữa (IQR - Interquatile Range). Các điểm dữ liệu nằm ngoài râu Minimum được xem xét là các điểm ngoại lai trái (Left outlier).

Các điểm dữ liệu nằm ngoài râu Maximum được xem xét là các điểm ngoại lai phải (Right outlier). Các điểm dữ liệu ngoại lai này được thể hiện bằng dấu chấm tròn trên biểu đồ Box-plot (hình 2.12) [54].



Hình 2.12: Hình dạng và giá trị của tập dữ liệu thể hiện trên biểu đồ Box – plot

d) Phương pháp xử lý dữ liệu ngoại lai

Các điểm dữ liệu ngoại lai có ảnh hưởng rất lớn đến độ chính xác của các mô hình, việc xử lý sao cho phù hợp với tập dữ liệu thường khó hơn rất nhiều so với việc phát hiện [55]. Có 2 nhóm phương pháp để xử lý dữ liệu ngoại lai gồm:

- Loại bỏ khỏi tập dữ liệu: Sau khi phát hiện các điểm ngoại lai ta thực hiện xóa các điểm này khỏi tập dữ liệu.

- Thay thế bằng một giá trị khác: Thay thế giá trị của các điểm ngoại lai bằng một giá trị khác phù hợp với tập dữ liệu [56].

2.4.1.5. Phương pháp chuẩn hóa dữ liệu

a) Khái niệm chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là quy trình xử lý dữ liệu chuyển đổi cấu trúc của các tập dữ liệu khác nhau thành định dạng dữ liệu chung. Là một phần của trường chuẩn bị dữ liệu, chuẩn hóa dữ liệu đề cập đến việc chuyển đổi tập dữ liệu sau khi dữ liệu được lấy từ hệ thống nguồn và trước khi nó được tải vào hệ thống đích. Trong quá trình thu thập, phân tích dữ liệu, ta thường gặp các quan sát mà giá trị của nó rất khác biệt so với giá trị của các quan sát khác, đây gọi là các bất thường (hay ngoại lai - Outlier). Do đó, phải thực hiện việc chuẩn hóa dữ liệu ban đầu, là bước chuẩn bị dữ liệu thường được yêu cầu trước khi thực hiện các thuật toán trong máy học, nhằm giúp thuật toán hiệu quả hơn [106].

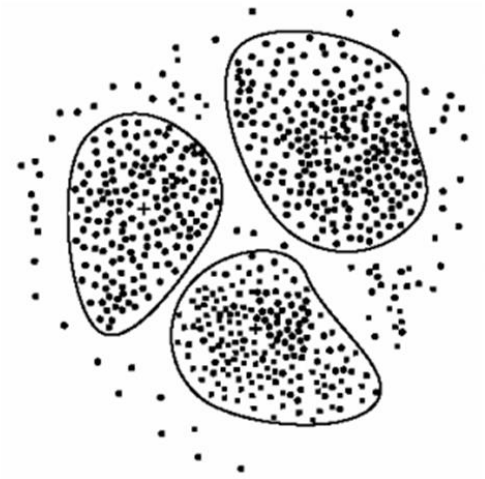
b) Các bước chuẩn hóa dữ liệu

Các bước chuẩn hóa dữ liệu gồm:

Bước 1: Làm sạch dữ liệu để loại bỏ nhiễu, hiệu chỉnh những phần dữ liệu không nhất quán.

Bước 2: Tích hợp dữ liệu (data integration): trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu;

Bước 3: Biến đổi dữ liệu (data transformation) hay chuẩn hoá dữ liệu (data normalization) [57].



Hình 2.13: Phân cụm dữ liệu để làm sạch

c) Các phương pháp chuẩn hóa dữ liệu

Các phương pháp chuẩn hóa dữ liệu trong học máy gồm:

- Trung tâm hóa dữ liệu: là phương pháp đưa các điểm dữ liệu trong tập dữ liệu về xoay quanh giá trị 0 thay vì xoay quanh giá trị trung bình của tập dữ liệu.

- Co giãn dữ liệu: là một phương pháp chuẩn hóa phạm vi của các đặc trưng dữ liệu và được thực hiện trong suốt quá trình tiền xử lý dữ liệu.

- Chuẩn hóa min-max: là phương pháp đơn giản nhất trong việc co giãn phạm vi của đặc trưng bằng việc co giãn chúng về phạm vi [0,1] hoặc [-1,1]. Công thức chung [57] như sau:

$$x' = \frac{x - \min(x)}{(x) - \min(x)} \qquad 2.4.2)$$

- Co giãn trung bình: là phương pháp co giãn xoay quanh trung bình có giá trị nằm trong khoảng [-0.5, 0.5] và được cho bởi công thức:

$$x' = \frac{x - \text{average}(x)}{(x) - \min(x)} \qquad (2.4.3)$$

- Chuẩn hóa tiêu chuẩn: Là thực hiện chính quy hóa dữ liệu giúp cho giá trị của mỗi đặc trưng có trung bình bằng 0 và phương sai bằng 1. Phương pháp này được sử dụng rộng rãi trong việc chuẩn hóa dữ liệu của nhiều thuật toán học máy (SVM, logistic regression và ANNs) theo công thức sau:

$$x' = \frac{x - average(x)}{std(x)} \qquad (2.4.4)$$

- Vec-tơ đơn vị: Là co giãn các thành phần của các véc-tơ đặc trưng là biến đổi sao cho véc-tơ đặc trưng sau khi biến đổi có độ dài bằng 1 theo:

$$x' = \frac{x}{\|x\|} \qquad (2.4.5)$$

- Mã hóa đặc trưng dạng nhóm: Biến đổi một đặc trưng dạng nhóm thành một hoặc nhiều đặc trưng dạng số. Một số phương pháp mã hóa đặc trưng dạng nhóm phổ biến hiện nay như: *mã hóa số, mã hóa one-hot* và *mã hóa nhị phân.*

Bảng 2.12: Mã hóa one - hot

Màu sắc	Bỏ đặc trưng nhóm	Đỏ	Vàng	Xanh
Đỏ		1	0	0
Vàng		0	1	0
Xanh		0	0	1

Bảng 2.13: Mã hóa nhị phân

Categori-cal features		Binary encode				
	→	=	x	x2	X3	X4
Cam	→	1	1	0	0	0
Tá	→	2	0	1	0	0
Xoài	→	3	0	0	1	0
Mít	→	4	0	0	0	1

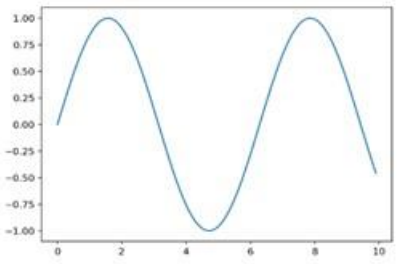
2.4.1.6. Phương pháp trực quan hóa dữ liệu

Khái niệm trực quan hóa (Data Visualization): Là việc mô tả dữ liệu một các đơn giản nhất dưới dạng các hình ảnh trực quan như bảng, biểu đồ, đồ thị, ... [48].

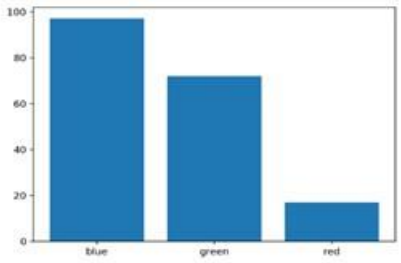
a) Các phương pháp để trực quan hóa dữ liệu cụ thể như sau:

- Biểu đồ đường: Line plot thường được sử dụng để biểu diễn dữ liệu có tính liên tục.

- Biểu đồ thanh: Bar chart thường sử dụng để biểu diễn số lượng tương đối cho các categories.



a)



b)

Hình 2.14: Trực quan hóa bằng biểu đồ đường (a) và biểu đồ thanh (b)

- Biểu đồ histogram: Histogram Plot thường được sử dụng để biểu diễn sự phân bố của một mẫu dữ liệu.

- Biểu đồ hộp: Box Plot thường được sử dụng để biểu diễn tóm tắt sự phân bố của các mẫu dữ liệu.

- Biểu đồ phân tán: Scatter Plot thường được sử dụng để biểu diễn tóm tắt sự phân bố của một hoặc nhiều cụm mẫu dữ liệu. Các điểm dữ liệu là sự kết hợp của 2 đặc trưng ở trục x - y.

b) Các công cụ hỗ trợ trực quan hóa dữ liệu

Các thư viện, công cụ Library/Tool trong ngôn ngữ lập trình Python hỗ trợ trực quan hóa dữ liệu trong công nghệ Machine Learning/Deep Learning cụ thể:

- Numpy/Scipy: thư viện tính toán số học cơ bản.
- Matplotlib: thư viện dùng để vẽ đồ thị, biểu đồ, …
- Jupyter Notebook: Web editor, có thể vừa thực thi vừa confirm kết quả.
- Pandas: thư viện xử lý lượng lớn data nhanh chóng.
- Scikit-Learn: thư viện chuẩn dành cho Machine Learning của Python.
- Gensim: thư viện xử lý ngôn ngữ tự nhiên chuyên biệt về topic model.
- TensorFlow: engine/library được phát triển bởi Google dành cho Deep Learning.

2.4.1.7. Phương pháp trình diễn/ biểu diễn dữ liệu

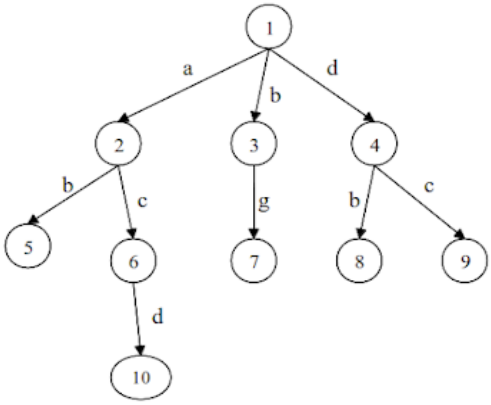
a) Khái niệm và mục đích của trình diễn/ biểu diễn dữ liệu

Trình diễn/ biểu diễn dữ liệu: là việc lựa chọn các cấu trúc dữ liệu phù hợp để lưu trữ dữ liệu lớn. Việc lựa chọn các cấu trúc dữ liệu phù hợp hỗ trợ việc quản lý, lưu trữ và xử lý thông tin hiệu quả. Biểu diễn dữ liệu tốt lựa chọn được phương pháp học máy hiệu quả [59]. Mục đích của biểu diễn dữ liệu gồm:

- Biểu diễn dữ liệu giúp phân tích rõ hơn về dữ liệu đầu vào bao gồm: sự phân bố dữ liệu, đặc tính và sự tương quan của các feature, trực quan các dữ liệu bị nhiễu, khuyết thiếu… Những sự hiểu biết này giúp ích rất lớn trong quá trình lựa chọn và traning mô hình.
- Biểu diễn/ trình diễn dữ liệu hỗ trợ việc lựa chọn một thuật toán, phương pháp học máy hiệu quả; hỗ trợ việc tối ưu hóa dung lượng bộ nhớ; hỗ trợ tăng tốc truy tìm dữ liệu lớn; hỗ trợ việc quản lý và lưu trữ dữ liệu.

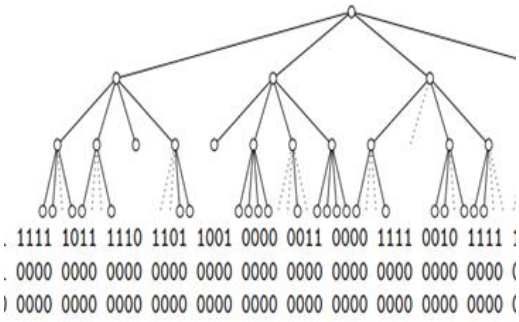
b) Các phương pháp trình diễn dữ liệu trong Machine Learning

Trình diễn Trie: Trie là một cấu trúc dữ liệu dùng để quản lý một tập hợp các xâu. Trie cho phép: Thêm một xâu, Xóa một xâu khỏi tập hợp và kiểm tra xâu có tồn tại trong tập hợp hay không. Trie gồm một gốc không chứa thông tin, trên mỗi cạnh lưu một ký tự, mỗi nút và đường đi từ gốc đến nút đó thể hiện 1 xâu, gồm các ký tự là các ký tự thuộc cạnh trên đường đi đó.



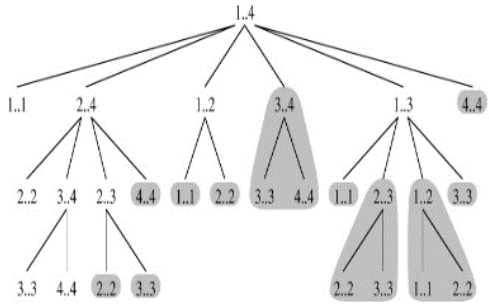
Hình 2.15: Trình diễn Trie

Succinct Trie: có cấu trúc dữ liệu cây Tries. Điều khác biệt là dữ liệu trên Succinct Tries là các dữ liệu nén. Điều này cho phép vừa hiệu quả trong việc truy tìm thông tin, mà vừa nén dữ liệu.



Hình 2.16: Trình diễn Succinct Trie

Dynamic Tree biểu diễn một cấu trúc dữ liệu trừu tượng cây với các nút cha và con như: parent(x); child(x, c); add(x, c);



Hình 2.17: Trình diễn Dynamic Tree

2.4.2. Các phương pháp trích rút các đặc trưng dữ liệu về hiện tượng KTTV

2.4.2.1. Phương pháp giảm chiều và số lượng dữ liệu

a) Khái niệm giảm chiều và giảm số lượng dữ liệu

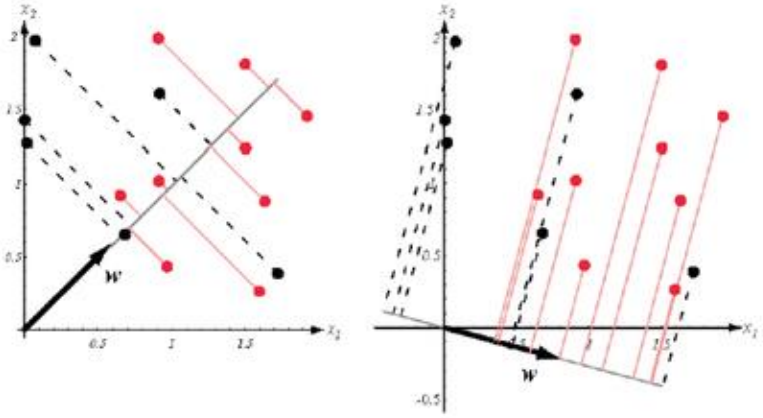
Giảm chiều dữ liệu là quá trình giảm số lượng các biến ngẫu nhiên hoặc các thuộc tính ngẫu nhiên. Thu nhỏ hay giảm chiều dữ liệu là một bước quan trọng khi thực thi các thuật toán khai phá dữ liệu và học máy. Một số phương pháp giảm số chiều thuộc tính của dữ liệu gồm: Principal Component Analysis (PCA); Principal Component Regression (PCR); Partial Least Squares Regression (PLSR); Sammon

Mapping; Multidimensional Scaling (MDS); Projection Pursuit; Linear Discriminant Analysis (LDA); Mixture Discriminant Analysis (MDA); Quadratic Discriminant Analysis (QDA); Flexible Discriminant Analysis (FDA) [34].

Giảm số lượng dữ liệu: là thay thế lượng dữ liệu ban đầu bởi biểu diễn dữ liệu mới nhỏ hơn. Các phương pháp giảm số lượng dữ liệu là có tham số hoặc không tham số. Với *phương pháp có tham số*, một mô hình được sử dụng để ước lượng dữ liệu, sao cho thông thường các tham số của dữ liệu được lưu trữ, thay vì dữ liệu ban đầu. P*hương pháp không tham số* lưu trữ giảm các biểu diễn của dữ liệu gồm *biểu đồ, phân cụm, lấy mẫu*, và *kết hợp khối dữ liệu*.

b) Phương pháp phân tích khác biệt tuyến tính (LDA)

Phân tích khác biệt tuyến tính (LDA) là một phương pháp giảm chiều dữ liệu cho bài toán phân loại và cũng có thể được coi là một phương pháp phân lớp, và cũng có thể được áp dụng đồng thời cho cả hai, tức giảm chiều dữ liệu sao cho việc phân lớp hiệu quả nhất [66]. Với LDA, dữ liệu thuộc hai lớp, được hiển thị bằng màu đỏ và đen, được chiếu lên một đường thẳng. Mục tiêu của ta là một phép chiếu trong đó dữ liệu được chiếu từ hai lớp không trùng nhau hết mức có thể, nên rõ ràng phép chiếu ở bên phải là thích hợp hơn (hình 2.18).

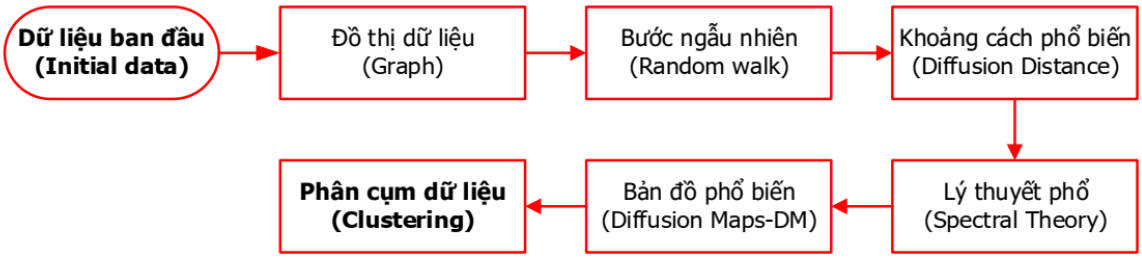


Hình 2.18: Phép chiếu trong phương pháp LDA

Các bước của LDA gồm: (i) Tính vec-tơ trung bình của 2 lớp; (ii) Tính ma trận phân tán của 2 lớp; (iii) Cộng 2 ma trận phân tán thành ma trận phân tán lớp S_W và tính nghịch đảo. Nhận được véc-tơ w bằng việc thực hiện nhân ma trận $S_W^{-1}(m_1 - m_2)$; (iv) Tính giá trị chiếu $y_i = w^t x_i$.

c) Phương pháp bản đồ phổ biến DM (Diffusion Maps) cho giảm chiều dữ liệu KTTV

Kỹ thuật cụ thể: Thực hiện giảm chiều không gian và thời gian cho bài toán dự đoán các giá trị của thời tiết. Các bước của phương pháp được minh họa như sau:



Hình 2.19: Các bước của phương pháp sử dụng các án xạ biểu đồ phổ biến

Bước 1: Dựng một đồ thị liên kết của các mẫu trong không gian đặc trưng khởi tạo S = {x1, x2, …, xn}.

Bước 2: Áp dụng kỹ thuật: bước ngẫu nhiên (random walk), tính khoảng cách phổ biến (diffusion distance), và các thuật toán khai phá dữ liệu để phân cụm dữ liệu.

Bước 3: Áp dụng phương pháp bản đổ phổ biến (DM) cho quá trình nén không gian và thời gian.

Bằng cách áp dụng DM trên tập dữ liệu, phương pháp DM cho kết quả khả quan trọng việc dự toán thông tin thời tiết. Kết quả thực nghiệm cho thấy với bài toán nén không gian, phương pháp cho kết quả tốt hơn thuật toán phân cụm k-means khi sử dụng khoảng cách Euclidean; và với bài toán nén thời gian, phương pháp cho kết quả tốt hơn PCA.

d) Phương pháp giảm chiều dữ liệu 3 bước để phân tích dữ liệu KTTV

Kỹ thuật cụ thể gồm ba bước: *Bước 1: Ước lượng số thành phần*: các nghiên cứu chỉ ra rằng không có một phương pháp tối ưu cho việc xác định các thành phần. Trong ràng buộc của sự độc lập về thời gian và các biến ngẫu nhiên của dữ liệu theo thời gian (time-series data), sự phân bố của các biến được xác định bằng cách sử dụng công cụ từ Random Matrix Theory (RMT). Tuy nhiên, RMT khó áp dụng trong phương pháp này cho đặc trưng dữ liệu. Từ đó, các tác giả đề xuất sử dụng phương pháp của Dray. *Bước 2*: Kiểm tra: kết hợp mỗi giá trị riêng sinh ra từ bước 1 để giải quyết bài toán phân phối rỗng (null distribution). *Bước 3*: Thu về các thành phần sau quá trình phân tích sau khi chuẩn hoá, bằng cách sử dụng Varimax. Thuật toán này được sử dụng rộng rãi trong cộng đồng các nhà khoa hoạ phân tích dữ liệu thời tiết, khí hậu để thu về các thành phần được dịch chuyển. Kết quả thực nghiệm trên bộ dữ liệu NCEP/NCAR cho thấy các thành phần được sinh ra từ phương pháp đề xuất có thể bao phủ nhiều chế độ thời tiết so với các phương pháp khác.

e) Phương pháp hồi quy trên các phần dữ liệu ngược

Áp dụng cho bài toán giảm chiều dữ liệu lượng mưa theo tháng trên vịnh sông Missouri (của Allison và cộng sự). Các tác giả chỉ ra rằng trong khi dự đoán nhiệt độ ngày đạt được kết quả khả quan, dự đoán lượng mưa theo ngày vẫn còn hạn chế và khó có thể cải thiện chất lượng của mô hình. Từ đó, các tác giả đề xuất phương pháp dự đoán trên tập dữ liệu lớn. Dữ liệu tại mỗi vị trí được đưa vào các mô hình không gian-thời gian truyền thống, sau đó được giảm số chiều với việc giảm số biến tại mỗi vị trí sử dụng các cải tiến của Slided Inverse Regression và bộ đánh giá Nadaraya-Watson (Nadaraya-Watson estimator). Kết quả thực nghiệm cho thấy phương pháp có thể được áp dụng trên một tập dữ liệu lớn. Độ chính xác của mô hình được cải thiện cho dự đoán lượng mưa ngày.

2.4.2.2. Phương pháp phân cụm dữ liệu

a) Khái niệm phân cụm dữ liệu

Là quá trình nhóm một tập các đối tượng vào nhiều nhóm hoặc cụm (clusters) sao cho các đối tượng trong cùng một cụm tương tự nhau, nhưng các đối tượng ở các cụm khác nhau thì có sự khác nhau. Sự không giống nhau và sự giống nhau được đánh giá dựa trên các giá trị thuộc tính mô tả các đối tượng và thường là các độ đo khoảng cách. Phân cụm là công cụ phân tích dữ liệu được sử dụng trong nhiều ứng dụng như sinh học, bảo mật, thông minh doanh nghiệp, hoặc tìm kiếm [35].

b) Các phương pháp phân cụm dữ liệu

Một số các nhóm phương pháp phân cụm được dùng phổ biến là:

- Phương pháp phân đoạn: Là tổ chức phân cụm các đối tượng vào tập các nhóm dữ liệu.
- Các phương pháp phân cụm phân nhóm cấu trúc: Là tổ chức phân cụm dữ liệu theo các nhóm có cùng cấu trúc.
- Các phương pháp dựa trên mật độ: Là tổ chức phân cụm dựa trên mật độ dữ liệu. Đây là phương pháp hiệu quả với Big data, có khả năng xử lý nhiễu tốt [67].
- Các phương pháp dựa trên lưới: Là tổ chức phân cụm dựa trên lưới dữ liệu.

c) So sánh các phương pháp phân cụm dữ liệu

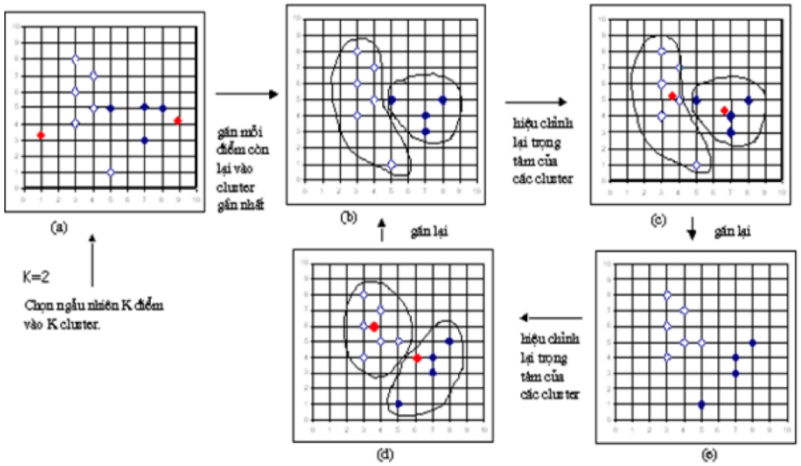
Các thuật toán phân cụm có sự phù hợp riêng cho từng bài toán và kiểu dữ liệu. Không có một thuật toán nào đáp ứng đầy đủ cho các loại bài toán. So sánh giữa các thuật toán tại Bảng 2.14.

Bảng 2.14: So sánh giữa các thuật toán phân cụm dữ liệu

Thuật toán	Độ phức tạp	Khả năng làm việc với

		dữ liệu có số chiều lớn
k-means	*O(Nkd)* (thời gian); *O(N+k)* (không gian)	Không
k-medoids	Như *k*-means	Không
DENCLUE	*O(NlogN)* (thời gian)	Có
STING	*O(N)* (thời gian)	Có
CLIQUE	Tuyến tính với số lượng đối tượng, bậc hai (quadratic) với số lượng chiều dữ liệu	Có

Như vậy, có thể thấy độ phức tạp và khả năng áp dụng của mỗi thuật toán là khác nhau, tuỳ thuộc điều kiện và bài toán cụ thể.



Hình 2.20: Thuật toán phân cụm dữ liệu k-means

2.4.2.3. Phương pháp phân hạng đặc trưng cho dữ liệu

a) Khái niệm phân hạng đặc trưng dữ liệu

Phân hạng đặc trưng (FR) là chọn số lượng biến (đặc trưng) quan trọng bằng cách xếp hạng các tính năng theo mức độ quan trọng của chúng trong mô hình. Mục đích của xếp hạng/ phân hạng đặc trưng (Feature Ranking) là để đo lường mức độ liên quan của các biến đặc trưng đến biến đích nhằm tìm ra các biến đặc trưng có khả năng phân loại tốt nhất [60]. Phân hạng đặc trưng cho dữ liệu KTTV có ý nghĩa đặc biệt quan trọng trong quyết định chọn ra những đặc trưng nào để xây dựng mô hình dự báo như dự báo bão, mưa hay lũ. Các phương pháp phân hạng đặc trưng gồm:

b) Phương pháp lọc - Xếp hạng dựa trên điểm số (Ranking based on scores)

Phương pháp lọc sẽ phân hạng các đặc trưng có liên quan bằng cách xem xét các thuộc tính vốn có của dữ liệu. Trong hầu hết các trường hợp, cách tiếp cận này áp dụng một phương pháp thống kê để tính độ đo quan trọng của mỗi thuộc tính. Các thuộc tính được xếp hạng theo điểm số và được chọn để giữ hoặc xóa khỏi bộ dữ liệu. Một số phương pháp lọc bao gồm kiểm tra bình phương chi, độ đo (gain) thông

tin và hệ số tương quan như: (i) Laplacian score; (ii) Chi-squared (Chi-bình phương); (iii) Fisher Score (FSCORE).

c) Các phương pháp đóng gói (Wrapper Methods)

Các phương pháp đóng gói thực chất thực hiện một dạng tối thiểu hóa hàm lỗi (risk minimization) trên tập $G = \{g_S : S \subseteq V_n\}$. Do đó, các phương pháp đóng gói tìm cách tính xấp xỉ S^+.

d) Các phương pháp nhúng (Embedded Methods)

Phương pháp nhúng được hình thành từ việc giải quyết bài toán tìm đặc trưng tối thiểu (min-feature). Bài toán này có thể được phát biểu dưới dạng một bài toán tối ưu nếu ta có thể xấp xỉ lỗi trung bình trên toàn không gian (expected risk) R, với lỗi trung bình trên tập dữ liệu huấn luyện (empirical risk).

2.4.2.4. Phương pháp lựa chọn đặc trưng cho dữ liệu

a) Các khái niệm về đặc trưng và lựa chọn đặc trưng dữ liệu

Đặc trưng (features): Là tập hợp các thuộc tính (attributes), thường biểu diễn như là một vector, kết hợp với một ví dụ. Trong bài toán phát hiện phân lớp ảnh khí tượng, một số các đặc trưng là: hướng gió, tốc độ gió, lượng mưa, ….[36].

Định nghĩa lựa chọn thuộc tính đặc trưng dữ liệu: Lựa chọn thuộc tính đặc trưng là một quá trình tìm ra một tập con các thuộc tính từ M tập thuộc tính của tập dữ liệu ban đầu. Trích chọn đặc trưng là một quá trình giảm chiều dữ liệu (dimentionality reduction process), ở đó các giá trị thô ban đầu được giảm chiều để tạo thành các đặc trưng cho quá trình xử lý trong khi vẫn phản ánh một cách chính xác dữ liệu ban đầu [70].

Mục đích của lựa chọn đặc trưng dữ liệu: Trong học máy, nhận dạng mẫu, và xử lý ảnh, trích rút đặc trưng (feature extraction) tạo ra các giá trị (features) from từ tập dữ liệu ban đầu sao cho các đặc trưng biểu diễn cho dữ liệu để sử dụng cho quá trình học máy và tổng quát hoá (generalization).

b) Các phương pháp lựa chọn đặc trưng dữ liệu

Về cơ bản việc trích rút các thuộc tính đặc trưng dữ liệu bao gồm hai phần: (i) Xây dựng các thuộc tính; (ii) Lựa chọn các thuộc tính đặc trưng.

Lựa chọn các thuộc tính đặc trưng: Là tìm ra những thuộc tính đại diện cho đối tượng, loại bỏ những thuộc tính thừa và gây nhiễu nhằm tăng hiệu suất của các thuật toán khai phá dữ liệu. Lựa chọn thuộc tính có thể tiến hành theo hai cách [71]:

- Cách 1: Xếp loại các thuộc tính theo một tiêu chuẩn nào đó và lấy ra k thuộc tính đầu tiên, do đó cách này là dựa vào ngưỡng để chọn thuộc tính.

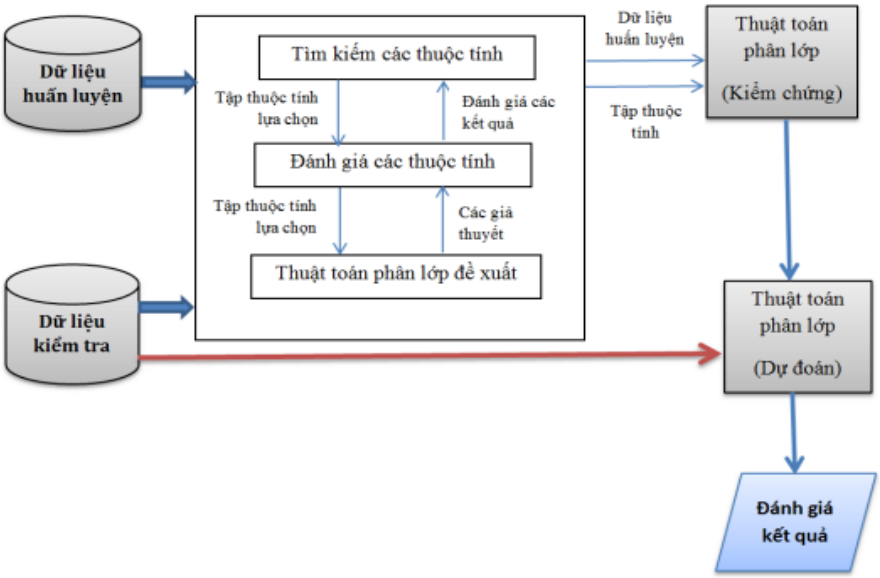
- Cách 2: Chọn ra tập con nhỏ nhất mà không làm giảm đi quá trình học, do đó với cách này tự động xác định số lượng thuộc tính.

Phương pháp lựa chọn thuộc tính là sự tổng hợp của ba thành phần: Chiến lược tìm kiếm, đánh giá chất lượng thuộc tính và ước lượng lựa chọn thuộc tính và mô hình lựa chọn thuộc tính.

(i). **Chiến lược tìm kiếm**: Lựa chọn thuộc tính có thể được xem như là một vấn đề tìm kiếm, trong đó mỗi bước trong không gian tìm kiếm xác định được một tập con thuộc tính liên quan.

(ii). **Ước lượng lựa chọn lựa thuộc tính:** Việc ước lượng lựa chọn lựa thuộc tính tối ưu gồm 2 nhiệm vụ: Một là, so sánh hai giai đoạn: trước và sau khi lựa chọn thuộc tính; Hai là, so sánh hai thuật toán lựa chọn thuộc tính [72].

(iii). **Các loại mô hình lựa chọn thuộc tính:** Về cơ bản có 02 phương pháp mô hình lựa chọn thuộc tính theo hai cách tiếp cận khác nhau là Filter và Wrapper: Mô hình Filter: Lựa chọn thuộc tính theo phương pháp đo lường khoảng cách và đo lường thông tin. Mô hình Wrapper: Lựa chọn thuộc tính bằng cách sử dụng độ chính xác của bộ phân lớp như một thước đo hiệu quả của bộ phân lớp (hình 2.21).

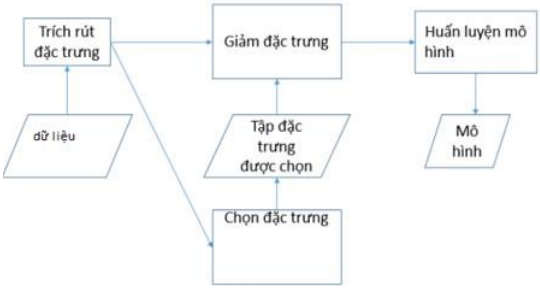


Hình 2.21: Mô hình lựa chọn thuộc tính đặc trưng Wrapper

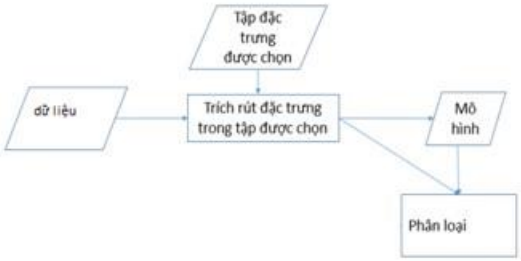
Mô hình Wrapper gồm 2 giai đoạn:

- Giai đoạn 1: Lựa chọn tập con thuộc tính tốt nhất dựa trên tiêu chí độ chính xác (của bộ dữ liệu tập huấn).

- Giai đoạn 2: Học và đánh giá mô hình (Learning and Testing), một bộ phân lớp sẽ học các tri thức từ dữ liệu tập huấn thông qua một tập các thuộc tính tốt nhất được chọn lựa, và được kiểm tra lại bằng một bộ dữ liệu test (hình 2.23).



Hình 2.22: Sơ đồ quá trình huấn luyện



Hình 2.23: Sơ đồ quá trình kiểm tra

2.4.2.5. Phương pháp phân đoạn dữ liệu theo thời gian

a) Các khái niệm dữ liệu theo thời gian

- Dữ liệu theo thời gian (time series data): là loại dữ liệu có sự biến đổi theo thời gian. Dữ liệu này xuất hiện nhiều trong thực tế và trong các bài toán KTTV[73].

- Dữ liệu thời gian thực hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận là thời gian.

- Phương pháp chuỗi thời gian sẽ dựa trên việc phân tích chuỗi quan sát của một biến duy nhất theo biến số độc lập là thời gian.

- Dữ liệu theo chuỗi thời gian: Trong các bài toán dự báo dựa trên dữ liệu KTTV, dữ liệu thường được biểu diễn dưới dạng chuỗi thời gian. Dữ liệu thời gian thực hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận là thời gian. Các dữ liệu chuỗi thời gian thường chia thành 4 thành phần [37]:

- Thành phần xu hướng dài hạn (long - term trend component): dùng để chỉ xu hướng tăng hay giảm của đại lượng X trong thời gian dài.

- Thành phần mùa (seasional component): Dùng để chỉ xu hướng tăng hay giảm của đại lượng X tính theo mùa trong năm.

- Thành phần chu kỳ (cyclical component): Thành phần này chỉ sự thay đổi của đại lượng X theo chu kỳ.

- Thành phần bất thường (irregular component): Dùng để chỉ sự thay đổi bất thường của các giá trị trong chuỗi thời gian, thành phần này không có tính chu kỳ.

Mặc dù các nghiên cứu ở Việt Nam cho các mô hình phân tích dữ liệu theo thời gian chủ yếu mới được áp dụng trong lĩnh vực tài chính, chứng khoán, tuy nhiên, các

kết quả khả quan trong lĩnh vực này là tiền đề tốt để áp dụng các kỹ thuật dự đoán theo thời gian cho dữ liệu KTTV.

b) Các phương pháp phân tích dữ liệu theo thời gian

Sử dụng các mô hình phân tích dữ liệu theo thời gian gồm:

(i). Quá trình trung bình trượt: Quá trình MA(1) mô tả quá trình y_t (giá tài sản tài chính, trái phiếu, cổ phiếu, tỷ giá...) theo thời gian phụ thuộc vào u_t (nhiễu trắng) nhưng không phụ thuộc vào biến trễ của nó.

$$y_t = \mu + u_t + \theta u_{t-1} \tag{2.4.6}$$

(ii). Quá trình tự hồi quy AR (Autoregressive) gồm:

- Quá trình tự hồi quy cấp 1 AR(1) không có hệ chặn có dạng sau:

$$y_t = \varphi y_{t-1} + u_t \tag{2.4.7}$$

- Quá trình tự hồi quy cấp 1 AR (1) có hệ chặn có dạng sau:

$$y_t = \alpha + \varphi y_{t-1} + u_t \tag{2.4.8}$$

- Quá trình tự hồi quy cấp p AR(p) có dạng sau:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + u_t \tag{2.4.9}$$

- Quá trình trung bình trượt tự hồi quy ARMA(p,q); là quá trình tích hợp của hai quá trình tự hồi quy AR và trung bình trượt MA, nó có dạng tổng quát sau:

m $$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} \tag{2.4.10}$$

- Dự báo quá trình AR(p): được thực hiện thông qua hàm sau:

$$Ey_t = \mu = \frac{\alpha}{1-\varphi} \Rightarrow \alpha = \mu(1-\varphi) \tag{2.4.11}$$

- Dự đoán quá trình MA(q) có dạng:

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} \tag{2.4.12}$$

- Dự đoán quá trình ARMA(p; q) có dạng:

$$\left(1 - \varphi_1 L - \varphi_2 L^2 - \ldots - \varphi_p L^p\right)(y_t - \mu) = \left(1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q\right) u_t \tag{2.4.13}$$

- Dự đoán quá trình ARIMA(p; d; q) có dạng:

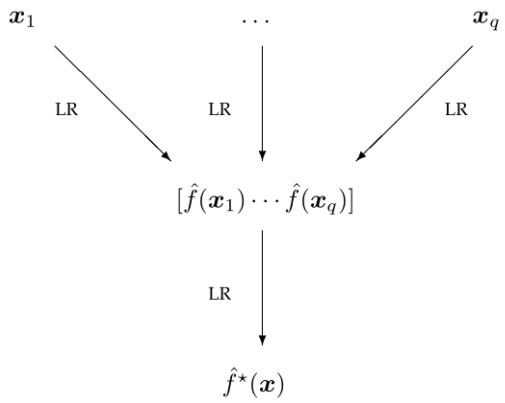
$$\left(1-\varphi_1 L-\varphi_2 L^2-\ldots-\varphi_p L^p\right)\left(y_t^*-\mu\right)=\left(1+\theta_1 L+\theta_2 L^2+\ldots+\theta_q L^q\right)u_t \quad (2.4.14)$$

(iii). Phương pháp hồi quy tuyến tính trong hai bước two-step LR

Hồi quy tuyến tính được áp dụng cho mỗi tập con đặc trưng $x_g$xg để được hàm hồi quy tuyến tính $\hat{f}(x_g)$ như sau:

$$\hat{f}(\boldsymbol{x}_g)=\boldsymbol{x}_g^T\hat{\boldsymbol{\beta}}_g, \quad (2.4.15)$$

Lược đồ của quá trình thực hiện two-step LR như sau:



Hình 2.24: Lược đồ của hồi quy tuyến tính hai bước - two-step LR

2.4.3. Các phương pháp xây dựng các mô hình ML hỗ trợ dự báo KTTV

2.4.3.1. Khái niệm học máy

Học máy - ML (M*achine Learning*) là một lĩnh vực của AI liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. ML có liên quan chặt chẽ đến thống kê, vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, ML tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán.

2.4.3.2. Mô hình học có giám sát - SL (Supervised Learning)

a) Khái niệm và các ứng dụng của mô hình học máy có giám sát

Học có giám sát (SL) là "huấn luyện" máy tính dựa trên những quan sát có dán nhãn. Những quan sát là những câu hỏi, và nhãn của chúng là những câu trả lời. Ý tưởng của SL là bằng việc ghi nhớ và tổng quát hóa một số quy tắc từ một tập câu hỏi có đáp án trước, máy tính sẽ có thể trả lời được những câu hỏi dù chưa từng gặp phải, nhưng có mối liên quan. Các ứng dụng chính của mô hình học có giám sát SL là xây dựng các công cụ phân loại và hồi quy dữ liệu.

b) Các thuật toán sử dụng trong mô hình học máy SL

Các thuật toán, phương pháp sử dụng trong mô hình học máy SL gồm:

(i) Thuật toán máy vecto hỗ trợ SVM (Support vector machine) [90]:

Máy véc tơ hỗ trợ hồi quy SVR là một phương pháp để cực tiểu hóa sự phức tạp mô hình bằng cách cộng thêm giá trị này vào hàm lỗi. Để minh họa ta xem xét một mô hình học máy tuyến tính dự báo cho bởi công thức:

$$f(x) = w^T x + b \tag{2.4.16}$$

Trong đó w là véc-tơ trọng số, b là độ dốc và x là véc-tơ đầu vào. Gọi x_m và y_m lần lượt là véc-tơ đầu vào, giá trị đầu ra thứ m của tập huấn luyện. Công thức tính hàm lỗi như sau:

$$J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}|y_m - f(x_m)|_\varepsilon \tag{2.4.17}$$

Số hạng thứ nhất của hàm lỗi chính là giá trị phạt độ phức tạp của mô hình, còn số hạng thứ hai là giá trị lỗi nhạy cảm với ε. Nếu hàm lỗi nhỏ hơn ε thì sẽ không phạt, đây là tham số được đưa thêm vào để điều chỉnh giảm độ phức tạp của mô hình. Vì vậy lời giải sẽ cực tiểu hóa hàm lỗi như công thức sau:

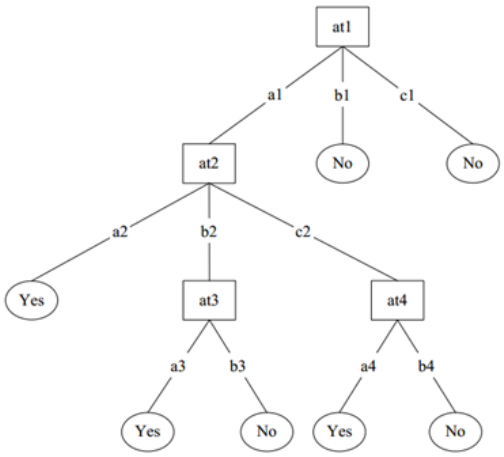
$$f(x) = \sum_{m=1}^{M}(\alpha_m^* - \alpha_m)x_m^T x + b \tag{2.4.18}$$

Trong đó α_m^*, α_m là nhân tử Lagrange. Véc-tơ huấn luyện đưa ra các số nhân Lagrange khác không được gọi là các véc-tơ hỗ trợ và đây là một khái niệm chính về lý thuyết SVR. Các véc-tơ không hỗ trợ không đóng góp trực tiếp vào lời giải và số lượng vectơ hỗ trợ là độ đo đo độ phức tạp của mô hình. Mô hình này được mở rộng cho trường hợp phi tuyến tính thông qua khái niệm nhân κ sinh ra công thức sau:

$$f(x) = \sum_{m=1}^{M}(\alpha_m^* - \alpha_m)\kappa(x_m^T x) + b \tag{2.4.19}$$

(ii) Thuật toán dựa trên hồi quy logic (cây quyết định) [90]:

Mô hình học máy có giám sát sử dụng thuật toán học hồi quy logic gồm: cây quyết định DT (Decision Tree) và phân lớp dựa trên tập luật. Thuật toán DT [90] là cây phân loại trường hợp bằng cách sắp xếp chúng dựa trên các giá trị đặc trưng. Mỗi một nút trong của cây tương ứng với một biến; cạnh nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự báo của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. DT có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên RF (random forest) sử dụng một số DT để có thể cải thiện tỉ lệ phân loại.



Hình 2.25: Mô hình thuật toán Cây quyết định

```
(1)  ComputerClassFrequency(T);
(2)  if OneClass or FewCases
return a leaf;
Create a decision node N;
(3)  ForEach Attribute A
ComputeGain(A);
(4)  N.test=AttributeWithBestGain;
(5)  if N.test is continuous
find Threshold;
(6)  ForEach T' in the splitting of T
(7)  if T' is Empty
Child of N is a leaf
else
(8)  Child of N=FormTree(T');
(9)  ComputeErrors of N;
return N
```

Hình 2.26: Giả mã của thuật toán Cây quyết định C4.5

(iii) Thuật toán K láng giềng gần nhất k-NN (k-Nearest Neighbor):

Là thuật toán để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp và tất cả các đối tượng trong tập dữ liệu. Một đối tượng được phân lớp dựa vào k láng giềng của nó, k là số nguyên dương được xác định trước khi thực hiện thuật toán [92]. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.Các bước thực hiện của thuật toán như sau: thực hiện cấu hình tham số k – số điểm lân cận; đánh giá 1 điểm mới của tập kiểm thử bằng cách xét k lân cận của nó; phân lớp cho điểm mới dựa trên nhãn của đa số mà k điểm trong tập huấn luyện gần nhất của nó được gán.

k-Nearest Neighbor

Classify (X, Y, **x**) // **X**: training data, **Y**: class labels of X, **x:** unknown sample **for i = 1 to m do**

Compute distance $\mathbf{d(X_{jt}\ x)}$

end for

Compute set / containing indices for the **k** smallest distances (X*. x).

return majority label for {Yj where **i** € /}

Hình 2.27: Thuật toán K láng giềng gần nhất k-NN

Minkowsky: $D(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^r\right)^{1/r}$
Manhattan: $D(x,y)=\sum_{i=1}^{m}\lvert x_i-y_i\rvert$
Chebychev: $D(x,y)=\max_{i=1}^{m}\lvert x_i-y_i\rvert$
Euclidean: $D(x,y)=\left(\sum_{i=1}^{m}\lvert x_i-y_i\rvert^2\right)^{1/2}$
Camberra: $D(x,y)=\sum_{i=1}^{m}\frac{\lvert x_i-y_i\rvert}{\lvert x_i+y_i\rvert}$
Kendall's Rank Correlation: $D(x,y)=1-\frac{2}{m(m-1)}\sum_{i=j}^{m}\sum_{j=1}^{i-1}sign(x_i-x_j)sign(y_i-y_j)$

Hình 2.28: Các hàm tính khoảng cách

(iv) Thuật toán dựa trên Perceptron nhiều lớp MLP (Multi-layer Perceptron):

- Thuật toán dựa trên Perceptron **một** lớp PLA (Perceptron Learning Algorithm): Thuật toán dựa trên Perceptron một lớp là chọn đường biên trước. Xét từng điểm một, nếu điểm đó bị phân lớp sai thì tiến đường biên về phía làm cho điểm đó được phân lớp đúng.

- Thuật toán dựa trên Perceptron **nhiều** lớp MLP (Multi-layer Perceptron): Thuật toán dựa trên Perceptron nhiều lớp là tập hợp của các perceptron chia làm nhiều nhóm, mỗi nhóm tương ứng với một layer. Hoạt động của chúng có thể được mô tả như sau tại tầng đầu vào các nơron nhận tín hiệu vào xử lý (tính tổng trọng số, gửi tới hàm truyền) rồi cho ra kết quả (là kết quả của hàm truyền); kết quả này sẽ được truyền tới các nơron thuộc tầng ẩn thứ nhất; các nơron tại đây tiếp nhận như là tín hiệu đầu vào, xử lý và gửi kết quả đến tầng ẩn thứ 2; quá trình tiếp tục cho đến khi các nơron thuộc tầng ra cho kết quả.

(v) Thuật toán học dựa trên thống kê:

Thuật toán Naives Bayes: là một nhóm các phân loại xác suất đơn giản dựa trên việc áp dụng định lý Bayes với các giả định độc lập giữa các đặc tính[41].

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A) \tag{2.4.20}$$

Trong đó: P(A|B) là xác suất có điều kiện A khi biết B, P(A) là xác suất giả thuyết A (tri thức có được về giải thuyết A trước khi có dữ liệu B), P(B|A) là xác suất có điều kiện B khi biết giả thuyết A,P(B) là xác suất của dữ liệu quan sát B không quan tâm đến bất kỳ giả thuyết A nào. Thuật toán này được áp dụng trong một số bài toán như: Đánh dấu một email là spam hay không; Phân loại bài viết tin tức thuộc lĩnh vực công nghệ, chính trị hay thể thao; Kiểm tra một đoạn văn bản mang cảm xúc tích cực/ tiêu cực; Sử dụng cho các phần mềm nhận diện khuôn mặt. …

(vi) Thuật toán hồi quy logistic:

Hồi quy logistic là một cách thống kê mạnh mẽ để mô hình hóa một kết quả nhị thức với một hoặc nhiều biến giải thích. Nó đo lường mối quan hệ giữa biến phụ thuộc phân loại và một hoặc nhiều biến độc lập bằng cách ước tính xác suất sử dụng một hàm logistic, là sự phân bố tích lũy logistic. Thuật toán này được sử dụng trong một số trường hợp: Điểm tín dụng (quyết định có cho khách hàng vay vốn hay không); Đo mức độ thành công của chiến dịch marketing; Dự đoán doanh thu của một sản phẩm nhất định; Dự đoán động đất ….

c) Nhận xét về các mô hình học máy có giám sát

Các thuật toán học máy có giám sát cho thấy khả năng dự báo vượt trội so với các kỹ thuật hiện đại nhất hiện nay, kết quả dự báo hứa hẹn đạt độ chính xác cao, ngoài ra hoàn toàn có thể tạo ra các mô hình lai ghép để tạo ra mô hình ổn định hơn

và tận dụng được ưu thế của từng mô hình đơn. Nghiên cứu có thể áp dụng và so sánh hiệu suất dự đoán của chuỗi Markov mở rộng và sáu thuật toán học máy phổ biến khác, cụ thể là: Lập trình di truyền, Vector hỗ trợ Hồi quy, Mạng nơron cơ sở xuyên tâm (RBF), Quy tắc M5, Cây mô hình M5 và k - láng giềng gần nhất.

Để hỗ trợ trong đánh giá toàn diện, cần chạy thử nghiệm bằng chuỗi thời gian diễn ra hình thái thời tiết bất thường trên các tập dữ liệu cho khu vực Bắc bộ, Bắc trung bộ, với đặc điểm khí hậu rất đa dạng. Các mô hình thống kê sẽ được áp dụng cho thấy các phương pháp học máy là có khả năng vượt trội so với kỹ thuật hiện đại nhất. Một đóng góp khác của nghiên cứu khi sử dụng các phương pháp học máy có giám sát này là phát hiện mối tương quan giữa các vùng khí hậu khác nhau và độ chính xác dự đoán. Do đó, những kết quả này cho thấy hiệu quả tích cực rằng hệ thống thông minh dựa trên máy học để dự đoán bài toán KTTV dựa trên độ chính xác dự đoán và với các mối tương quan tối thiểu hiện có trên các vùng khí hậu.

Trong phạm vi nghiên cứu của đề tài chúng tôi chỉ sử dụng các mô hình học máy có giám sát.

2.4.3.3. Cơ sở lý thuyết và phương pháp nghiên cứu

Đối với bài toán dự báo lũ, ARIMA là một trong những mô hình tuyến tính hiệu quả và được biết đến rộng rãi nhất áp dụng cho dự báo dữ liệu theo chuỗi thời gian Trong nghiên cứu của Birylo và cộng sự (2018) nhóm tác giả đã sử dụng mô hình ARIMA để dự đoán mực nước ngầm sử dụng 3 tham số, lượng mưa, dòng chảy bề mặt và thoát hơi nước. Kết quả dự báo cho 12 tháng đã chỉ ra rằng mô hình ARIMA có hiệu năng tốt. Mirzavand và Ghazavi đã áp dụng 5 mô hình theo chuỗi thời gian là mô hình trượt trung bình (MA) mô hình trượt trung bình tự hồi quy (ARMA) mô hình trượt trung bình kết hợp tự hồi quy (ARIMA) và mô hình ARIMA theo mùa (SARIMA) và kết hợp với một vài loại mô hình theo chuỗi thời gian để dự báo mực nước ngầm. Các kết quả thử nghiệm chỉ ra rằng việc kết hợp các mô hình chuỗi thời gian cải thiện độ chính xác của dự báo mực nước ngầm. Valipour et al. (2013) đã so sánh hiệu năng dự báo của ARMA ARIMA và mạng thần kinh trí tuệ nhân tạo hồi quy (ANN) trong việc dự báo dòng chảy của đập DEz. Kết quả đã chỉ ra rõ rằng mô hình ARIMA tốt hơn ARMA trong thử nghiệm 12 tháng nhưng mô hình này không tốt bằng nơ-ron nhân tạo (ANN) trong thử nghiệm 60 tháng. Yu et al. (2017) đã nghiên cứu mô hình ARIMA để dự báo mực nước tại 3 trạm tại vùng trung lưu của sông Trường Giang. Kết quả cho thấy độ chính xác của mô hình ARIMA giảm khi kích thước dự báo tăng lên. Phương pháp này cho kết quả tốt với dự báo ngắn hạn nhưng không tốt khi dự báo mực nước dài hạn. Tác giả cũng chỉ ra rằng tính chất phi tuyến tính và tính dừng của chuỗi thời gian có thể dẫn đến sự không

chắc chắn khi sử dụng trực tiếp mô hình ARIMA. Vì vậy cần thiết phải kết hợp các loại mô hình khác nhau để nâng cao hiệu năng dự báo.

Trong những năm gần đây, các phương pháp học máy thống kê (không tham số) (ML) đã góp phần rất lớn vào sự phát triển vượt bậc của hệ thống dự báo, cung cấp các giải pháp với chi phí thấp nhưng hiệu quả mà chỉ cần sử dụng dữ liệu mực nước trong quá khứ. Nguyen et al. (2015) so sánh hiệu năng của 3 mô hình học máy là LASSO, Random Forest (RF) và Support Vector Regression (SVR) trong việc dự báo mực nước của sông MeKong. Mô hình SVR cho kết quả tốt với sai số bình quân 0.486 (m) trong 5 ngày (giá trị sai số chấp nhận của mô hình dự báo lũ lụt là 0.5m-0,75m . Garcia et al. (2016) đã nghiên cứu thuật toán RF để dự báo mực nước của 2 trạm của lưu vực sông Cagaya ò Philipines. Mối tương quan giữa kết quả dự báo và dữ liệu thật chỉ ra rằng hiệu quả dự báo tốt của phương pháp sử dụng ML dự báo mực nước có thể được áp dụng cho các trạm khác trên các lưu vực sông lớn trên khắp Philipines. Pasupa và Jungjareantrat (2016) thực hiện việc dự báo mực nước trong sông Chao Phyra ở Thái Lan sử dụng các phương pháp thống kê học máy là hồi quy tuyến tính (LR), hồi quy hàm nhân (KL), SVR, K-láng giềng gần nhất (KNN) và rừng ngẫu nhiên (RF). Mô hình SVR sử dụng hàm nhân cơ bản Radial và dữ liệu theo chuỗi thời gian 72 tiếng đã đưa ra được kết quả tốt nhất với ít lỗi nhất khi so sánh với phương pháp Quân đội hoàng gia Thái Lan sử dụng.

Trong một số nghiên cứu khác, SVR cũng chứng minh khả năng dự đoán dòng chảy của sông (Garsole và Rajurkar, 2015; Adnan và cộng sự, 2018; Bafitlhile và Li, 2019). Yang và cộng sự. (2017) đã tiến hành dự báo chuỗi thời gian mực nước trên hồ chứa Taiwans Shimen áp dụng năm thống kê phương pháp học máy (mạng RBF, Kstar, KNN, RF và cây ngẫu nhiên). Kết quả thí nghiệm cho thấy RF cho kết quả dự báo chính xác hơn các phương pháp khác. RF cũng đạt được kết quả tốt hơn hơn các phương pháp học máy thống kê khác như SVM, ANN, cây quyết định (DT) khi dự đoán mực nước hàng ngày (Wang et al., 2018). Hipni và cộng sự. (2013) đã áp dụng SVM để dự báo mực nước đập hàng ngày của cửa Klang tại Malaysia so với hệ thống ANFIS. Kết quả đã chứng minh rõ ràng rằng SVM tốt hơn ANFIS. Trong một nghiên cứu khác, (Khan và Coulibaly, 2006) đã dự đoán dài hạn đoán mực nước hồ sử dụng SVM, mạng nơ-ron nhiều lớp (MLP) và mô hình tự hồi quy theo mùa (SAR)). Kết quả cho thấy cách tốt nhất trong ba cách tiếp cận này là SVM.

Mạng nơ-ron nhân tạo (ANNs) là một trong những phương pháp học máy được sử dụng rộng rãi trong lĩnh vực thủy văn như mô hình hóa dòng chảy, đánh giá chất lượng nước và dự báo mực nước. Trong những năm gần đây ANNs được áp dụng để dự báo mực nước (Kim and Seo (2015), Kasiviswanathan et al. (2016)). Phương

pháp học sâu dựa trên ANNs đang phát triển nhanh, cho kết quả chính xác nên thu hút được sự quan tâm của rất nhiều các nghiên cứu trong cả lĩnh vực công nghiệp lẫn giáo dục (LeCun et al 2015). Mạng LSTM là một trong những mô hình học sâu được áp dụng thành công trong lĩnh vực thủy văn như là dự báo lũ lụt (Le et al (2019)) hoặc là dự báo mực nước hồ Hrnjica and Bonacci (2019); Xu et al. (2019)

Mô hình lai có khả năng khai thác được lợi thế của từng mô hình riêng lẻ giúp nâng cao độ chính xác và linh hoạt của mô hình (Zhang (2003). Vì thế một số nghiên cứu đã thực hiện kết hợp các loại mô hình dự báo để cải thiện khả năng dự báo mực nước. (Mousavi-Mirkalaei and Banihabib (2019); Seo et al. (2015); Zhong et al. (2017, 2019); Xu et al.(2019)). Trong nghiên cứu này, trước hết chúng tôi đưa ra hướng kết hợp ARIMA với các phương pháp học máy không tuyến tính và học sâu LSTM cho dự báo mực nước theo chuỗi thời gian. Sự mới lạ của phương pháp này là mô hình hóa thành phần tuyến tính và phi tuyến tính riêng biệt với mô hình thống kê tuyến tính ARIMA và mô hình máy học phi tham số phi tuyến tính LSTM. Kết quả của phương pháp mới này được thử nghiệm với bộ dữ liệu thu được tại sông Hồng

Trong nghiên cứu giải quyết bài toán dự báo nước biển dâng do bão chúng tôi nhận thấy trước đây, cách tiếp cận thông thường để dự báo nước dâng do bão là sử dụng mô hình dự báo số trị, tuy nhiên các mô hình này đòi hỏi mất nhiều năng lực tính toán. Một cách tiếp cận khác là sử dụng các thuật toán học máy như mạng nơ-ron [118] để dự đoán các mối quan hệ giữa mực nước dâng và các đặc trưng tương ứng như là mực nước biển, gió, khí áp trên mặt biển và các đặc tính của cơn bão nhiệt đới. Người ta đã xây dựng mô hình dự báo nước dâng do bão sử dụng một số mô hình AI [119] để dự báo mực nước dâng cao nhất sử dụng các tham số của cơn bão nhiệt đới: áp suất tâm bão, bán kính gió lớn nhất,.. Kết quả cho thấy việc dùng mạng nơ-ron nhân tạo cho kết quả tốt hơn so với máy hỗ trợ véc-tơ. Các kết quả đã chỉ ra rằng phương pháp sử dụng trí tuệ nhân tạo và khung lưới tự do hoàn toàn đáp ứng được độ chính xác với tốc độ dự báo nhanh. So sánh với các mô hình thông thường các mô hình dựa trên mạng nơ-ron có thời gian tính toán nhanh trong khoảng 10 phút sẽ cho ra kết quả dự báo sau khi huấn luyện xong mô hình. Tuy nhiên mô hình dựa trên mạng nơ-ron này là dạng hộp đen vì vậy rất khó để giải thích chúng hơn nữa các mô hình loại này thường không đạt được khả năng ước lượng tại các cao điểm điều này rất quan trọng khi dự báo nước dâng do bão.

2.4.4. Các phương pháp xây dựng mô hình AI hỗ trợ dự báo KTTV

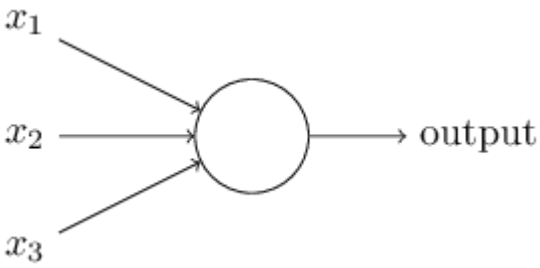
2.4.4.1. Các khái niệm trong mạng Nơ-ron– NN

Mạng nơ-ron NN (Neural Network) là những mô hình được lập trình tối ưu. Thông thường, để lập trình, ta “nói” cho máy tính biết phải làm gì, được xác định

chính xác để máy tính thực hiện [62]. Ngược lại, trong một NN, chúng ta không "nói" cho máy tính biết cách giải quyết vấn đề. Thay vào đó, NN học hỏi từ dữ liệu quan sát, tìm ra giải pháp riêng cho vấn đề hiện tại. Năm 2006 là việc phát hiện ra các kỹ thuật học trong một loại cấu trúc mạng mới gọi là mạng nơ-ron sâu (DN-deep network). Những kỹ thuật này hiện được gọi là học sâu (DL- deep learning). Ngày nay, DN và DF đạt được hiệu suất vượt trội trong nhiều vấn đề quan trọng về thị giác máy tính, nhận dạng giọng nói và xử lý ngôn ngữ tự nhiên.

Kỹ thuật DL đang được triển khai trên quy mô lớn bởi các công ty như Google, Microsoft và Facebook.

Mạng nơ-ron (NN) là một loại tế bào thần kinh nhân tạo (ANN) được gọi là perceptron. Ví dụ về hoạt động của perceptionron như hình 2.29.

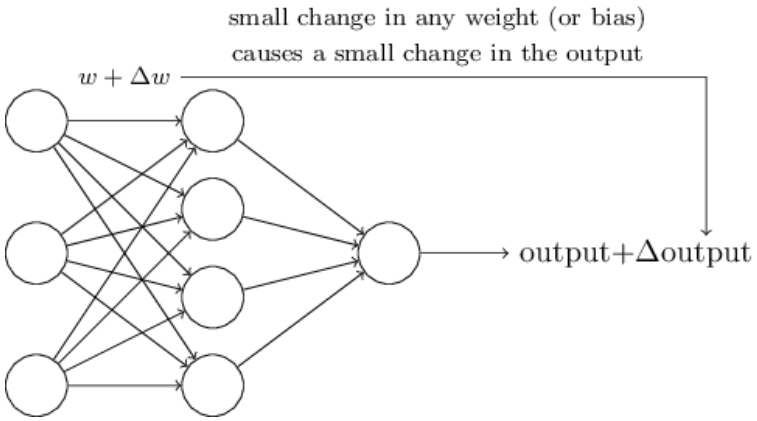


Hình 2.29: Mô hình hoạt động của Perceptron

Trong ví dụ trên, perceptron có ba đầu vào, x1, x2, x3 (Smith n.d.). Một quy tắc đơn giản để tính toán đầu ra bằng các trọng số, w1, w2, w3. Đầu ra của nơ-ron, 00 hoặc 11, được xác định bởi liệu tổng trọng số ∑jwjxj nhỏ hơn hoặc lớn hơn một số giá trị ngưỡng như trong công thức 2.4.21:

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{ threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{ threshold} \end{cases} \tag{2.4.21}$$

Thuật toán học cho mạng NN rất quan trọng để NN có thể hoạt động cho mục đích nào đó. Ví dụ: các đầu vào vào mạng có thể là dữ liệu pixel thô từ hình ảnh được viết tay, được quét bằng chữ số. Mục tiêu là muốn NN tìm hiểu các trọng số để đầu ra từ mạng phân loại chính xác chữ số. Để xem cách học có thể hoạt động được không, giả sử chúng ta thực hiện một thay đổi nhỏ trong một số trọng số (hoặc sai lệch) trong mạng. Những gì chúng ta muốn là cho sự thay đổi nhỏ về trọng số này chỉ gây ra một thay đổi nhỏ tương ứng trong đầu ra từ mạng.



Hình 2.30: Mô hình huấn luyện mạng NN

Do vậy, cần cho mạng NN là các Nơ-ron kích hoạt. Nghĩa là đầu ra của Nơ-ron sẽ được đi qua một hàm kích hoạt như hàm Sigmoid sau.

$$\sigma(z) \equiv \frac{1}{1+e^{-z}}.$$

2.4.22)

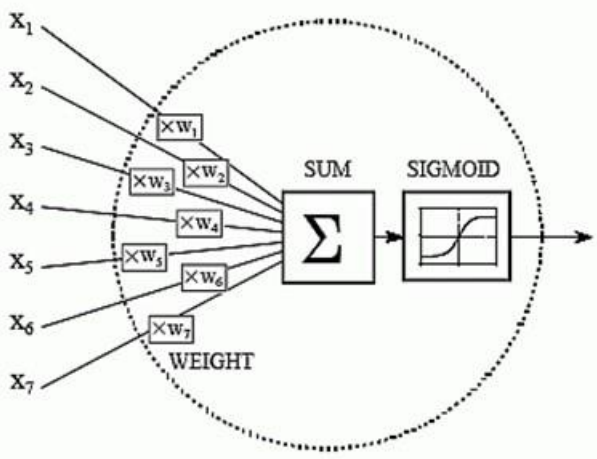
Thay x_1, x_2, x_3 và trọng số w_1, w_2, w_3, và bias b vào công thức (2.4.28) ta có đầu ra của nơ-ron là:

$$\frac{1}{1+\exp(-\sum_j w_j x_j - b)}.$$

(2.4.23)

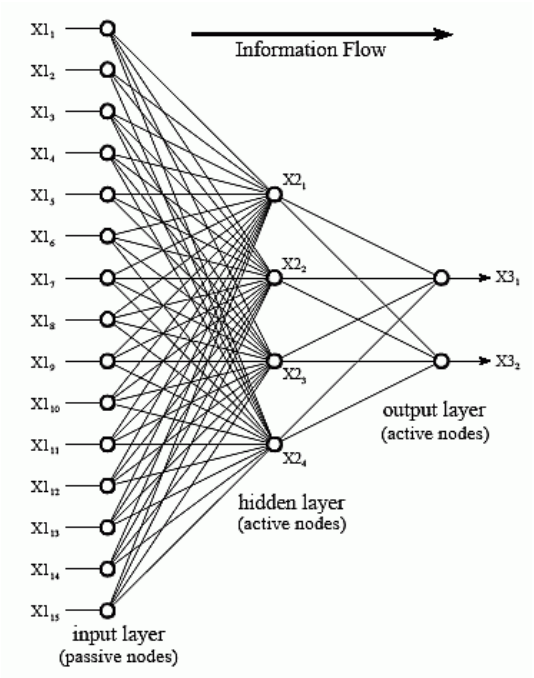
Để hiểu sự tương đồng với mô hình perceptron, giả sử z ≡ w*x + b là một số dương lớn. Khi đó $e^{-z} \approx 0$ và do đó σ (z) ≈ 1. Nói cách khác, khi z = w*x + b là dương và rất lớn, đầu ra từ một nơ-ron sigmoid là xấp xỉ 1, giống như nó đã có cho một perceptron.

Mặt khác, giả sử z = w*x + b rất âm. Khi đó $e^{-z} \rightarrow$ vô cùng và σ (z) ≈ 0. Vì vậy, khi z = w*x + b rất nhỏ, hành vi của một nơ-ron sigmoid cũng gần đúng với một perceptron. Chỉ khi w*x + b lớn thì có nhiều sai lệch so với mô hình perceptron. Do vậy, gắn hàm kích hoạt vào đầu ra của mỗi nơ-ron, lúc đó, mỗi một nơ-ron sẽ có dạng như hình 2.31.



Hình 2.31: Cấu trúc một Nơ-ron

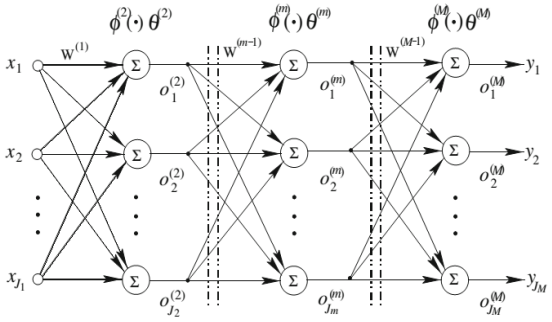
Hình 2.32 mô tả luồng đi của thông tin ở các tầng ẩn và tầng đầu ra của mạng nơ-ron. Mỗi giá trị đầu vào (x_i) được nhân với một trọng số (w_i) sau đó cộng lại cho qua hàm kích hoạt sẽ được giá trị đầu ra cho Nơ-ron. Sau đó mạng nơ-ron sẽ có kiến trúc tổng thể gồm ba lớp và mỗi Nơ-ron ở lớp ẩn và lớp đầu ra có kiến trúc như đã được mô tả (Hình 2.32).



Hình 2.32: Kiến trúc của mạng Nơ-ron

2.4.4.2. Mô hình mạng Nơ-ron hướng tiến FNN

MLP (Multi-layer Perceptrons) là một mạng Nơ-ron hướng tiến FNN (FeedForward Neural Networks) có một hay nhiều lớp neuron ẩn giữa tầng/ lớp vào và tầng/lớp ra[74]. Kiến trúc của MLP được mô tả như trong hình 2.33.



Hình 2.33: Mạng Nơ-ron hướng tiến FNN

Giả thuyết có mạng Nơ-ron có M lớp, trong đó mỗi lớp có J_m, m=1,..., M neuron. Trọng số của các khớp nối từ tầng thứ m-1, đến tầng m, được ký hiệu bằng ma trận W^{m-1}, độ lệch (bias), đầu ra, hàm kích hoạt (activation function) của neuron thứ i trong tầng m được ký hiệu tương ứng là: $\theta_i^{(m)}$, $o_i^{(m)}$, $\phi_i^{(m)}$. Đầu ra của mạng MLP dựa trên đầu vào là mẫu huấn luyện p, được định nghĩa theo phương trình sau[75]:

$$\hat{\boldsymbol{y}}_p = \boldsymbol{o}_p^{(M)}, \quad \boldsymbol{o}_p^{(1)} = \boldsymbol{x}_p, \tag{2.4.24}$$

Vector đầu ra mạng tầng thứ m, trên mẫu huấn luyện p, được tính theo:

$$\boldsymbol{net}_p^{(m)} = \left[\mathbf{W}^{(m-1)}\right]^T \boldsymbol{o}_p^{(m-1)} + \boldsymbol{\theta}^{(m)} \tag{2.4.25}$$

Trong đó:

$$\boldsymbol{net}_p^{(m)} = \left(net_{p,1}^{(m)}, \ldots, net_{p,J_m}^{(m)}\right)^T \quad (2.4.26)$$

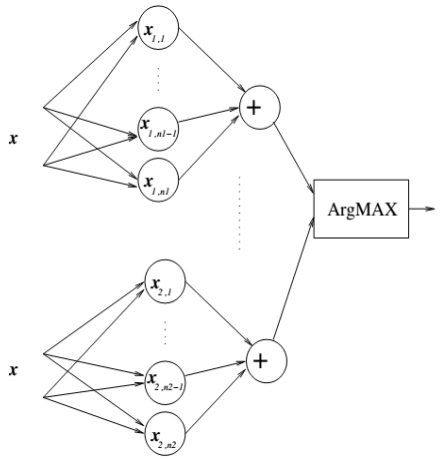
$W^{(m-1)}$ là ma trận cỡ $J^{m-1} \times J^m$, θ^m là vector độ lệch. Đầu ra của mạng sẽ được đưa vào hàm kích hoạt để tính toán đầu ra của các Neuron trong tầng thứ m theo:

$$\boldsymbol{o}_p^{(m)} = \phi^{(m)}\left(net_p^{(m)}\right) \quad (2.4.27)$$

Trong đó hàm kích hoạt $\phi^{(m)}$ có thể ở một số dạng hàm như hàm đồng nhất, hàm tangent, hàm sigmoid. Trong các tầng ẩn thường các neuron sẽ có hàm kích hoạt giống nhau và là các hàm trơn khả vi (nhiều lần) như hàm sigmoid.

2.4.4.3. Mô hình mạng Nơ-ron xác suất PNN

Mạng Nơ-ron xác suất PNN (Probabilistic Neural Networks) là dạng sơ khởi của mạng cơ sở bán kính RBF (Radial Basis Function) sử dụng các cửa sổ Pazen cho bài toán xấp xỉ các hàm mật độ xác suất, và về lý thuyết có thể đạt được hiệu suất phân lớp tối ưu theo nghĩa Bayes [74]. Cấu trúc của các mạng PNN gồm có 4 tầng.



Hình 2.34: Kiến trúc mạng Nơ-ron xác suất PNN

Tầng 1 là tầng đầu vào, tầng thứ 2 là các hàm Gaussian sử dụng tập dữ liệu huấn luyện là các tâm hàm, tầng 3 là tầng tổng hợp tín hiệu đầu ra từ tầng 2, và tầng 4 là tầng đầu ra (giá trị hàm thực với bài toán hồi quy, và số hiệu lớp với bài toán phân lớp). Cấu trúc và quá trình học của PNN tương tự như mạng RBF. Cả mạng RBF và mạng Nơ-ron đa lớp MLP đều có thể dùng để giải quyết các bài toán học có giám sát là phân lớp (classification) và hồi quy (regression). Trong khi mạng MLP sử dụng các lớp Neuron để tính toán với hàm chuyển thì mạng RBF sử dụng các hàm cơ sở với đầu vào là khoảng cách từ điểm dữ liệu đến các tâm của các hàm cơ sở này [75].

2.4.4.4. Mô hình mạng Nơ-ron mờ

Trong thực tế tính toán và lập luận rất nhiều trường hợp thông tin không thể biểu diễn một cách tường minh (bằng các con số) hay chắc chắn mà thường biểu diễn bởi các biến ngôn ngữ. Trong trường hợp như vậy phương pháp tốt nhất để biểu diễn

thông tin có độ bất định loại này là dựa trên lý thuyết về tập mờ (Fuzzy Sets) và logic mờ (Fuzzy Logic) được đưa ra bởi Zadeh.

a) Định nghĩa Tập mờ

Tập mờ (Fuzzy Sets). Một tập mờ *A* trong χ được định nghĩa như sau [74]:

$$\mathcal{A} = \{(x, \mu_{\mathcal{A}}(x)) | \, x \in \mathcal{X}\}, \tag{2.4.28}$$

Trong đó $\mu_A(x) \in [0,1]$ được gọi là hàm thành viên/ hàm thuộc (membership function) của *A*. Khi $\mu_A(x) = 1$, x thuộc *A*, khi $\mu_A(x) = 0$, x không thuộc *A*. Hoặc tập mờ A cũng có thể được định nghĩa như sau:

$$\mathcal{A} = \begin{cases} \sum_{x_i \in \mathcal{X}} \frac{\mu_{\mathcal{A}}(x_i)}{x_i}, & \text{if } \mathcal{X} \text{ is discrete} \\ \int_{\mathcal{X}} \frac{\mu_{\mathcal{A}}(x)}{x}, & \text{if } \mathcal{X} \text{ is continuous} \end{cases}. \tag{2.4.29}$$

b) Trích rút luật mờ từ mạng NN đa lớp MLP

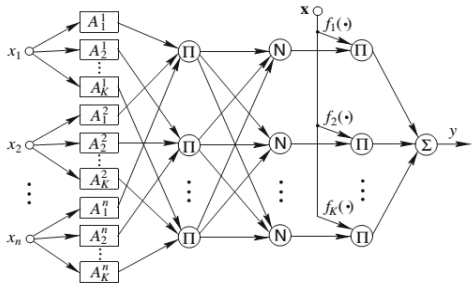
Một trong những điểm yếu của mạng NN thường bị phê phán là hoạt động như một hộp đen và không có khả năng giải thích. Do đó, đã có nhiều nghiên cứu để nhằm bạch hóa hoạt động của mạng NN, một trong những hướng triển vọng chính là trích rút luật mờ từ các mạng NN. Điều này cũng giúp minh chứng cho sự tương đương về khả năng học và giải quyết vấn đề của mạng NN và Hệ mờ. Toán tử mờ iteractive-or (i-or) được định nghĩa nhờ vào việc áp dụng khái niệm f-đối ngẫu của hàm logistics [75]. Toán tử i-or được định nghĩa như sau:

$$a \otimes b = \frac{a \cdot b}{(1-a) \cdot (1-b) + a \cdot b}. \tag{2.4.30}$$

Sử dụng toán tử-or có thể thiết lập cách thức trích rút trí thức học được từ dữ liệu trong mạng NN 3 lớp MLP.

c) Mô hình Neuro fuzzy ANFIS

ANFIS là mô hình Neurofuzzy phổ biến nhất kết hợp giữa mạng NN đa lớp MLP và logic, hệ mờ. Mô hình ANFIS có 6 lớp (n-nK-K-K-K-1), biểu diễn đồ thị cho mô hình hệ TSK. Ký tự N trong hình tròn ký hiệu cho toán tử chuẩn hóa, và x= (x1,x2,...,xn)T.

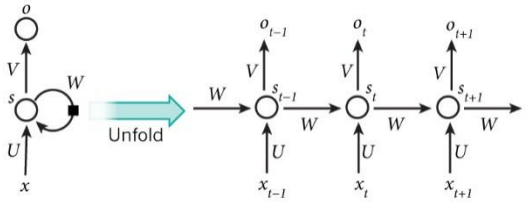


Hình 2.35: Biểu diễn mô hình ANFIS

2.4.4.5. Mô hình mạng Nơ-ron truy hồi RNN

a) Định nghĩa mạng Nơ-ron truy hồi RNN

Mạng Nơ-ron truy hồi RNN (Recurrent Neural Network) khi thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó [76]. Nói cách khác, RNN có khả năng nhớ các thông tin được tính toán trước đó. Một mạng RNN cơ bản có dạng như sau:



Hình 2.36: Phép triển khai của một mạng Nơ-ron hồi quy RNN

Việc tính toán bên trong RNN được thực hiện như sau:

- x_t là đầu vào tại bước t;

- s_t là trạng thái ẩn tại bước t, là bộ nhớ của mạng, s_t được tính toán dựa trên cả các trạng thái ẩn phía trước và đầu vào tại bước đó:

$$s_t = f(Ux_t + Ws_{t-1}) \tag{2.4.37}$$

- Hàm f thường là một hàm phi tuyến tính.

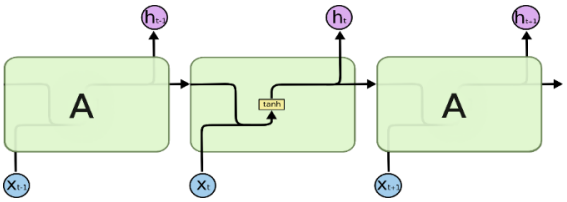
- o_t là đầu ra tại bước t.

b) Các khả năng ứng dụng của RNN

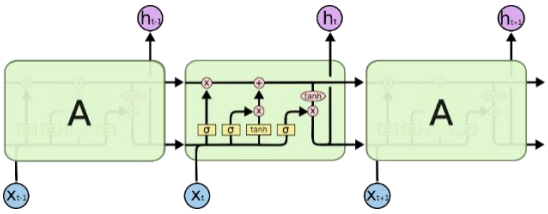
Các ứng dụng của RNN trong xử lý ngôn ngữ tự nhiên gồm: Mô hình hóa ngôn ngữ và sinh văn bản; dịch máy; nhận dạng giọng nói; mô tả hình ảnh.

c) Mạng RNN mở rộng

Các mạng RNN mở rộng để xử lý các nhược điểm của mô hình RNN truyền thống như sau: Mạng hồi quy RNN 2 chiều (Bidirectional RNN); Mạng hồi quy RNN 2 chiều sâu (Deep Bidirectional RNN); Mạng bộ nhớ dài-ngắn LSTM (Long Short Term Memory Networks). Các mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-đun này có cấu trúc rất đơn giản, thường là một tầng. LSTM các mô đun có cấu trúc 4 tầng tương tác đặc biệt với nhau.



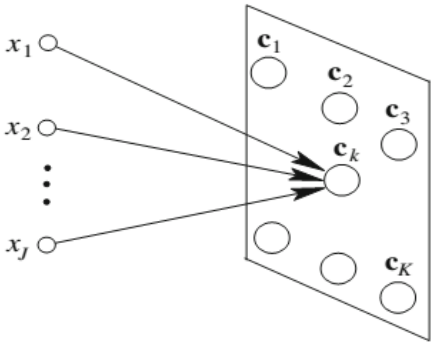
Hình 2.37: Mô đun lặp một tầng đơn của mạng RNN chuẩn



Hình 2.38: Mô đun lặp 4 tầng của mạng LSTM

2.4.4.6. Mô hình mạng tự tổ chức SOM

Mạng Kohonen là một mạng NN hướng tiến J-K với các nút được kết nối đủ. Tầng ra được gọi là tầng Kohonen, các nút đầu vào được kết nối đầy đủ với tầng neuron đầu ra với các trọng số kết nối. Mạng Kohonen sử dụng mô hình học tương tranh, các mẫu được nạp vào mạng tuần tự theo thời gian và không cần chỉ ra đầu ra lý tưởng.



Hình 2.39: Cấu trúc mạng tự tổ chức Kohonen

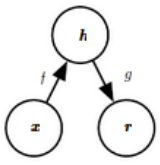
Mạng SOM (Self Organizing Maps) cơ bản là một mạng NN dùng kiến trúc của mạng Kohonen. Với mỗi đầu vào là vector dữ liệu **x**, SOM tính khoảng cách Euclide từ các neuron k tới **x** và tìm ra neuron có prototype gần **x** nhất theo:

$$\|\boldsymbol{x}_t - \boldsymbol{c}_w\| = \min_{k \in \mathcal{A}} \|\boldsymbol{x}_t - \boldsymbol{c}_k\|, \tag{2.4.31}$$

Trong đó w ký hiệu chỉ số của neuron thắng cuộc, được gọi là exictation center, là tâm của một nhóm vector nằm gần nhất với $\mathbf{c}_w$.

2.4.4.7. Mô hình mạng Auto-Encode (Bộ mã hóa tự động AE)

Bộ mã hóa tự động Auto Encoder - AE (hình 2.40) là một mạng NN được huấn luyện để cố gắng sao chép đầu vào đến đầu ra của nó. Mạng có thể được xem như bao gồm hai phần: một chức năng mã hóa (encoder) h = f (x) và một bộ giải mã (decoder) để tạo ra một bản khôi phục lại đầu vào r = g(h).

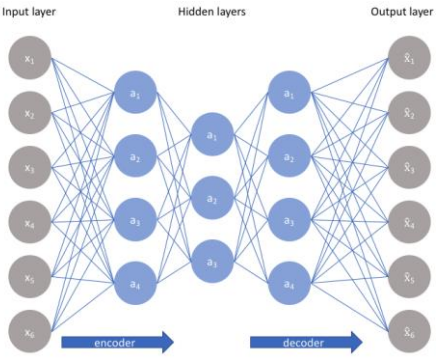


Hình 2.40: Cấu trúc AE

Nếu một AE thành công chỉ đơn giản là học cách thiết lập g(f (x) = x) với tất cả các mẫu x. AE được thiết kế để không thể học cách sao chép hoàn hảo (LeCun

2015). Các AE hiện đại có khả năng tổng quát hóa ý tưởng của cả encoder và decoder từ các hàm xác định đến các ánh xạ ngẫu nhiên $p_{encoder}(h|x)$ và $p_{decoder}(x|h)$.

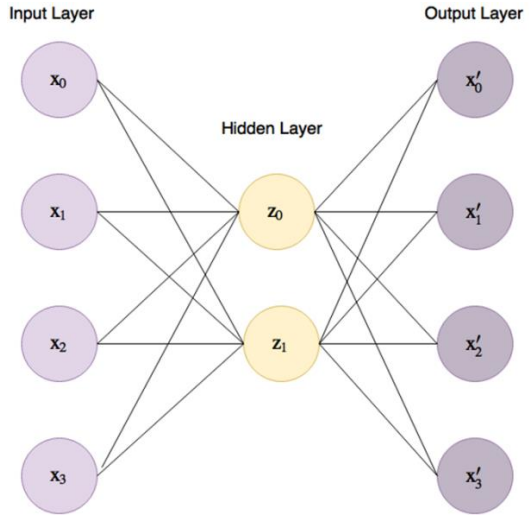
Một AE có hai thành phần là encoder f và decoder g. Một số loại AE phổ biến sau: (i) AE không đầy đủ (Undercomplete AE); (ii) AE hiệu chỉnh (Regularized AE); (iii) AE thưa (Sparse AE); (iv) AE với nhiễu (Denoising AE); (v) AE hiệu chỉnh bằng hàm phạt (Regularizing by Penalizing Derivatives).



Hình 2.41: Minh họa về cấu trúc mạng Undercomplete AE

a) Ứng dụng AE trong giảm chiều dữ liệu

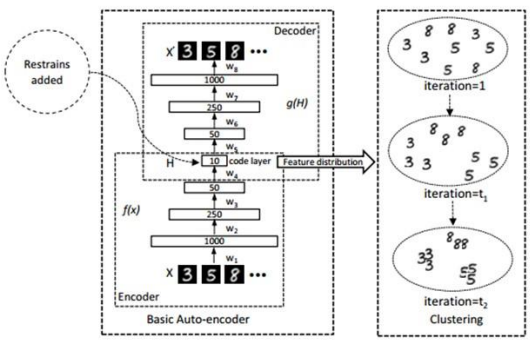
AE là một ứng dụng để ánh xạ dữ liệu sang một không gian biểu diễn khác. Khi ta giới hạn số node ẩn ở biến ẩn (m) phải lớn hơn số node biểu diễn dữ liệu đầu vào (n) thì kết quả tương tự như sparse AE. Khi cấu trúc m nhỏ hơn n thì kết quả nhận được một biểu diễn nén lại của dữ liệu đầu vào vì dữ liệu đầu vào được biểu diễn ở không gian có số chiều ít hơn. Đó chính là khả năng giảm chiều dữ liệu của AE.



Hình 2.42: Giảm chiều dữ liệu sử dụng AE

b) Ứng dụng AE trong phân cụm dữ liệu

AE là một phương pháp tốt để phân cụm dữ liệu. Nó cung cấp các ánh xạ phi tuyến bằng cách học encoder và decoder. Encoder của AE thực sự là một hàm ánh xạ phi tuyến và decoder khôi phục dữ liệu gốc từ biến ẩn là kết quả của encoder. Quá trình này lặp lại để đảm bảo các hàm ánh xạ hiệu quả nhất để biểu diễn dữ liệu gốc.



Hình 2.43: Phân cụm dữ liệu sử dụng AE

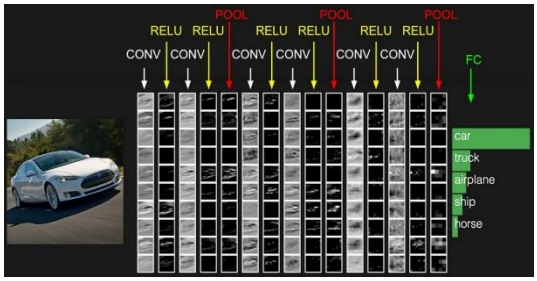
2.4.4.8. Mô hình mạng Nơ-ron tích hợp chập CNN

Mạng NN tích chập CNN (Convolutional Neural Networks) có một số điểm tương tự như mạng NN thông thường. Đó là được tạo bởi các nơ-ron có tham số học là các trọng số và các bias. Mỗi nơ-ron nhận một vài đầu vào, thực hiện tính tích vô hướng giữa trọng số và đầu vào, rồi theo sau bởi một hàm phi tuyến. Toàn bộ mạng nơ-ron sẽ cho ra một hàm điểm số (score function) khả vi: Từ các điểm ảnh của ảnh sẽ cho ra các điểm số của ảnh theo mỗi lớp phân loại và có hàm chi phí (loss function) nằm ở lớp cuối cùng (lớp fully-connected).

Mạng NN thông thường: Nhận một đầu vào (một vector) và truyền qua các lớp ẩn. Mỗi lớp ẩn tạo bởi một tập các nơ-ron, với mỗi nơ-ron kết nối đầy đủ (fully-connected) với tất cả các nơ-ron ở lớp trước và các nơ-ron trong cùng một lớp thì độc lập với nhau, không chia sẻ bất cứ kết nối nào. Lớp fully-connected cuối cùng gọi là lớp output và lớp này thể hiện điểm số cho mỗi lớp phân loại. Mạng NN thông thường không phù hợp với những bức ảnh lớn.

Khối nơ-ron 3 chiều ConvNets (Convolutional Neural Networks): có các nơ-ron được sắp xếp theo 3 chiều: weight, height, depth. Một nơ-ron trong một lớp sẽ chỉ kết nối với một vùng nhỏ trong lớp trước, thay vì tất cả nơ-ron trong cấu trúc fully-connected.

Mạng CNN có 4 loại lớp chính: Convolutional Layer, ReLU Layer, Pooling Layer, Fully-Connected Layer. Bố trí các lớp này sẽ tạo thành một cấu trúc ConvNet. Ví dụ một kiến trúc ConvNet cho bộ dữ liệu CIFAR-10 và mạng ConvNet có kiến trúc INPUT-CONV-RELU-POOL-FC:



Hình 2.44: Một kiến trúc CNN

- INPUT [32x32x3]: chứa các điểm ảnh của bức ảnh.
- CONV: Mỗi nơ-ron kết nối với một vùng trên lớp trước đó
- RELU: Lớp này sẽ áp dụng hàm kích hoạt lên từng phần tử.
- POOL: Lớp này sẽ giảm kích thước theo chiều width và height.
- FC: Lớp này sẽ tính điểm cho mỗi lớp phân loại.

Theo cách này, ConvNet sẽ chuyển từ ảnh gốc qua các lớp, từ các giá trị điểm ảnh tới class score cuối cùng. Một số lớp sẽ có tham số (parameters) và một số lớp không có tham số.

2.4.4.9. Phương pháp để tự động tối ưu mạng NN

a) Khái niệm tự động tối ưu mạng NN:

Là tự động hóa học máy (AutoML), từ việc lựa chọn thuật toán phù hợp đến tự động lựa chọn đặc trưng và điều chỉnh tham số mạng NN. Một số phương pháp, công cụ để tự động tối ưu mạng NN gồm AutoML, TPOT, GA [77].

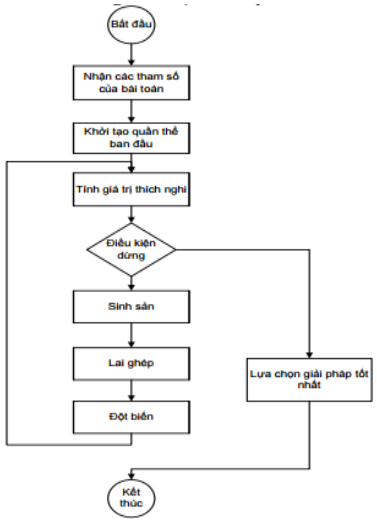
b) Phương pháp Giải thuật di truyền GA (Genetic Algorithm):

Giải thuật di truyền (GA) là thuật toán bắt chước sự chọn lọc tự nhiên và di truyền. Những mô hình tính toán của GA được lấy cảm hứng từ cơ chế tự nhiên tiến hóa, bao gồm cả sự thích nghi, sinh sản, và đột biến. Thuật toán di truyền là một tìm kiếm heuristic và một phương pháp tối ưu hóa lấy cảm hứng từ quá trình chọn lọc tự nhiên. Nó được sử dụng rộng rãi để tìm một giải pháp gần tối ưu cho các vấn đề tối ưu hóa với không gian tham số lớn [77].

Các hoạt động cơ bản tạo thành GA như: toán tử lựa chọn, toán tử lai ghép, toán tử đột biến.

Các thành phần cơ bản của GA gồm:

- Mã hóa (Encoding);
- Khởi tạo quần thể ban đầu (Innitial population generation);
- Hàm thích nghi (Fitness Function);
- Lựa chọn (Selection);
- Lai ghép (Crossover);
- Đột biến (Mutation);
- Chiến lược thay thế (Replacement Strategy);
- Tiêu chuẩn kết thúc (Termination Criteria).



Hình 2.45: Sơ đồ Giải thuật di truyền

Ưu, nhược điểm của GA

Ưu điểm: Khả năng song song của thuật toán; GA duyệt qua không gian tìm kiếm sử dụng nhiều cá thể và ít mắc phải cực trị địa phương; khi đã có thuật toán gen cơ bản, chỉ cần viết một NST mới để xử lý bài toán khác; dễ thể hiện và thay đổi hàm thích nghi với 1 cách mã hóa.

Nhược điểm: Nhược điểm chính của GA là thời gian tính toán; GA có thể chậm hơn các thuật toán khác; có thể kết thúc tính toán bất cứ lúc nào.

2.4.5. Các phương pháp xác định độ tin cậy của hệ thống AI dự đoán KTTV

2.4.5.1. Phương pháp và tiêu chí xác định độ tin cậy của mô hình dự đoán

a) Độ tin cậy của các mô hình dự đoán

Để đánh giá chất lượng dự báo (có thể là của một mô hình) Murphy (1993) đã đưa ra 3 khái niệm để chỉ mức độ "tốt" của một dự báo là:

- **Độ chắc chắn** của dự báo: là mức độ phù hợp giữa dự báo và kiến thức hiểu biết của dự báo viên;
- **Chất lượng** của dự báo: mức độ phù hợp giữa dự báo với thực tế xảy ra;
- **Ý nghĩa hay giá trị** của dự báo: là mức độ mà dự báo có thể giúp cho người ra quyết định thấy rõ hoặc đạt được lợi ích nào đó về kinh tế hoặc khác [78].

b) Phương pháp đánh giá độ tin cậy của các mô hình học máy

Có hai phương pháp để đánh giá độ tin cậy cho một mô hình học máy:

- Phương pháp 1: Sử dụng các thang đo: là phương pháp định lượng, giúp cho việc đánh giá mô hình một cách chính xác.
- Phương pháp 2: Trực quan hóa và hình ảnh hóa các kết quả của các mô hình, đây là phương pháp định tính.

Cần kết hợp cả 2 cách trên để có những đánh giá tốt nhất đối với một mô hình học máy nói chung và các mô hình dự đoán, dự báo KTTV nói riêng.

2.4.5.2. Các tiêu chí đánh giá chất lượng dự báo của các mô hình học máy

Murphy (1993) đã đưa ra 9 thuộc tính tiêu chí để đánh giá chất lượng dự báo như sau:

- **Độ lệch (Bias)**: Là mức độ phù hợp giữa trung bình quan trắc và trung bình dự báo.
- **Tính liên kết (Association)**: Là mức độ chặt chẽ về quan hệ tuyến tính giữa dự báo và quan trắc (ví dụ như hệ số tương quan).
- **Độ chính xác (Accuracy)**: Là thuật ngữ chung chỉ mức độ phù hợp giữa dự báo và thực tế (xác định bởi quan trắc). Sai khác giữa giá trị quan trắc và giá trị dự báo được gọi là sai số.
- **Kỹ năng dự báo (Skill):** Kỹ năng hay độ chính xác tương đối là độ chính xác của một dự báo so với một dự báo tham chiếu nào đó.
- **Độ tin cậy (Reliability):** Là sự phù hợp trung bình giữa các giá trị dự báo và các giá trị quan trắc. Thông thường độ tin cậy được cải tiến nhờ Bias.

- **Độ phân giải (Resolution):** Là năng lực của dự báo có thể phân loại tập các sự kiện thành các tập con có phân bố tần suất khác nhau. Nghĩa là phân bố nhận được khi "A" được dự báo khác với phân bố nhận được khi "B" được dự báo.

- **Độ sắc nét (hay độ nhạy bén - Sharpness):** Là xu hướng của dự báo có thể dự báo được các giá trị cực trị. Độ nhọn là một thuộc tính của dự báo, tương tự như độ phân giải.

- **Độ phân biệt (hay độ phân lớp - Discrimination**): Là khả năng của dự báo có thể tách biệt các quan trắc thành những trường hợp có tần suất dự báo cao hơn, tức là có khả năng phân lớp.

- **Độ biến động (Uncertainty):** Là sự dao động của các giá trị quan trắc trong tập mẫu đánh giá và không phụ thuộc vào giá trị dự báo. Đại lượng này liên quan đến độ "khó khăn" của dự báo.

Theo truyền thống, thì đánh giá chất lượng dự báo tập trung nhấn mạnh vào độ chính xác và kỹ năng dự báo.

2.4.5.3. Kiểm định độ tin cậy của mô hình bằng phương pháp thống kê

a) Các chỉ số thống kê để kiểm tra độ tin cậy của mô hình dự đoán

Để kiểm tra độ tin cậy của các mô hình dự đoán người ta sử dụng hai chỉ số thống kê là: Hệ số Cronbach Alpha và Hệ số tương quan biến tổng.

- **Hệ số Cronbach Alpha:** là hệ số cho phép đánh giá xem nếu đưa các biến quan sát nào đó thuộc về một biến nghiên cứu (biến tiềm ẩn, nhân tố) thì nó có phù hợp không. Hair et al (2006) đưa ra quy tắc đánh giá như sau:

  + 0.6. Thang đo nhân tố là không phù hợp (có thể trong môi trường nghiên cứu đối tượng không có cảm nhận về nhân tố đó).
  + 0.6 - 07: Chấp nhận được với các nghiên cứu mới.
  + 0.7 - 0.8: Chấp nhận được.
  + 0.8 - 0.95: tốt.
  + >= 0.95: Chấp nhận được nhưng không tốt, nên xét các biến quan sát có thể có hiện tượng "trùng biến".

- **Hệ số tương quan biến tổng** là hệ số cho biến mức độ "liên kết" giữa một biến quan sát trong nhân tố với các biến còn lại. Tiêu chuẩn để đánh giá một biến có thực sự đóng góp giá trị vào nhân tố hay không là hệ số tương quan biến tổng phải lớn hơn **0.3**. Nếu biến quan sát có hệ số tương quan biến tổng nhỏ hơn 0.3 thì phải loại nó ra khỏi nhân tố đánh giá [78].

b) Các độ đo sử dụng trong bài toán phân loại (Classification)

Khi xây dựng một mô hình Machine Learning, cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh khả năng của các mô hình. Có rất nhiều cách đánh giá một mô hình phân lớp. Tuỳ vào những bài toán khác nhau mà chúng ta sử dụng các phương pháp khác nhau. Các phương pháp thường được sử dụng là: Accuracy score, Confusion matrix, True-False Positive-Negative, Đường cong đặc trưng, Area Under the Curve, Precision và Recall. F1-score.

c) Các độ đo sử dụng trong bài toán hồi quy (Regression)

Các phương pháp đánh giá độ tin cậy các dự đoán với bài toán hồi quy gồm:

- Các tiêu chí theo **giá trị sai biệt tuyệt đối** gồm: (i) MAE: Trung bình của sai biệt tuyệt đối; (ii) MEDAE: Trung vị của sai biệt tuyệt đối; (iii) SAE: Tổng sai biệt tuyệt đối; (iv) MAPE Mean Absolute Percentage Error.

- Các tiêu chí theo **bình phương sai số** gồm: (i) MSE: trung bình bình phương sai số; (ii) MEDSE: Trung vị bình phương sai số; (iii) SSE: Tổng bình phương sai số; (iv) RMSE: Căn bậc 2 của trung bình bình phương sai số; (v) MSLE Mean Squared Logarithmic Error; (vi) RMSLE: Root Mean Squared Logarithmic Errors.

- Các tiêu chí theo khảo sát **sai biệt tương đối** của mô hình: (i) RRSE Root Relative Squared Error; (ii) RAE: Relative Absolute Error; (iii) Rsq: hệ số xác định.

- Tiêu chí đánh giá mối **tương quan tuyến tính** giữa giá trị thực và tiên lượng: (i) Kendall's Tau (thứ hạng); (ii) Rho của Spearman (phi tham số); (iii) Pearson'r.

d) Các độ đo sử dụng trong bài toán phân nhóm (clustering) gồm:

Bao gồm: (i) Chỉ số số BIC; (ii) Chỉ số Calinski-Harabasz; (iii) Chỉ số Davies-Bouldin index (DBI); (iv) Chỉ số Silhouette;

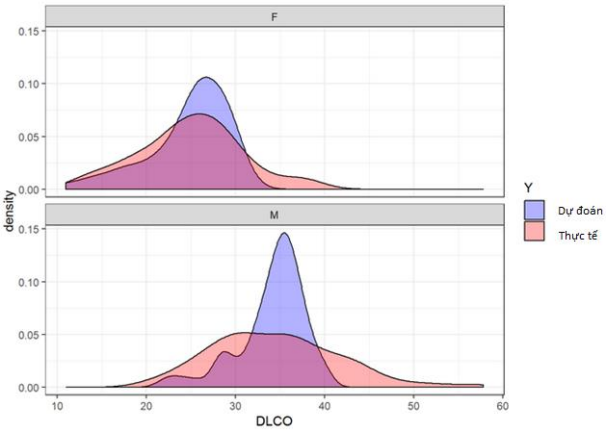
Và một số chỉ số đánh giá phân cụm mờ: Chỉ số đánh giá PC, Chỉ số đánh giá CE (Classication Entropy); Chỉ số đánh giá XB; Dunn's Index (DI); Alternative Dunn Index (ADI); Chỉ số SC.

2.4.5.4. Kiểm định độ tin cậy mô hình dự đoán bằng trực quan biểu đồ

Có thể kiểm tra độ tin cậy của mô hình một cách trực quan bằng hình ảnh mà không cần đến các chỉ số thống kê, đây là phương pháp dễ hiểu và hiệu quả. Các phương pháp kiểm định cụ thể dưới đây [78]:

a) So sánh mật độ phân bố giữa thực tế và tiên lượng

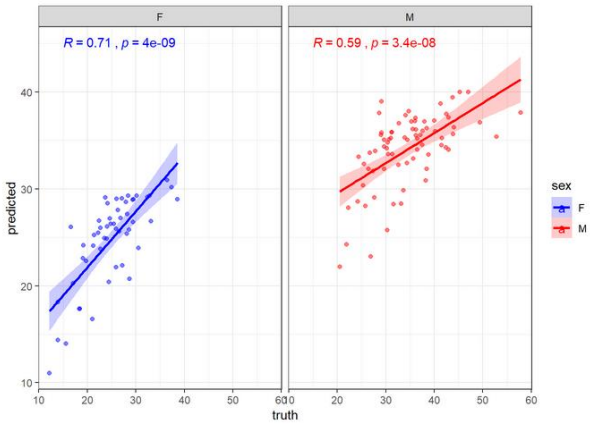
Các mô hình hồi quy đều hoạt động dựa trên một giả định về phân bố của biến kết quả ngẫu nhiên, nếu mô hình chính xác, hình ảnh phân bố kết quả của nó sẽ đồng dạng và trùng lắp với phân bố thực tế của đại lượng cần ước tính. Trên biểu đồ, phẩm chất của mô hình được đánh giá bằng mức độ đồng dạng giữa 2 phân bố, vị trí trung tâm. Sai sót của mô hình thể hiện ở phần diện tích không chồng lắp.



Hình 2.46: Biểu đồ phẩm chất của mô hình dự đoán

b) Tương quan tuyến tính giữa giá trị thực tế và tiên lượng

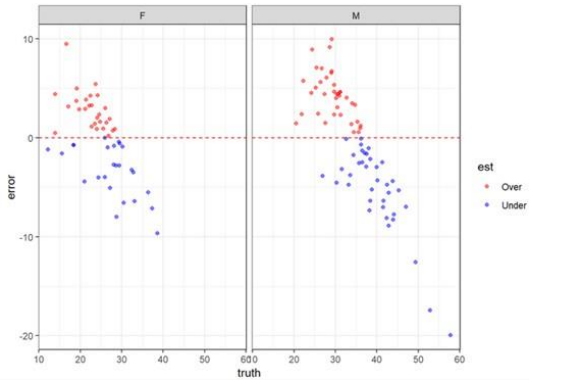
Áp dụng một biểu đồ tán xạ và một đồ thị tuyến tính giữa 2 vectors: predicted và truth cho phép khảo sát tương quan tuyến tính giữa kết quả của mô hình và quan trắc thực tế.



Hình 2.47: Tương quan tuyến tính giữa mô hình dự đoán và quan sát thực tế

c) Khảo sát trực tiếp sai biệt giữa thực tế và tiên lượng

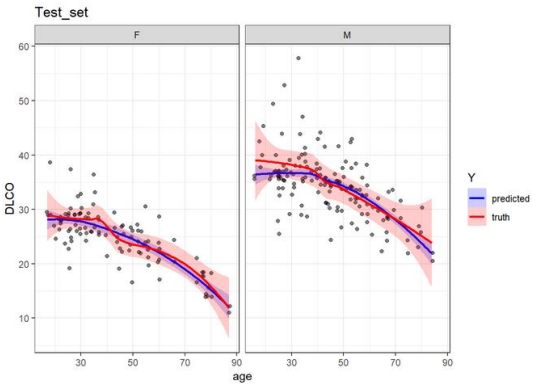
Khảo sát trực tiếp khuynh hướng sai biệt của mô hình (over hay underes-timation), và sai biệt này là ngẫu nhiên hay có hệ thống: (Dạng biểu đồ này tương tự như Bland-Altman plot).



Hình 2.48: Khảo sát khuynh hướng sai biệt của mô hình và thực tế quan sát

d) Kiểm tra tính hợp lý của nội dung mô hình

Sử dụng dạng biểu đồ để kiểm tra đồng thời tính chính xác và hợp lý của mô hình so với quan sát thực.



Hình 2.49: Kiểm tra tính hợp lý của nội dung mô hình dự báo

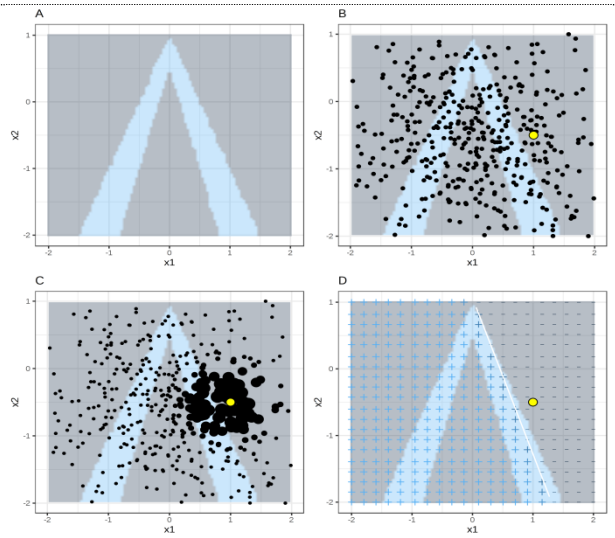
2.4.5.5. Phương pháp giải thích cho các dự đoán của hệ thống AI/ML

Có hai phương pháp để giải thích dự đoán cho các mô hình học máy, đó là: mô hình LIME, và mô hình SHAP. LIME chưa được giải quyết khi sử dụng cho dữ liệu dạng bảng, SHAP thì chậm và bỏ qua sự phụ thuộc thuộc tính. Do đó cần sử dụng cả 2 mô hình này một cách hợp lý, và linh hoạt.

a) Giải thích theo mô hình cục bộ - LIME

Mô hình cục bộ LIME (Local Interpretable Model-Agnostic Explanations) - thực hiện theo quy trình sau: Đầu tiên, LIME lấy thông tin về đặc tính phân phối của feature (Data input) dựa vào training dataset và nội dung (feature được dùng) trong model; thông tin này được lưu trữ trong một object gọi là explainer. Việc diễn giải sẽ được áp dụng cho 1 cá thể mới (unseen case). Sau đó, LIME áp dụng mô hình cho toàn bộ những điểm trong không gian nhiễu này, đồng thời tính khoảng cách giữa các điểm mô phỏng đến điểm trung tâm là trường hợp được xét. Khoảng cách này sẽ được chuyển thành thang điểm (score). Tiếp theo, LIME chọn một số lượng M features tiêu biểu nhất cho phép mô tả tốt nhất khoảng cách nói trên. Cuối cùng, LIME dựng một mô hình rất đơn giản cho các điểm mô phỏng, sử dụng M features được chọn làm predictor, để giải nghĩa cho outcome của model. Mô hình này có dạng tuyến tính, hoặc mô hình Decision Tree [80]. Ưu thế lớn nhất của LIME, đó là tính độc lập với thuật toán (Algorithm). LIME không phụ thuộc vào bản chất của Algorithm, thậm chí nó không dùng đến (và không cần biết) cơ chế bên trong của mô hình. Do đó, LIME có thể áp dụng cho các algorithm như Naive Bayes, KNN, Random Forest, Neural network. Tính độc lập này cho phép LIME hoạt động tại mọi thời điểm, cho mọi version của mô hình.

A) Dự đoán rừng ngẫu nhiên cho các thuộc tính x1 và x2. Các lớp dự đoán: 1 (tối) hoặc 0 (sáng). B) Trường hợp quan tâm (dấu chấm lớn) và dữ liệu được lấy mẫu từ một phân phối bình thường (dấu chấm nhỏ). C) Gán trọng số cao hơn cho các điểm gần quan tâm. D) Dấu hiệu của lưới cho thấy sự phân loại của mô hình được học tại địa phương từ các mẫu có trọng số. Đường trắng đánh dấu ranh giới quyết định (P (class = 1) = 0,5).



Hình 2.50: Thuật toán LIME cho dữ liệu dạng bảng

b) Giải thích theo mô hình SHAP

SHAP (SHPley exitive exPlanations) là một phương pháp để giải thích các dự đoán riêng lẻ. Mục tiêu của SHAP là giải thích dự đoán của một thể hiện x bằng cách tính toán sự đóng góp của từng thuộc tính vào dự đoán. Phương pháp giải thích SHAP tính toán các giá trị Shapley từ lý thuyết trò chơi liên minh. Các giá trị thuộc tính của một thể hiện dữ liệu đóng vai trò là người chơi trong liên minh. Một cải tiến mà SHAP mang đến cho bảng là giải thích giá trị Shapley được biểu diễn dưới dạng phương pháp phân bổ thuộc tính bổ sung, mô hình tuyến tính. Quan điểm đó kết nối các giá trị LIME và Shapley. SHAP chỉ định giải thích là:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2.4.32}$$

Trong đó g là mô hình giải thích, z thuộc khoảng {0,1}^M là vectơ liên minh, M là kích thước liên kết tối đa và **ϕj** thuộc {R} là thuộc tính thuộc tính cho một thuộc tính j, các giá trị Shapley. Vectơ liên minh của Vectơ, được gọi là các thuộc tính đơn giản hóa.

2.4.6. Các mô hình ra quyết định thống kê trong mô phỏng quá trình dự báo KTTV

Ra quyết định thống kê, dự báo thống kê là một trong những bài toán quan trọng trong khai phá dữ liệu và học máy nhằm tìm ra các quy luật của dữ liệu. Nguyên lý của ra quyết định thống kê, dự báo thống kê là dựa vào dữ liệu lịch sử để xây dựng mô hình thống kê nhằm dự báo các dữ liệu trong tương lai để đưa ra quyết

định. Có nhiều các phương pháp ra quyết định thống kê, dự báo thống kê khác nhau tùy thuộc vào từng đặc thù của các bài toán cụ thể và nhiệm vụ khai phá dữ liệu.

2.4.6.1. Lý thuyết ra quyết định thống kê trong học máy

Các phương pháp ra quyết định thống kê có thể sử dụng trong mô hình học máy như sau: (i) Quyết định theo lý thuyết quyết định Bayes; (ii) Quyết định theo dựa vào xác suất tiên nghiệm (Priori probabilities); (iii) Quyết định theo dựa vào xác suất có điều kiện của từng lớp; (iv) Quyết định theo dựa vào xác suất hậu nghiệm (Posterior Probabilities); (v) Quyết định theo xác suất lỗi (Probability of Error); (vi) Quyết định theo Rủi ro có điều kiện (Conditional Risk); (vii) Quyết định theo theo tỷ lệ lỗi phân loại nhỏ nhất (Min - Error - Rate Classification); (viii) Quyết định theo hàm phân bố xác suất; (ix) Quyết định theo phân bố Gauss; (x) Quyết định theo phân loại Minimax; (xi) Quyết định theo tiêu chí Neyman - Pearson; (xii) Quyết định theo mô hình cây quyết định DT (Decision Trees) và cây quyết định ID3 [40].

2.4.6.2. Mô hình ra quyết định thống kê với đa tiêu chuẩn

a) Khái niệm về mô hình ra quyết định thống kê đa tiêu chuẩn

Mô hình ra quyết định đa tiêu chuẩn MCDM dựa trên cơ sở lý thuyết tập mờ (Zadeh, 1965) là một công cụ hiệu quả để giải quyết các vấn đề lựa chọn phức tạp bao gồm nhiều tiêu chuẩn và lựa chọn, đặc biệt đối với các biến mang tính định tính. Các tiêu chuẩn định tính thường có đặc điểm mơ hồ, khó phân định chuẩn xác, gây khó khăn cho việc tổng hợp kết quả đánh giá theo các tiêu chí và việc đưa ra quyết định. Phương pháp MCDM sẽ lượng hóa các tiêu chí này, tính toán tổng điểm của các đối tượng đánh giá theo trọng số của mỗi tiêu chí và giúp người ra quyết định có được một cơ sở chắc chắn và chuẩn xác hơn [82].

b) Các mô hình ra quyết định thống kê đa tiêu chuẩn

Một số mô hình được sử dụng phổ biến hiện nay như:

***AHP*:** Là một kỹ thuật có cấu trúc để tổ chức và phân tích các quyết định phức tạp, dựa trên toán học và tâm lý học. AHP có thể kiểm tra tính nhất quán trong cách đánh giá của người ra quyết định. Quy trình phân tích theo thứ bậc dễ hiểu, xem xét nhiều tiêu chí nhỏ và phân tích cả yếu tố định tính lẫn định lượng.

ANP: Được ứng dụng trong xây dựng vấn đề ra quyết định từ việc chỉnh hợp các mục tiêu, tiêu chí, lựa chọn khác nhau, và so sánh cặp đôi các tiêu chí để đưa ra phương án lựa chọn tốt nhất.

***DEA*:** Được sử dụng để đo lường thực nghiệm độ hiệu quả của các đơn vị ra quyết định (DMU). DEA có liên kết chặt chẽ với lý thuyết kinh tế trong sản xuất,

công cụ này cũng được sử dụng trong cho điểm chuẩn trong hoạt động quản lý để đánh giá khả năng sản xuất và hoạt động dịch vụ.

***FTOPSIS*:** Là kĩ thuật cung cấp thứ tự ưu tiên của các lựa chọn thay thế tương đương với giải pháp lí tưởng. Tính toán mô hình sẽ cho ra 2 điểm gọi là: giải pháp lí tưởng tích cực (PIS) và giải pháp lí tưởng tiêu cực (NIS). Sau đó, tính khoảng cách của từng giải pháp từ 2 điểm cố định trên. Sự lựa chọn tốt nhất là sự lựa chọn có khoảng cách ngắn nhất với PIS và xa nhất với NIS.

***PROMETHEE*:** Thay vì chỉ ra một quyết định "đúng", phương pháp Promethee giúp các nhà sản xuất quyết định tìm sự thay thế phù hợp nhất với mục tiêu của họ và sự hiểu biết của họ về các vấn đề. Nó cung cấp một khuôn khổ toàn diện và hợp lý cho cấu trúc vấn đề đưa ra quyết định, xác định và định lượng các cuộc xung đột và hiệp lực của mình, các cụm hành động, và làm nổi bật các lựa chọn thay thế chính và các lý luận cấu trúc phía sau.

2.4.6.3. Mô hình ra quyết định thống kê dựa trên nhiều chuyên gia (MPDM)

Mô hình ra quyết định dựa trên nhiều chuyên gia (MPDM) với cấu trúc ưu tiên khác nhau dựa trên hai tiêu chí đồng thuận: 1) một biện pháp xấp xỉ chỉ ra sự thỏa thuận giữa các chuyên gia; 2) một biện pháp xấp xỉ để tìm chỉ ra một ý kiến cá nhân là từ ý kiến nhóm. Cơ chế phản hồi này dựa trên đơn giản và quy tắc dễ dàng để giúp các chuyên gia thay đổi ý kiến của họ để có được một mức độ đồng thuận càng cao càng tốt [83].

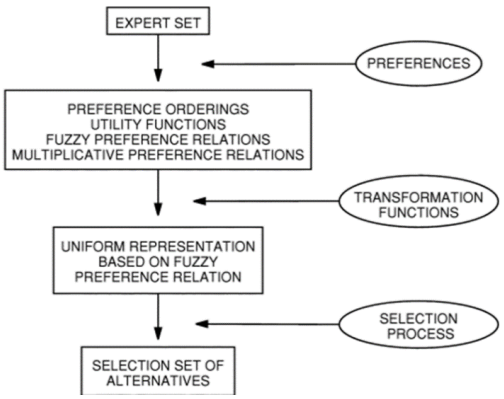
$$X = \{x_1, \ldots, x_n\} \tag{2.4.49}$$

Đăt X là một tập hợp hữu hạn những lựa chọn thay thế phải được phân loại từ tốt nhất để tới xấu nhất, sử dụng thông tin được cung cấp bởi một nhóm chuyên gia.

$$E = \{e_1, e_2, \ldots, e_m\}. \tag{2.4.50}$$

E là tập các ý kiến chuyên gia, mỗi chuyên gia có ý tưởng, thái độ, động lực, tính cách riêng và có các ý kiến riêng khác nhau.

Quá trình này quá trình được thể hiện trong hình 2.51. Quá trình giải quyết này được phát triển trong hai bước là đồng bộ thông tin và áp dụng một quy trình lựa chọn. Một mô hình đồng thuận được xác định cho các vấn đề MPDM với các cấu trúc ưu tiên khác nhau, trong đó định nghĩa: giá trị cao nhất là sự nhất trí và thấp nhất là sự bất đồng. Mô hình này có các đặc điểm chính:

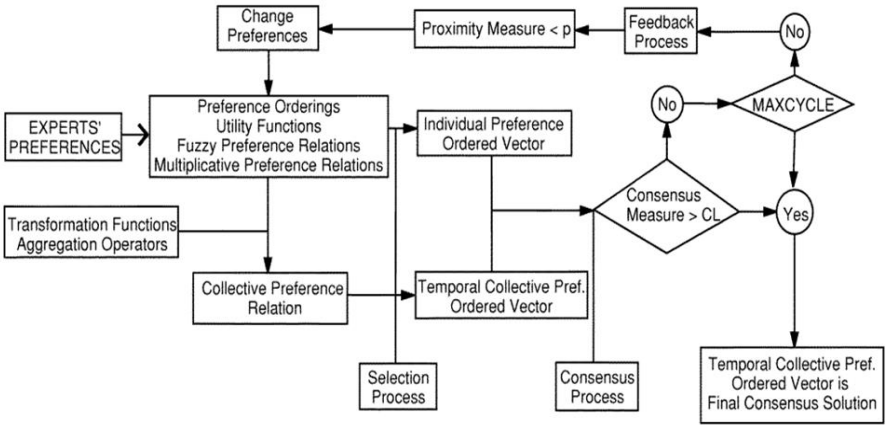


Hình 2.51: Quá trình giải quyết ra quyết định nhiều người

- Dựa trên hai tiêu chí đồng thuận mềm: một sự đồng thuận biện pháp và một biện pháp gần.

- Cả hai tiêu chí đồng thuận được xác định bằng cách so sánh giải pháp cá nhân với giải pháp tập thể sử dụng như tiêu chí so sánh các vị trí của các lựa chọn thay thế trong từng giải pháp.

- Một hệ thống hỗ trợ đồng thuận được xác định bằng cách sử dụng ở trên tiêu chuẩn đồng thuận và một cơ chế phản hồi đó là có thể thay thế các hành động của người điều hành trong sự đồng thuận đạt quá trình.



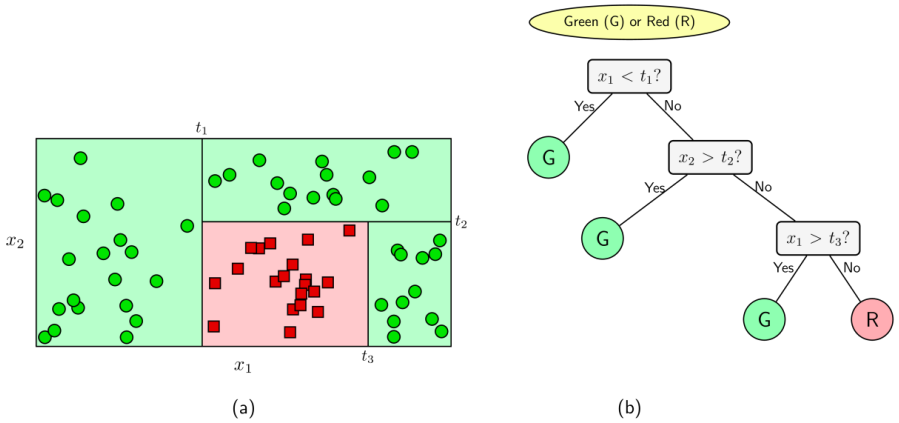
Hình 2.52: Mô hình đồng thuận

Theo mô hình này, các chuyên gia không đồng ý với ý kiến của họ, để đạt được sự đồng thuận phải qua các quá trình lặp đi lặp lại, nghĩa là thỏa thuận chỉ đạt được sau nhiều vòng xem xét. Để tránh sự đồng thuận không hội tụ sau vài vòng thảo luận, đặt số vòng tối đa thảo luận.

2.4.6.1. Ra quyết định thống kê theo mô hình cây quyết định

a) Mô hình cây quyết định DT (Decision Trees)

Machine Learning có một mô hình ra quyết định dựa trên các câu hỏi. Mô hình này có tên là *Cây quyết định DT* (*Decision Tree*) [40]. Decision Tree là một mô hình Supervised Learning, có thể được áp dụng vào cả hai bài toán Classification và Regression. Việc xây dựng một Decision Tree trên dữ liệu huấn luyện cho trước là việc đi xác định các *câu hỏi* và *thứ tự của chúng*. Một điểm đáng lưu ý của Decision Tree là nó có thể làm việc với các đặc trưng (thường được gọi là *thuộc tính – attribute*) dạng *categorical*, thường là rời rạc và không có thứ tự. Ví dụ, *mưa, nắng* hay *xanh, đỏ*, v.v. Decision tree cũng làm việc với dữ liệu có vector đặc trưng bao gồm cả thuộc tính dạng categorical và liên tục (*numeric*). Một điểm đáng lưu ý nữa là Decision Tree ít yêu cầu việc chuẩn hoá dữ liệu.

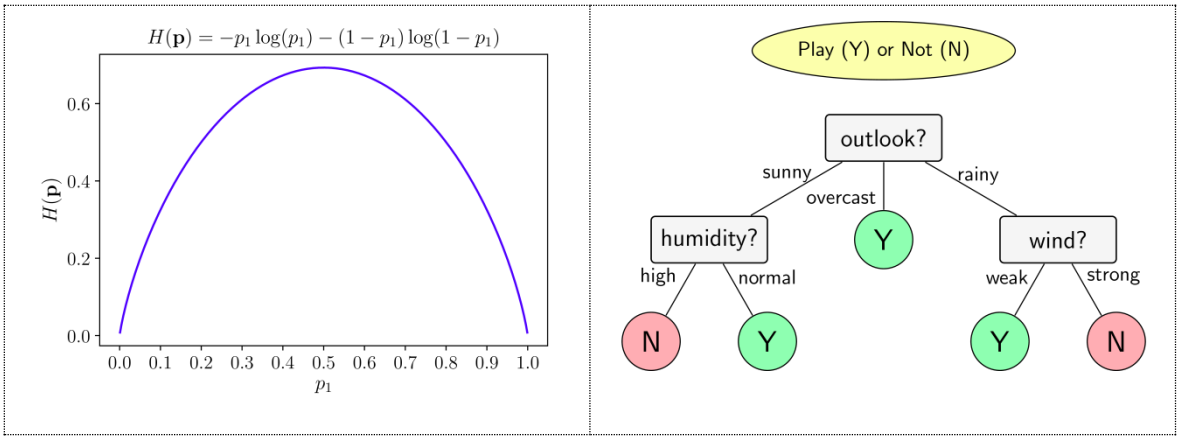


Hình 2.53: Mô hình Cây quyết định

b) Mô hình cây quyết định ID3

ID3 là một mô hình thuật toán Decision tree được áp dụng cho các bài toán phân chia theo hạng (Classification) mà tất cả các thuộc tính đều ở dạng phân loại (Categorical). ID3 còn được gọi là *entropy-based decision tree* và sử dụng hàm phân phối Entropy sau:

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log(p_i) \qquad 2.4.52$$



Hình 2.54: Đồ thị hàm Entropi	Hình 2.55: Mô hình Cây quyết định ID3

2.5. Kết chương 2

Các nội dung Chương 2 trên đã trình bày về phạm vi, đối tượng, số liệu phục vụ nghiên cứu. Phân tích các phương pháp công nghệ Big data gồm: (i) Thành phần, kỹ thuật của giải pháp Big data, (ii) Giải pháp, yếu tố đảm bảo vận hành hệ thống Big data hỗ trợ công nghệ AI dự đoán KTTV. Đi sâu phân tích các phương pháp công nghệ học máy, AI để nhận dạng, hỗ trợ dự báo KTTV gồm: (i) Các phương pháp lưu trữ và tiền xử lý dữ liệu về hiện tượng KTTV; (ii) Các phương pháp trích rút các đặc trưng dữ liệu KTTV; (iii) Các phương pháp xây dựng mô hình AI/ ML hỗ trợ dự báo KTTV; (iv) Phương pháp xác định độ tin cậy của mô hình AI/ML dự đoán KTTV; (v) Các phương pháp quyết định thống kê trong mô phỏng quá trình dự báo KTTV bằng AI.

Trong Chương tiếp theo, chúng tôi tiến hành: (i) Phát triển các công cụ, phương pháp học máy, AI để hỗ trợ dự báo KTTV; (ii) Xây dựng Big data phục vụ dự báo KTTV; (iii) Xây dựng mô hình AI để nhận dạng, dự báo KTTV nguy hiểm; (iv) Xây dựng Framework tích hợp các module hỗ trợ dự báo KTTV.

3. CHƯƠNG 3: PHÁT TRIỂN CÔNG CỤ, PHƯƠNG PHÁP HỌC MÁY, AI VÀ XÂY DỰNG BIG DATA KTTV

Chương này sẽ trình bày các kết quả: (i) Phát triển các công cụ, phương pháp học máy, AI để xây dựng mô hình AI hỗ trợ dự báo KTTV; (ii) Xây dựng Big data phục vụ dự báo KTTV bằng công nghệ AI. Trên cơ sở đó thiết lập và triển khai hệ thống AI để hỗ trợ dự báo các hiện tượng KTTV nguy hiểm, cụ thể.

3.1. Phát triển các công cụ, phương pháp học máy, AI để hỗ trợ dự báo KTTV

3.1.1. Công cụ để phát hiện, xử lý các dữ liệu KTTV mất mát

3.1.1.1. Phát triển công cụ phát hiện, xử lý các dữ liệu KTTV mất mát

a) Sử dụng phương pháp giá trị trung bình

Một kỹ thuật phổ biến là sử dụng phương pháp giá trị trung bình của những quan sát mà sự mất mát không xảy ra. Phương pháp áp dụng trong các trường hợp khi mà số lượng quan sát có mất mát dữ liệu là nhỏ. Tuy nhiên, khi số lượng giá trị bị mất mát là lớn việc sử dụng giá trị trung bình làm mất đi phương sai của dữ liệu. Dưới đây là minh họa đoạn code sử dụng phương pháp giá trị trung bình (mean).

```
from sklearn.datasets import fetch_california_housing
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import mean_squared_error
from math import sqrt
import random
import numpy as np
random.seed(0)

#Fetching the dataset
import pandas as pd
dataset = fetch_california_housing()
train, target = pd.DataFrame(dataset.data), pd.DataFrame(dataset.target)
train.columns = ['0','1','2','3','4','5','6','7']
train.insert(loc=len(train.columns), column='target', value=target)

#Randomly replace 40% of the first column with NaN values
column = train['0']
print(column.size)
missing_pct = int(column.size * 0.4)
i = [random.choice(range(column.shape[0])) for _ in range(missing_pct)]
column[i] = np.NaN
print(column.shape[0])

#Impute the values using scikit-learn SimpleImpute Class
from sklearn.impute import SimpleImputer
```

b) Sử dụng phương pháp giá trị xuất hiện thường xuyên

Kỹ thuật này sử dụng giá trị thường xuyên xuất hiện nhất trong mẫu để thay thế cho các giá trị bị mất. Đoạn code thay thế giá trị mất bởi giá trị thường xuyên xuất hiện trong mẫu.

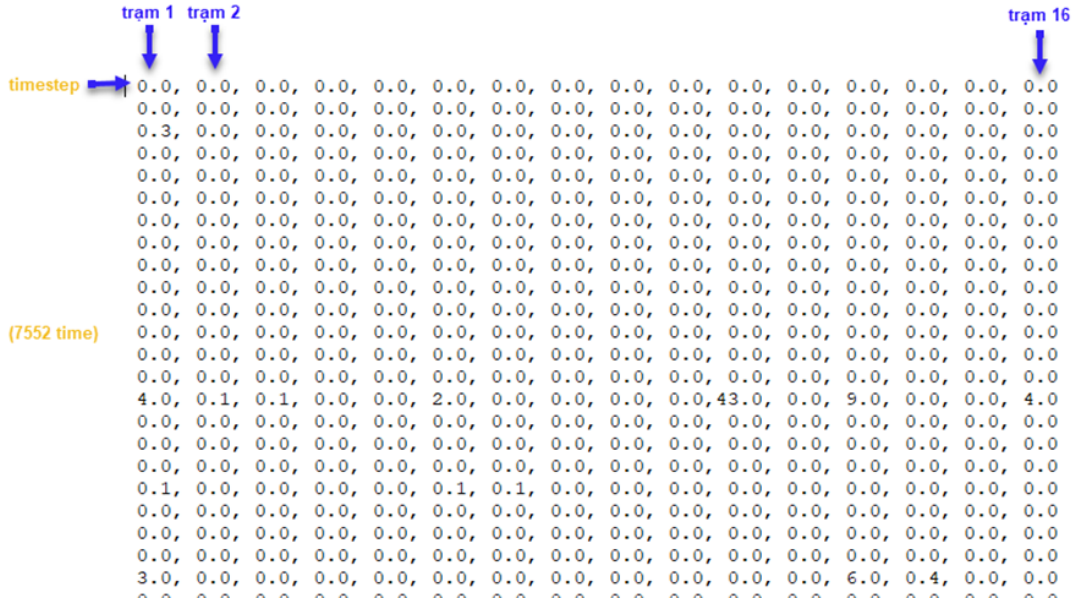
```
#Impute the values using scikit-learn SimpleImpute Class

from sklearn.impute import SimpleImputer
imp_mean = SimpleImputer( strategy='most_frequent')
imp_mean.fit(train)
imputed_train_df = imp_mean.transform(train)
```

3.1.1.2. Thử nghiệm công cụ trên số liệu lượng mưa

Sử dụng kỹ thuật trên để chạy dữ liệu thực tế thu được từ các trạm quan trắc. Danh sách trạm test Rainfall:

1) 48800 - MUONG LAY 22.066667556762695, 103.1500015258789
2) 48803 - LAO CAI 22.5, 103.96666717529297
3) 48805 - HA GIANG 22.816667556762695, 104.96666717529297
4) 48806 - SON LA 21.33333396911621, 103.9000015258789
5) 48808 - CAO BANG 22.66666603088379, 106.25
6) 48811 - DIEN BIEN 21.366666793823242, 103.0
7) 48812 - TUYEN QUANG 21.816667556762695, 105.21666717529297
8) 48818 - HOA BINH 20.816667556762695, 105.33333587646484
9) 48823 - NAM DINH 20.433332443237305, 106.1500015258789
10) 48825 - HA DONG 20.96666717529297, 105.75
11) 48826 - PHU LIEN 20.799999237060547, 106.63333129882812
12) 48830 - LANG SON 21.83333396911621, 106.76667022705078
13) 48833 - BAI CHAY 20.96666717529297, 107.06666564941406
14) 48837 - TIEN YEN 21.33333396911621, 107.4000015258789
15) 48838 - MONG CAI 21.516666412353516, 107.96666717529297
16) 48842 - HOI XUAN 20.366666793823242, 105.08333587646484

Thời gian lấy dữ liệu test: 1/6/2016 đến 1/1/2019. Tổng: 944 ngày ~ 944*8 (8 obs/ ngày) = 7552 dòng.



Hình 3.1: Dữ liệu thực tế về lượng mưa

Thực hiện xóa ngẫu nhiên một số dữ liệu một số thời điểm ngẫu nhiên để biến đổi dữ liệu thu được đầy đủ từ trạm quan trắc thành dữ liệu thiếu. Chú ý dữ liệu thiếu là các khoảng trắng.

Hình 3.2: Dữ liệu lượng mưa bị mất mát

Nội suy lại các dữ liệu mất mát này. Sau xử lý, dữ liệu được khôi phục sau mất mát là đúng với dữ liệu thực tế ban đầu thu được từ các trạm.

Hình 3.3: Dữ liệu lượng mưa sau khi phát hiện, xử lý mất mát

3.1.2. Công cụ để phát hiện, xử lý các dữ liệu KTTV ngoại lai

3.1.2.1. Phát triển công cụ phát hiện, xử lý các dữ liệu KTTV ngoại lai

Sử dụng ngôn ngữ lập trình Python, mã nguồn được viết trên hệ thống Google Colab, sử dụng 3 thư viện nguồn mở bao gồm Pandas, Matplotlib và Seaborn để đọc và trích xuất dữ liệu quan trắc KTTV. Thực hiện phương pháp tính giá trị Z-Score theo công thức 2.4.1 đã trình bày trong phần 2.4.1.4 của các điểm trong tập dữ liệu để phát hiện và xử lý cho dữ liệu mất mát và bất thường. Để xác định ngoại lai, lựa chọn ngưỡng có giá trị 5 để lọc các điểm xem xét ngoại lai.

3.1.2.2. Thử nghiệm công cụ trên số liệu nhiệt độ

a) Sử dụng tập dữ liệu nhiệt độ tại các trạm quan trắc 8 obs/ngày

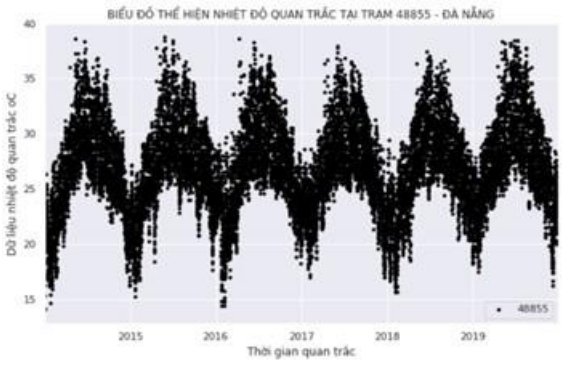
Các trạm quan trắc 8 obs/ngày, mỗi lần cách nhau 3 giờ tại các thời điểm 00h, 03h, 06h, 09h, 12h, 15h, 18h, 21h theo giờ GMT tương ứng với 01h, 04h, 07h, 10h, 13h, 16h, 19h, 22h giờ Việt Nam. Danh sách 43 trạm thu thập để thực nghiệm trong bảng dưới đây.

Bảng 3.1: Liệt kê một số trạm quan trắc 8 obs/ngày

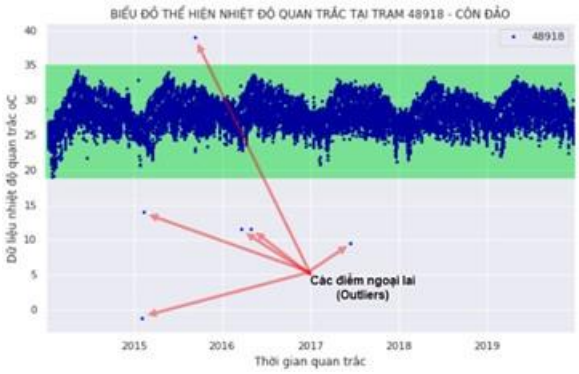
STT	Mã trạm	Tên trạm	STT	Mã trạm	Tên trạm	STT	Mã trạm	Tên trạm
1	48800	Mường Lay	16	48823	Nam Định	31	48920	Trường Sa
2	48811	Điện Biên	17	48842	Hồi Xuân	32	48890	Phan Rang
3	48806	Sơn La	18	48840	Thanh Hóa	33	48887	Phan Thiết
4	48818	Hòa Bình	19	48845	Vinh	34	48889	Phú Quý
5	48803	Lào Cai	20	48846	Hà Tĩnh	35	48866	Pleiku
6	48805	Hà Giang	21	48/86	Kỳ Anh	36	48875	Buôn Ma Thuật
7	48812	Tuyên Quang	22	48848	Đồng Hới	37	48894	Nhà Bè

STT	Mã trạm	Tên trạm	STT	Mã trạm	Tên trạm	STT	Mã trạm	Tên trạm
8	48808	Cao Bằng	23	48852	Huế	38	48903	Vũng Tàu
9	48830	Lạng Sơn	24	48860	Hoàng Sa	39	48918	Côn Đảo
10	48838	Móng Cái	25	48855	Đà Nẵng	40	48910	Cần Thơ
11	48837	Tiên Yên	26	48863	Quảng Ngãi	41	48917	Phú Quốc
12	48833	Bãi Cháy	27	48/96	Hoài Nhơn	42	48907	Rạch Giá
13	48826	Phù Liễn	28	48870	Quy Nhơn	43	48914	Cà Mau
14	48839	Bạch Long Vĩ	29	48873	Tuy Hòa			
15	48825	Hà Đông	30	48877	Nha Trang			

Dữ liệu thu thập từ các trạm được lưu trữ trong CSDL MongoDB, thực hiện việc kết nối tới máy chủ CSDL và truy xuất các thông số nhiệt độ của 43 trạm trong khoảng thời gian từ 01h ngày 01/01/2014 tới 22h ngày 31/12/2019. Đây là tập dữ liệu gốc (dữ liệu thô - Raw dataset) được tổng hợp khi các trạm gửi về, quá trình thu thập dữ liệu, truyền nhận và lưu trữ có thể do các nguyên nhân chủ quan và khách quan dẫn đến dữ liệu có thể bị mất mát (missing), bị sai lệch…. Do đó trước khi sử dụng cho những mục đích cụ thể, các số liệu này cần phải được xử lý các điểm dữ liệu ngoại lai cho 43 trạm này.



Hình 3.4: Minh họa tập dữ liệu không chứa dữ liệu ngoại lai



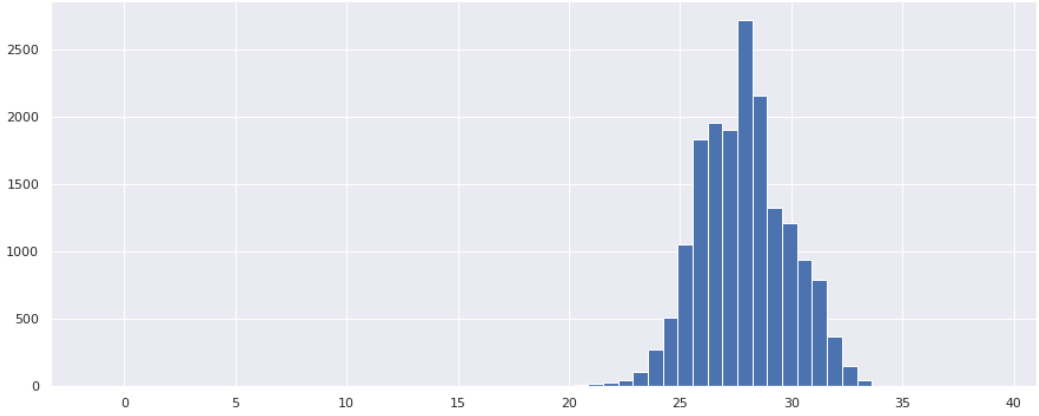
Hình 3.5: Minh họa tập dữ liệu chứa các điểm dữ liệu ngoại lai

Tiến hành đọc và trích xuất dữ liệu quan trắc của trạm 48918. Khái quát về tập dữ liệu của trạm này như sau.

Bảng 3.2: Thông số tập dữ liệu quan trắc của trạm 48918

Thời điểm bắt đầu dữ liệu	01:00:00 01-01-2014
Thời điểm kết thúc dữ liệu	22:00:00 31-12-2019
Tổng số điểm dữ liệu	17 528
Số điểm có dữ liệu	17 495
Số điểm dữ liệu thiếu	33
Giá trị trung bình của tập dữ liệu	27.8478
Độ lệch chuẩn của tập dữ liệu	2.0407

Giá trị cực tiểu	-1.3
Tứ phân vị thứ nhất (Q1)	26.4
Tứ phân vị thứ hai (Q2)	27.8
Tứ phân vị thứ ba (Q3)	29.1
Giá trị cực đại	39.0



Hình 3.6: Biểu đồ histogram của tập dữ liệu nhiệt độ trạm 48918

b) Sử dụng Z-Score phát hiện dữ liệu nhiệt độ ngoại lai của trạm 48918:

Mã nguồn và kết quả sử dụng phương pháp Z-Score để xác định ngoại lai cho trạm 48918 như trong hình dưới đây:

```
[15]  1 #Tính Zscores cho các điểm dữ liệu của trạm 48918 lưu vào cột zscores
      2 df_48918['zscore'] = (df_48918['48918'] - df_48918['48918'].mean())/df_48918['48918'].std(ddof=0)
      3 #So sánh Zscore với một ngưỡng (=5), điểm nào có zscore>5 xem xét như một ngoại lai
      4 #Cột outlier cho biết giá trị zscore có lớn hơn ngưỡng hay không (0: Không | 1: Có)
      5 df_48918['outlier'] = (abs(df_48918['zscore'])>5).astype(int)
      6 #Hiển thị những thời điểm xem xét ngoại lai (Zscore>5)
      7 df_48918.loc[df_48918.outlier==1]
```

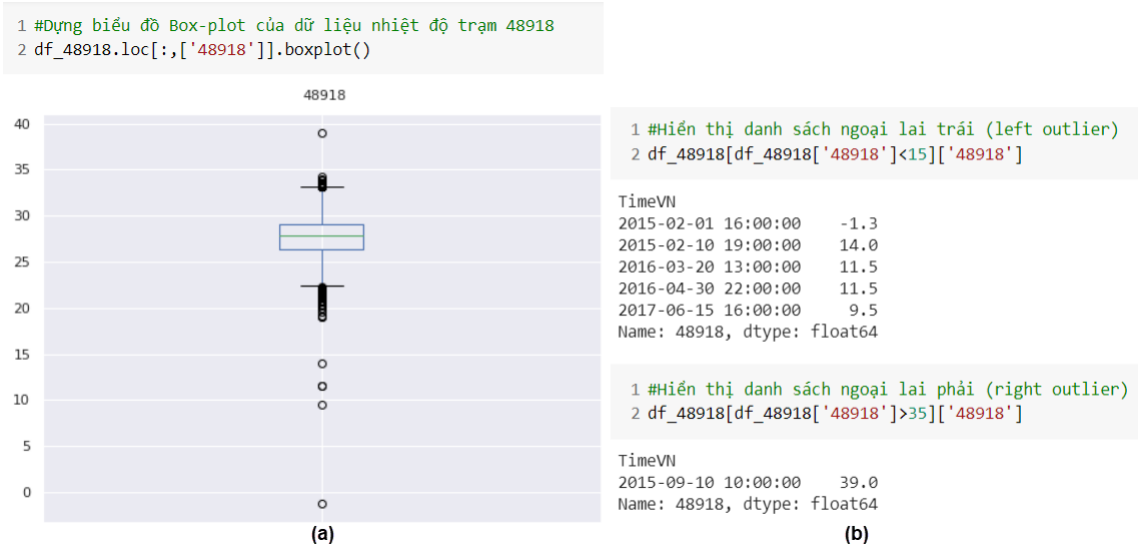
TimeVN	48918	zscore	outlier
2015-02-01 16:00:00	-1.3	-14.283356	1
2015-02-10 19:00:00	14.0	-6.785862	1
2015-09-10 10:00:00	39.0	5.464945	1
2016-03-20 13:00:00	11.5	-8.010942	1
2016-04-30 22:00:00	11.5	-8.010942	1
2017-06-15 16:00:00	9.5	-8.991007	1

Hình 3.7: Sử dụng Z-Score phát hiện dữ liệu ngoại lai của trạm 48918

Như vậy, theo phương pháp Z-Score với ngưỡng lọc chọn bằng 5 **có tất cả 6 điểm dữ liệu được xem xét là ngoại lai**. Trong đó có 5 điểm ngoại lai trái (zscore < 0) và 1 điểm ngoại lai phải (zscore>0).

c) Sử dụng biểu đồ Box-plot phát hiện dữ liệu nhiệt độ ngoại lai của trạm 48918

Theo như biểu đồ Box-plot trong hình 3.8 a, có khá nhiều điểm dữ liệu nằm trên và dưới hai râu Min và Max của biểu đồ, về nguyên tắc các điểm này đều là các điểm dữ liệu ngoại lai. Tuy nhiên, trong phần xử lý, chỉ xem xét các điểm dữ liệu có mức độ sai khác lớn, tách biệt hoàn toàn khỏi tập dữ liệu. Từ biểu đồ Box-plot, có thể lọc các điểm ngoại lai trái với ngưỡng 15^0C, các điểm ngoại lai phải với ngưỡng 35^0C. Kết quả tách các điểm ngoại lai trái - phải (hình 3.8b).



Hình 3.8: Phát hiện (a) và hàm tách (b) các điểm dữ liệu ngoại lai

d) Kiểm chứng các điểm dữ liệu ngoại lai được phát hiện

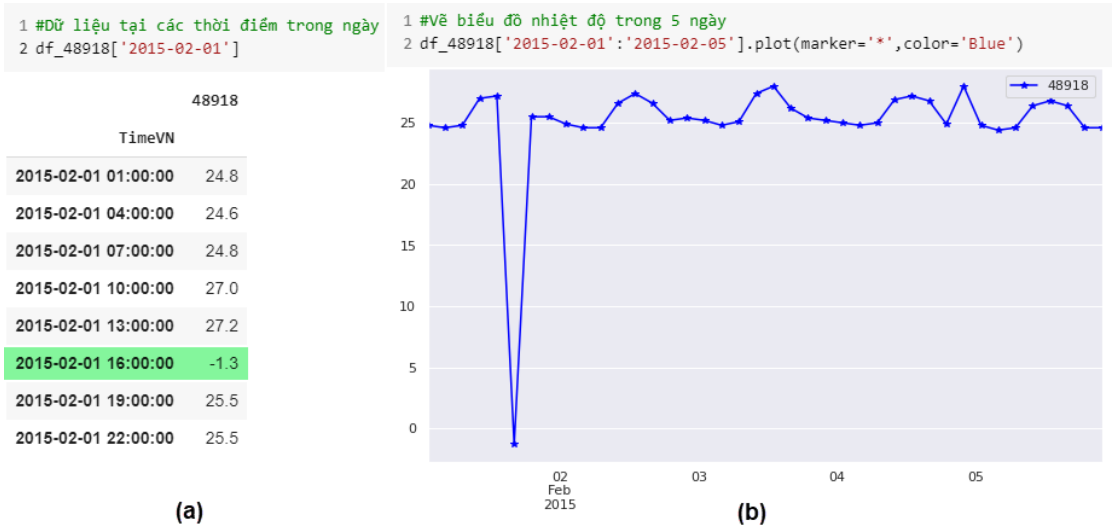
Kết quả thu được từ hai phương pháp Z-Score và Biểu đồ Box-plot đều cho cùng một danh sách 6 điểm dữ liệu ngoại lai, chi tiết như trong Bảng 3.3.

Bảng 3.3: Thời điểm và giá trị quan trắc xem xét ngoại lai của trạm 48918

STT	Thời điểm	Giá trị ghi nhận
1	2015-02-01 16:00:00	-1.3
2	2015-02-10 19:00:00	14.0
3	2015-09-10 10:00:00	39.0
4	2016-03-20 13:00:00	11.5
5	2016-04-30 22:00:00	11.5
6	2017-06-15 16:00:00	9.5

Dữ liệu nhiệt độ thu nhận được từ các trạm quan trắc có dạng chuỗi thời gian, với bước số liệu là 3h/ lần số liệu. Do vậy, để khẳng định đây là các điểm ngoại lai, cần kiểm chứng dữ liệu này trong một chuỗi dữ liệu tương ứng với điểm đó tại hai thời điểm có giá trị nhỏ nhất và lớn nhất để minh họa.

Với kết quả kiểm chứng dữ liệu nhiệt độ giá trị nhỏ nhất của trạm 48918 tại thời điểm 16h ngày 01/02/2015 như thể hiện trong hình 3.9a, có thể khẳng định dữ liệu thu thập được tại thời điểm này là hoàn toàn sai lệch. Nhiệt độ ghi nhận tại thời điểm 16h phải có mối tương quan với nhiệt độ tại thời điểm trước đó (lúc 13h) và sau đó (lúc 19h). Ngoài ra, đồ thị biểu diễn nhiệt độ của trạm trong khoảng thời gian 5 ngày từ 1h ngày 01/02/2015 đến 22h ngày 05/02/2015 (hình 3.9b) cũng thể hiện rõ mức độ sai khác dữ liệu tại thời điểm này.



Hình 3.9: Kết quả (a) và trình diễn kết quả (b) kiểm chứng điểm dữ liệu ngoại lai có giá trị thấp nhất tại trạm 48918

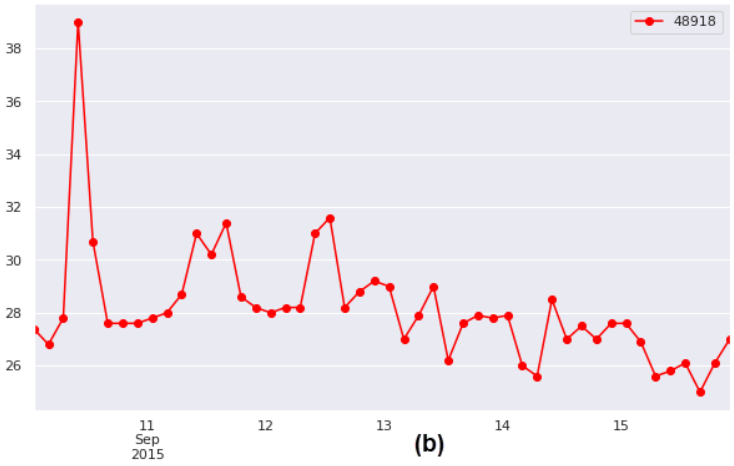
Tương tự như vậy, hình 3.10 thể hiện kết quả kiểm chứng dữ liệu ngoại lai có giá trị cao nhất tại thời điểm 10h ngày 10/09/2015. Hình 3.10a hiển thị toàn số liệu quan trắc trong ngày 10/09/2015. Hình 3.10b thể hiện đồ thị nhiệt độ quan trắc trong khoảng thời gian 5 ngày từ 1h ngày 10/09/2015 đến 22h ngày 15/09/2015. Chúng ta cũng dễ dàng nhận thấy dữ liệu quan trắc tại thời điểm 10h ngày 10/09 có mức độ sai khác tương đối lớn so với mặt bằng chung của các điểm đo. Hơn nữa, nhiệt độ tại thời điểm 10h có giá trị là 39^0C cao hơn nhiệt độ lúc lúc 13h là $8,3^0$C. Điều này trong thực tế là phí lý khi thời điểm Tmax trong ngày không phải là 3h như bình thường.

```
#Dữ liệu tại các thời điểm trong ngày
df_48918['2015-09-10']
```

TimeVN	48918
2015-09-10 01:00:00	27.4
2015-09-10 04:00:00	26.8
2015-09-10 07:00:00	27.8
2015-09-10 10:00:00	39.0
2015-09-10 13:00:00	30.7
2015-09-10 16:00:00	27.6
2015-09-10 19:00:00	27.6
2015-09-10 22:00:00	27.6

(a)

```
#Vẽ biểu đồ nhiệt độ trong 5 ngày
df_48918['2015-09-10':'2015-09-15'].plot(marker='o',color='Red')
```



(b)

Hình 3.10: Kết quả (a) và trình diễn kết quả (b) kiểm chứng điểm dữ liệu ngoại lai có giá trị thấp nhất tại trạm 48918

Từ các kết quả kiểm chứng có thể khẳng định các điểm này là các điểm dữ liệu ngoại lai, có giá trị sai khác rất lớn so với giá trị thực tế. Do vậy, cần phải được xử lý trước khi sử dụng cho bất kỳ mục đích nào.

e) Xử lý các dữ liệu ngoại lai sau khi phát hiện

Do dữ liệu này là dữ liệu dạng chuỗi thời gian, do vậy, chọn phương pháp thay thế các điểm ngoai lai về giá trị rỗng (None), coi như đó là một thời điểm dữ liệu thiếu (missing data). Sau đó, sẽ sử dụng phương pháp xử lý dữ liệu thiếu chung cho toàn bộ tập dữ liệu.

```
#Xử lý dữ liệu ngoại lai của trạm 48918
df_48918.loc['2015-02-01 16:00:00',['48918']] = None
df_48918.loc['2015-02-10 19:00:00',['48918']] = None
df_48918.loc['2016-03-20 13:00:00',['48918']] = None
df_48918.loc['2016-04-30 22:00:00',['48918']] = None
df_48918.loc['2017-06-15 16:00:00',['48918']] = None
df_48918.loc['2015-09-10 10:00:00',['48918']] = None
```

Hình 3.11: Xử lý dữ liệu ngoại lai trạm 48918

3.1.3. Công cụ để chuẩn hóa dữ liệu KTTV

3.1.3.1. Phát triển công cụ chuẩn hóa dữ liệu KTTV

a) Phát triển các thư viện mã nguồn mở để thực hiện chuẩn hóa dữ liệu

Phát triển các module NumPy, SciPy và Pandas trong thư viện mã nguồn mở của Python thực hiện chức năng biên dịch nhanh cho các thao tác toán học và số để chuẩn hóa dữ liệu trong học máy. Trong đó, NumPy tính toán hiệu quả các mảng và ma trận đa chiều; SciPy cung cấp các chức năng hữu ích cho tối ưu, hồi quy, biến đổi Fourier và nhiều kĩ thuật khác; Thư viện Pandas thực hiện các thao tác dữ liệu:

- Đọc/ ghi dữ liệu giữa bộ nhớ và nhiều định dạng file: csv, text, excel, sql database, hdf5;

- Liên kết dữ liệu thông minh, xử lý dữ liệu bị thiếu. Tự động đưa dữ liệu phi cấu trúc về dạng có cấu trúc; thay đổi bố cục của dữ liệu;

- Tích hợp cơ chế trượt, lập chỉ mục, lấy ra tập con từ tập dữ liệu lớn; thêm, xóa các cột dữ liệu;

- Tập hợp, thay đổi dữ liệu với group by, cho phép thực hiện các toán tử trên tập dữ liệu;

- Lập chỉ mục theo các chiều của dữ liệu (cao chiều và dữ liệu thấp chiều).

b) Phát triển các công cụ thuật toán chuẩn hóa dữ liệu trong học máy gồm:

Phát triển các công cụ thuật toán chuẩn hóa dữ liệu trong học máy gồm:

- Thuật toán chuẩn hóa theo phương pháp Trung tâm hóa dữ liệu (Centering data);

- Thuật toán chuẩn hóa theo phương pháp Co giãn dữ liệu (Scaling data);

- Thuật toán chuẩn hóa theo phương pháp Chuẩn hóa min-max (rescaling);

- Thuật toán chuẩn hóa theo phương pháp Co giãn trung bình (mean normalization);

- Thuật toán chuẩn hóa theo phương pháp Chuẩn hóa tiêu chuẩn (standardisation);

- Thuật toán chuẩn hóa theo phương pháp Vec-tơ đơn vị;

- Thuật toán chuẩn hóa theo phương pháp Mã hóa đặc trưng dạng nhóm;

- Thuật toán chuẩn hóa theo phương pháp Mã hóa số - Numeric Encoding;

- Thuật toán chuẩn hóa theo phương pháp Mã hóa One-Hot.

```
print('Mean normalisation')
mmin = data[:,0:8].min(axis=0)
mmax = data[:,0:8].max(axis=0)
maverage = data[:,0:8].mean(axis=0)
newdata = data.copy()
newdata[:,0:8] = (newdata[:,0:8] - maverage)
newdata[:,0:8] = np.divide(newdata[:,0:8], np.array(mmax - mmin))
print(newdata)
```

a)

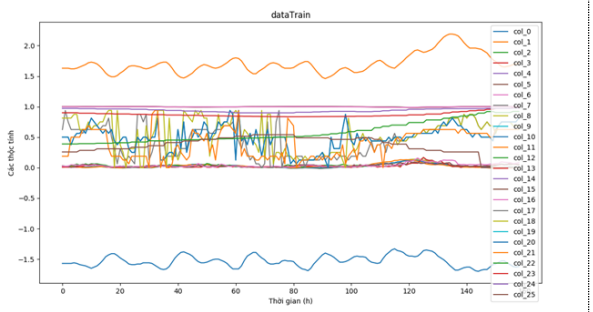
```
print('Min max scaling')
from sklearn import preprocessing as pp
mms = pp.MinMaxScaler()
data_mms = mms.fit_transform(data)
print(data_mms)
```

b)

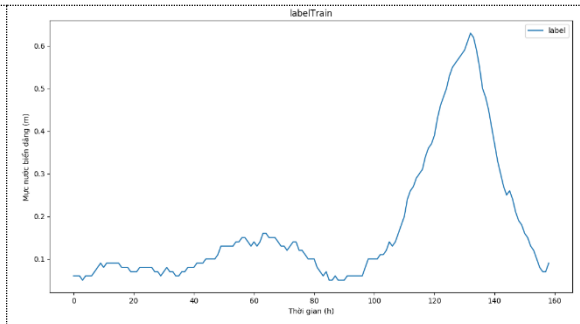
Hình 3.12: Thuật toán chuẩn hóa theo phương pháp co giãn trung bình (a) và phương pháp Chuẩn hóa min-max (b)

3.1.3.2. Thử nghiệm công cụ trên dữ liệu nước biển dâng

Thử nghiệm công cụ chuẩn hóa dữ liệu nước biển dâng theo phương pháp Chuẩn hóa tiêu chuẩn (standardisation) kết quả cụ thể như sau:



a)



b)

Hình 3.13: Dữ liệu mực nước biển trước (a) và sau khi chuẩn hóa (b)

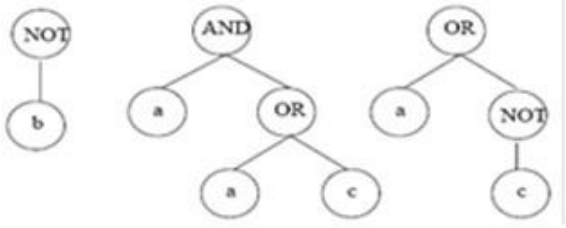
3.1.4. Thuật toán xây dựng mô hình dự báo nước biển dâng do bão

3.1.4.1. Lập trình di truyền

Lập trình di truyền (Genetic Programming - GP) là một sơ đồ tiến hóa để tìm ra lời giải bài toán. Khả năng của GP là tự học định nghĩa của một hàm từ các mẫu điều này giúp GP là một sự lựa chọn tuyệt vời cho bài toán phân tích hồi quy ký hiệu [120]. Chính vì vậy GP được sử dụng rộng rãi để xây dựng các mô hình hồi quy cho các ứng dụng thực tế. Chẳng hạn như mô hình dự đoán giá cổ phiếu sử dụng GP để tạo ra một chiến lược đầu tư sinh lãi [121]. Trong [122] GP được sử dụng để xây dựng mô hình dự báo sóng thời gian thực. Các kết quả của các nghiên cứu trên đã chỉ ra rằng GP là một công cụ đầy hứa hẹn cho các ứng dụng dự báo cho dữ liệu các vùng biển. Trong nghiên cứu [123] GP được sử dụng để dự báo độ xói mòn ống xảy ra ở lòng sông và kết quả cho thấy việc sử dụng GP có kết quả khả thi hơn so với sử dụng phương trình hồi quy và hệ thống nơ-ron nhân tạo trong việc mô hình hóa dự đoán độ sâu xói mòn xung quanh các "ống". Tuy nhiên GP đã và chưa được áp dụng trong dự báo nước dâng do bão vì vậy trong bài báo này tác giả đề xuất nghiên cứu áp dụng GP để xây dựng mô hình "hộp trắng" (một dạng mô hình dễ hiểu) cho việc dự báo nước biển dâng. Lập trình di truyền (Genetic Programming - GP) ra đời vào năm 1992 [120] với tham vọng nhằm đưa ra một quần thể các chương trình mà chúng có thể tiến hóa một cách tự động trên những dữ liệu huấn luyện. Với nghĩa này, GP được xem như là một phần của học máy. Dựa trên lý thuyết tiến hóa của Darwinian, GP đưa ra các chương trình mã hóa dưới dạng các chuỗi di truyền thông qua quá trình tiến hóa và chọn lọc tự nhiên để tìm được chuỗi di truyền (chương trình) tốt đáp ứng được yêu cầu bài toán.

Biểu diễn chương trình

Chương trình trong GP được biểu diễn dưới dạng cây, trong đó mỗi nút được gán nhãn là một ký hiệu thuộc tập hàm (F) hay tập kết (T).



Hình 3.14: Biểu diễn chương trình GP

Toán tử di truyền

Toán tử lai ghép (crossover)

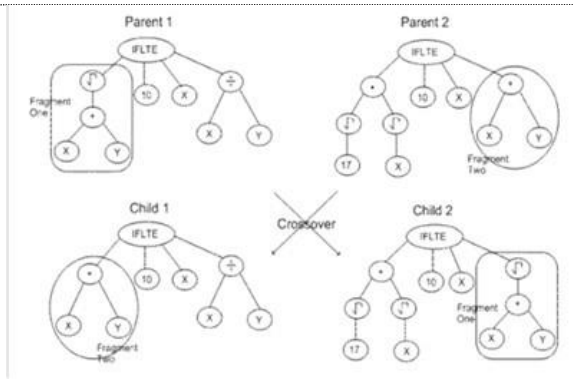
Thể hiện quá trình trao đổi nhiễm sắc thể giữa hai cây bố mẹ. Toán tử gồm các bước sau:

- Chọn một nút ngẫu nhiên trên mỗi cây bố mẹ;
- Hoán đổi hai cây con có gốc tại hai nút vừa chọn và tráo đổi chúng cho nhau.

Toán tử đột biến (Mutation)

Là quá trình đột biến của một bộ nhiễm sắc thể được tạo ra. Gồm các bước sau:

- Chọn ngẫu nhiên một nốt bất kì trên cây cha (mẹ);
- Xóa cây con thuộc nốt được chọn;
- Sinh ngẫu nhiên một cây con mới vào vị trí vừa xóa.



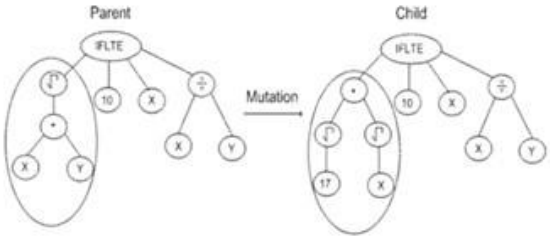
Hình 3.15: Toán tử lai ghép

Tái sinh (reproduction)

Nếu một cá thể được tái sinh chúng sẽ được sao chép y nguyên vào quần thể, hay nói cách khác là sẽ có hai cá thể giống nhau trong quần thể.

Đánh giá độ tốt (fitness)

Mỗi một chương trình được gán một giá trị được gọi là độ tốt, giá trị này sẽ có ảnh hưởng quan trọng đến việc cá thể có được lựa chọn để thực hiện các toán tử di truyền hay không.



Hình 3.16: Toán tử biến dị

Như vậy các bước để chạy một thuật toán GP:

Bước 1: Khởi tạo ngẫu nhiên một quần thể (thế hệ 0) các cá thể được tạo ra từ tập hàm và tập kết.

Bước 2: Thực hiện lặp (các thế hệ) theo các bước phụ sau cho đến khi thỏa mãn điều kiện kết thúc (tìm thấy lời giải tối ưu hoặc đạt đến số thế hệ nào đó):

- Đánh giá độ tốt của các cá thể.
- Chọn 1 hoặc 2 cá thể từ quần thể với xác suất phụ thuộc vào độ tốt của chúng để tham gia vào các toán tử di truyền c.
- Tạo các cá thể mới cho quần thể bằng việc áp dụng các phép toán di truyền sau với một xác suất đã định: (i) Tái sinh; (ii) Lai ghép; (iii) Đột biến.

Sau khi kết thúc quá trình tiến hóa, cá thể tốt nhất của toàn bộ quá trình chạy được coi như là kết quả của quá trình chạy.

Bên cạnh các phương pháp truyền thống: cây quyết định, tập luật quyết định, hàm thống kê và mạng nơron các nghiên cứu đã cho thấy rằng GP cũng là một phương pháp giải bài toán dự báo với độ chính xác cao bằng cách tiến hóa ra cây biểu thức. Một trong những lý do cho phép ta tin tưởng điều này là quá trình tìm kiếm của GP có kết quả tốt đối với những bài toán có không gian tìm kiếm lớn.

3.1.4.2. Lập trình di truyền cho bài toán dự báo nước biển dâng do bão

Việc sử dụng lập trình di truyền (GP) để dự báo nước biển dâng sau bão gần đây cũng đã được một số nghiên cứu áp dụng. Sahoo và Bhaskaran [124] đã đề xuất sử dụng GP để dự đoán nước dâng do bão và ngập lụt do các cơn bão nhiệt đới. Các thí nghiệm được thực hiện trên các bộ dữ liệu từ bờ biển Odisha đến tiếp giáp với Vịnh Bengal. Các kết quả đã chỉ ra rằng cả mạng nơ-ron nhân tạo (ANN) và GP đều dự báo rất tốt so với dữ liệu thực tế. Tuy nhiên, GP đã không được nghiên cứu sâu hơn nữa về các mô hình để dự báo sau khi thực hiện với thời gian dự báo khác nhau. Hơn nữa, tính linh hoạt của GP để tự động chọn các đặc trưng để xây dựng các mô hình có thể hiểu được để dự báo nước dâng do bão cũng chưa được nghiên cứu. Do đó, bài viết này tiếp tục nghiên cứu khả năng của GP để xây dựng các mô hình dự báo mức độ nước dâng sau bão. Trong nước hầu như chưa có nghiên cứu nào về bài toán dự báo nước biển dâng do bão, và hầu chắc chắn rằng chưa có nghiên cứu nào sử dụng công cụ học máy để dự báo nước biển dâng.

3.1.5. Thuật toán mô hình dự báo mưa lớn diện rộng và không khí lạnh

3.1.5.1. Thuật toán Gradient Boosting (GB)

a) Giới thiệu

Thuật toán GB là một kỹ thuật học máy cho bài toán phân loại và hồi quy, nó sinh ra mô hình dự đoán theo cách kết hợp các mô hình dự đoán yếu, điển hình như các cây ra quyết định. Nó xây dựng mô hình từng bậc như những phương pháp Boosting khác thực hiện, và nó khái quát hóa chúng bởi cho phép tối ưu hàm tổn thất khả vi (differentiable loss function) bất kỳ.

Ý tưởng của Gradient Boosting bắt nguồn từ quan sát bởi Leo Breiman, xem Breiman (1997)[125], ở đó Boosting được hiểu như một thuật toán tối ưu trên một hàm giá thích hợp. Các thuật toán hồi quy hiện Gradient Boosting được phát triển bởi Friedman (1999) đồng thời với hướng tiếp cận Gradient Boosting hàm tổng quát hơn của Mason và cộng sự (1999)[126]. Mason và cộng sự giới thiệu cái nhìn lý thuyết về các thuật toán boosting như là thuật toán lặp hướng giảm hàm (functional gradient descent). Đó là những thuật toán làm tối ưu một hàm giá trên không gian hàm bằng cách chọn lặp lại một hàm (giả thuyết yếu) mà nó chỉ trong hướng gra-đi-ăng âm. Cách nhìn gra-đi-ăng hàm của Boosting dẫn đến sự phát triển của các thuật toán Boosting trong nhiều lĩnh vực học máy và thống kê xa hơn là hồi quy và phân loại. Trong phần này chúng tôi trình bày lại Gradient Boosting giới thiệu bởi Li [131].

Giống như các phương pháp Boosting khác, GB kết hợp những "learners" yếu thành một "learner" mạnh duy nhất theo một cách lặp lại. Cách dễ dàng nhất để giải thích trong bối cảnh hồi quy bình phương tối tiểu, ở đó mục đích là "dạy" một mô hình F để dự đoán giá trị theo công thức $\hat{y} = F(x)$ bằng cách tối tiểu lỗi bình phương trung bình $(\hat{y} - y)^2$, được lấy trung bình qua một vài tập dữ liệu huấn luyện của các giá trị thực sự của biến đầu ra y.

Ở mỗi bước $m, 1 \leq m \leq M$ của GB, giả sử rằng có một mô hình không hoàn hảo F_m (ở bước đầu tiên, một mô hình rất yếu chỉ dự đoán giá trị trung bình y trong dữ liệu huấn luyện có thể được sử dụng). Thuật toán GB cải tiến trên F_m bằng cách xây dựng một mô hình mới bằng cách cộng thêm một ước lượng h để đưa ra mô hình tốt hơn

$$F_{m+1}(x) = F_m(x) + h(x). \tag{3.1.1}$$

Để tìm h, giải pháp GB bắt đầu với quan sát mà một h hoàn hảo kéo theo

$$F_{m+1}(x) = F_m(x) + h(x) = y \tag{3.1.2}$$

hoặc tương đương,

$$h(x) = y - F_m(x) \tag{3.1.3}$$

Do đó, GB sẽ khớp h với phần dư $y - F_m(x)$. Giống như những biến thể khác của boosting, mỗi F_{m+1} học để hiệu chỉnh hàm dự đoán trước F_m. Khái quát hóa ý tưởng này với các hàm mất mát khác không chỉ với bình phương lỗi, cho các vấn đề

xếp hạng và phân loại bắt nguồn từ quan sát rằng phần dư $y - F(x)$ cho một mô hình nào đó là gradient âm (đối với $F(x)$) của hàm mất mất lỗi bình phương $\frac{1}{2}\big(y - F(x)\big)^2$. Do đó GB là một thuật toán hướng giảm (gradient descent); và tổng quát hóa, nó dẫn đến sự ghép nối vào một hàm mất mát khác và gradient của nó.

b) Thuật toán

Trong nhiều vấn đề học có giám sát, chúng ta có một biến đầu ra y và một vectơ của các biến đầu vào x liên kết qua một phân bố xác suất $P(x, y)$. Sử dụng một tập huấn luyện $\{(x_1, y_1), \dots, (x_n, y_n)\}$ của các giá trị x và y tương ứng đã biết, mục đích là tìm một xấp xỉ $\hat{F}(x)$ của hàm $F(x)$ làm tối tiểu giá trị kỳ vọng của một hàm mất mát $L(y, F(x))$ đã được chỉ rõ:

$$\hat{F} = E_{x,y}\big[L\big(y, F(x)\big)\big] \qquad 3.1.4$$

Phương pháp GB giả sử y nhận giá trị thực và tìm kiếm một xấp xỉ $\hat{F}(x)$ dưới dạng một tổng có trọng số của các hàm $h_i(x)$ từ một lớp H, được gọi là cơ sở các hàm học (yếu):

$$F(x) = \sum_{i=1}^{M} \quad \gamma_i h_i(x) + const. \qquad 3.1.5$$

Phù hợp với nguyên lý cực tiểu rủi ro thực nghiệm, phương pháp cố gắng tìm một xấp xỉ $\hat{F}(x)$ làm tối thiểu giá trị trung bình của hàm mất mát trên dữ liệu huấn luyện. Nó làm như vậy bằng cách bắt đầu với một mô hình, gồm một hàm hằng $F_0(x)$, và mở rộng tăng dần theo phương thức tham lam:

$$F_0(x) = \sum_{i=1}^{n} \quad L(y_i, \gamma), \qquad 3.1.6$$

$$F_m(x) = F_{m-1}(x) + \sum_{i=1}^{n} \quad L(y_i, F_{m-1}(x_i) + h_m(x_i)), \qquad 3.1.7$$

trong đó $h \in H$ là một hàm học cơ sở. Đáng tiếc, sự lựa chọn hàm h tốt nhất ở mỗi bước đối với hàm mất mát bất kỳ L nói chung là vấn đề tính toán tối ưu không khả thi. Do đó chúng ta sẽ đơn giản hóa bằng cách áp dụng hướng giảm dốc nhất đối với bài toán tối thiểu này. Nếu chúng ta xem xét trường hợp liên tục nghĩa là H là tập các hàm khả vi bất kỳ trên R, chúng ta sẽ cập nhật mô hình theo những phương trình sau:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \quad \nabla_{F_{m-1}} L\big(y_i, F_{m-1}(x_i)\big), \qquad 3.1.8$$

$$\sum_{i=1}^{n} \quad L\Big(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L\big(y_i, F_{m-1}(x_i)\big)\Big), \qquad 3.1.9$$

Ở đó đạo hàm được lấy tương ứng với các hàm F_i với $i \in \{1, \dots, m\}$. Tuy nhiên trong trường hợp rời rạc, nghĩa là tập H là rời rạc, chúng ta sẽ lựa chọn hàm ứng viên h gần với gradient của L nhất, đối với trường hợp này hệ số γ có thể được tính toán với sự hỗ trợ của tìm kiếm trên đường thẳng đối với những phương trình trên. Chú ý

rằng hướng tiếp cận này là tiên nghiệm, do đó không thu được lời giải chính xác đối với bài toán đã cho, nghĩa là không thu được xấp xỉ thỏa đáng.

Giả code của phương pháp GB tổng quát:

Đầu vào: Tập huấn luyện $\{(x_i, y_i)\}_{i=1}^{n}$, một hàm mất mát khả vi $L(y, F(x))$, số lần lặp M.

Thuật toán:

1) Khởi tạo mô hình với giá trị hằng:

$$F_0(x) \sum_{i=1}^{n} \quad L(y_i, \gamma). \tag{3.1.10}$$

2) Với $m = 1$ đến M:

- Tính toán giả phần dư:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \text{ với } i = 1, \dots, n. \tag{3.1.11}$$

- Khớp hàm học cơ sở (nghĩa là cây) $h_m(x)$ với các giả phần dư, nghĩa là huấn luyện chúng sử dụng tập huấn luyện $\{(x_i, r_{im})\}_{i=1}^{n}$.

- Tính toán bội số γ_m bằng cách giải bài toán tối ưu một chiều sau:

$$\gamma_m \sum_{i=1}^{n} \quad L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)). \tag{3.1.12}$$

- Cập nhật mô hình:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{3.1.13}$$

c) Gradient tree boosting

GB điển hình thường được sử dụng với cây ra quyết định (đặc biệt cây hồi quy và phân loại - CART trees) với kích thước cố định như là các hàm học cơ sở. Với trường hợp đặc biệt này, Friedman đưa ra một thay đổi đối với phương pháp GB để cải tiến chất lượng khớp của mỗi hàm học cơ sở [127].

GB khái quát ở bước thứ m sẽ khớp cây ra quyết định $h_m(x)$ với các giả phần dư. Gọi J_m là số lá của cây. Cây đó sẽ được chia không gian đầu vào thành J_m miền rời nhau $R_{1m}, \dots, R_{J_m m}$ và dự đoán một giá trị hằng trên mỗi miền. Sử dụng khái niệm hàm đặc trưng, đầu ra của $h_m(x)$ cho đầu vào x có thể được viết thành tổng:

$$h_m(x) = \sum_{j=1}^{J_m} \quad b_{jm} 1_{R_{jm}}(x), \tag{3.1.14}$$

trong đó b_{jm} là giá trị dự đoán trong miền R_{jm}. Tiếp theo các hệ số b_{jm} được nhân với giá trị γ_m nào đó, giá trị này được lựa chọn bằng cách sử dụng tìm kiếm tuyến tính sao cho tối tiểu hàm mất mát, và mô hình được cập nhật như sau:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad \gamma_m \sum_{i=1}^{n} \quad L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)). \quad 3.1.15$$

Friedman đưa thay đổi thuật toán này sao cho nó lựa chọn giá trị tối ưu phân tách γ_{jm} cho mỗi các miền của cây, thay cho một giá trị γ_m cho cả cây. Ông ta gọi thuật toán cải tiến này là "TreeBoost". Các hệ số b_{jm} từ thủ thục khớp cây có thể đơn giản loại bỏ va quy luật cập nhật mô hình trở thành:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \quad \gamma_{jm} 1_{R_{jm}}(x), \quad \gamma_{jm} \sum_{x_i \in R_{jm}} \quad L(y_i, F_{m-1}(x_i) + \gamma) \; 3.1.16$$

Kích thước của các cây:

J, số node kết thúc trong các cây, là tham số của phương pháp có thể điều chỉnh cho một tập dữ liệu một cách thủ công. Nó điều khiển mức độ cho phép tối đa sự tương tác giữa các biến trong mô hình. Với $J = 2$ (Cây quyết định mộ cấp - Decision Stump), không cho phép sự tương tác giữa các biến. Với $J = 3$, mô hình có thể bao gồm các ảnh hưởng đến 2 biến, …. Hastie và cộng sự (2009) nhận xét rằng $4 \leq J \leq 8$ thường hoạt động tốt cho Boosting và kết quả thường tương đối không nhạy cảm đối với sự lựa chọn của J trong khoảng này, $J = 2$ là không đủ cho nhiều ứng dụng và $J > 10$ là không cần thiết [128].

d) Chỉnh hóa

Khớp tập huấn luyện quá khít có thể dẫn đến giảm khả năng khái quát của mô hình. Một vài kỹ thuật gọi là chỉnh hóa giảm ảnh hưởng học quá bằng cách ép buộc thủ tục khớp. Một tham số chỉnh hóa tự nhiên là số lần lặp M trong GB (nghĩa là số cây trong mô hình khi các hàm học cơ sở là cây ra quyết định). Tăng M làm giảm lỗi trên tập huấn luyện, nhưng đặt nó quá cao có thể dẫn đến học quá. Một giá trị M tối ưu thường được lựa chọn bằng cách kiểm tra lỗi dự đoán trên một tập dữ liệu phê chuẩn riêng rẽ (validation data set). Bên cạnh điều khiển giá trị M, một vài kỹ thuật chỉnh hóa khác cũng được sử dụng.

Co (Shrinkage)

Phần quan trọng của phương pháp GB là chỉnh hóa bằng phương pháp co trong biến đổi luật cập nhật như sau:

$$F_m(x) = F_{m-1}(x) + \nu . \gamma_m h_m(x), \; 0 \leq \nu \leq 1, \qquad 3.1.17$$

trong đó tham số ν được gọi là "tốc độ học" ("learning rate").

Thực nghiệm chỉ ra rằng sử dụng các tốc độ học nhỏ (sao cho $\nu \leq 0.1$) thu được những sự cải thiện đáng kể về khả năng khái quát hóa của mô hình so với GB mà không sử dụng phương pháp co ($\nu = 1$). Tuy nhiên, nó dẫn đến tăng thời gian

tính toán trong cả huấn luyện và xác nhận, tốc độ học thấp yêu cầu nhiều vòng lặp hơn.

GB ngẫu nhiên (Stochastic gradient boosting)

Ngay sau giới thiệu về GB, Friedman đã đề xuất một biến đổi nhỏ với thuật toán này bằng phương pháp Bagging của Breiman, xem Friedman (1999). Cụ thể, ông ta đã đề xuất rằng ở mỗi bước của thuật toán, một hàm học cơ sở sẽ khớp trên một mẫu con được lấy mẫu không thay thế từ tập huấn luyện. Chú ý rằng điều này khác với Bagging, phương pháp lấy mẫu có thay thế bởi vì nó sử dụng các mẫu có cùng kích thước với tập dữ liệu huấn luyện. Friedman đã quan sát thấy sự cải tiến đáng kể về độ chính xác của GB với sự thay đổi này.

Kích thước mẫu con là một phần (không đổi) f của kích thước dữ liệu huấn luyện. Khi $f = 1$, thuật toán là xác định và trùng với miêu tả ở trên. Giá trị f nhỏ hơn đưa sự ngẫu nhiên vào thuật toán và tránh hiện tượng học quá, đóng vai trò như một loại chỉnh hóa. Thuật toán cũng trở nên nhanh hơn, bởi các cây hồi quy phải khớp tập dữ liệu nhỏ hơn ở mỗi vòng lặp. Friedman (1999) thu được rằng $0.5 \leq f \leq 0.8$ dẫn đến kết quả tốt cho các tập dữ liệu có kích thước nhỏ và trung bình. Do đó f thường đặt bằng 0.5, nghĩa là một nửa tập huấn luyện được sử dụng để xây dựng mỗi hàm học cơ sở.

Cũng giống như trong Bagging, lấy tập con mẫu cho phép xác định tỉ lệ lỗi ngoài mẫu con (out-of-bag error) của cải tiến hiệu năng dự đoán bằng cách đánh giá các dự đoán trên những quan sát không được sử dụng trong xây dựng hàm cơ sở học tiếp theo. Ước lượng ngoài mẫu con giúp tránh phải có một tập dữ liệu xác nhận độc lập, nhưng thường đánh giá thấp sự cải tiến hiệu năng thực sự và số lần lặp tối ưu.

Số quan sát trong các lá

Thực hiện gradient tree boosting cũng thường sử dụng chỉnh hóa bằng giới hạn số lượng nhỏ nhất của các quan sát trong các node kết thúc của cây (tham số này được gọi là n.minobsinnode trong gói gbm của R, xem Ridgeway (2007)) [132]. Nó được sử dụng trong quá trình xây dựng cây bằng cách bỏ qua những phân tách dẫn đến các node chứa số mẫu huấn luyện nhỏ hơn giá trị này. Áp đặt giới hạn này giúp giảm biến đổi trong dự đoán ở các lá.

Phạt độ phức tạp của cây

Một kỹ thuật chỉnh hóa hữu ích khác cho gradient tree boosting là phạt độ phức tạp mô hình của mô hình học, xem Tianqi [133]. Độ phức tạp mô hình có thể xác định như là tỉ lệ số lá trong các cây được học. Cùng tối ưu hàm mất mát và độ phức tạp mô hình tương ứng với thuật toán cắt tỉa sau đó để loại bỏ những nhánh mà

không làm giảm hàm mất mát bởi một ngưỡng nào đó. Những kiểu chỉnh hóa khác như phạt l_2 trên những giá trị lá cũng có thể thêm vào để tránh học quá.

e) Sử dụng

GB có thể được sử dụng trong lĩnh vực học xếp hạng. Máy tìm kiếm web thương mại Yahoo và Yandex sử dụng biến thể của GB trong các máy xếp hạng học máy của họ.

f) Tên gọi

Phương pháp này được gọi bởi nhiều tên gọi khác nhau. Friedman đã giới thiệu kỹ thuật hồi quy của ông là "Máy GB" ("Gradient Boosting Machine" (GBM)), xem Friedman (1999). Mason và cộng sự (1999) miêu tả lớp lý thuyết tổng quát của các thuật toán như là "GB hàm" (Functional Gradient Boosting"). Một ứng dụng phổ biến mã nguồn mở cho R gọi nó là "Mô hình GB" ("Generalized Boosting Model"). Các ứng dụng thương mại từ hệ thống Salford sử dụng các tên "Cây hồi quy cộng tính bội" (Multiple Additive Regression Trees" - MART) và "Mạng cây" ("TreeNet"), cả hai đã được đăng ký thương mại.

3.1.5.2. Thuật toán nhận dạng và dự báo không khí lạnh, mưa lớn diện rộng

Bước 1: Tối ưu mô hình phân loại GBM sử dụng tối ưu bayes, ở đây do thời gian tối ưu rất lâu nên chọn một bộ tham số tối ưu.

Bước 2: Chia ngẫu nhiên dữ liệu huấn luyện thành các phần sao cho tỉ lệ có/không có mưa lớn diện rộng (hoặc không khí lạnh) = 1 trong đó lớp số ít có thể được lặp lại, sau đó kết hợp các mô hình GBM tương ứng với các lần chia bằng phương pháp voting cho nhãn và lấy trung vị cho xác suất mưa lớn diện rộng (hoặc không khí lạnh).

Dữ liệu: thời gian lấy dữ liệu từ 1/1/2014 đến 31/12/2019, cụ thể danh sách các trạm trong file csv đính kèm:

- Dữ liệu nhiệt độ tại 43 trạm, tần suất quan trắc 3h một lần.
- Dữ liệu lượng mưa tại 187 trạm, tần suất quan trắc 6h một lần.

Kết quả: Ở đây chúng tôi sử dụng dữ liệu nhiệt độ, và lượng mưa tại trong 36h trước 19h hàng ngày để dự báo nhiệt độ trung bình,và tổng lượng mưa của ngày kế tiếp. Kết quả cụ thể khi học thuật toán trên cho dữ liệu huấn luyện từ 2014 đến hết năm 2018, và kiểm tra đánh giá trên tập dữ liệu năm 2019 như sau:

Với dự báo rét đậm với nhiệt độ trung bình trong ngày dưới 15 độ C (tên trực tiếp của các trạm trong danh sách đính kèm):

trạm: 48800, auc: 0.9958217270194986	trạm: 48830, auc: 0.96132744547036
trạm: 48811, auc: 0.9971988795518206	trạm: 48838, auc: 0.9787063720781299
trạm: 48806, auc: 0.9832657200811359	trạm: 48837, auc: 0.9719405003380662
trạm: 48818, auc: 0.9873949579831933	trạm: 48833, auc: 0.9842154131847725
trạm: 48803, auc: 0.9954481792717087	trạm: 48826, auc: 0.980037140204271
trạm: 48805, auc: 0.9932394366197184	trạm: 48839, auc: 0.9923822714681441
trạm: 48812, auc: 0.9856442577030813	trạm: 48825, auc: 0.9767873723305479
trạm: 48808, auc: 0.9812804782129936	*auc trung bình các trạm:* ***0.9798546099808678***

Với dự báo mưa lớn (tổng lượng mưa trong ngày lớn hơn 50mm)

trạm: 48/34 , auc: 0.7692537313432836	trạm: 48/92, auc: 0.8365384615384616
trạm: 48838, auc: 0.7701551566778218	trạm: 48/93, auc: 0.8925714285714287
trạm: 48/52, auc: 0.761904761904762	trạm: 48/94, auc: 0.8535100286532951
trạm: 48/50, auc: 0.8810991268618387	trạm: 48863, auc: 0.9229340761374187
trạm: 48836, auc: 0.7656054931335831	trạm: 48/95, auc: 0.9126760563380282
trạm: 48846, auc: 0.842292089249493	trạm: 48/83, auc: 0.7981733524355301
trạm: 48/86, auc: 0.7683809523809523	trạm: 48895, auc: 0.7396389524535978
trạm: 48852, auc: 0.8016901408450704	trạm: 48899, auc: 0.6960049937578028
trạm: 48/91, auc: 0.8944761904761905	auc trung bình các trạm: **0.8090980442926845**

3.1.6. Thuật toán xây dựng mô hình thống kê bão

Dữ liệu gồm tốc độ gió tại tâm bão, vị trí tâm bão thực và dự báo từ các trung tâm dự báo của 214 cơn bão diễn ra từ 2010 đến 2019. Chúng tôi thống kê chỉ ra trung tâm nào dự báo chính xác tâm bão, tốc độ gió lớn nhất của bão trong bán kính và khoảng thời gian tương ứng với các tháng quan tâm.

3.1.7. Thuật toán xây dựng mô hình dự báo lũ

Dữ liệu dạng chuỗi thời gian luôn bao gồm phần tuyến tính cũng như phi tuyến (Zhang, 2003). Thực tế là không có một mô hình tổng quát nào có thể xử lý được thành công cả thành phần tuyến tính và phi tuyến. Mô hình thống kê tuyến tính như ARIMA có thể mô hình hóa thành phần phi tuyến của dữ liệu theo chuỗi thời gian không được tốt nhưng nó lại hiệu quả khi mô hình hóa thành phần tuyến tính (Ömer Faruk, 2010; Valenzuela et al., 2008). Trong khi đó các mô hình máy học thống kê không tham số như SVM, RF, KNN hoặc LSTM có thể mô hình hóa bất kỳ thành phần phi tuyến nào. Do đó để có thể có kết quả dự báo tốt hơn, nhóm nghiên cứu đề xuất kết hợp để tạo ra mô hình lai dựa trên ý tưởng mô hình hóa riêng biệt các thành phần tuyến tính và phi tuyến của dữ liệu theo chuỗi thời gian.

3.1.7.1. Lựa chọn thuận toán các mô hình thành phần

Có một số mô hình thống kê tuyến tính có thể mô hình hóa dữ liệu theo chuỗi thời gian đáng chú ý là các mô hình họ ARIMA như ARIMA, ARMA, Seasonal

ARMA….) Gegenbauer ARMA (GARMA (Woodward et al., 2017)). Các loại mô hình này giả định rằng các quy trình là cố định có nghĩa là giá trị trung bình của các chuỗi của bất kỳ mô hình nào trong loại này và tương quan phương sai giữa các giá trị quan trắc là không thay đổi theo thời gian. Tuy nhiên mô hình ARIMA có thể phù hợp với dữ liệu có chuỗi thời gian không cố định dựa vào mô hình ARMA, bằng một quy trình khác biệt giúp chuyển đổi hiệu quả từ dữ liệu không cố định sang dữ liệu cố định.

Hơn nữa thì Valipour và các cộng sự (2013) đã chỉ ra rằng ARIMA có hiệu suất tốt hơn ARMA bởi vì có thể tạo ra được chuỗi thời gian cố định trong giai đoạn huấn luyện cũng như dự báo. Thêm vào đó trong quá trình huấn luyện của ARIMA, chúng tôi luôn quan tâm đến các yếu tố theo mùa như là một đặc tính dữ liệu dự báo lũ.

PARMA (ARMA theo chu kỳ) sẽ mô hình hóa dữ liệu chuỗi thời gian có hàm trung bình, phương sai và hiệp phương sai thay đổi theo mùa (Anderson và các cộng sự 2013). Chúng thường được sử dụng cho dữ liệu có thuộc tính chu kỳ. Như vậy PARMA có thể xử lý ở một mức độ nào đó dữ liệu có chuỗi thời gian không cố định nếu chuỗi thời gian đó trùng với mùa. Tất cả các bộ dữ liệu sử dụng đều có thành phần mùa làm cho dữ liệu không cố định.

Bảng 3.4: Kết quả dự báo trung bình 12 giờ tới của các phương pháp ARIMA, GARMA và PARMA

Size	Method	Sim	MAE	RMSE	FSD	R	NSE
12h	GARMA	0.46	38.03	39.99	0.95	0.76	-35.4
	PARMA	0.61	17.79	19.73	1.6	0.78	-4.36
	ARIMA	**0.65**	**15.07**	**17.3**	**1.01**	**0.61**	**-3.08**

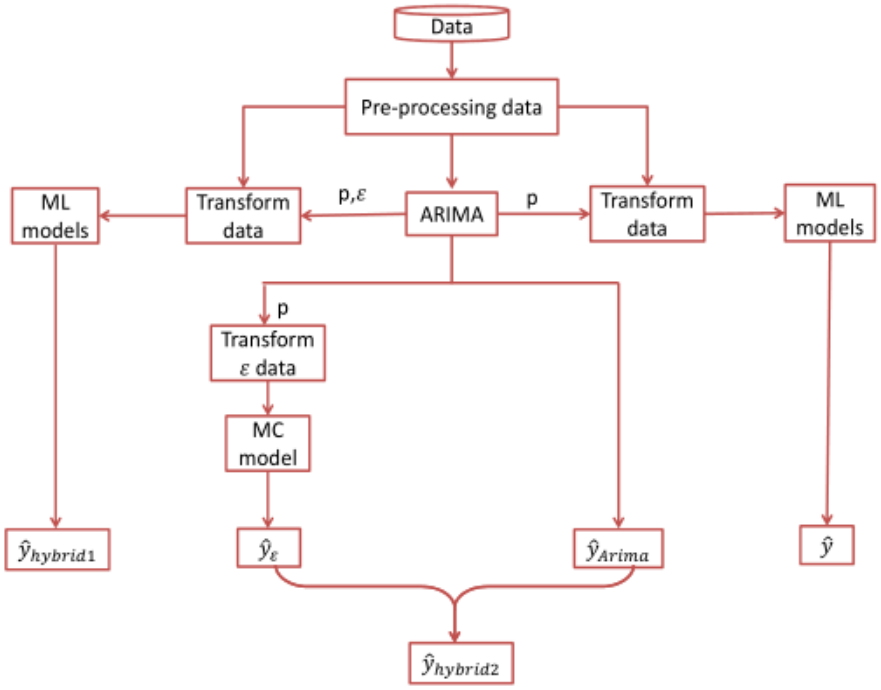
Vì vậy ARIMA, PARMA và GAMRMA là những ứng cử viên phù hợp để mô hình thành phần phi tuyến tính dữ liệu chuỗi thời gian trong đó ARMIA về lý thuyết là phù hợp nhất. Trong ví dụ ở bảng 3.4 mô tả lỗi trung bình của các phương pháp ARIMA, PARMA, GAMRMA của dự báo trung bình 12 giờ tới tại Hà Nội. Rõ ràng rằng ARMIA tốt hơn đáng kể so với các phương pháp khác do đó trong bài này sẽ chọn ARIMA cho phương pháp lai ghép của chúng tôi để mô hình hóa thành phần phi tuyến tính. Nhóm nghiên cứu chọn các phương pháp RF, SVM, KNN, LSTM vì chúng là các phương pháp học thống kê không tham số có nền tảng lý thuyết chắc chắn và đều là đại diện cho các phương pháp học máy thống kê phổ biến hiện nay thứ 3 là các phương pháp này đã được áp dụng thành công trong việc mô hình hóa dữ liệu thủy văn theo chuỗi thời gian.

3.1.7.2. Chi tiết về thuật toán của mô hình đề xuất

Hình 3.17 mô tả các bước của phương pháp nhóm nghiên cứu đề xuất để dự báo dữ liệu lũ theo chuỗi thời gian. Mục tiêu của phương pháp là tập trung vào các lợi ích của mô hình dự báo khác nhau cụ thể là theo loại (tuyến tính và phi tuyến) theo độ phức tạp (mô hình đơn mô hình lai). Một chuỗi thời gian có thể coi là sự kết hợp giữa 2 thành phần tuyến tính và phi tuyến.

$$y(t) = L(t) + N(t) \quad 3.1.18$$

Trong đó L(t) và N(t) là thành phần tuyến tính và phi tuyến tương ứng. Cả 2 thành phần này phải được ước tính từ dữ liệu theo chuỗi thời gian.



Hình 3.17: Sơ đồ mô hình huấn luyện dự báo lũ

Phương pháp đề xuất bao gồm 3 giai đoạn: (i) Mô hình tuyến tính; (ii) Mô hình phi tuyến; (iii) Dự báo giá trị tương lai. Trong giai đoạn đầu tiên sẽ áp dụng mô hình ARIMA để trích xuất phần tuyến tính của chuỗi thời gian (bao gồm cả dữ liệu lịch sử có phân đoạn theo chu kỳ biến đổi khí hậu). Sau đó ở giai đoạn 2 thì thành phần còn lại và các giá trị trễ của chuỗi thời gian sẽ được sử dụng làm đầu vào cho giai đoạn huấn luyện mô hình. Cuối cùng ở giai đoạn 3 thì các giá trị trong tương lai sẽ được ước tính từ các mô hình lai khác nhau. Chi tiết các bước này được mô tả như sau:

- Mô hình tuyến tính ARIMA được áp dụng trên toàn bộ giá trị chuỗi thời gian để thu được các giá trị dự đoán $\hat{L}(t)$ và phần còn lại, phần này sẽ được sử dụng là môt phần đầu vào của bước thứ 2 (mô hình lai).

- Đặc biệt trong nghiên cứu sử dụng tham số thứ tự p thu được từ mô hình ARIMA để xác định thời độ trễ tối ưu cho đầu vào của mô hình lai. Nó giúp có thể chuyển đổi dữ liệu chuỗi thời gian 1 chiều thành dữ liệu p chiều. Vì vậy phương pháp học máy thống kê có thể áp dụng được để dự báo dữ liệu chuỗi thời gian. Cần chú ý rằng thay vì chọn theo kinh nghiệm thì tham số p có được dựa trên phương pháp hướng dữ liệu từ quá trình xây dựng mô hình ARIMA (hàm ACF (hàm tự tương quan, hàm tự tương quan một phần) đươc thực hiện để xác định độ trễ thời gian khác nhau.

Mô hình tuyến tính - phi tuyến

Thành phần còn lại của bước 1 đóng vai trò quan trọng trong việc lựa chọn các mô hình tuyến tính thích hợp vì một mô hình không hoàn toàn tuyến tính nếu nó vẫn chưa thành phần phi tuyến trong đó. Thông thường chúng tôi không thể phát hiện bất kỳ thành phần phi tuyến nào khi phân tích phần dư và trên thực tế không phương pháp thông kê nào cho phép xác định các mối quan hệ hồi quy tự động không tuyến tính. Vì thế chúng tôi xử lý phần dư bằng phương pháp học máy thống kê. Phần dư từ mô hình ARIMA được tính bằng:

$$\epsilon(t) = y(t) - \hat{L}(t) \tag{3.1.19}$$

Trong đó ϵ(t) là phần dư và $\hat{L}(t)$ là giá trị dự báo của mô hình ARIMA tại thời điểm t. Để tìm được mối quan hệ phi tuyến của chuỗi thời gian thì ϵ(t) có thể được tính dựa vào mô hình học máy thống kê như là KNN, SVM, RF, và LSTM. Trong nghiên cứu này nhóm nghiên cứu xây dựng 2 loại mô hình lai.

Mô hình 1:

$$\hat{y}_{hybrid1}(t) = f(y(t-1), ..., y(t-p), \epsilon(t-1), \epsilon(t-2)) \tag{3.1.20}$$

Trong đó f là hàm phi tuyến thu được từ mô hình học máy. Thực hiện một vài bước tiền thực nghiệm trong tập dữ liệu sử dụng để tìm ra điểm dừng của ϵ phù hợp với các giá trị khác nhau (1,2,3….p). Kết quả chỉ ra rằng khi giá trị past window lớn hơn 2 thì giá trị dự báo được không cải thiện thêm và yêu cầu nhiều thời gian tính toán hơn. Vì vậy giá trị past window của ϵ theo kinh nghiệm nên là 2.

Mô hình 2:

Hàm thuật toán dự báo dữ liệu phi tuyến như sau.

$$\hat{y}_{\epsilon}(t) = f(\epsilon(t-1), ..., \epsilon(t-p)) \tag{3.1.21}$$

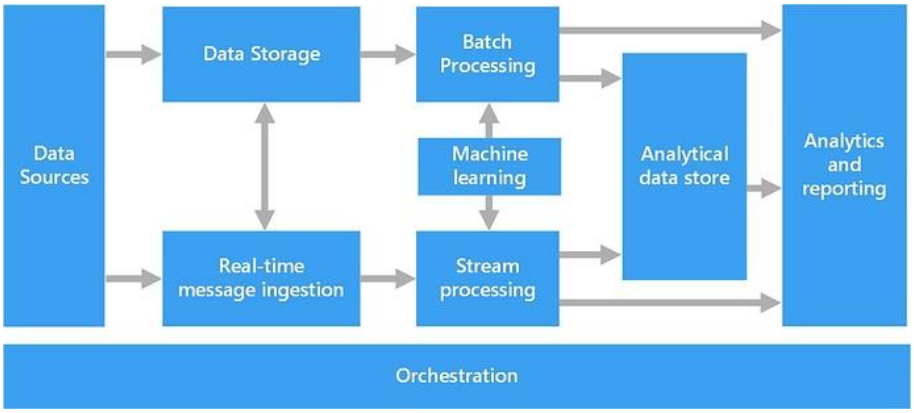
Trong đó f là môt hàm phi tuyến thu được từ mô hình học máy thống kê và ϵ là phần còn lại thu được từ bước 1. Vì thế giá trị dự báo tổng hợp cuối cùng là:

$$\hat{y}_{hybrid2}(t) = \hat{y}_{\epsilon}(t) + \hat{y}_{arima}(t) \quad 3.1.22$$

3.2. Xây dựng Bigdata phục vụ dự báo KTTV theo công nghệ AI

3.2.1.1. Cấu trúc của Big data

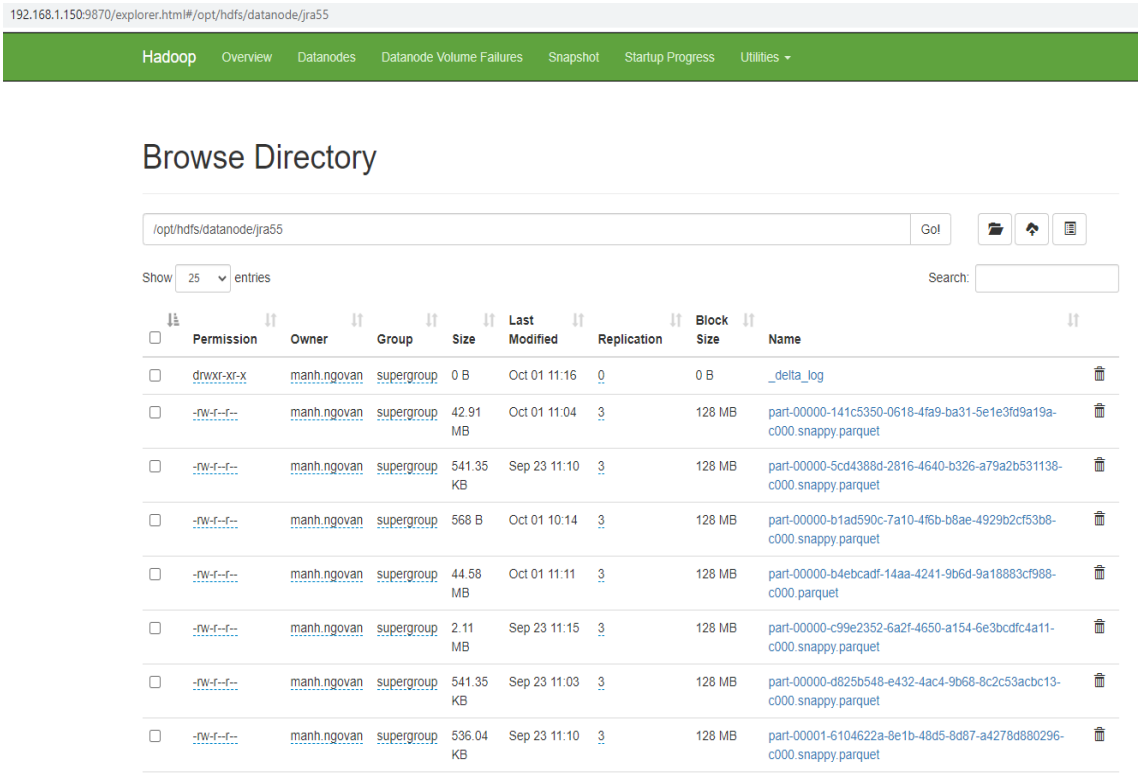
Kiến trúc của hệ thống Big data thường bao gồm các thành phần sau[19]:



Hình 3.18: Kiến trúc Big data

- **Data Sources**: **dữ liệu nguồn** - là nơi dữ liệu được sinh ra, bao gồm dữ liệu có cấu trúc, dữ liệu phi cấu trúc và dữ liệu bán cấu trúc. Dữ liệu từ nhiều nguồn khác nhau như từ các ứng dụng, cơ sở dữ liệu quan hệ (giao dịch mua/ bán, giao dịch ngân hàng, …) hay dữ liệu thời gian thực từ các thiết bị IoT (hình ảnh theo dõi từ camera, cảm biến nhiệt độ, độ ẩm, ….).

- **Data Storage**: **nơi lưu trữ dữ liệu** - được thiết kế để lưu trữ lại khối lượng rất lớn các loại dữ liệu với các định dạng khác nhau được sinh ra bởi dữ liệu nguồn (**Data Source**) trong mô hình xử lý dữ liệu theo lô. Hiện tại, **Apache Hadoop HDFS** đang được sử dụng phổ biến để triển khai thành phần này trong các hệ thống Big Data.



Hình 3.19: Dữ liệu thô được lưu trữ trên HDFS

- **Batch Processing**: **Xử lý dữ liệu theo lô** - thành phần này cho phép xử lý một lượng lớn dữ liệu thông qua việc đọc dữ liệu từ các file nguồn, lọc dữ liệu theo các điều kiện nhất định, tính toán trên dữ liệu, và ghi kết quả xuống 1 file đích.

- **Real-time Message Ingestion**: **Thu thập dữ liệu thời gian thực** - Hệ thống Big data có thể thu thập và lưu trữ các loại dữ liệu trong thời gian thực phục vụ cho việc xử lý dữ liệu theo luồng (Streaming Processing).

- **Stream Processing**: **Xử lý dữ liệu theo luồng** - tương tự như xử lý dữ liệu theo lô, sau khi thu thập dữ liệu thời gian thực, dữ liệu cần phải được lọc theo các điều kiện nhất định, tính toán trên dữ liệu, và ghi kết quả dữ liệu sau khi được xử lý.

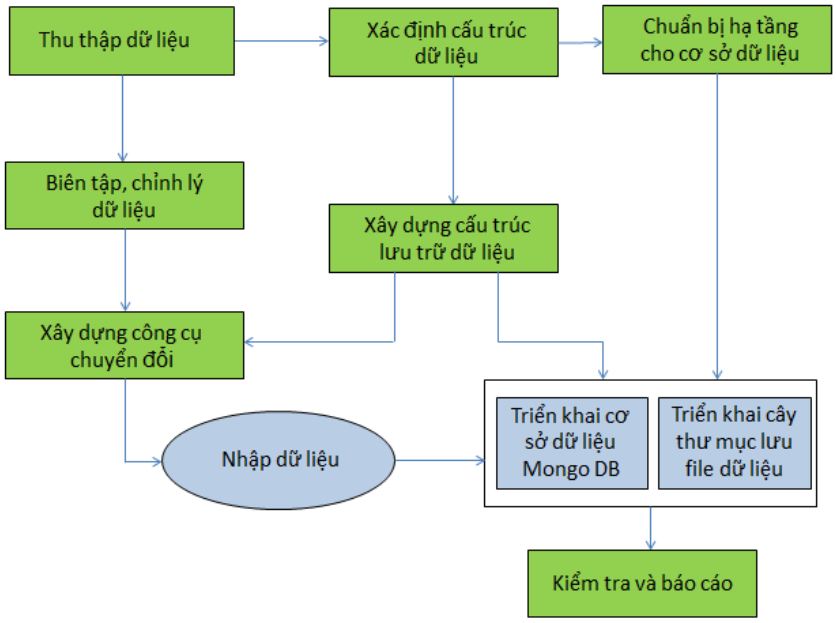
- **Analytical Data Store**: **lưu trữ dữ liệu phân tích** - lưu trữ dữ liệu đã được xử lý theo định dạng có cấu trúc để phục vụ cho các công cụ phân tích dữ liệu (BI Tools).

- **Analysis and Reporting: Phân tích và báo cáo** - thành phần này đáp ứng việc tự khai thác dữ liệu data self-service. Cho phép người dùng cuối trực quan hóa dữ liệu (data visualization), phân tích dữ liệu, cũng như kết xuất các báo cáo khác nhau.

- **Orchestration: Điều phối** - thành phần điều phối các công việc trong một hệ thống Big Data để đảm bảo luồng xử lý dữ liệu được thông suốt, từ việc thu thập dữ liệu, lưu trữ dữ liệu đến lọc, tính toán trên dữ liệu [19].

3.2.2. Thiết lập Big data KTTV

Để thiết lập Big Data phục vụ mô hình AI dự đoán KTTV, đề tài tiến hành các bước tạo lập và chạy thử nghiệm kho dữ liệu theo quy trình sau (hình 3.20).



Hình 3.20: Sơ đồ triển khai thiết lập Big data KTTV

3.2.2.1. Dữ liệu thu thập

Xác định dữ liệu đầy đủ, tin cậy là nền tảng để có thể giúp các hệ thống AI/ML phân tích và dự báo hoạt động và đưa ra kết quả chính xác. Ngay trong giai đoạn đầu tiên, đề tài đã tiến hành điều tra, khảo sát, thu thập số liệu KTTV và tài liệu liên quan về bão, mưa lớn diện rộng, không khí lạnh, lũ (trên hệ thống sông Hồng), nước biển dâng do bão (ven biển Bắc Bộ và Bắc Trung Bộ). Dữ liệu được thu thập gồm:

- Dữ liệu khí tượng bề mặt (độ dài 10 năm) gồm: Nhiệt độ trung bình ngày, nhiệt độ tối thấp ngày, nhiệt độ tối cao ngày, lượng mưa ngày, độ ẩm, gió (hướng và tốc độ), khí áp của tất cả các trạm quan trắc trên toàn quốc.

- Dữ liệu thủy văn (10 năm) và hải văn (10 năm).

- Dữ liệu tái phân tích của các yếu tố khí tượng hải văn và dữ liệu vệ tinh trong 10 năm gần đây.

- Dữ liệu về bão ảnh hưởng đổ bộ vào Việt Nam trong 10 năm gần đây.

- Dữ liệu về lũ, hồ thủy điện và các dữ liệu liên quan trên hệ thống sông Hồng trong 10 năm gần đây.

- Dữ liệu về mưa lớn diện rộng trong 10 năm gần đây.

- Dữ liệu về không khí lạnh trong 10 năm gần đây.

3.2.2.2. Biên tập, chỉnh lý dữ liệu

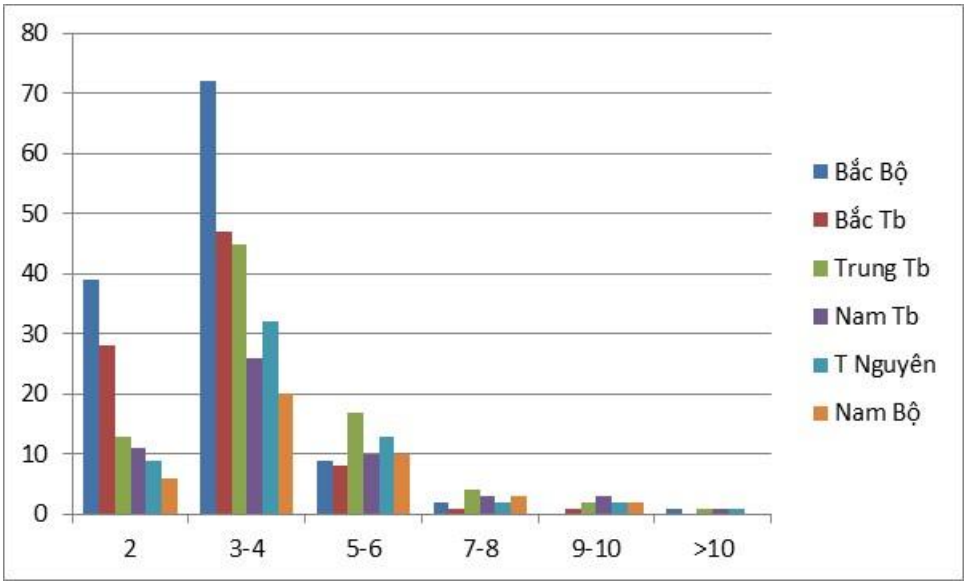
Dữ liệu được thu thập từ nhiều nguồn khác nhau, gồm các nguồn dữ liệu trong và ngoài nước, với độ chuẩn hóa rất khác nhau. Việc biên tập, chỉnh lý các loại dữ liệu đều tuân thủ các quy định của ngành KTTV, do các cán bộ đúng chuyên môn thực hiện và được báo cáo chi tiết trong các Báo cáo công việc của đề tài. Cụ thể:

- Các dữ liệu như khí tượng bề mặt và thủy văn trong nước đã được chuẩn hóa khá tốt nhờ công tác chỉnh biên, lưu trữ theo quy định của ngành KTTV, với các loại dữ liệu này, đề tài đã tiến hành rà soát chuỗi số liệu, xác định các trường hợp thiếu số liệu và đánh giá khả năng sử dụng, tiến hành một số chỉnh sửa định dạng dữ liệu cho phù hợp.

- Các dữ liệu khác như các bản tin cảnh báo bão quốc tế, do tính chất tức thời trong cảnh báo dự báo, có cấu trúc thay đổi và dữ liệu không đồng nhất, đề tài đã tiến hành đánh giá khả năng sử dụng dữ liệu, khả năng bổ sung các khoảng thiếu trong chuỗi số liệu và tiến hành các phương pháp chuyển đổi cấu trúc dữ liệu nhất định.

- Các dữ liệu số lưu trữ theo dịnh đạng không ký tự như dữ liệu tái phân tích định dạng GRIB hay dữ liệu vệ tinh định dạng SATAID được chạy kiểm tra qua các công cụ hiển thị dữ liệu miễn phí như các thư viện GRIB trên môi trường LINUX hay phần mềm hiển thị SATAID do JMA cung cấp. Các trường lưu trữ trong các dữ liệu này đều được phân tích chi tiết.

- Các dữ liệu có liên quan trực tiếp đến kết quả phân tích lũ trên hệ thống sông Hồng và nước biển dâng do bão ven biển Bắc Bộ và Bắc Trung Bộ như dữ liệu về mưa lớn diện rộng, không khí lạnh được các chuyên gia về KTTV phân tích và đánh giá kỹ, từ đó tạo tiền đề cho thiết kế các phân tích máy AI sau này. Ví dụ như biểu đồ phân tích trong báo cáo thu thập dữ liệu về mưa lớn diện rộng của đề tài (hình 3.21).



Hình 3.21: Biểu đồ về tổng số ngày mưa lớn diện rộng kéo dài theo đợt trên các khu vực dự báo từ năm 2008-2017

3.2.2.3. Xác định cấu trúc dữ liệu

Hoạt động này có thể được tiến hành song song với hoạt động biên tập, chỉnh lý dữ liệu. Đây là tiền đề để có thể lưu trữ được dữ liệu trong Big data. Các dữ liệu thu thập sẽ được phân tích số lượng các trường thông tin lưu bên trong và xác định các tính chất bổ sung của dữ liệu như cách thức lưu trữ các giá trị rỗng (null), số phần thập phân lẻ sau dấu phẩy của các giá trị, định dạng dấu thời gian (timestamp) cho các giá trị Date Time, ...

Bảng 3.5: Phân tích cấu trúc dữ liệu tái phân tích

Mã	Thông số trường	Đơn vị	Mực
1	Áp suất	Pa	Mặt nước hoặc mặt đất
83	Độ mấp mô của bề mặt	m	Mặt nước hoặc mặt đất
118	Nhiệt độ phát sáng +	K	Mặt nước hoặc mặt đất
71	Độ che phủ mây tổng *	%	90 - 1100 hPa
75	Độ che phủ mây cao *	%	90 - 500 hPa
74	Độ che phủ mây trung *	%	500 - 850 hPa
73	Độ che phủ mây thấp *	%	850 - 1100 hPa
2	Áp suất quy về mực nước biển	Pa	Mực nước biển trung bình
11	Nhiệt độ	K	2m
18	Độ lệch điểm sương	K	2m
51	Độ ẩm riêng	$kg\ kg^{-1}$	2m
52	Độ ẩm tương đối	%	2m
33	Thành phần u của gió	$m\ s^{-1}$	10m
34	Thành phần v của gió	$m\ s^{-1}$	10m

3.2.2.4. Xây dựng cấu trúc lưu trữ dữ liệu

Căn cứ trên cấu trúc sẵn có của dữ liệu, đề tài tiến hành xây dựng cấu trúc lưu trữ dữ liệu. Có thể lược bỏ một số thông số dữ liệu hoặc phân tích một thông số để lưu trữ trong hai hay nhiều trường, cắt vùng số liệu như dữ liệu tái phân tích hay dữ liệu vệ tinh để giảm thiểu dung lượng lưu trữ. Một số các cấu trúc lưu trữ dữ liệu được đề xuất để đảm bảo có thể lưu trữ được tập dữ liệu lớn của để tài gồm:

- *Danh mục*: gồm các danh mục trạm và danh mục các thông số. Các tập dữ liệu này có tính chất tĩnh, ít thay đổi.

- *Chuỗi số liệu thời gian*: đặc trưng là các chuỗi số liệu quan trắc, có đặc điểm đồng nhất và biến thiên đều theo thời gian.

- *Chuỗi số liệu biến đổi*: đặc trưng là các bản tin cảnh báo, dự báo đường đi của bão của các trung tâm lớn trên thế giới. Các số liệu này có đặc điểm không xác định thời gian bắt đầu và kết thúc, cũng như số trường thông tin trong tập số liệu có thể thay đổi.

- *Lưới số liệu*: đặc trưng là số liệu tái phân tích với giá trị thông số được lưu trong một mảng các giá trị phân bố theo kinh độ và vĩ đô.

- *File số liệu*: số liệu vệ tinh định dạng SATAID được dạng file trong cây thư mục với các đường dẫn số liệu được lưu trong CSDL chính để thực hiện truy xuất dữ liệu hiệu quả từ phần mềm.

3.2.2.5. Xây dựng công cụ chuyển đổi

- **Công cụ chuyển đổi**: là phần mềm tự động được xây dựng riêng trong khuôn khổ đề tài để chuyển đổi tự động tập dữ liệu đã được biên tập, chỉnh lý sang Big data đã được thiết kế.

- **Nền tảng công nghệ**: Các công cụ được xây dựng bằng ngôn ngữ lập trình C#, Python, các thủ tục nội của các cơ sở dữ liệu SQL sẵn có.

- **Nguyên lý hoạt động**: là chuyển đổi dữ liệu sang dạng JSON, trước khi có thể dùng các phương thức (methods) sẵn có của MongoDB để nhập các dữ liệu này.

Ví dụ một đoạn code chuyển đổi dữ liệu SQL sang định dạng JSON sử dụng thủ tục của SQL Server tại hình 3.22.

```
select RTRIM(Storm_List.Ten) as _StormQT_Name, RTRIM(Storm_List.TenVN) as
_StormVN_Name, RTRIM(Storm_List.DataType) as FC_Index,
        Storm_List.FirstDate as FirstDate,
        Storm_List.LastDate as LastDate,
         (select  Track_Contents.CTime as Obs, Cast((SUBSTRING(Track_Contents.CDate,1,4)+'-
'+SUBSTRING(Track_Contents.CDate,5,2)+'-'+
         SUBSTRING(Track_Contents.CDate,7,2)) as datetime) as Track_Date,
          Track_Contents.[CTime] as CTime,Track_Contents.[KinhDo] as Long
    ,Track_Contents.[ViDo] as Lat,Track_Contents.[PMin] as [PMin],Track_Contents.[VMax] as
[VMax],Track_Contents.[R35] as [R35]
         ,Track_Contents.[R50] as [R50] from JMA as Track_Contents where
Track_Contents.ID=Storm_List.ID
          FOR XML AUTO, TYPE, ELEMENTS)
        from ThongTin as Storm_List FOR XML AUTO
```

Hình 3.22: Đoạn thủ tục chuyển đổi dữ liệu dự báo của JMA

3.2.2.6. Chuẩn bị hạ tầng CNTT cho hệ thống cơ sở dữ liệu

Chuẩn bị hạ tầng CNTT cho hệ thống CSDL phục vụ vận hành các mô hình AI dự báo KTTV gồm máy chủ, kết nối máy chủ theo cụm, kết nối với thiết bị lưu trữ SAN, cụ thể như sau:

- Máy chủ: Máy chủ phải có cấu hình bảo đảm hiệu năng cao, đảm nhiệm được khối lượng tính toán lớn, xử lí đồ họa chuyên nghiệp, có không gian lưu trữ dữ liệu lớn, độ an toàn cao và khả năng truy xuất nhanh chóng. Hệ thống tản nhiệt hiệu quả cao, đảm bảo hoạt động liên tục trong thời gian dài. Thông số kỹ thuật chủ yếu như sau: Dual Socket, Intel® Xeon®; 128 GB Dual Rank x 4 DDR4; Intel® Xeon® Processor; 08 Slot Hot Plug Hard Drives; 10 TB HDD; DVD-RW; 2 Power Supply; OS hỗ trợ.

- Về thiết bị lưu trữ SAN: Thiết bị lưu trữ thế hệ mới, đảm bảo lưu trữ dữ liệu lâu dài, sử dụng các công nghệ giao tiếp và lưu trữ mới nhất hiện nay, đảm bảo tính bảo mật và toàn vẹn dữ liệu, có khả năng khôi phục khi xảy ra lỗi. Thông số kỹ thuật chủ yếu như sau: Dung lượng: Up to 211TB raw and 821TB; Điều khiển lưu trữ: 40 TB HDD; Hỗ trợ: (6) Maximum, Expansion Shelves supported; Giao diện máy chủ: Fibre and iSCSI; Điều khiển lưu trữ: Redundant storage controllers; Tính năng sẵn có: Triple+ Parity RAID for data protection (Triple drive parity plus intra-drive parity); Tương thích hệ thống: Microsoft Windows/ Linux

3.2.2.7. Cài đặt cơ sở dữ liệu MongoDB

Cài đặt CSDL MongoDB trên máy chủ Ubuntu thực hiện theo các bước sau:

- Cài đặt các thư viện trợ giúpsudo apt-get install libcurl3 openssl

- Download file MongoDB .tar.gz trên trang chủ mongo: https://www. mongodb. com/download-center?jmp=homepage#community.

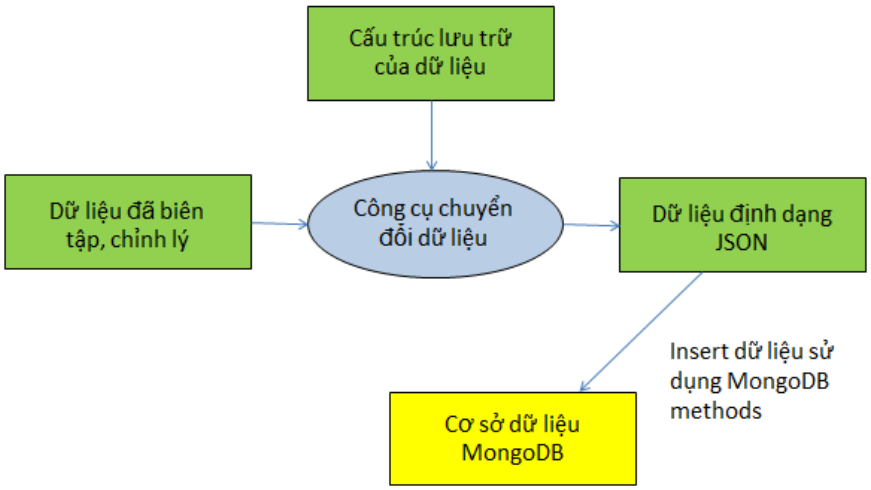
- Giải nén file vừa tải về: tar -zxvf mongodb-linux-*-4.0.0.tgz;

- Tạo biến môi trường: export PATH=<mongodb-install-directory>/bin: $PATH. Trong đó, mongodb-install-directory là path thư mục vừa giải nén.

- Kết nối với MongoDB: mongo --host ip:post

3.2.2.8. Triển khai lưu trữ dữ liệu vào cơ sở dữ liệu MongoDB

Trong bước này, các dữ liệu đã được biên tập, chỉnh lý được các công cụ chuyển đổi dữ liệu tự động sang định dạng JSON. Sau đó các dữ liệu này được nhập (insert) vào trong CSDL MongoDB theo các phương thức sẵn có của MongoDB. Sơ đồ triển khai lưu trữ dữ liệu và CSDL MongoDB như hình sau.



Hình 3.23: Hoạt động nhập dữ liệu bán tự động vào MongoDB

3.2.3. Lưu trữ số liệu KTTV trong Bigdata

3.2.3.1. Lưu trữ số liệu quan trắc bề mặt

Để phục vụ nghiên cứu, phát triển thử nghiệm và đánh giá các phương pháp học máy, nhận dạng, hỗ trợ dự báo một số hiện tượng KTTV nguy hiểm như bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão, đề tài đã tiến hành thu thập, thống kê, phân tích các loại số liệu khí tượng bề mặt. Cụ thể như sau:

a) Thời gian số liệu và khối lượng trạm thu thập: Thu thập số liệu trong vòng 10 năm (từ năm 2008-2017); khối lượng là 187 trạm khí tượng bề mặt thuộc Tổng cục KTTV trên phạm vi cả nước.

b) Yếu tố lưu trữ gồm: Nhiệt độ trung bình ngày, nhiệt độ tối thấp ngày, nhiệt độ tối cao ngày, lượng mưa ngày, độ ẩm, gió (hướng và tốc độ), khí áp.

c) Dạng và cấu trúc số liệu lưu trữ: Số liệu ở dạng bảng excel minh họa tại Bảng 3.6 và cấu trúc lưu trữ tại **phụ lục 01**.

Bảng 3.6: Minh họa biểu số liệu nhiệt độ tối cao trạm Phủ Liễn năm 2014

Trạm: Phủ Liễn Tỉnh:Hải Phòng			KẾT QUẢ TÍNH TOÁN, CHỈNH LÝ, BIÊN TẬP TÀI LIỆU **NHIỆT ĐỘ TỐI CAO**						Kinh độ: 106.38 Vĩ độ: 20.48 Đơn vị: oC			
Năm/ Ngày	***Tháng***											
2014	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	21.5	23.3	20.8	26.6	25.2	33.1	30.9	33.9	27.9	32.7	30.4	23.0
2	22.1	20.7	25.4	24.4	29.9	33.7	30.8	33.5	30.2	31.9	28.3	16.6
3	23.4	24.1	20.7	26.3	28.9	34.0	33.2	33.6	30.7	33.4	22.8	18.6
4	23.5	26.0	20.5	25.9	30.0	32.9	33.8	32.8	32.3	32.6	24.7	18.0
5	22.7	21.8	17.7	21.9	23.1	33.8	29.6	32.4	32.4	27.5	27.3	18.5
6	20.2	21.6	19.8	25.6	25.1	30.2	29.9	30.3	32.7	30.7	24.9	18.9

7	21.7	23.2	19.8	25.7	26.6	31.0	32.0	27.7	32.7	31.6	23.7	17.2
8	25.0	19.8	20.0	23.8	27.8	33.0	34.4	30.7	31.2	31.5	20.2	20.6
9	17.6	19.2	18.3	24.2	33.9	34.2	33.2	30.2	31.1	31.2	20.3	20.4

d) Về nhập và kiểm tra dữ liệu: Nhập tự động vào cơ sở dữ liệu lớn Big Data một lần bằng công cụ phần mềm được phát triển riêng cho đề tài và được kiểm tra chuỗi thời gian trên cơ sở dữ liệu một lần nữa để đảm bảo tính chính xác cao.

e) Lưu trữ danh mục trạm quan trắc bề mặt: Tại Collection T_StationList.

f) Lưu trữ số liệu nhiệt độ trung bình ngày: Tại Collection TT.

g) Lưu trữ số liệu nhiệt độ tối thấp ngày: Tại Collection Tm.

h) Lưu trữ số liệu nhiệt độ tối cao ngày: Tại Collection Tx.

i) Lưu trữ số liệu lượng mưa ngày: Tại Collection RainFall.

j) Lưu trữ số liệu độ ẩm: Tại Collection Humidity.

k) Lưu trữ số liệu gió (hướng và tốc độ): Tại Collection Wind.

l) Lưu trữ số liệu khí áp: Tại Collection P

m) Số liệu mưa từ các trạm/ điểm đo mưa tự động:

- Số lượng trạm/ điểm đo mưa: khoảng hơn 1000 trạm;
- Cấu trúc số liệu mưa 1 giờ được lưu tại Collection Rain1h;
- Cấu trúc số liệu mưa 6 giờ được lưu tại Collection Rain6h;
- Cấu trúc số liệu mưa 24 giờ được lưu tại Collection Rain24h;
- Cờ chất lượng cho biết có bao nhiêu giá trị quan trắc đã được cộng dồn để có số liệu lượng mưa 1h/ 6h/ 24h, từ đó xác định độ tin cậy của số liệu.

3.2.3.2. Lưu trữ số liệu bão

- Định dạng số liệu: được chuyển đổi và lưu trữ từ nhiều nguồn khác nhau với các định dạng không thống nhất như Excel, Access, cơ sở dữ liệu MS SQL server.
- Số liệu bão của Việt Nam được lưu trữ theo cấu trúc tại Collection dbStorms.
- Số liệu nước ngoài: các bản tin cảnh báo bão của JMA, bao gồm cả số liệu bão thực tế và số liệu cảnh báo bão từ các trung tâm lớn trên thế giới như Mỹ, Nhật, Hồng Kông. Cấu trúc lưu trữ số liệu trong Collection Warning.

3.2.3.3. Lưu trữ số liệu mưa lớn diện rộng

Lưu trữ số liệu mưa lớn diện rộng trong cơ sở dữ liệu lớn Big Data được lưu tại Collection RainFall.

3.2.3.4. Lưu trữ số liệu không khí lạnh

Lưu trữ số liệu không khí lạnh trong Big data được căn cứ theo nhiệt độ trung bình, nhiệt độ cao nhất, nhiệt độ thấp nhất của trạm tương ứng. Số liệu được lưu tại các Collection TT, Tx, Tm.

3.2.3.5. Lưu trữ số liệu thủy văn và lũ

Để phục vụ cho nghiên cứu, phát triển thử nghiệm mô hình nhận dạng hình thế và hỗ trợ dự báo, cảnh báo lũ trên hệ thống sông Hồng, đề tài đã tiến hành thu thập, chỉnh lý, biên tập, phân tích, đánh giá, tổng hợp bộ số liệu thủy văn, cụ thể như sau:

- Loại số liệu thu thập: Mực nước, lưu lượng;
- Khối lượng trạm thu thập: 260 trạm;
- Phạm vi số liệu: các trạm trên hệ thống sông Hồng - Thái Bình và các trạm thủy văn trên phạm vi cả nước;
- Thời gian số liệu: 10 năm (từ năm 2008-2017);
- Dạng và cấu trúc dữ liệu: dạng file excel có cấu trúc **tại phụ lục 01**.
- Nhập dữ liệu vào Big data: Nhập tự động bằng phần mềm.

Bảng 3.7: Minh họa biểu số liệu mực nước trạm Mù Cang Chải năm 2011

Trạm: Mù Cang Chải Sông : Nậm Kim				KẾT QUẢ TÍNH TOÁN, CHỈNH LÝ, BIÊN TẬP TÀI LIỆU MỰC NƯỚC						Kinh độ: 104.85 Vĩ độ: 21.85 Đơn vị: mm		
Năm/ Ngày	***Tháng***											
2011	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
1	0	0	0	23422	31230	31233	23440	23435	23427	23427	31233	23424
2	0	0	0	23422	31230	31233	23437	31239	31236	23426	23425	23424
3	0	0	0	23422	18744	31233	31239	31239	31236	23426	23425	23424
4	0	0	0	23422	13391	31233	31242	31247	31238	23426	23425	23424
5	0	0	0	23422	15619	31233	31238	31239	31237	23426	31233	23424
6	0	0	0	23422	31233	31238	31238	31238	31237	23426	23425	23424
7	0	0	0	23424	18741	31237	31237	31239	31236	23426	23425	23424
8	0	0	0	23423	31235	31233	18748	31238	31236	23426	31233	23424
9	0	0	0	23422	31235	31234	31242	31238	31236	23426	23424	23424

- Danh mục trạm thủy văn được lưu trong Collection T_HydroStationList.
- Số liệu mực nước được lưu trong CollectionWaterLevel.
- Số liệu lưu lượng nước được lưu trong Collection Discharge.
- Lưu trữ số liệu lũ, lụt trong Big data được căn cứ theo mực nước của trạm tương ứng. Số liệu được lưu trong CollectionWaterLevel.

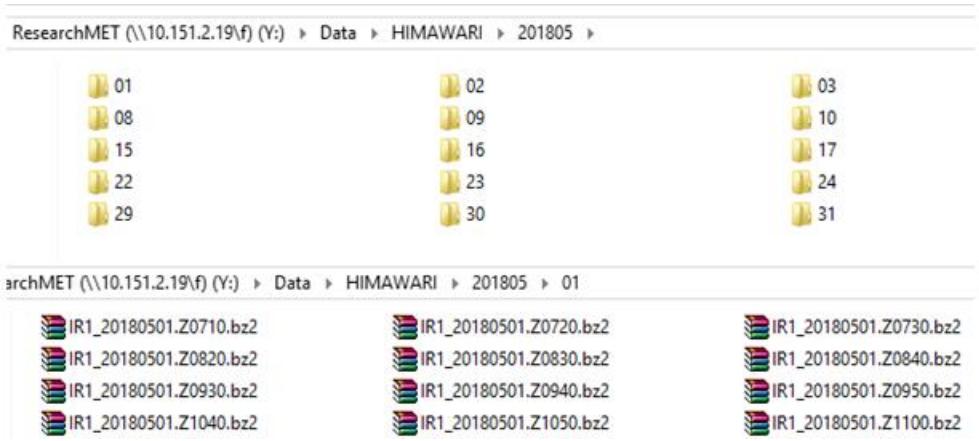
3.2.3.6. Lưu trữ số liệu hải văn và nước biển dâng do bão

Để phục vụ cho nghiên cứu, phát triển thử nghiệm hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Bộ) bằng mô hình sử dụng AI, đề tài đã tiến hành thu thập, chỉnh lý, biên tập, phân tích, đánh giá, tổng hợp bộ số liệu hải văn, cụ thể như sau:

- Loại số liệu thu thập: độ cao sóng và chu kỳ sóng;
- Khối lượng trạm thu thập: 06 trạm;
- Phạm vi số liệu: của các trạm vùng biển từ Quảng Ninh đến Thanh Hóa;
- Thời gian số liệu: 4 năm (từ năm 2014-2017);
- Dạng và cấu trúc dữ liệu: dạng file excel có cấu trúc tại **phụ lục 01**.
- Nhập dữ liệu vào Big data: Nhập tự động bằng phần mềm.
- Danh mục trạm hải văn được lưu chung với danh mục trạm thủy văn trong Collection T_HydroStationList.
- Số liệu độ cao sóng được lưu trong Collection WaveHeight.
- Số liệu chu kỳ sóng được lưu trong Collection WaveFreq.
- Lưu trữ số liệu nước dâng do bão trong Big data căn cứ theo mực nước của trạm triều tương ứng và thời gian bão trong số liệu bão Việt Nam. Số liệu được lưu trong Collection WaterLevel và Collection dbStorms.

3.2.3.7. Lưu trữ số liệu viễn thám

- Số liệu viễn thám được lưu trong CSDL filebase và có thể truy xuất thông qua đường dẫn được lưu trong MongoDB (Hình 3.24).



Hình 3.24: Lưu trữ dữ liệu vệ tinh Himawari trên SAN

- Cấu trúc lưu trữ đường dẫn file số liệu vệ tinh tại Collection Satellite.

3.2.3.8. Lưu trữ số liệu tái phân tích

- Số liệu tái phân tích được đề tài sử dụng là số liệu JRA-55 của JMA;

- Các loại số liệu tái phân tích thu thập gồm: Gió (u, v) độ cao 10 m; độ ẩm tương đối và tuyệt đối độ cao 2m; áp suất bề mặt; nhiệt độ điểm địa lý 2 m; nhiệt độ mặt nước biển; hơi nước cột tổng.

- Giai đoạn phân tích số liệu là: 55 năm, từ năm 1958 – 2012;

- Công nghệ phân tích số liệu: Công nghệ đồng hóa dữ liệu.

- Tên trường cho các file số liệu tái phân tích JRA55 là anl_surf.

- Mảng các giá trị số liệu theo tọa độ sau khi cắt vùng gồm 8480 phần tử.

Bảng 3.8: Mảng lưu trữ giá trị JRA55 của các thành phần số liệu

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	lat	Double	x	kinh độ
2	lon	Double	x	vĩ độ
3	value	Double	x	giá trị các thành phần số liệu

- Cấu trúc lưu trữ số liệu tái phân tích gồm:

  + Gió u và v được lưu trong Collection: JRA55_10m_Uwind_component và JRA55_10m_Vwind_component.

  + Độ ẩm tương đối và tuyệt đối được lưu trong Collection: JRA55_2m_relative_humidity và JRA55_2m_specific_humidity.

  + Áp suất được lưu trong Collection JRA55_Pressure_reduced_MSL.

  + Nhiệt độ điểm địa lý được lưu trong Collection JRA55_2m_temperature.

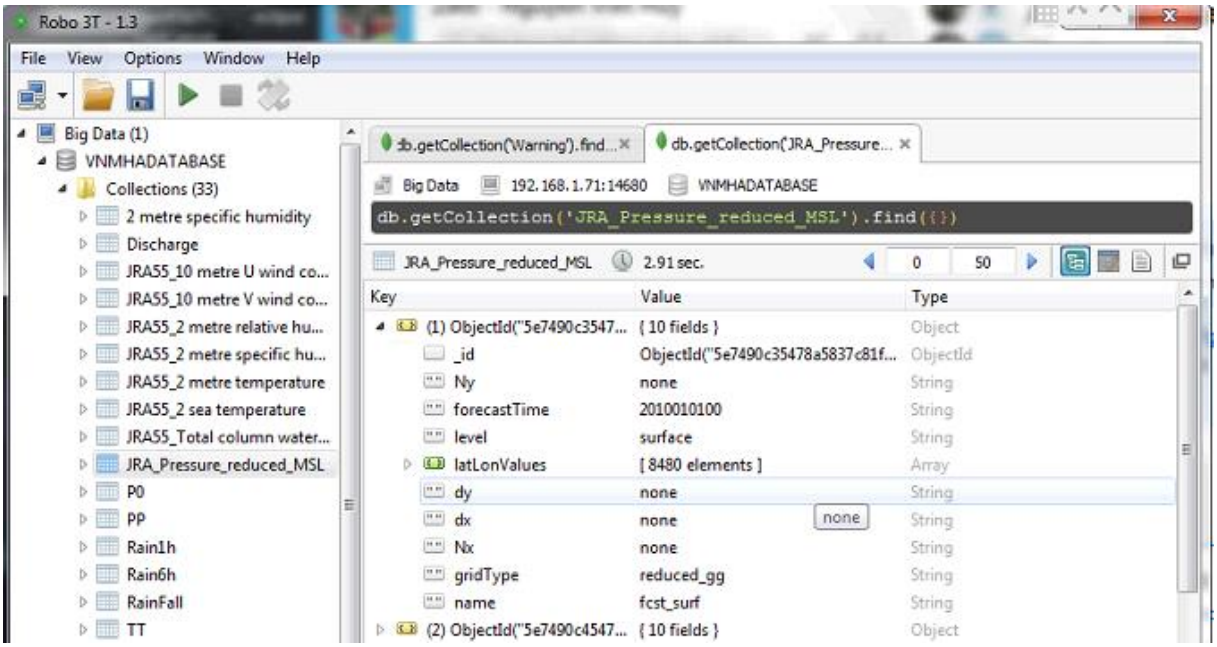
  + Nhiệt độ mặt nước biển lưu trong Collection JRA55_sea_temperature.

  + Hơi nước cột tổng lưu trong Collection JRA55_total_column_water_vapour.

3.2.4. Khả năng, phương thức truy truy xuất dữ liệu từ Big data

Hệ thống Big data của đề tài được thiết lập phải đảm bảo về khả năng kết nối và phương thức truy xuất dữ liệu từ Big data với các mô hình AI để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng KTTV nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam, cụ thể là bão, mưa lớn diện rộng, không khí lạnh ở khu vực Bắc Bộ, lũ trên hệ thống sông Hồng và nước biển dâng do bão (ven biển Bắc Bộ và Bắc Trung Bộ). Khả năng kết nối và phương thức truy xuất dữ liệu đảm bảo tính sẵn sàng cao, đáp ứng được các yêu cầu của ứng dụng. Phương thức truy cập CSDL sử

dụng các công cụ sẵn có của MongoDB sẽ cho phép truy xuất các dữ liệu với tốc độ nhanh hơn so với các hệ quản trị cơ sở dữ liệu quan hệ.



Hình 3.25: Dữ liệu tái phân tích áp suất quy về mực nước biển từ Big Data

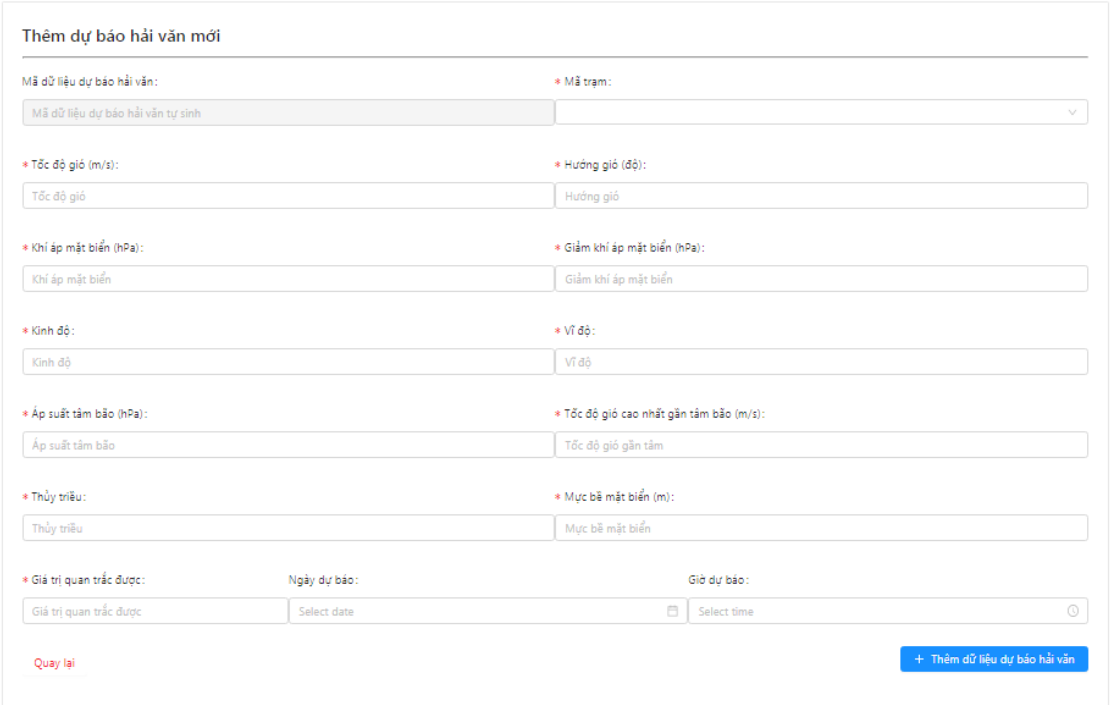
3.2.5. Khả năng cập nhật của Big Data

Hệ thống Big Data được cập nhật theo 2 phương thức, thủ công và tự động. Hầu hết các dữ liệu như dữ liệu bão, dữ liệu tái phân tích, dự liệu vệ tinh ... được cập nhật thời gian thực sử dụng các module chạy ngầm.

```
select RTRIM(Storm_List.Ten) as _StormQT_Name, RTRIM(Storm_List.TenVN) as _StormVN_Name,
RTRIM(Storm_List.DataType) as FC_Index,
        Storm_List.FirstDate as FirstDate,
        Storm_List.LastDate as LastDate,
         (select  Track_Contents.CTime as Obs, Cast((SUBSTRING(Track_Contents.CDate,1,4)+'-
'+SUBSTRING(Track_Contents.CDate,5,2)+'-'+
         SUBSTRING(Track_Contents.CDate,7,2)) as datetime) as Track_Date,
          Track_Contents.[CTime] as CTime,Track_Contents.[KinhDo] as Long
    ,Track_Contents.[ViDo] as Lat,Track_Contents.[PMin] as [PMin],Track_Contents.[VMax] as
[VMax],Track_Contents.[R35] as [R35]
         ,Track_Contents.[R50] as [R50] from JMA as Track_Contents where
Track_Contents.ID=Storm_List.ID
          FOR XML AUTO, TYPE, ELEMENTS)
from ThongTin as Storm_List FOR XML AUTO
```

Hình 3.26: Đoạn thủ tục chuyển đổi dữ liệu dự báo của JMA

Một số các dữ liệu quan trắc thời gian thực (như dữ liệu quan trắc hải văn) được cập nhật thông qua giao diện web.



Hình 3.27: Giao diện thêm mới dữ liệu hải văn

3.3. Kết chương 3

Các nội dung Chương 3 đã trình bày các nội dung nghiên cứu về phát triển công cụ, phương pháp và xây dựng mô hình hệ thống học máy/ AI gồm: (i) Nghiên cứu phát triển các công cụ, phương pháp và xây dựng mô hình hệ thống học máy, AI để hỗ trợ dự báo KTTV; (ii) Xây dựng Big data phục vụ dự báo KTTV theo công nghệ AI.

Chương tiếp theo sẽ trình bày kết quả nghiên cứu xây dựng và triển khai các hệ thống dự báo KTTV theo công nghệ AI gồm: (i) Xây dựng mô hình AI để nhận dạng, dự báo hiện tượng KTTV nguy hiểm; (ii) Triển khai hệ thống AI để hỗ trợ dự báo bão khu vực Bắc Bộ; (iii) Triển khai hệ thống AI để hỗ trợ dự báo mưa lớn diện rộng khu vực Bắc Bộ; (iv) Triển khai hệ thống AI để hỗ trợ dự báo không khí lạnh khu vực Bắc Bộ; (v) Triển khai hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng; (vi) Triển khai hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ; (vii) Xây dựng và triển khai hệ thống Framework tích hợp các module AI hỗ trợ dự báo KTTV.

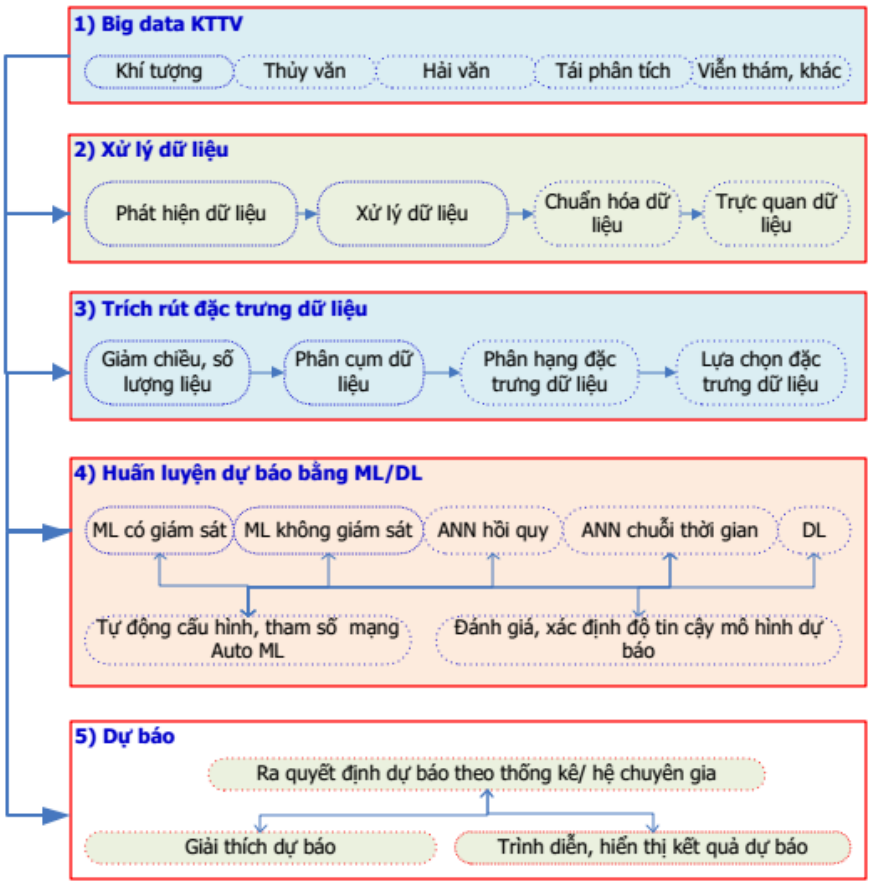
4. CHƯƠNG 4: XÂY DỰNG VÀ TRIỂN KHAI HỆ THỐNG AI HỖ TRỢ DỰ BÁO CÁC HIỆN TƯỢNG KTTV NGUY HIỂM

Trên cơ sở phát triển công cụ, phương pháp và xây dựng mô hình hệ thống học máy/ AI ở Chương 3, trong Chương này sẽ trình bày quá trình thiết lập và triển khai các hệ thống AI hỗ trợ dự báo bão, mưa lớn diện rộng, không khí lạnh khu vực Bắc Bộ; hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng vàhệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ và hệ thống Framework tích hợp các module AI hỗ trợ dự báo KTTV.

4.1. Xây dựng mô hình AI để nhận dạng, hỗ trợ dự báo KTTV

4.1.1. Kiến trúc mô hình AI để nhận dạng, dự báo KTTV

Sơ đồ kiến trúc của mô hình AI nhận dạng, hỗ trợ dự báo KTTV gồm các khối sau: (i) khối Big data KTTV; (ii) khối xử lý dữ liệu; (iii) khối trích rút đặc trưng dữ liệu; (iv) Khối huấn luyện mô hình dự báo; (v) khối dự báo KTTV (Hình 4.1).



Hình 4.1: Sơ đồ kiến trúc mô hình AI nhận dạng, hỗ trợ dự báo KTTV

- Khối Big data KTTV gồm các module: dữ liệu khí tượng, dữ liệu thủy văn, dữ liệu hải văn, dữ liệu tái phân tích, dữ liệu viễn thám.

- Khối xử lý dữ liệu gồm các module: phát hiện dữ liệu, xử lý dữ liệu, chuẩn hóa dữ liệu và trực quan dữ liệu.

- Khối trích rút đặc trưng dữ liệu gồm các module: giảm chiều/ số lượng dữ liệu, phân cụm dữ liệu, phân hạng đặc trưng và lựa chọn đặc trưng dữ liệu.

- Khối huấn luyện mô hình dự báo gồm các module: các mô hình ML/ AI huấn luyện nhận dạng và dự báo; tự động tối ưu cấu hình và tham số mô hình; đáng giá và xác định độ tin cậy của mô hình huấn luyện dự báo KTTV.

- Khối dự báo KTTV gồm các module: Ra quyết định dự báo thống kê/ hệ chuyên gia; giải thích dự báo; hiển thị, trình diễn kết quả dự báo của mô hình.

4.1.2. Nguyên lý hoạt động của mô hình AI KTTV

Nguyên lý hoạt động của mô hình AI hỗ trợ dự báo KTTV cụ thể như sau:

Bước 1 - Xử lý dữ liệu đầu vào của mô hình: Dữ liệu truy xuất từ Big data sẽ được thực hiện các bước tiền xử lý dữ liệu gồm: phát hiện và xử lý các dữ liệu mất mát, không chắc chắn, ngoại lai; tiến hành chuẩn hóa dữ liệu, loại bỏ các dữ liệu thừa, trùng hợp.

Bước 2 - Trích rút đặc trưng dữ liệu đầu vào ban đầu: Dữ liệu sau xử lý sẽ được đưa tới khối trích rút đặc trưng dữ liệu để thực hiện giảm chiều/ số lượng dữ liệu; phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu để làm đầu vào cho các mô hình AI/ML huấn luyện dự báo.

Bước 3 - Huấn luyện nhận dạng và dự báo: Trên cơ sở dữ liệu đầu vào đã được xử lý, trích rút đặc trưng, các mô hình học máy DL/AI sẽ thực hiện các bước tự động nhận dạng dữ liệu đầu vào về hiện tượng thời tiết nguy hiểm, tự động cấu hình, tham số và triển khai huấn luyện dự báo trên dữ liệu đầu vào đã được nhận dạng và tiến hành đánh giá xác định độ tin cậy dự báo của các mô hình. Các mô hình học sâu/ AI được triển khai trên hệ thống tính toán hiệu năng cao Cray XC-40 và các máy chủ nghiệp vụ cấu hình cao.

Bước 4 - Thực hiện dự báo: Căn cứ kết quả huấn luyện dự báo từ các mô hình AI với các tham số khác nhau, hệ thống sẽ thực hiện ra quyết định và giải thích quá trình dự báo.

Bước 5 - Trình diễn hiển thị kết quả dự báo: Hệ thống sẽ sử dụng một số các thư viện như Numpy, Matplotlib, Jupyter notebook, Pandas ... để trình diễn, hiển thị kết quả dự đoán phù hợp với yếu tố và mô hình dự báo (theo các dạng biểu đồ đường, biểu đồ cột, biểu đồ phân tán .....).

Kết quả của các bước xử lý sẽ được thực hiện lưu trữ tại Big data.

4.1.3. Xây dựng cơ chế xử lý dữ liệu đầu vào ban đầu

4.1.3.1. Cơ chế phát hiện và xử lý dữ liệu đầu vào ban đầu cho mô hình AI

Cơ chế thuật toán để phát hiện dữ liệu KTTV đầu vào trong mô hình AI được thực hiện theo các phương pháp:

- Sử dụng điểm tiêu chuẩn Z-Score để phát hiện dữ liệu đầu vào bất thường, ngoại lai.

- Sử dụng kỹ thuật logic mờ Fuzzy C-Mean clustering với Type -2 fuzzy logic system để phát hiện dữ liệu dị thường và nhóm dữ liệu KTTV đầu vào mong muốn để cải thiện độ chính xác và hiệu quả của mô hình AI.

- Sử dụng đồ thị, biểu đồ Box-plot để phát hiện dữ liệu KTTV đầu vào mô hình AI bị mất mát hoặc thiếu.

Cơ chế thuật toán xử lý dữ liệu KTTV đầu vào cho mô hình AI được thực theo các phương pháp:

- Loại bỏ khỏi tập dữ liệu: Đây là cách xử lý dữ liệu đơn giản và dễ thực hiện. Nhưng phương pháp này chỉ áp dụng cho tập dữ liệu chứa các biến độc lập.

- Thay thế bằng một giá trị khác: Thay thế giá trị của các điểm dữ liệu mất mát, bất thường bằng một giá trị khác phù hợp với tập dữ liệu.

4.1.3.2. Lập trình phát hiện và xử lý dữ liệu

Lựa chọn ngôn ngữ lập trình Python, mã nguồn được viết trên hệ thống Google Colab, sử dụng các thư viện nguồn mở bao gồm Pandas, Matplotlib và Seaborn.

4.1.4. Xây dựng cơ chế huấn luyện dự báo các hiện tượng KTTV

4.1.4.1. Cơ chế huấn luyện dự báo

Cơ chế huấn luyện dự báo theo hồi quy dữ liệu và xử lý chuỗi thời gian trong các mô hình ML/ AI thực hiện theo các phương pháp học máy sau:

- LR - Hồi quy tuyến tính (Linear Regression);
- Two Step LR - Hồi quy tuyến tính 2 bước (Two Step Linear Regression);
- LDA 2 bước - Phân tích khác biệt tuyến tính (Linear Discriminant Analysis);
- GP - Lập trình di truyền (Genetic Programming);
- SVM - Máy vec-tơ hỗ trợ (Support Vector Machine);
- DCT - Cây quyết định (Decision Tree);
- k-NN - k hàng xóm gần nhất (k-Nearest Neighbor);

- MLP - Mạng Perceptron nhiều lớp (Multi-layer Perceptron);
- ARIMA - Mô hình thống kê tuyến tính (ARIMA, PARMA và GAMRMA);
- LSTM - Mạng bộ nhớ dài ngắn (Long Short Term Memory Networks);
- RF - Phương pháp rừng ngẫu nhiên (Random Forest).

4.1.4.2. Thuật toán triển khai huấn luyện dự báo

Ví dụ về thuật toán huấn luyện dự báo các dữ liệu không gian - thời gian là sử dụng hồi quy tuyến tính 2 bước (LR-2 bước) gồm các bước sau:

- Bước 1: Các đặc trưng được chia thành các nhóm con và sau đó sử dụng hồi quy tuyến tính để có được giá trị hồi quy với mỗi đặc trưng của các nhóm con.
- Bước 2: Áp dụng thuật toán hồi quy cho các giá trị hồi quy thu được ở bước 1 để tạo ra giá trị hồi quy cuối cùng.

4.1.4.3. Thuật toán đánh giá chất lượng mô hình dự báo

Thuật toán đánh giá, xác định chất lượng và hiệu năng của mô hình AI dự báo sử dụng các chỉ số Sim, MAE, RMSE, R score, FSD và NSE. Cụ thể:

- Sim: Xác định % tương tự giữa giá trị dự báo và giá trị thực đo.
- MAE: Sai số tuyệt đối trung bình giữa giá trị dự báo và giá trị thực đo.
- RMSE: Sai số bình phương tối thiểu là giá trị trung bình của bình phương sai số giữa giá trị dự báo và giá trị thực đo tương ứng.
- R score: Hệ số tương quan giữa 2 biến dự báo và thực đo. Chỉ số này nói lên chất lượng của mô hình dự báo.
- FSD: Độ lệch chuẩn, FSD tiến tới 0 thì phương pháp này là hoàn hảo.
- NSE: Hệ số hiệu quả mô hình NashTHER để đánh giá khả năng dự báo của mô hình thủy văn. NSE càng cao thì giá trị dự báo càng sát giá trị quan trắc.

4.2. Thiết lập và triển khai Deep Learning trên hệ thống tính toán hiệu năng cao

4.2.1. Thiết lập nền tảng hỗ trợ triển khai học sâu DL trên Cray XC-40

Thiết lập các nền tảng để hỗ trợ triển khai học sâu trên hệ thống tính toán hiệu năng cao gồm:

TensorFlow: là một lớp phần mềm mã nguồn mở cung cấp một bộ các quy tắc tính toán số hiệu suất cao.

gRPC: là một lớp lớp gọi thủ tục từ xa mã nguồn mở (Remote Procedure Call layer - RPC) ban đầu được phát triển bởi Google.

Horovod: là một khung DL phân tán, mã nguồn mở cho TensorFlow của Uber. Horovod sử dụng giao diện truyền thông điệp MPI để thiết lập một hạ tầng phân tán cho TensorFlow. Các phiên bản gần đây của Horovod sử dụng các lớp truyền tin NCCL và NCCL2 của NVIDIA để tối ưu hóa hiệu suất truyền tin trên các hệ thống hiện đại với nhiều GPU trên mỗi nút.

4.2.2. Thiết lập Cray PE DL Plugin

Tính toán hiệu năng cao (High Performance Computing - HPC) với số lượng nodes tính toán lớn được sử dụng để giải quyết các vấn đề hạn chế của DL trong bài toán dữ liệu lớn. Hãng Cray đã cung cấp một module cắm (Cray Programming Environments DL Plugin - Cray PE DL Plugin) cho phép lập trình DL trên môi trường song song cho tính toán hiệu năng cao.

Module cắm Cray PE DL Plugin giải quyết vấn đề học song song thông qua một kết hợp các cải tiến thuật toán và tối ưu hóa cao hoạt động giao tiếp dựa trên giao diện truyền thông điệp (Message Passing Interface - MPI). So với các khung DL tính toán các đạo hàm trung bình toàn cục chỉ dựa trên một tính toán MPI chung Allreduce thì Cray PE DL Plugin vượt trội hơn hẳn. Nghiên cứu này mô tả các giải pháp được sử dụng trong Cray PE DL Plugin và làm thế nào các giải pháp này tạo ra hiệu suất tối ưu trên nền tảng Cray. Nghiên cứu thảo luận về việc áp dụng Cray PE DL Plugin vào TensorFlow, một khung DL phổ biến và đánh giá các cải tiến hiệu suất trên nền tảng Cray-XC40 với Bộ xử lý Intel KNL. Kết quả cho thấy hiệu suất thời gian tính toán giảm khoảng 10 (khi sử dụng 8 nodes tính toán) cho bài toán dự báo tốc độ gió sử dụng DL dựa trên hồi quy (Long Short Term Memory - LSTM).

4.2.3. Thiết lập và triển khai áp dụng song song hóa trên Cray XC-40

4.2.3.1. Giải pháp thiết kế

Nhiều khung song song hóa cho DL, chẳng hạn như gRPC trong TensorFlow, gồm hai lớp xử lý. Các hoạt động xử lý máy chủ tham số (Parameter Server – PS) thu thập biến thiên từ các hoạt động xử lý máy trạm, tính toán biến thiên trung bình toàn thể, cập nhật các tham số mạng và gửi các giá trị tham số mới tới các máy trạm. Thông thường người dùng có thể chọn số lượng hoạt động xử lý PS. Chỉ có một hoặc một số giới hạn các hoạt động xử lý PS trên một số lượng lớn máy trạm sẽ gặp phải các vấn đề về hiệu suất và hạn chế quy mô. Cấu hình kiểu này thiết lập một mẫu giao tiếp nhiều-đến-ít, gây tắc nghẽn hầu hết các mạng. Một số lượng hạn chế các hoạt

động xử lý PS cũng sẽ gặp khó khăn trong việc gửi các giá trị tham số cập nhật đủ nhanh để theo kịp nhu cầu của máy trạm. Tăng số lượng các hoạt động xử lý PS có thể làm giảm các nút thắt cổ chai trong truyền tin và cập nhật tham số. Tuy nhiên, sử dụng quá nhiều các hoạt động xử lý PS lại dẫn đến các mẫu giao tiếp nhiều-nhiều, sẽ không đáp ứng số lượng lớn các nút. Xác định số lượng tối ưu các hoạt động xử lý PS sẽ rất mất công sức của người dùng. Dùng gRPC trong TensorFlow, người dùng còn phải cung cấp tên nút và số cổng, như vậy lại nảy sinh các vấn đề về khả năng sử dụng. Cray PE DL Plugin xử lý các vấn đề về khả năng sử dụng và hiệu suất mở rộng trong TensorFlow và các khung DL tương tự. Không có hoạt động xử lý PS khi sử dụng Cray PE DL Plugin. Mỗi xử lý là một máy trạm, và một hoạt động giảm thiểu toàn thể tùy chỉnh thay thế cho hoạt động tính toán biến thiên trung bình của toàn bộ các hoạt động xử lý PS. Mỗi máy trạm sau đó có thể dễ dàng tính toán cập nhật tham số mạng, việc này thường chỉ tốn một phần nhỏ của thời gian thực hiện. Thuật toán trong Hình 4.2 mô tả sơ bộ cách học song song dữ liệu sử dụng Cray PE DL Plugin. Giảm thiểu tùy chỉnh được tối ưu hóa cụ thể cho hoạt động DL và có thể thấy hiệu suất cao hơn 35% so với MPI Iallreduce mặc định có sẵn trong Cray MPICH khi kích thước thông điệp và vị trí xử lý là tương đương. Ngoài ra để cải thiện hiệu suất truyền tin ở một quy mô lớn hơn, thiết lập giảm thiểu tùy chỉnh cũng cung cấp khả năng tuyệt vời chồng lớp truyền tin/ tính toán. Khả năng tạm ẩn truyền tin trong pha tính toán biến thiên trung bình đóng vai trò chính trong việc cải thiện thời gian cho đào tạo phân tán.

Require: N = total number of epochs
Require: n = total number of training samples
Require: k = number of MPI ranks
Require: b = number of training samples in a batch per process
1: **for** epoch = $1 \cdots N$ **do**
2: **for** step = $1 \cdots n/(bk)$ **do**
3: $g_{step} \leftarrow$compute_gradients(local_batch$_{step}$)
4: $G_{step} \leftarrow$mc.gradients(g_{step})
5: $loss_{step} \leftarrow$apply_gradients(G_{step})

Hình 4.2: Thuật toán huấn luyện song song dữ liệu

Code giả lập cho thuật toán huấn luyện song song dữ liệu. Cray PE DL Plugin được trình bày bằng mc, và hàm trung bình đạo hàm là hàm mc.gradients. Trong đó:

- N: tổng số chu kỳ;
- n: số lượng mẫu huấn luyện;
- k: số cấp MPI;
- b: số mẫu huấn luyện trong một lô dữ liệu trong một lần xử lý.

Không cần điều chỉnh TensorFlow để sử dụng Cray PE DL Plugin cho song song hóa. Tính năng TenorFlow Op được sử dụng để thêm các bước truyền tin cần thiết vào đồ thị thực hiện một cách tối ưu (tài liệu có tại https://www.tensorflow.org /extend/adding_an_op). Người dùng có thể bắt đầu với một TensorFlow nối tiếp hoặc một tập lệnh client khung khác rồi gọi thêm các thành phần cần thiết để khởi tạo, truyền tin và kết thúc. Đối với các tình huống yêu cầu nhiều giảm thiểu biến thiên cùng lúc, các nhóm luồng (thread) giảm thiểu được sử dụng để tăng tốc độc lập với một vài hoạt động gọi đơn giản. Giao diện C / C ++ và Python 2/3 có sẵn trong Cray PE DL Plugin. Cray PE DL Plugin đã có sẵn trong gói Cray Developer Toolkit - CDT được cài cho các hệ thống Cray XC. Bản hiện tại CDT 19.09 hỗ trợ Keras, TensorFlow 1.3.1 và kiến trúc dựa trên Intel®CPU và GPU NVIDIA.

4.2.3.2. Thiết lập bộ tập lệnh trong Cray PE DL plugin

Các bước cần có trong Cray PE DL Plugin bao gồm:

- Khởi tạo Cray PE DL Plugin.
- Khởi tạo các giá trị tham số mô hình ban đầu: Chỉ định số lượng nhóm, luồng, kích thước mô hình.
- Sử dụng Cray PE DL Plugin để truyền các biến thiên sau khi tính toán các biến thiên và thực thi mô hình.
- Kết thúc Cray PE DL Plugin.

Trong phần này sẽ trình bày chi tiết cho việc áp dụng cho một tập lệnh bằng Python cho phép sử dụng Keras thực thi các mô hình học máy DL trên Cray-XC40.

***(i).* Khởi tạo**: Bước đầu tiên là khởi tạo Cray PE DL Plugin. Điều này được thực hiện bằng cách trước tiên import module rồi thiết lập môi trường ban đầu (hình 4.3):

```
from __future__ import print_function
import keras
from keras.datasets import mnist
from keras.models import Sequential
from keras.layers import Dense, Dropout, Flatten
from keras.layers import Conv2D, MaxPooling2D
from keras import backend as K
# BO SUNG CHO CRAY
import math
import tensorflow as tf
import dl_comm.keras as cdl

config = tf.ConfigProto()
#config.gpu_options.allow_growth = True
#config.gpu_options.visible_device_list = str(cdl.local_rank())
#config.gpu_options.per_process_gpu_memory_fraction = 0.8
K.set_session(tf.Session(config=config))
# KET THUC BO SUNG CHO CRAY
```

Hình 4.3: Khởi tạo Cray PE DL Plugin

Cray PE DL Plugin sử dụng cấu hình cho cả CPU và GPU, để thực hiện sử dụng GPU cho tính toán cần thiết lập tham số cấu hình ban đầu như hình 4.4:

```
config = tf.ConfigProto()
#config.gpu_options.allow_growth = True
#config.gpu_options.visible_device_list = str(cdl.local_rank())
#config.gpu_options.per_process_gpu_memory_fraction = 0.8
```

Hình 4.4: Thiết lập tham số cấu hình ban đầu

(ii). Khởi tạo các tham số mô hình ban đầu: Với Keras, cần thiết lập các thông số ban đầu của mô hình sẽ sử dụng như hình 4.5:

```
model = Sequential()
model.add(Conv2D(32, kernel_size=(3, 3),
                 activation='relu',
                 input_shape=input_shape))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(num_classes, activation='softmax'))
```

Hình 4.5: Thiết lập các thông số ban đầu

(iii). Tổng hợp biến thiên: Hoạt động truyền tin và tính toán chuyên sâu này được tối ưu hóa cao trong Plugin Cray PE DL Plugin. Hoạt động này được đặt giữa hoạt động tính toán biến thiên và cập nhật mô hình, được cấu hình như hình 4.6.

```
# BO SUNG CHO CRAY
opt = keras.optimizers.Adadelta(1.0 * cdl.get_nranks())

# Wrap the optimizer to use the Plugin
opt = cdl.DistributedOptimizer(opt)
# KET THUC BO SUNG CHO CRAY
```

Hình 4.6: Tối ưu hóa trong Cray

Một tập lệnh học nối tiếp thường sử dụng một phương thức tối thiểu hóa minimize() của một đối tượng tối ưu hóa optimizer. Phương thức này tính toán biến thiên và cập nhật mô hình với các biến thiên này. Việc cập nhật kết quả tính toán của các bước song song được thực hiện bởi hàm callback như sau:

```
# BO SUNG CHO CRAY
# Add callback to broadcast initial variables
callbacks = [cdl.BroadcastVariablesCallback(0, K)]
```

Thực hiện học máy của mô hình và đánh giá kết quả của mô hình được thực hiện như hình 4.7.

```
model.fit(x_train, y_train,
          batch_size=batch_size,
# BO SUNG CHO CRAY
          callbacks=callbacks,
# KET THUC BO SUNG CHO CRAY
          epochs=epochs,
          verbose=verbose_level,
          validation_data=(x_test, y_test))
score = model.evaluate(x_test, y_test, verbose=0)
```

Hình 4.7: Học máy và đánh giá kết quả mô hình

(iv). Kết thúc: Bước bắt buộc cuối cùng để chuyển đổi một tập lệnh đào tạo nối tiếp là kết thúc Cray PE DL Plugin, tương tự như việc kết thúc MPI. Khi rank của node trả về là 0 khi đó quá trình tính toán song song kết thúc như hình 4.8.

```
# BO SUNG CHO CRAY
if cdl.get_rank() == 0:
  print('Test loss:', score[0])
  print('Test accuracy:', score[1])
# KET THUC BO SUNG CHO CRAY
```

Hình 4.8: Kết thúc Cray PE DL Plugin

4.2.3.3. Triển khai, cài đặt Cray PE DL Plugin trên Cray-XC40

Hiệu suất tốt nhất đạt được với một cấp MPI trên mỗi nút. Cray PE DL Plugin phải được cấu hình để sử dụng 2-4 luồng (thread) truyền tin. Trong vài trường hợp với các nút GPU, hiệu suất có thể được cải thiện bằng cách sử dụng lên đến 8 luồng. Đối với đào tạo MKL và MKL-DNN, quan trọng là không đặt OMP_NUM_THREADS quá cao, nếu không các core có thể bị đăng ký vượt mức. Ví dụ: nếu có 36 core vật lý trên một nút, hiệu suất tối ưu đạt được với OMP_NUM_THREADS=34, nên để lại 2 core/ luồng để liên lạc với Cray PE DL Plugin. Ngoài ra, với TensorFlow và ví dụ tf_cnn_benchmarks, num_intra_threads nên được đặt để phù hợp với giá trị OMP_NUM_THREADS, và num_inter_threads thường được đặt từ 1-3 luồng tùy thuộc vào số lượng HyperThreads có trên mỗi core. Đối với các KNL CPU, tốt nhất để lại một HyperThread rảnh trên mỗi core. Biến môi trường KMP_BLOCKTIME có thể giúp cải thiện hiệu suất đôi chút nếu được đặt là 0 hoặc 30.

Đối với các nút GPU, số lượng luồng CUDA được sử dụng để nhớ đệm dữ liệu đến chủ thể tính toán có thể được điều chỉnh thông qua biến môi trường ML_COMM_NUM_CUDA_STREAMS và số lượng bản sao mà mỗi luồng thực hiện có thể được thay đổi với biến môi trường ML_COMM_CPY_PER_ CUDA_STREAM. Các cài đặt mặc định, 2 và 8, tương ứng, theo thử nghiệm là tốt nhất cho gần như tất cả các tình huống.

4.2.4. Thực nghiệm với thuật toán sử dụng mạng LSTM trên Cray -XC40

Hệ thống Cray -XC40 để triển khai thực nghiệm mô hình cụ thể như sau: Thông số máy chủ thường CPU (2 Multi cores (28 thread) Intel Xeon E5-2690 v4s), RAM (64 GB). Đối với máy Cray-XC40, 1 node có CPU (2 cores (36 thread), Intel Xeon E5-2697 v4 18C 2.3 GHz), RAM (16 GB). Cấu hình thực thi LSTM trong thực nghiệm trên Cray-XC40 sử dụng 8 nodes, mỗi node sử dụng 34 thread.

Đề tài đã sử dụng mô hình học sâu DL LSTM (Long Short Term Memory) để thử nghiệm dự báo tốc độ gió trong 18 giờ tiếp theo trên hệ thống tính toán hiệu năng cao Cray - XC40.

Dữ liệu thực nghiệm là dữ liệu quan trắc tốc độ gió từ 01/01/2014 đến 31/12 /2019, với tần suất quan trắc 3h một lần, tại 13 trạm quan trắc: Hà Giang, Cao Bằng, Tuyên Quang, Hòa Bình, Nam Định, Hà Đông, Phú Liễn, Lạng Sơn, Bãi Cháy, Tiên Yên, Móng Cái, Bạch Long Vĩ, Hội Xuân. Số lượng mẫu học là 15,200 mẫu, số lượng mẫu kiểm tra là 500 mẫu. Thực nghiệm sẽ thực thi LSTM với các epoc =500, 800.

Thiết lập mô hình LSTM trên máy chủ thông thường và máy chủ Cray-XC40 sử dụng Cray PE DL Plugin.

```
"""##Xây dựng mô hình LSTM"""
#Stacked LSTM (Nhiều lớp LSTM)
model = Sequential()
model.add(LSTM(50, activation='relu',
               return_sequences=True,
               batch_input_shape=(None,8, 13)))
model.add(LSTM(50,
               return_sequences=False,
               activation='relu'))
model.add(Dropout(0.3))
model.add(Dense(13))
model.compile(optimizer='adam',
              loss='mse',
              metrics=['accuracy'])
model.summary()
"""##Huấn luyện mô hình"""
#Fit model Vanilla LSTM
start_time = time.time()
epochTime = 500
model.fit(x_train, y_train, batch_size=3, epochs=epochTime, validation_split=validation_split)
end_time = time.time()
```

a)

```
"""##Xây dựng mô hình LSTM"""
validation_split = 0.05 #Tỷ lệ % tách tập Validation trong tập Test
#epochTimes = 12.0
epochTimes = 500
# CRAY ADDED
epochs = int(math.ceil(epochTimes / cdl.get_nranks()))
# END CRAY ADDED
#Stacked LSTM (Nhiều lớp LSTM)
model = Sequential()
model.add(LSTM(50, activation='relu',
               return_sequences=True,
               batch_input_shape=(None,8, 13)))
model.add(LSTM(50,
               return_sequences=False,
               activation='relu'))
model.add(Dropout(0.3))
model.add(Dense(13))
# CRAY ADDED
opt = keras.optimizers.Adadelta(1.0 * cdl.get_nranks())
# Wrap the optimizer to use the Plugin
opt = cdl.DistributedOptimizer(opt)
#END CRAY ADDED
model.compile(optimizer = opt,
              loss='mse',
              metrics=['accuracy'])
model.summary()
# CRAY ADDED
# Add callback to broadcast initial variables
callbacks = [cdl.BroadcastVariablesCallback(0, K)]
```

b)

Hình 4.9: Cấu hình LSTM trên máy chủ thường (a) và trên máy Cray XC-40 (b)

4.2.5. Kết quả và phân tích đánh giá

Trong phần này sẽ phân tích và đánh giá kết quả đạt được khi thực thi thực nghiệm trên máy chủ thường và Cray-XC40. Kết quả dự đoán với tính chính xác của LSTM khi chạy trên máy chủ thường và Cray-XC40 là tương đương nhau với

khoảng sai số trung bình giữa máy chủ thường và Cray-XC40 khoảng 0.03. Các bảng sau cung cấp kết quả chi tiết và kết quả sai số trung bình.

Bảng 4.1: Sai số dự báo của LSTM

STT	Station	Trên Cray-XC40		Trên máy chủ thường	
		MAE (Epoc = 500)	MAE (Epoc = 800)	MAE (Epoc = 500)	MAE (Epoc = 800)
1	48805 - HA GIANG	0.81523	0.9741	0.81484	0.82883
2	48808 - CAO BANG	0.77397	0.65291	0.71858	0.85846
3	48812 - TUYEN QUANG	0.55088	0.67463	0.56044	0.66563
4	48818 - HOA BINH	0.66411	0.52507	0.58697	0.67589
5	48823 - NAM DINH	0.73743	0.72945	0.72503	0.75761
6	48825 - HA DONG	0.42744	0.72918	0.46304	0.55307
7	48826 - PHU LIEN	0.85284	0.71751	0.83042	0.91817
8	48830 - LANG SON	0.58352	0.53875	0.58961	0.53911
9	48833 - BAI CHAY	0.74615	0.74463	0.82476	0.79877
10	48837 - TIEN YEN	0.81491	0.86019	0.84871	0.86146
11	48838 - MONG CAI	0.85648	0.88636	0.80696	0.8723
12	48839 - BACH LONG VY	0.89402	0.68479	0.74131	0.73887
13	48842 - HOI XUAN	0.82788	0.88395	0.8547	0.97105
14	All Station	0.73422417	0.738584	0.72041706	0.77225093

Đối với hiệu suất về thời gian thực thi mô hình LSTM thì trên Cray-XC40 cho kết quả nhanh trung bình gấp 10 lần so mới máy chủ thông thường (trong trường hợp Cray-XC40 sử dụng 8 nodes, mỗi node sử dụng 34 thread và máy chủ thường có 28 thread). Kết quả chi tiết ở bảng 4.2.

Bảng 4.2: Thời gian thực thi mô hình LSTM

STT	Cấu hình	Thời gian (phút)
1	Cray-XC40, 8 nodes, epoc = 500	34.8
2	Cray-XC40, 8 nodes, epoc = 800	54.1
3	Máy chủ cứng, epoc = 500	382.68
4	Máy chủ cứng, epoc = 800	577.04

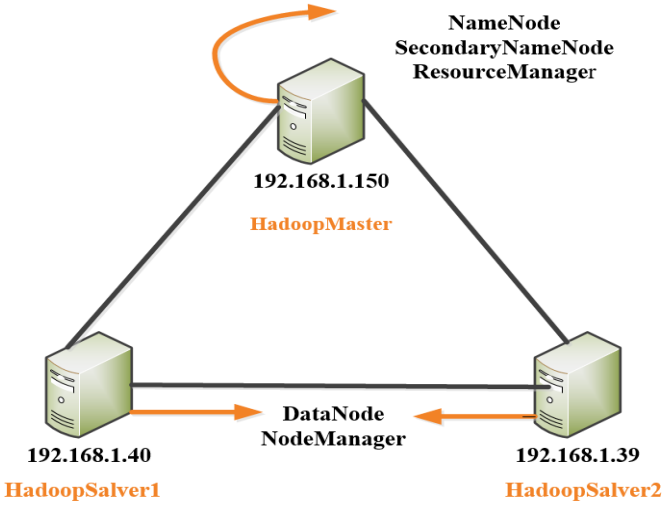
Với kết quả về tính chính xác giống nhau khi sử dụng LSTM trên môi trường máy chủ thường và Cray-XC40 và hiệu năng về thời gian học của Cray-XC40 tốt hơn hẳn máy chủ thường (giảm 10 lần, nếu tăng số lượng nodes tiếp thì thời gian tiếp tục giảm). Với thời gian như trong bảng 4.2, LSTM hoàn toàn có thể được sử dụng được vào trong nghiệp vụ dự báo KTTV. Đây chính là ý nghĩa khi sử dụng Cray PE DL cho các bài toán DL trong nghiệp vụ thực tế.

4.3. Triển khai hệ thống Big data

Triển khai hệ thống Big data phục vụ các mô hình AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm gồm:

4.3.1. Mô hình giải pháp và triển khai hạ tầng vật lý

Mô hình giải pháp triển khai hạ tầng vật lý của hệ thống Bigdata như sau:



Hình 4.10: Mô hình giải pháp triển khai hạ tầng vật lý hệ thống Big data

Hệ thống được triển khai trên 04 máy chủ có thông số kỹ thuật và kết nối mạng như sau:

Bảng 4.3: Thông số máy chủ và kết nối mạng hệ thống Big data

STT	Thông số vật lý	Địa chỉ IP	Ghi chú
I	**BigData**		
1	CPU 16xIntel(R) Xeon(R) CPU X5675 @ 3.07GHz; RAM: 16GB; HDD: 500GB	192.168.1.150	- DataName - master
2	CPU 16xIntel(R) Xeon(R) CPU X5675 @ 3.07GHz; RAM: 16GB; HDD: 500GB	192.168.1.39	- DataNode - bigdata2 (slave1)
3	CPU 16xIntel(R) Xeon(R) CPU X5675 @ 3.07GHz; RAM: 16GB; HDD: 500GB	192.168.1.40	- DataNode - bigdata1 (slave2)
II	**Server NoSQL database**		
4	CPU 56xIntel(R) Xeon(R) CPU E5-	192.168.1.71	- Cài CSDL NoSQL

	2690 v4 @ 2.60GHz/ RAM 64GB - HDD: 1TB		mongodb - mongodb-1 node
5	CPU Intel® Xeon® E5-2620 v3 2.4GHz,15M Cache,8.00GT/s QPI,Turbo,HT,6C/12T (85W) Max Mem 1866MHz Memory Capacity 4x8GB RDIMM, 2133MT/s, Dual Rank, x8 Data Width - HDD: 4TB	192.168.1.11	- Cài CSDL NoSQL mongodb - mongodb-2 node

4.3.2. Cài đặt các phần mềm, Hadoop và NoSQL

4.3.2.1. Cài đặt Java

- Install Java: sudo apt-get install openjdk-8-jdk;
- Install SSH: sudo apt install ssh;
- Cấu hình ssh cho máy master có thể ssh đến slave1, slave2.
- Cấu hình hosts trên các nodes;
- Cấu hình thư mục: Tạo thư mục trên các máy theo cấu trúc;
- Cấu hình biến môi trường: sudo nano /etc/profile.

4.3.2.2. Cài đặt Hadoop

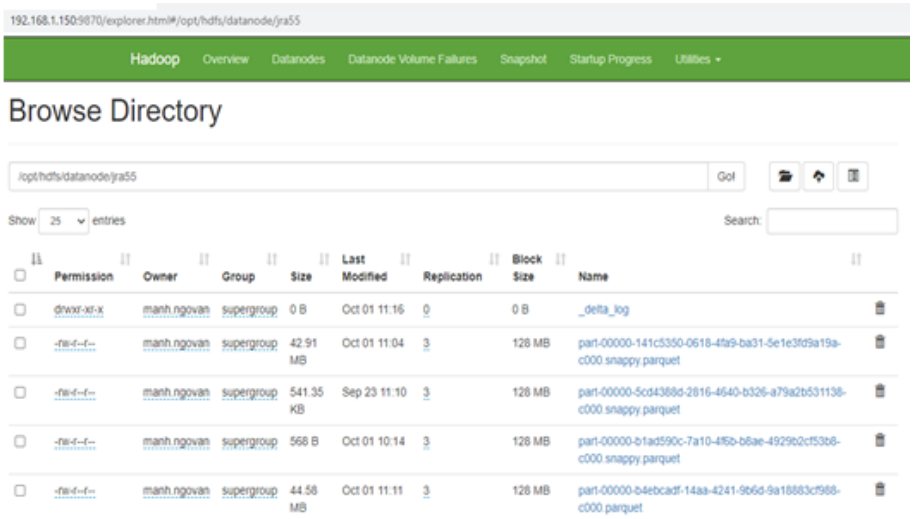
- Tải Hadoop V3.2.1 tại link: wget -c -O hadoop.tar.gz http://www-eu.apache.org/dist/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz;
- Giải nén: sudo tar -xvf hadoop.tar.gz --directory=/opt/hadoop --strip 1;
- Cấu hình biến môi trường cho các nodes: /etc/hadoop/hadoop-env.sh;
- Cấu hình Hadoop gồm: (i) Cấu hình trên Master: hdfs-site.xml; core-site.xml; yarn-site.xml; mapred-site.xml; Workers; (ii) Cấu hình trên Slave: hdfs-site.xml; core-site.xml; mapred-site.xml; yarn-site.xml.
- Start hệ thống tại Master: /opt/hadoop/sbin.

4.3.2.3. Cài đặt NoSQL

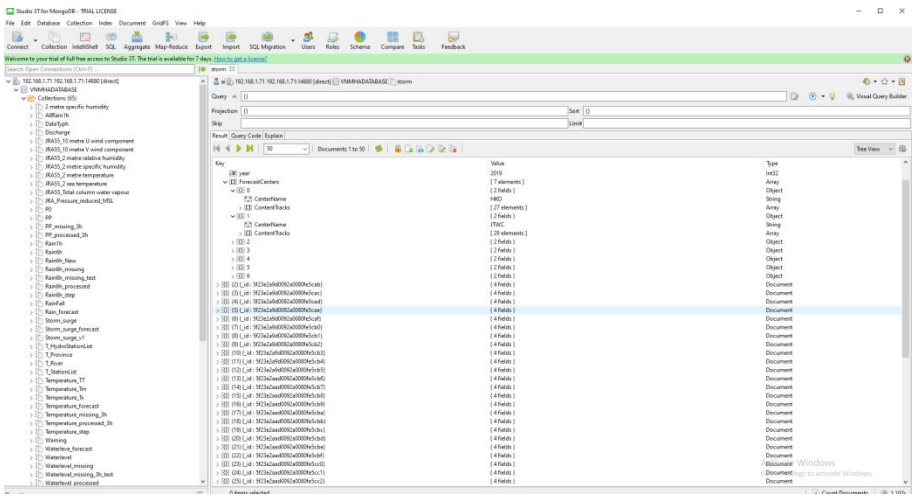
- Install SSH và cấu hình ssh bằng authentication cho 2 máy;
- Cấu hình hosts trên các nodes;
- Cài đặt mongodb trên 2 máy;
- Cấu hình file config của mongodb: sudo nano /etc/ mongod.conf.

4.3.3. Kết quả triển khai

Minh họa kết quả triển khai Big data cụ thể tại các hình sau:



Hình 4.11. Dữ liệu thô được lưu trữ trên HDFS



Hình 4.12. Dữ liệu lưu trữ trong MongoDB

4.4. Triển khai hệ thống AI để hỗ trợ dự báo bão khu vực Bắc Bộ

4.4.1. Dữ liệu cho hệ thống AI hỗ trợ dự báo bão

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo bão khu vực Bắc Bộ sử dụng bộ dữ liệu quan trắc của các trạm KTTV, cụ thể như sau:

- Dữ liệu quan trắc của 16 trạm khí tượng khu vực Bắc Bộ: Mường Lay, Lào Cai, Hà Giang. Sơn La, Cao Bằng, Điện Biên, Tuyên Quang, Hòa Bình, Nam Định, Hà Đông, Phủ Liễn, Lạng Sơn, Bãi Cháy, Tiên Yên, Móng Cái, Hồi Xuân.Tần suất dữ liệu quan trắc: 8 obs/ ngày; 3 giờ/ lần.

- Số lượng cơn bão là: **103 cơn bão**; trung bình khoảng 9 cơn/ năm.

- Thời gian dữ liệu bão là: 10 năm, từ năm 2008 - 2018.

Bảng 4.4: Thống kê số lượng các cơn bão đổ bộ vào Việt Nam năm 2018

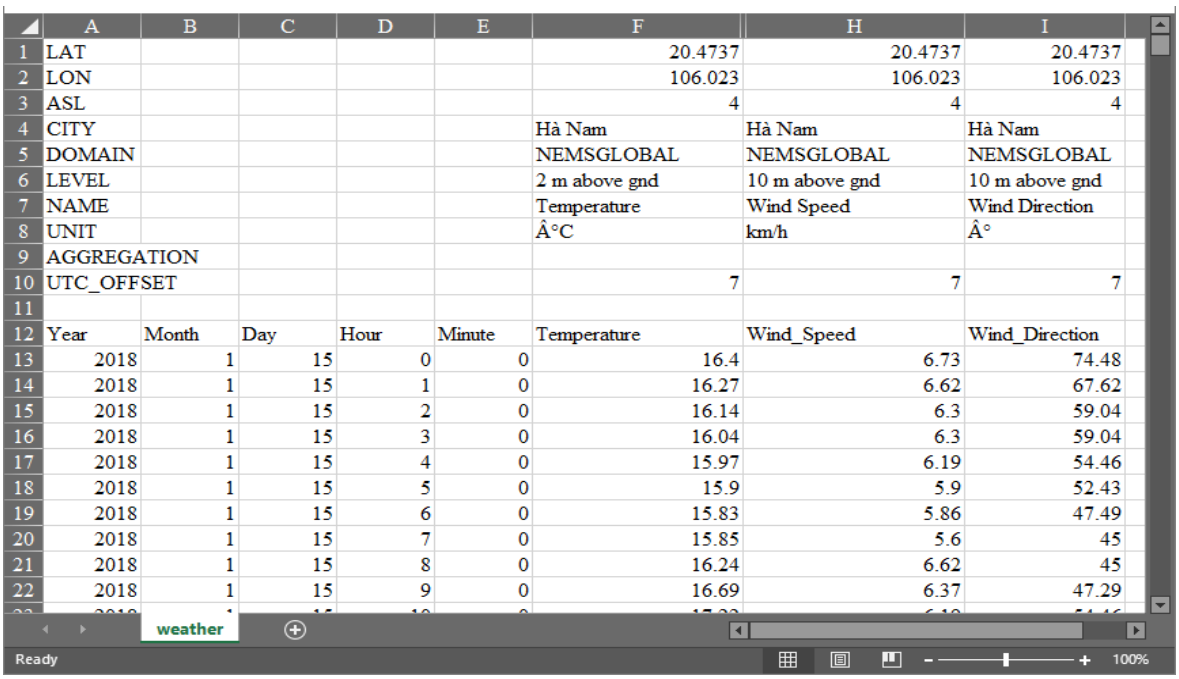
TT	Năm	Tháng	Loại	Tên bão và ATNĐ			Cường độ		Số ngày tồn tại	Phạm vi hoạt động	
				Số hiệu QT	Tên QT	Số hiệu VN	Pmin	Vmax		Nơi phát sinh	Nơi kết thúc
1	2018	1	TS	1801	BOLAVEN	Bão số 1	1002	35	1	9.8-117.4	12.1-110.0
2	2018	6	TS	1804	EWINIAR	Bão số 2	996	40	3	16.5-110.8	23.1-112.4
3	2018	7	TS	1809	SON-TINH	Bão số 3	990	45	2	19.0-120.6	19.1-104.6
4	2018	8	*STS*	1816	BEBINCA	Bão số 4	985	50	4	20.6-112.5	19.6-105.9
5	2018	9	STY	1822	MANGKHUT	Bão số 6	905	110	10	12.2-166.3	23.7-107.3
6	2018	9	TS	1823	BARIJAT	Bão số 5	998	40	2	21.3-120.8	21.7 - 109.5
7	2018	10	STY	1826	YUTU	Bão số 7	905	115	12	08.5-158.1	20.8-116.6
8	2018	11	TS	1827	TORAJI	Bão số 8	1004	35	1	10.2-111.7	10.9-110.3
9	2018	11	*STS*	1829	USAGI	Bão số 9	985	55	3	10.7-117.4	10.3-107.2

Danh sách 103 cơn bão đổ bộ vào Việt Nam trong vòng 10 năm (2008-2017) tại Báo cáo công việc số 92. Thông tin chi tiết về cơn bão cụ thể như sau:

Bảng 4.5: Thông tin chi tiết của 1 cơn bão

STT	QT Name	VN Name	Long	Lat	Pmin	Vmax	R1	R2	R3	Thời gian
1	MANGKHUT	SIEU BAO	139,8	14	915	105	30	110	200	9/11/2018 0:00
2	MANGKHUT	SIEU BAO	138,6	13,7	915	105	30	110	200	9/11/2018 0:00
3	MANGKHUT	SIEU BAO	137,5	13,9	915	105	30	110	200	9/11/2018 0:00
4	MANGKHUT	SIEU BAO	136,2	13,9	905	110	50	120	240	9/12/2018 0:00
5	MANGKHUT	SIEU BAO	135,2	14	905	110	50	150	250	9/12/2018 0:00
6	MANGKHUT	SIEU BAO	134	14,3	905	110	50	150	250	9/12/2018 0:00
7	MANGKHUT	SIEU BAO	133,2	14,4	905	110	50	150	250	9/12/2018 0:00
8	MANGKHUT	SIEU BAO	132,5	14,4	905	110	30	150	250	9/12/2018 0:00
9	MANGKHUT	SIEU BAO	131,4	14,5	905	110	30	120	250	9/13/2018 0:00
10	MANGKHUT	SIEU BAO	130,6	14,7	905	110	30	120	240	9/13/2018 0:00
11	MANGKHUT	SIEU BAO	130	14,7	905	110	30	120	240	9/13/2018 0:00
12	MANGKHUT	SIEU BAO	129,5	14,9	905	110	30	120	230	9/13/2018 0:00

Các yếu tố ảnh hưởng tới một cơn bão thường bao gồm thông số về vận tốc, hướng gió và khí áp. File weather.csv dưới đây minh họa các tham số liên quan tới một cơn bão tại trạm đo đặt tại Hà Nam, có tọa độ (20,4737; 106,023). Số liệu quan trắc theo từng giờ.



	A	B	C	D	E	F	H	I
1	LAT					20.4737	20.4737	20.4737
2	LON					106.023	106.023	106.023
3	ASL					4	4	4
4	CITY					Hà Nam	Hà Nam	Hà Nam
5	DOMAIN					NEMSGLOBAL	NEMSGLOBAL	NEMSGLOBAL
6	LEVEL					2 m above gnd	10 m above gnd	10 m above gnd
7	NAME					Temperature	Wind Speed	Wind Direction
8	UNIT					Â°C	km/h	Â°
9	AGGREGATION							
10	UTC_OFFSET					7	7	7
11								
12	Year	Month	Day	Hour	Minute	Temperature	Wind_Speed	Wind_Direction
13	2018	1	15	0	0	16.4	6.73	74.48
14	2018	1	15	1	0	16.27	6.62	67.62
15	2018	1	15	2	0	16.14	6.3	59.04
16	2018	1	15	3	0	16.04	6.3	59.04
17	2018	1	15	4	0	15.97	6.19	54.46
18	2018	1	15	5	0	15.9	5.9	52.43
19	2018	1	15	6	0	15.83	5.86	47.49
20	2018	1	15	7	0	15.85	5.6	45
21	2018	1	15	8	0	16.24	6.62	45
22	2018	1	15	9	0	16.69	6.37	47.29

Hình 4.13: Số liệu quan trắc các tham số của một cơn bão

4.4.2. Phương pháp dự báo bão

4.4.2.1. Dữ liệu đầu vào

Dữ liệu của 103 cơn bão (2010 - 2019): Tại thời điểm 00 và dữ liệu dự báo lấy: Lat, Lon, Vmax, Pressure, bán kính gió ảnh hưởng R30KT, R50KT, R64KT.

Dữ liệu dự báo bão của 6 trung tâm: JMA - Nhật Bản, NMC - Trung Quốc, JTWC - Hải quân Mỹ, KMA - Hàn Quốc, HKO - Hồng Kông, CWB - Đài Loan.

4.4.2.2. Phương pháp dự báo

Bước 1: Phân vùng dự báo:

- Phân vùng dữ liệu thành các ô lưới hình vuông cạnh R (km);
- Tự động rò tìm R để đảm bảo đủ dữ liệu của các cơn bão đưa vào tính toán;
- Phân dữ liệu theo mùa (3 tháng một, tương ứng với 4 mùa trong năm).

Bước 2: Xác định trung tâm dự báo tốt nhất:

- Xác định trung tâm có kết quả dự báo tốt nhất cho từng yếu tố vị trí (Lat, Lon), Vmax;
- Xác định trung tâm có kết quả dự báo tốt nhất cho cả 2 yếu tố vị trí và Vmax.

Bước 3: Tham khảo ý kiến chuyên gia:

- Xác định trọng số ưu tiên cho từng trung tâm;
- Tự khởi tạo các trọng số cho từng trung tâm ban đầu theo ý kiến chuyên gia;

- Cho phép chuyên gia cấu hình và chỉnh sửa trọng số.

4.4.3. Dự báo bão bằng mô hình AI với tập dữ liệu huấn luyện

4.4.3.1. Công cụ thực hiện

a) Phương pháp, công cụ:

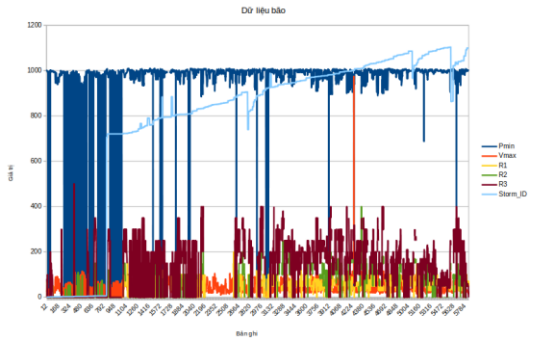
Sử dụng phương pháp hồi quy tuyến tính KNN và kỹ thuật phân lớp SVM, LSTM xử lý chuỗi thời gian để huấn luyện dự báo bão.

b) Mô tả dữ liệu đầu vào

Các trường dữ liệu sử dụng và thực hiện mô phỏng dữ liệu bão như sau.

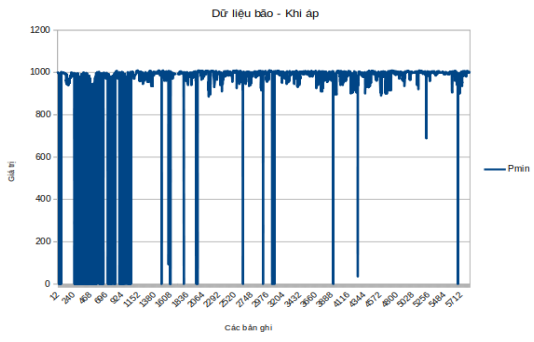
STT	Tên trường	Kiểu dữ liệu	Mô tả
1	StormVN Name	Nvarchar	Tên cơn bão(Việt Nam)
2	StormQT ID	ISODate	Mã cơn bão(quốc tế)
3	StormQT Name	Nvarchar	Tên quốc tế
4	Track Date	ISODate	Ngày
5	Obs	Real	Giờ
6	Long	Real	Kinh độ
7	Lat	Real	Vỹ độ
8	Pmin	Real	Giá trị khí áp tại tâm (Hpa)
9	Vmax	Real	Giá trị gió lớn nhất của cơn bão
10	R1	Real	Sai số tâm bão(Km)
11	R2	Real	Bán kính gió cấp 6 (Km)
12	R3	Real	Bán kính gió cấp 10 (Km)

Hình 4.14: Các trường dữ liệu bão

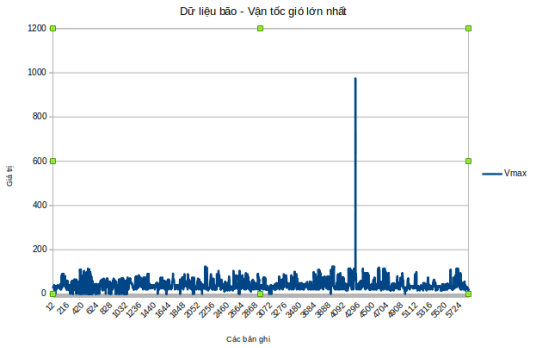


Hình 4.15: Mô phỏng dữ liệu bão

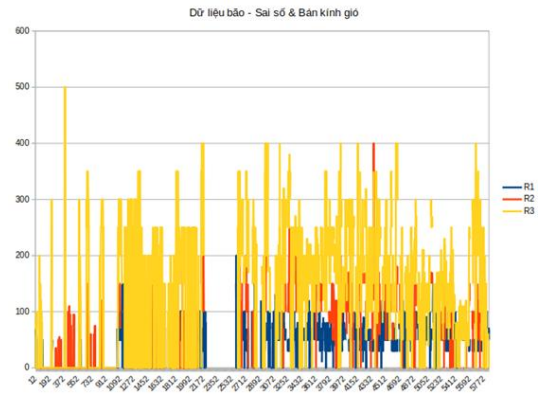
Mô tả dữ liệu báo theo từng trường như sau.



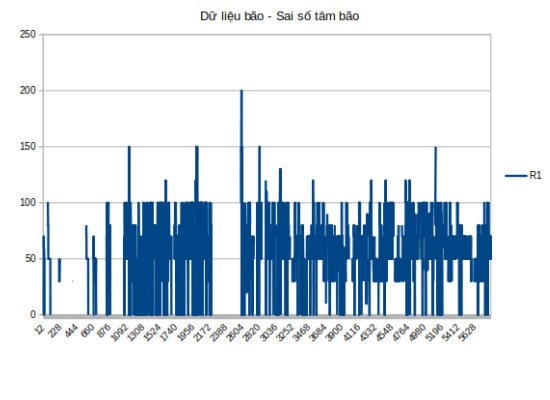
Hình 4.16: Trường dữ liệu bão - khí áp



Hình 4.17: Trường dữ liệu bão - Vmax



Hình 4.18: Trường sai số và bán kính gió



Hình 4.19: Trường dữ liệu sai số tâm bão

c) Đọc và chuẩn hóa dữ liệu bão bằng hồi quy dữ liệu

Dữ liệu bão được phân chia theo 2 tập: huấn luyện (70%), thử nghiệm (30%). Thực hiện đọc và chuẩn hóa dữ liệu bão với các đoạn code được minh họa như sau:

```
def readData(filePath=filePath):
        pandasData = pd.read_csv(filePath, sep=',')
        # data= pandasDa-
ta['Long','Lat','Pmin','Vmax','R1','R2','R3']
        # data = pandasData['Lat']
        # dataMatrix = pandasData.as_matrix()
        # dataAll = dataMatrix[:, 1:8]
        # print("pandasData: ",pandasData)
        # print('dataAll: ', dataAll)
        # pandasData = pandasData.replace(0, 'NaN')
        # pandasData = pandasData.replace(0, 'infini-
ty')
        dataGio = pandasDa-
ta[['Long','Lat','Pmin','R1','R2','R3']]
        dataGio = dataGio.as_matrix()
        labelGio = pandasData['Vmax']
```

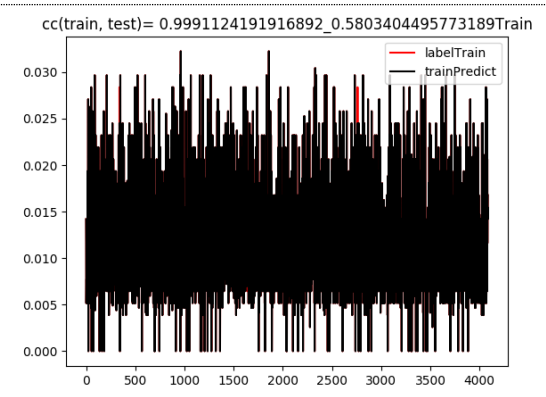
Hình 4.20: Đoạn code đọc dữ liệu bão

```
def filterData(data):
    for i in range(len(data)):
        for j in range(len(data[0])):
            if np.isnan(data[i][j]) or np.isinf(data[i][j]):
                data[i][j] = 0
    return data
def normalData(data):
    numRow = len(data)
    numCol = len(data[0])
    maxCol = []
    minCol = []
    for j in range(numCol):
        tempCol = []
        for i in range(numRow):
```

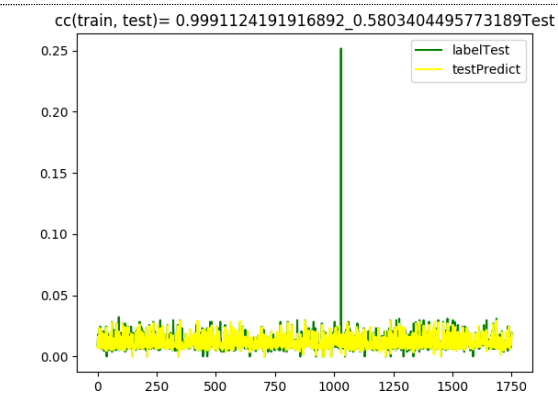
Hình 4.21: Đoạn code chuẩn hóa dữ liệu bão

4.4.3.2. Kết quả thực hiện

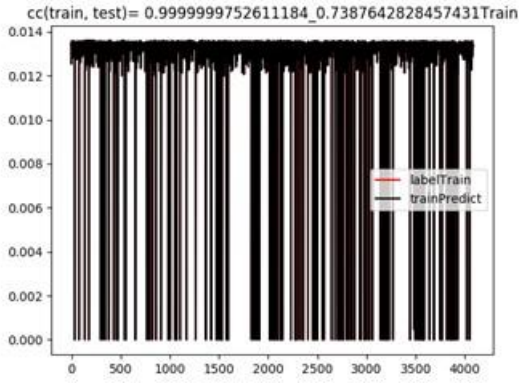
Kết quả dự báo vận tốc gió lớn nhất (Vmax) trong tâm bão với KNN hồi quy tuyến tính như sau:



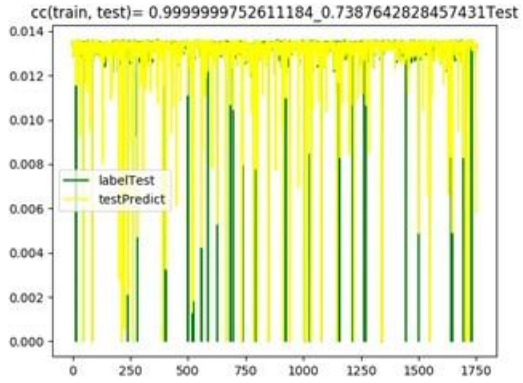
Hình 4.22: Kết quả dự báo Vmax với tập huấn luyện theo KNN hồi quy tuyến tính



Hình 4.23: Kết quả dự báo Vmax với tập thử nghiệm theo KNN hồi quy tuyến tính



Hình 4.24: Kết quả dự báo áp suất tại tâm bão với tập huấn luyện theo KNN



Hình 4.25: Kết quả dự báo áp suất tại tâm bão với tập dữ liệu kiểm tra theo KNN

Nhận xét: Triển khai sử dụng các kỹ thuật học máy có giám sát với bài toán hồi quy để huấn luyện dự báo bão. Cụ thể, huấn luyện dự báo áp suất trong tâm bão, và vận tốc gió trong tâm bão cho kết quả rất tốt. Hệ số CC trên tập thử nghiệm là lớn nhất là 70%. Có thể dùng kết quả này trong bài toán dự báo áp suất và vận tốc gió lớn nhất trong tâm bão. Kết quả dự báo áp suất lớn nhất tại tâm bão tốt hơn so với kết quả dự báo vận tốc lớn nhất tại tâm bão.

4.4.3.3. Kết quả thực hiện bằng công cụ xử lý chuỗi thời gian

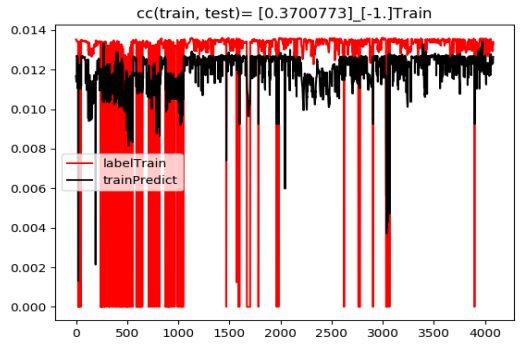
Đoạn code và kết quả huấn luyện dự báo Vmax theo LSTM như sau:

```
def doLSTM(dataTrain, dataTest, labelTrain, labelTest):
        num_output = 1
        num_dimensions = len(dataTrain[0])
        batchsize = 128
        drop_out = 0.05
        timestep = 1
        dataTrain, labelTrain = Con-
vert2Flow.convert2Flow(dataTrain, labelTrain, timestep)
        dataTest, labelTest =
……
```

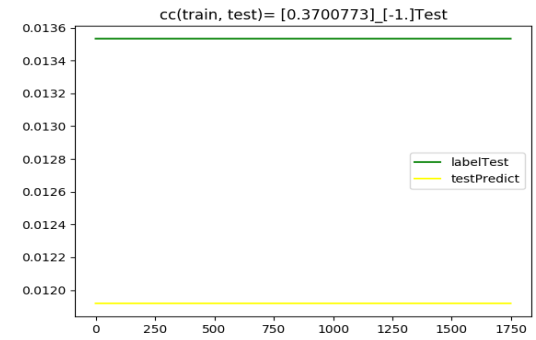
Hình 4.26: Đoạn code huấn luyện dự báo Vmax theo LSTM

```
128/4082 [.....] - ETA: 0s - loss: 3.2376e-06
1536/4082 [==========>.....] - ETA: 0s - loss:
6.8663e-06
2560/4082 [==================>.....] - ETA: 0s -
loss: 6.6760e-06
3712/4082 [==========================>...] -
ETA: 0s - loss: 6.7189e-06
4082/4082 [==============================] - 0s
45us/step - loss: 6.5920e-06
Epoch 4999/5000
……
```

Hình 4.27: Minh họa kết quả huấn luyện dự báo Vmax theo LSTM



Hình 4.28: Trực quan hóa kết quả dự báo Vmax bằng LSTM theo tập huấn luyện



Hình 4.29: Trực quan hóa kết quả dự báo Vmax bằng LSTM theo tập kiểm tra

Kết quả dự báo bão sử dụng học máy với bài toán xử lý chuỗi thời gian cho kết quả không tốt. Nguyên nhân chính do dữ liệu mỗi cơn bão cụ thể có ít bản ghi (từ 10 tới 15 bản ghi thông tin cơn bão). Do đó, hướng tiếp cận bài toán xử lý chuỗi thời gian không cho kết quả tốt. Để tăng hiệu quả của việc dự đoán, cần bổ sung thêm dữ liệu của từng cơn bão, tăng số lượng bản ghi và kết hợp phương pháp khác.

4.4.4. Trình diễn kết quả dự báo bão

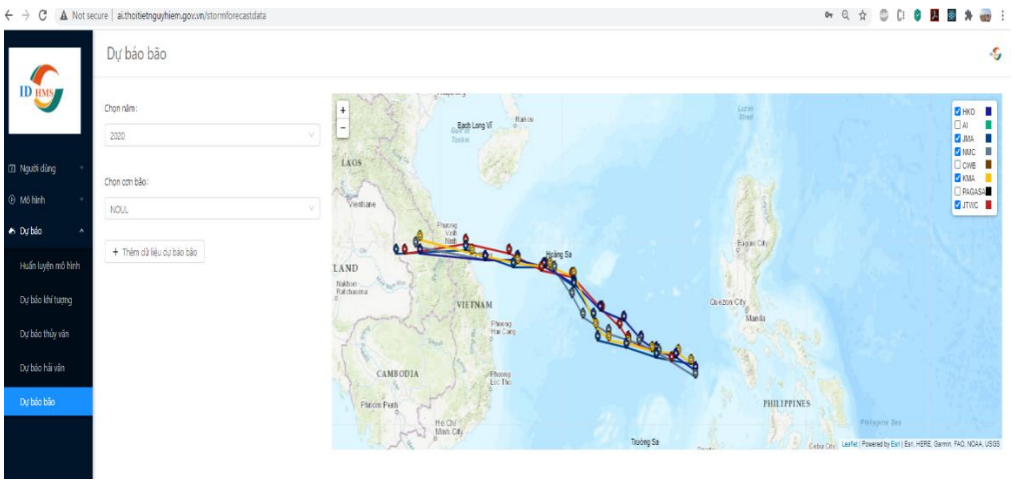
4.4.4.1. Công cụ thực hiện

Sử dụng các phương pháp, công cụ để thực hiện trình diễn dữ liệu bão gồm:

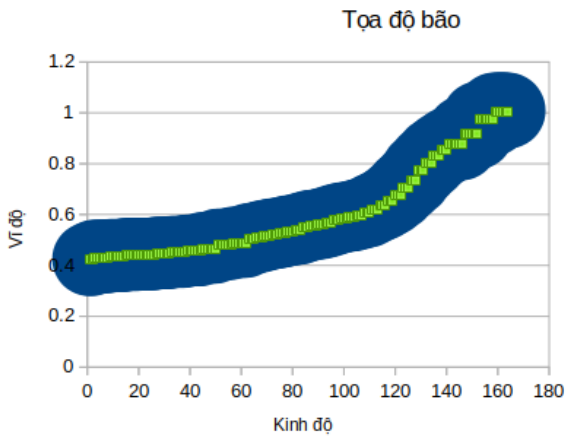
- Phương pháp sử dụng mô hình số trị;
- Phương pháp trình diễn dữ liệu sử dụng bản đồ;
- Phương pháp sử dụng Box - plot;
- Phương pháp mô hình cây;
- Phương pháp trình diễn biểu đồ cột và đường;
- Công cụ trình diễn Trie;
- Công cụ trình diễn Succinct Tries.

4.4.4.2. Kết quả triển khai

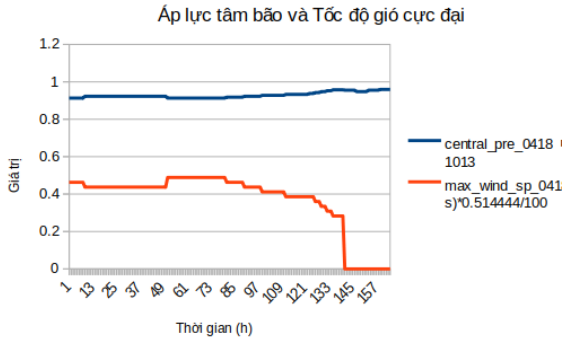
Sử dụng các phương pháp, công cụ nêu trên để trình diễn dữ liệu quan trắc và dự báo bão, kết quả cụ thể như sau:



Hình 4.30: Đường đi cơn bão NOUL từ 15/9/2020 - 18/9/2020



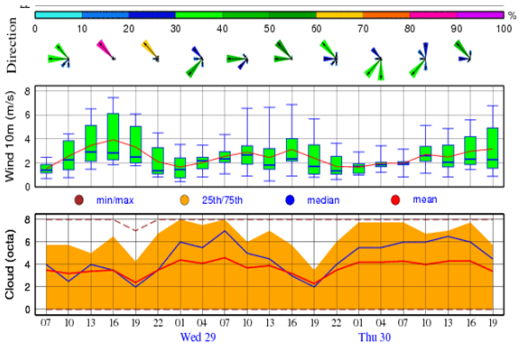
Hình 4.31: Trình diễn kết quả dự báo tọa độ tâm bão bằng biểu đồ đường



Hình 4.32: Trình diễn kết quả dự báo áp suất tâm bão và tốc độ gió cực đại của bão bằng biểu đồ đường

4.4.5. Kết luận

Kết quả triển khai sử dụng phương pháp dự báo đường đi của bão dựa trên số liệu của một số đài trên thế giới cho kết quả khá tốt. Để tăng hiệu quả của việc dự đoán, cần nghiên cứu thêm về đặc điểm đường đi của bão để kết hợp phương pháp AI dự báo chính xác hơn.



Hình 4.33: Trình diễn kết quả dự báo vận tốc, hướng gió trong bão bằng Box-plot

4.5. Triển khai hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ

4.5.1. Dữ liệu phục vụ hệ thống AI dự báo nước biển dâng

4.5.1.1. Triển khai dữ liệu quan trắc nước biển dâng do bão

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo nước biển dâng do bão khu vực Bắc Bộ và Bắc Trung Bộ cụ thể như sau:

- Dữ liệu quan trắc của các trạm khí tượng hải văn trên vùng biển Quảng Ninh đến Thanh Hóa.

- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

- Các yếu tố hải văn gồm: độ cao sóng, chu kỳ sóng, nước biển dâng do bão, mực nước thủy triều.

- Dữ liệu bão để thử nghiệm là dữ liệu nước dâng do bão đo các trạm quan trắc (Bảng 4.6).

Bảng 4.6: Danh sách một số cơn bão đo tại trạm Hòn Dấu, Hòn Ngư

STT	Tên bão	Thời điểm bắt đầu	Thời điểm kết thúc
1)	Bão số 14 (Haiyan)	05/11/2013	11/11/2013
2)	Bão số 1	13/06/2014	17/06/2014
3)	Bão Rammasun	12/07/2014	21/07/2014
4)	Bão số 1 (Kujira)	19/06/2015	25/06/2015
5)	Bão số 4 (Mujigae)	01/10/2015	05/10/2015
6)	Bão số 1 (Mirinae)	25/07/2016	28/07/2016
7)	Bão số 2 (NIDA)	28/7/2016	03/08/2016
8)	Bão số 3 (DIANMU)	15/08/2016	19/08/2016
9)	Bão số 7(Sarika)	13/10/2016	19/10/2016
10)	Bão số 8 (HAIMA)	15/10/2016	23/10/2016
11)	Bão số 6 (Hato)	20/08/2017	24/08/2017
12)	Bão Talim	10/09/2017	18/9/2017

*4.5.1.2. Các thuộc tính dữ liệu **nước biển dâng***

a) Các thuộc tính dữ liệu sử dụng

Dữ liệu nước biển dâng do bão sau khi được trích chọn các đặc trưng được mô tả như ở dưới bảng sau:

Bảng 4.7: Mô tả các đặc trưng dữ liệu nước biển dâng do bão

STT	Tên trường	Ý nghĩa	Kiểu giá trị
1	Date, Time	Ngày Giờ thu thập dữ liệu	String
2	water_level	Mực nước dâng	Float
3	wind_spe	Tốc độ gió	Float
4	wind_dir	Hướng gió	Float
5	latitude	Kinh độ của bão	Float
6	longitude	Vĩ độ của bão	Float
7	central_press	Áp lực tâm bão	Float
8	max_win_spe	Vận tốc gió tối đa	Float
9	sea_surface_level	Mực nước dâng trên bề mặt	Float
10	storm_surge	mực nước dâng do bão	Float

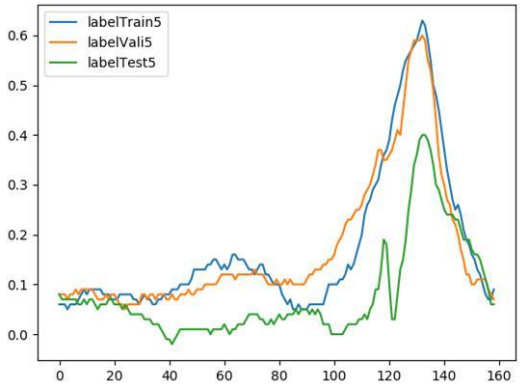
b) Phân chia dữ liệu

Phân chia dữ liệu nước biển dâng do bão thành 3 tập, tập huấn luyện, tập đánh giá huấn luyện, và tập thử nghiệm. Để dự báo nước biển dâng thời hạn đến 5h, 12h,

và 24h; sau khi gán nhãn dữ liệu nước biển dâng, số lượng bản ghi trong mỗi tập huấn luyện, thử nghiệm, và xác nhận như sau:

Bảng 4.8: Số lượng bản ghi dữ liệu nước biển dâng do bão

STT	Số lượng bản ghi các tập dữ liệu		
	Huấn luyện	Xác nhận	Thử nghiệm
5h	159	159	159
12h	152	152	152
24h	140	140	140



Hình 4.34: Mô phỏng nhãn mực nước biển dâng trên các tập huấn luyện, xác nhận và thử nghiệm

4.5.2. Xử lý dữ liệu nước biển dâng

4.5.2.1. Công cụ thực hiện

- Về xử lý dữ liệu **nước biển dâng thiếu/ mất mát**: Sử dụng phương pháp trung bình;

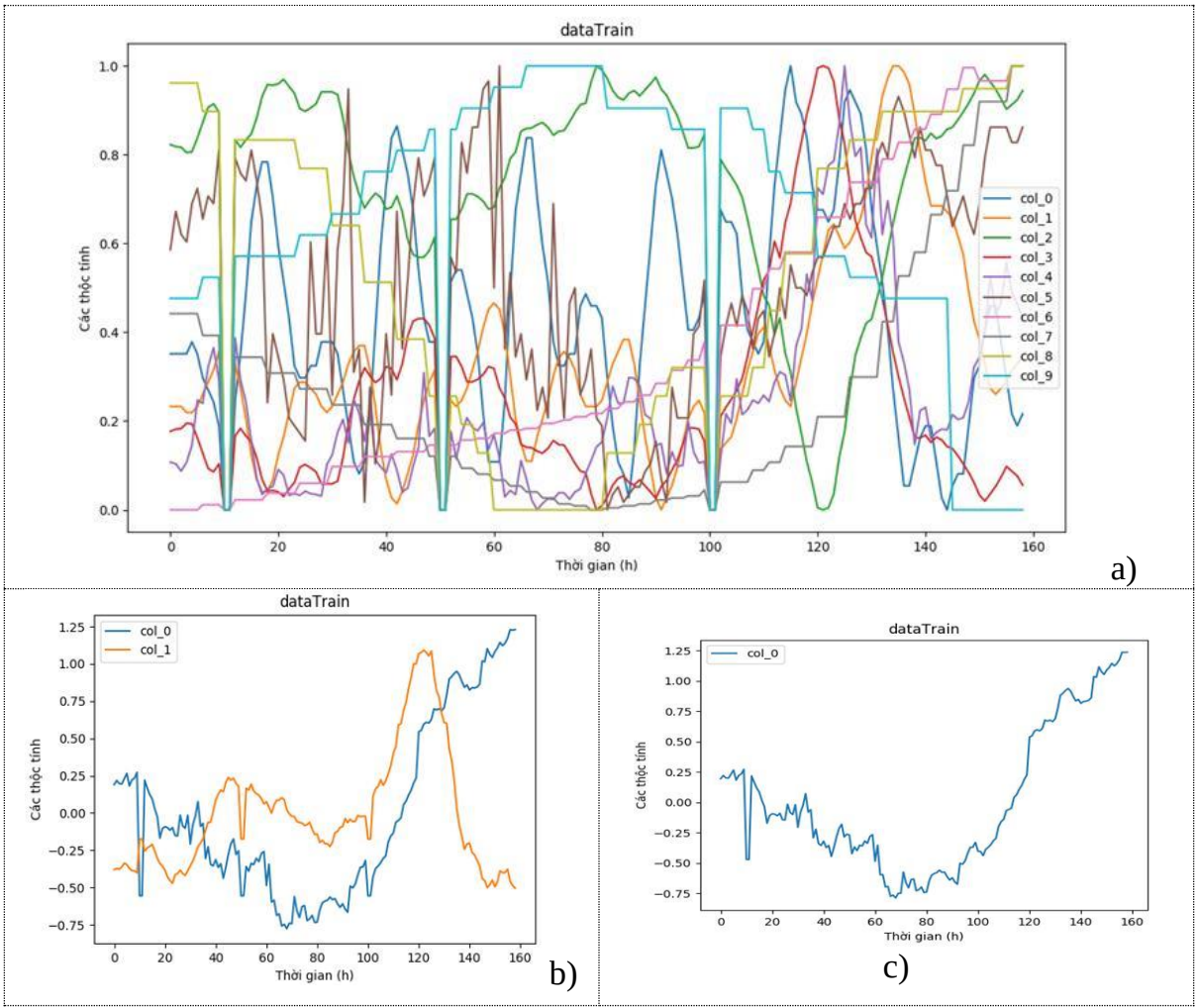
- Về xử lý dữ liệu **nước biển dâng không chắc chắn:** Sử dụng thuật toán phân cụm Gaussian Mixture Models;

- Về phát hiện và xử lý dữ liệu **nước biển dâng ngoại lai/ bất thường** (outlier): sử dụng phương pháp đồ thị, phương pháp phân tích chuỗi thời gian, phương pháp giá trị trung bình và phương pháp phân cụm.

4.5.2.2. Xử lý các dữ liệu nước biển dâng mất mát

Bước 1: Kiểm tra dữ liệu gốc với 10 thuộc tính dữ liệu, với t=10,t=11,50,100 kết quả dữ liệu trên tập huấn luyện đều cho giá trị là 0 đối với tất cả các thuộc tính. Đây là dữ liệu mất mát do không nhận được giá trị từ trạm đo (Hình 4.35a). Co giảm thuộc tính bằng PCA để kiểm tra dữ liệu ta có (Hình 4.35b). Kiểm tra với PCA bằng 1 (Hình 4.35c).

Dữ liệu mất mát do không nhận được giá trị từ trạm đo (Hình 4.35a). Co giảm thuộc tính bằng PCA để kiểm tra dữ liệu ta có (Hình 4.35b). Kiểm tra với PCA bằng 1 (Hình 4.35c).

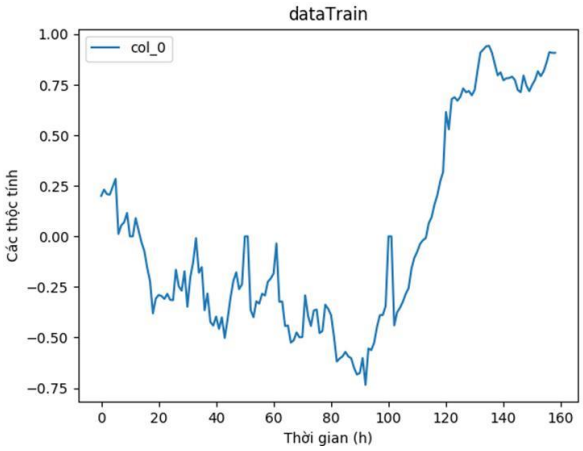


Hình 4.35: Xử lý dữ liệu nước biển dâng mất mát

Bước 2: Xử lý dữ liệu nước biển dâng mất mất bằng phương pháp trung bình. Dữ liệu sau khi xử lý missing như sau:

```
import numpy as np
from sklearn.preprocessing import Imputer
model = Imputer(missing_values=0, strategy="mean")
dataTrain5 = model.fit_transform(dataTrain5,len(dataTrain5[0]))
```

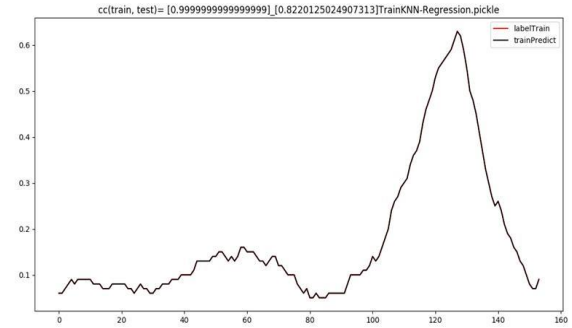
Hình 4.36: Thuật toán xử lý dữ liệu nước biển dâng mất mát



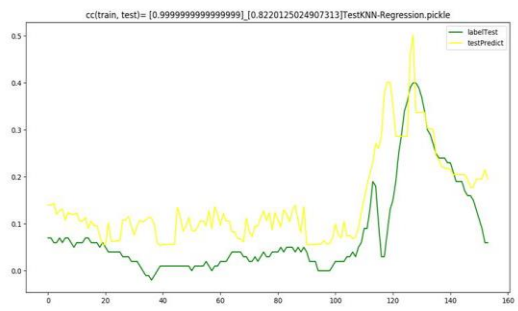
Hình 4.37: Dữ liệu nước biển dâng sau khi xử lý mất mát

*4.5.2.3. Xử lý dữ liệu **nước biển dâng** không chắc chắn*

Sử dụng thuật toán trong Gaussian Mixture Model để xử lý dữ liệu nước biển dâng không chắc chắn, kết quả xử lý như sau:



Hình 4.38: Xử lý dữ liệu nước biển dâng bất thường với tập huấn luyện

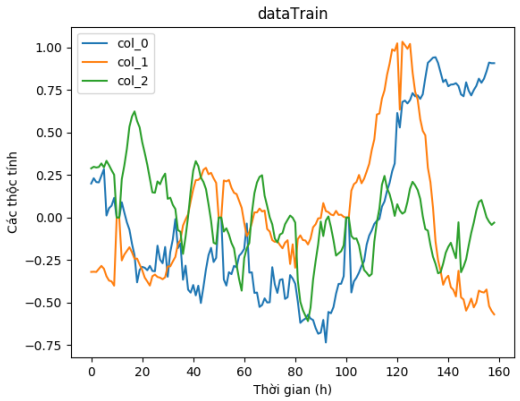


Hình 4.39: Xử lý dữ liệu nước biển dâng bất thường với tập kiểm tra

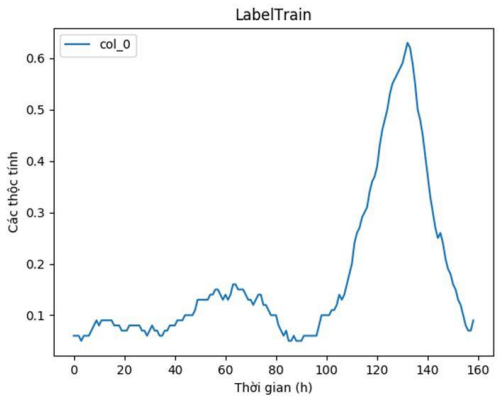
*4.5.2.4. Xử lý dữ liệu **nước biển dâng** bất thường (ngoại lai)*

a) Phát hiện dữ liệu nước biển dâng bất thường

Phát hiện dữ liệu bất thường bằng phương pháp quan sát đồ thị và nhãn của dữ liệu huấn luyện. Kết quả như sau:



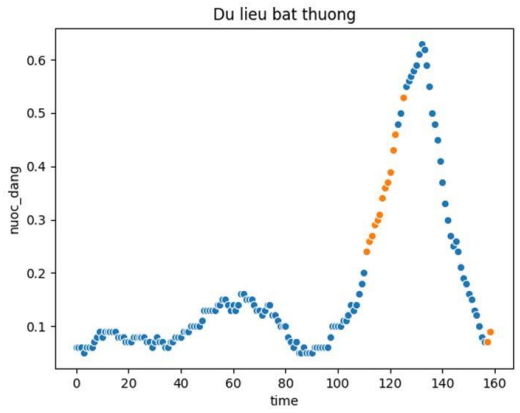
Hình 4.40: Phát hiện dữ liệu nước biển dâng bất thường bằng quan sát đồ thị



Hình 4.41: Phát hiện dữ liệu nước biển dâng bất thường bằng quan sát nhãn

Phát hiện dữ liệu bất thường bằng phương pháp thống kê, sau khi trừ đi độ lệch chuẩn, dữ liệu từ t=100 đến t=400 có bất thường. Cụ thể, dữ liệu bất thường năm ở các thời điểm outliers (114, 118, 112, 115, 116, 117, 113, 119, 125, 111, 158, 120, 121, 157, 122), dữ liệu bất thường được mô phỏng màu cam. Minh họa kết quả phát hiện dữ liệu bất thường như sau:

```
from sklearn.gaussian_process import GaussianPro-
cess
……….
g = mixture.GMM(n_components=2)
g.fit(dataTrain5)
……..
dataProb.append([logProb[i], X_name[i]])
dataProb = sorted(dataProb, key=itemgetter(0))
# print  dataProb
outliers = []
percentageCutOff = 10.
numOutlier = percentageCutOff / 100 * numSample;
numOutlier = int(numOutlier)
for i in range(numOutlier):
outliers.append(dataProb[i][1])
print('outliers',outliers)
```

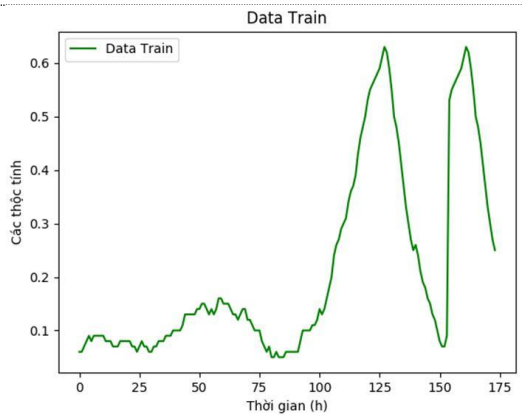


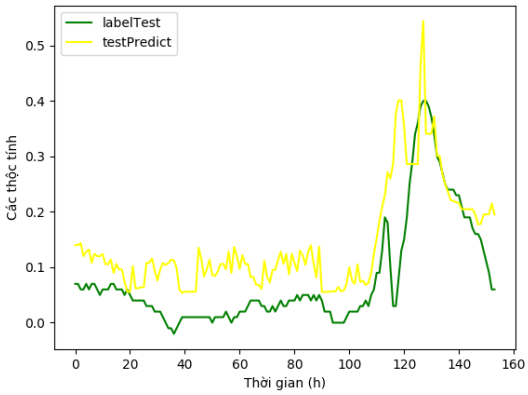
Hình 4.42: Thuật toán phát hiện dữ liệu nước biển dâng bất thường

Hình 4.43: Phát hiện dữ liệu nước biển dâng bất thường bằng phương pháp thống kê

b) Xử lý dữ liệu nước biển dâng bất thường

Kết quả xử lý dữ liệu bất thường bằng KNN-Regression như sau:





Hình 4.44: Xử lý dữ liệu dữ liệu nước biển dâng bất thường với tập huấn luyện

Hình 4.45: Xử lý dữ liệu dữ liệu nước biển dâng bất thường với tập thử nghiệm

4.5.3. Chuẩn hóa dữ liệu nước biển dâng

4.5.3.1. Công cụ thực hiện

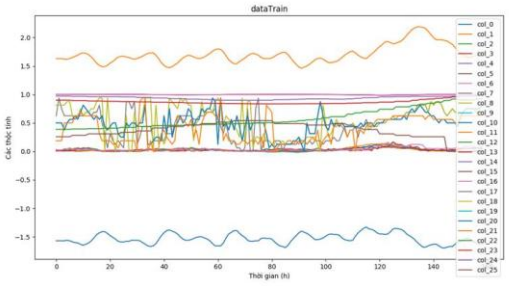
Sử dụng công cụ Pandas trong thư viện mã nguồn mở của Python để thực hiện chuẩn hóa dữ liệu nước biển dâng trong bão.

4.5.3.2. Kết quả triển khai

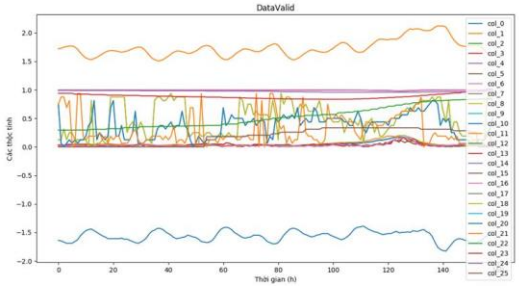
Sử dụng thuật toán để hiển thị dữ liệu nước biển dâng trong bão tại các trạm quan trắc trước khi chuẩn hóa.

```
def drawDataTrain(data, title):
    title_param = title
    plt.title(title_param)
    for j in range(len(data[0])):
        col = data[:,j]
        plt.plot(col, label= 'col_' +
str(j))
    plt.legend()
    plt.show()
```

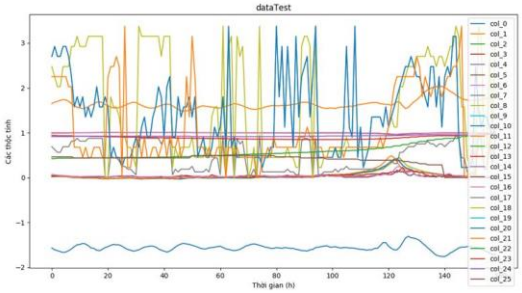
Hình 4.46: Code hiển thị dữ liệu nước biển dâng trước khi chuẩn hóa



Hình 4.47: Dữ liệu nước biển dâng trạm Hòn Dấu trước khi chuẩn hóa với tập huấn luyện

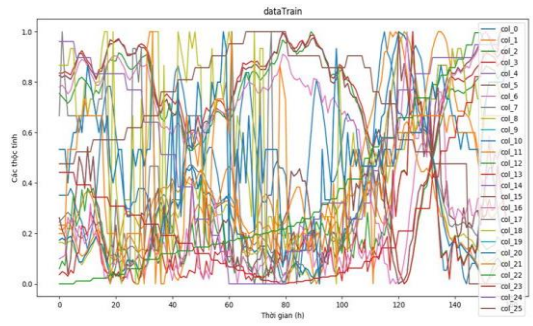


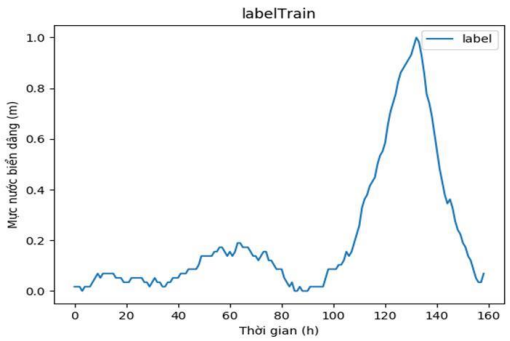
Hình 4.48: Dữ liệu nước biển dâng trạm Hòn Dấu trước chuẩn hóa với tập xác nhận



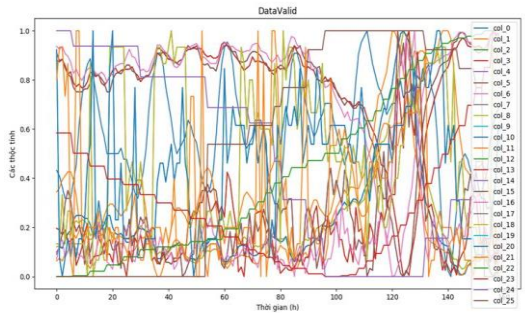
Hình 4.49: Dữ liệu nước biển dâng trạm Hòn Dấu trước chuẩn hóa với tập thử nghiệm

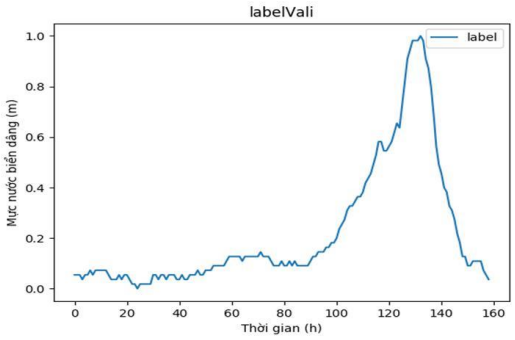
Chuẩn hóa dữ liệu nước biển dâng trong bão theo phương pháp Chuẩn hóa tiêu chuẩn (standardisation) với các **tập huấn luyện**, tập xác nhận và tập thử nghiệm, kết quả cụ thể như sau:



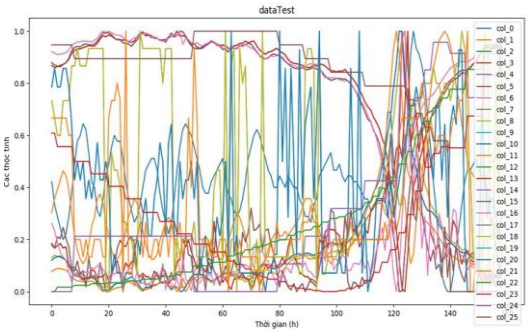


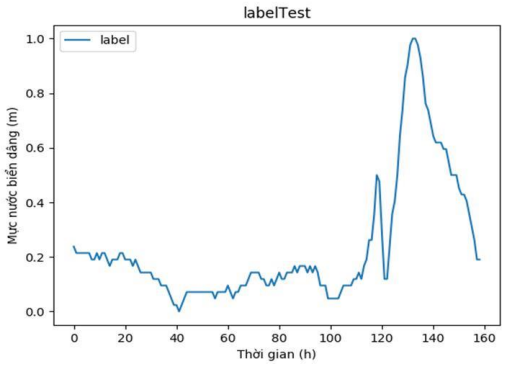
Hình 4.50: Dữ liệu nước biển dâng tại trạm Hòn Dấu sau chuẩn hóa với tập huấn luyện





Hình 4.51: Dữ liệu nước biển dâng trạm Hòn Dấu sau chuẩn hóa với tập xác nhận





Hình 4.52: Dữ liệu nước biển dâng trạm Hòn Dấu sau chuẩn hóa với tập xác nhận

Nhận xét: Thực hiện các thuật toán để chuẩn hóa tập dữ liệu thu thập. Kết quả, đã chuẩn hóa: 1013 dữ liệu khí áp trên mặt biển, 100 dữ liệu độ giảm khí áp trong bão trên mặt biển, 100 dữ liệu tốc độ gió, 360 dữ liệu hướng gió, 150 dữ liệu kinh độ bão, 50 dữ liệu vĩ độ của bão, 1013 dữ liệu áp suất tâm bão. Dữ liệu được chuẩn hóa sẽ nâng cao chất lượng của các mô hình dự báo nước biển dâng do bão.

4.5.4. Dự báo nước biển dâng bằng mô hình AI với tập dữ liệu huấn luyện

4.5.4.1. Công cụ thực hiện

- Sử dụng mô hình lập trình di truyền GP để huấn luyện dự báo nước biển dâng do bão. Các tham số khi cài đặt mô hình GP tại Bảng 4.9.

Bảng 4.9: Các tham số khi cài đặt GP

Tham số	Giá trị
Tập hàm	+, –, x, /, sin, cos, ln, √
Tập kết	Biến thuộc tính
Kích thước quần thể	1000
Thuật toán khởi tạo	Ramped half–and–half
Độ cao lớn nhất của cây	15

Số thế hệ	200
Xác suất thực hiện lai ghép	0,9
Xác suất thực hiện đột biến	0,1
Phương pháp chọn lựa	Tranh đấu kích thước bằng 3

- Ngoài ra, nhóm thực hiện đề tài còn sử dụng các mô hình học máy phổ biến khác để huấn luyện dự báo nước biển dângđể thực hiện so sánh đánh giá gồm:

+ SVR - Hồi quy vector hỗ trợ (Support Vector Regression);

+ DCT - Cây quyết định (Decision Tree);

+ kNN - Thuật toán hàng xóm K gần nhất (K Nearest Neighbors);

+ MLP - Mạng Perceptron nhiều lớp (Multi-layer Perceptron);

+ RF - Rừng ngẫu nhiên (Random Forest).

4.5.4.2. Kết quả triển khai

a) Dữ liệu huấn luyện dự báo

Dữ liệu thử nghiệm là dữ liệu nước dâng của các cơn bão đo các trạm quan trắc khu vực ven biển Bắc Bộ và Bắc Trung Bộ (trạm Hòn Dấu, Hòn Ngư) trước thời điểm nước dâng cao nhất 24h.

Dữ liệu các tham số đầu vào bao gồm:

- Tham số khí tượng: tốc độ gió (WS) (m/s), hướng gió (WD) (độ), khí áp trên mặt biển (hPa) và độ giảm khí áp trong bão trên mặt biển (DSLP) (1013 hPa).

- Tham số hải văn: mực nước bề mặt biển (SS), thủy triều (SSL).

- Tham số theo cơn bão: kinh độ (LG), vĩ độ (LT) (độ), áp suất tâm bão (CAP) (hPa) và tốc độ gió cao nhất gần tâm bão (HWS) (m/s).

Giá trị đầu ra: là giá trị nước biển dâng do bão. Các giá trị dữ liệu thu thập sẽ được chuẩn hóa theo công thức sau:

- $\eta_i^t = \tilde{\eta}_i^t$ với giá trị mực nước dâng;

- $v_{SSL} = \tilde{v}_{SSL}$ với giá trị mực nước thủy triều;

- $v_{SLP} = \tilde{v}_{SLP}/1013\ hPa$ cho khí áp trên mặt biển;

- $v_{DSLP} = \tilde{v}_{DSLP}/100\ \ hPa$ cho độ giảm khí áp trong bão trên mặt biển;

- $v_{WS} = \tilde{v}_{WS}/100\ m/s$ với tốc độ gió;

- $v_{WD} = \tilde{v}_{WD}/360\ deg$ với hướng gió;

- $v_{LG} = \tilde{v}_{LG}/150^oE$ với kinh độ của bão;
- $v_{LT} = \tilde{v}_{LT}/50^oN$ với vĩ độ của bão;
- $v_{CAP} = \tilde{v}_{CAP}/1013\ hPa$ với áp suất tâm bão;
- $v_{HWS} = \tilde{v}_{HWS}$ với tốc độ gió lớn nhất gần tâm bão.

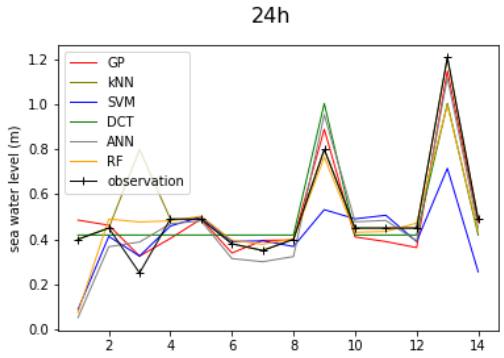
Trong đó dấu (~) bên phải của các phương trình thể hiện giá trị gốc của các tham số.

b) Kết quả huấn luyện dự báo nước biển dâng

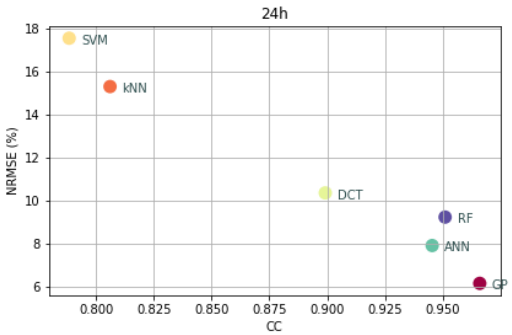
Hàm đánh giá độ tốt của mỗi cá thể sử dụng hàm NRMSE (Root Mean Square Error). Thực hiện chạy GP 30 lần độc lập, mỗi lần chạy với giá trị khởi tạo khác nhau, sau mỗi lần chạy ta sẽ nhận được một lời giải tốt nhất. Sau 30 lần chạy, sẽ có 30 lời giải tương ứng, sắp xếp các lời giải đó theo thứ tự tăng dần giá trị độ phù hợp, lựa chọn lời giải trung vị (median) của dãy đó dùng làm mô hình cuối cùng.

Sử dụng hai độ đo là NRMSE (Normal Root Mean Squared Error) là RMSE chuẩn hóa tính theo phần trăm; và CC (Correlation Coefficient) là hệ số tương quan để so sánh hiệu suất của GP với các phương pháp khác.

So sánh kết quả huấn luyên dự báo lũ giữa mô hình GP và các mô hình kNN, SVM, DCT, ANN, RF như các hình dưới đây.



Hình 4.53: So sánh giá trị dự báo của các mô hình và dữ liệu quan trắc thực tế tại trạm Hòn Dấu của 12 cơn bão



Hình 4.54: So sánh giá trị NRMSE và CC của các mô hình dự báo với dữ liệu tại Hòn Dấu

Hình 4.53: So sánh các giá trị dự báo của 6 mô hình dự báo và giá trị quan trắc mực nước thực tế, kết quả cho thấy mô hình kết quả của GP (màu xanh) bám sát nhất với đường màu đen (giá trị quan trắc thực tế) đặc biệt tại các điểm cao. Điều đó cho thấy rằng mô hình dự báo GP có khả năng dự đoán gần đúng nhất các điểm dữ liệu thực tế.

Hình 4.54. Giá trị NRMSE của 6 phương pháp dự báo nằm trong khoảng từ 6% đến 18%. Còn giá trị CC nằm trong khoảng từ 0,75 đến 0,97. Phương pháp GP vừa cho kết quả giá trị NRMSE nhỏ (sai số ít nhất) và CC lớn nhất (gần gũi với giá trị thực nhất kể cả các điểm cao) trong số 6 phương pháp. Như vậy, trên tập dữ liệu thực đo của 12 cơn bão khác nhau, GP cho kết quả dự báo tốt nhất so với các phương pháp còn lại. Kết quả khẳng định độ tin cậy của GP vượt trội so với các mô hình dự báo khác.

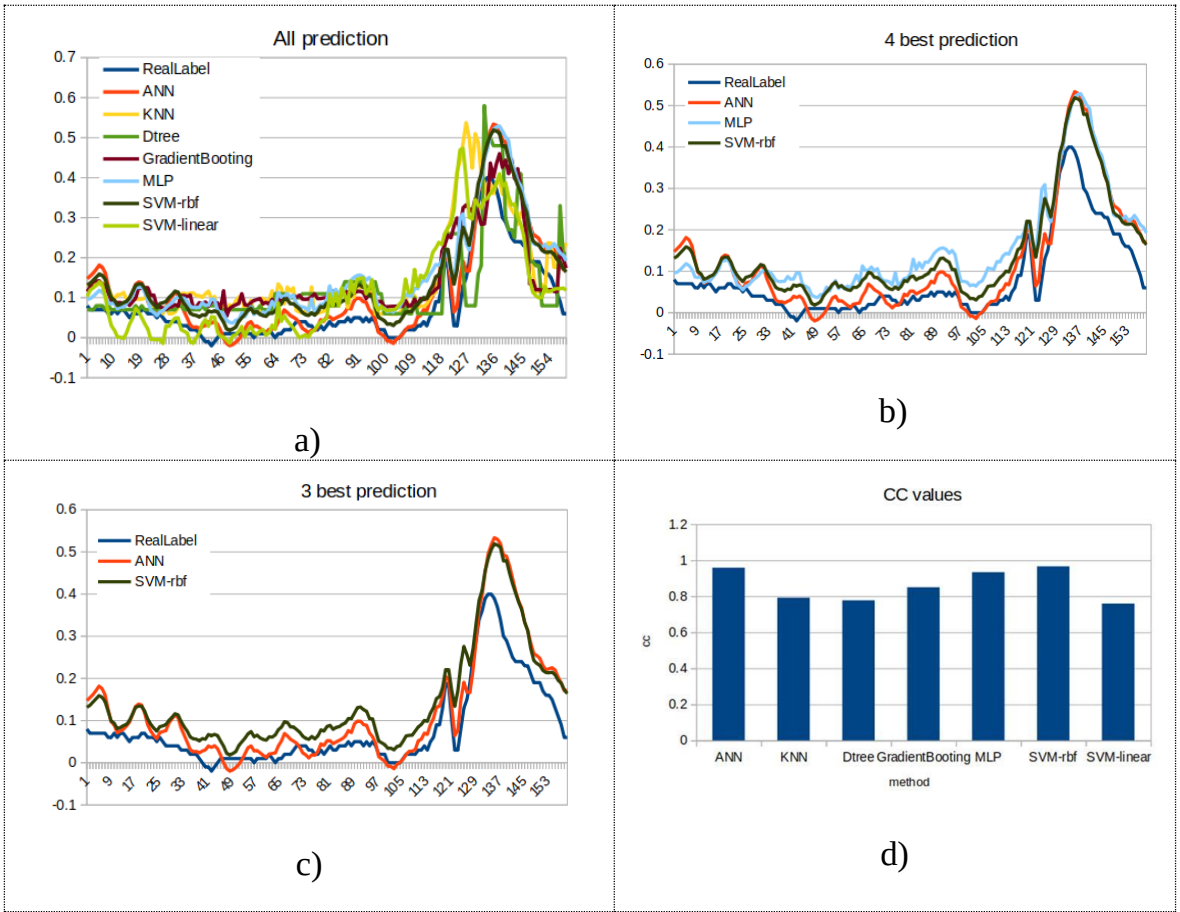
4.5.5. Trình diễn kết quả dự báo nước biển dâng

4.5.5.1. Công cụ thực hiện

Sử dụng phương pháp biểu diễn cây Python trong ngôn ngữ lập trình Python để trình diễn kết quả dự báo nước biển dâng do bão(đồ thị graph, biểu đồ,…).

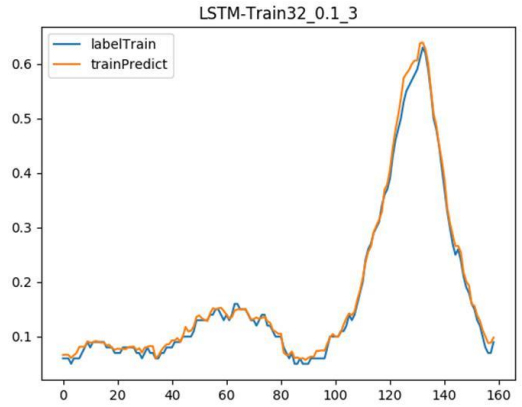
4.5.5.2. Kết quả triển khai

Trình diễn kết quả dự báo nước dâng do bão bằng các mô hình dự báo khác nhau như sau:

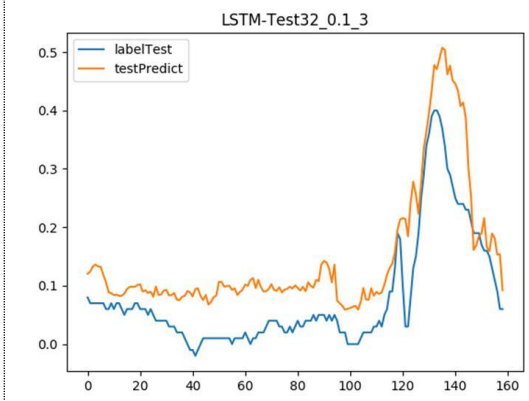


Hình 4.55: Trình diễn kết quả dự báo nước biển dâng do bão bằng các mô hình Tất cả các mô hình dự báo ; b) 4 Mô hình có kết quả dự báo tốt nhất, c) 3 Mô hình dự báo tốt nhất, d) trình diễn tương quan dự bảo của các mô hình

Trình diễn kết quả dự báo nước dâng do bão bằng mô hình LSTM như sau:



a)



b)

Hình 4.56: Trình diễn kết quả dự báo nước biển dâng do bão bằng LSTM
a) Với tập huấn luyên ; b) Với tập thử nghiệm

4.5.6. Kết luận

Kết quả sử dụng GP để dự báo nước biển dâng do bão cho thấy GP vượt trội hơn về hiệu năng so với các phương pháp dự báo khác (MLP, SVM, kNN, DCT, RF). Vì vậy, trong tương lai, nhóm nghiên cứu sẽ tiếp tục cải tiến GP để thu được kết quả dự báo tốt hơn nữa. Ngoài ra,cần tiếp tục sử dụng GP để thử nghiệm áp dụng cho dữ liệu tại các trạm quan trắc khác, với các cơn bão khác và với thời gian dự báo trước ngắn hơn (12h, 5h) để có được kết quả dự báo phù hợp với yêu cầu thực tế.

4.6. Triển khai hệ thống AI để hỗ trợ dự báo mưa lớn diện rộng khu vực Bắc Bộ

4.6.1. Dữ liệu cho hệ thống AI hỗ trợ dự báo mưa

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo mưa lớn diện rộng khu vực Bắc Bộ sử dụng dữ liệu đo mưa của 26 trạm KTTV, cụ thể như sau:

1) 48/86 - KY ANH 18. 100000381469727, 106.26667022705078
2) 48800 - MUONG LAY 22.066667556762695, 103.1500015258789
3) 48803 - LAO CAI 22.5, 103.96666717529297
4) 48805 - HA GIANG 22.816667556762695, 104.96666717529297
5) 48806 - SON LA 21.33333396911621, 103.9000015258789
6) 48808 - CAO BANG 22.66666603088379, 106.25
7) 48811 - DIEN BIEN 21.366666793823242, 103.0
8) 48812 - TUYEN QUANG 21.816667556762695, 105.21666717529297
9) 48818 - HOA BINH 20.816667556762695, 105.33333587646484
10) 48823 - NAM DINH 20.433332443237305, 106.1500015258789

11) 48825 - HA DONG 20.96666717529297, 105.75
12) 48826 - PHU LIEN 20.799999237060547, 106.63333129882812
13) 48830 - LANG SON 21.83333396911621, 106.76667022705078
14) 48833 - BAI CHAY 20.96666717529297, 107.06666564941406
15) 48837 - TIEN YEN 21.33333396911621, 107.4000015258789
16) 48838 - MONG CAI 21.516666412353516, 107.96666717529297
17) 48839 - BACH LONG VY 20.133333206176758, 107.71666717529297
18) 48840 - THANH HOA 19.75, 105.78333282470703
19) 48842 - HOI XUAN 20.366666793823242, 105.08333587646484
20) 48845 - VINH 18.66666603088379, 105.68333435058594
21) 48846 - HA TINH 18.350000381469727, 105.9000015258789
22) 48848 - DONG HOI 17.46666717529297, 106.61666870117188
23) 48852 - HUE 16.433332443237305, 107.58333587646484
24) 48866 - PLEIKU 13.966666221618652, 108.01667022705078
25) 48875 - BUON MA THUAT 12.666666984558105, 108.05000305175781
26) 48887 - PHAN THIET 10.933333396911621, 108.0999984741211

- Tần suất dữ liệu đo mưa: 4 obs/ ngày; 6 giờ/ lần.
- Thời gian dữ liệu mưa là: 10 năm, từ năm 2008 - 2018.

4.6.2. Xử lý dữ liệu mưa

4.6.2.1. Công cụ thực hiện

Sử dụng thuật toán *fillna*. *fillna* sau để phát hiện dữ liệu mưa mất mát. Cụ thể như sau:

```
import pandas as pd

# Set display options
pd.set_option('display.max_columns', 10)

# read data from csv file
dataset = pd.read_csv('forGMMMS.csv', header=0)

# print first 5 rows of data set
print(dataset.head())
print(dataset.describe())
dataset[dataset.date.isnull()]
```

Hình 4.57: Đoạn code xác định dữ liệu mưa mất mát

	date	wp1	wp2	wp3	wp4	wp5	wp6	wp7
3	NaN	0.060	0.085	0.109	0.022	0.010	0.013	0.0
17	NaN	0.000	0.016	0.000	0.000	0.056	0.000	0.0
18	NaN	0.005	0.032	0.000	0.000	0.030	0.000	0.0
19	NaN	0.000	0.037	0.000	0.017	0.020	0.013	0.0
20	NaN	0.000	0.048	0.000	0.050	0.000	0.013	0.0
23	NaN	0.015	0.127	0.000	0.000	0.010	0.013	0.0
48	NaN	0.201	0.386	0.415	0.000	0.000	0.026	0.0

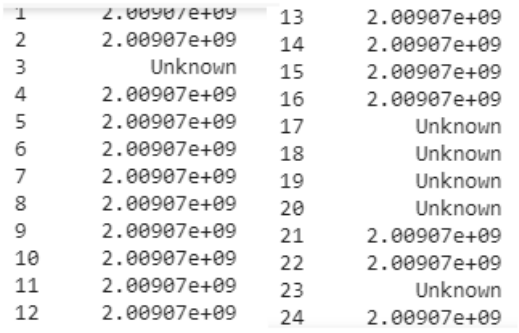
Hình 4.58: Kết quả phát hiện dữ liệu mưa mất mát

4.6.2.2. Kết quả triển khai

Kết quả triển khai sử dụng thuật toán *fillna*. *fillna* để xử lý dữ liệu mưa mất mát, thay thế giá trị mất mát “NaN” bằng"Unknown"như sau:

	date	wp1	wp2	wp3	wp4	wp5	wp6	wp7
3	NaN	0.060	0.085	0.109	0.022	0.010	0.013	0.0
17	NaN	0.000	0.016	0.000	0.000	0.056	0.000	0.0
18	NaN	0.005	0.032	0.000	0.000	0.030	0.000	0.0
19	NaN	0.000	0.037	0.000	0.017	0.020	0.013	0.0
20	NaN	0.000	0.048	0.000	0.050	0.000	0.013	0.0
23	NaN	0.015	0.127	0.000	0.000	0.010	0.013	0.0
48	NaN	0.201	0.386	0.415	0.000	0.000	0.026	0.0

Hình 4.59: Xác định dữ liệu mưa mất mát



Hình 4.60: Thay thế dữ liệu mưa mất mát "NaN" bằng "Unknown"

Sử dụng thư viện Sklearn, tạo hàm *score_dataset(X_train, X_test, y_train, y_test)* để so sánh chất lượng của cách tiếp cận khác nhau khi xử lý điểm dữ liệu mưa mất mát. Xác định điểm từ cột bỏ đi do có giá trị dữ liệu mưa bị thiếu, thêm các cột có giá trị thiếu bằng giá trị dữ liệu mưa dự đoán mới. Kết quả như sau:

```
from sklearn.impute import SimpleImputer
my.imputer = SimpleImputer()
imputed.X.train = my.imputer.fit_transform(X_tram) imput-
ed.X.test = my.imputer.transform(X_test) prIntCMean Absolute
Error from Imputation:')
pnnt(score_dataset(imputed.X.train, imputed.X.test, y.train,
y.test))
```

Hình 4.61: Đoạn code xác định điểm dữ liệu mưa mất mát

```
imputed.X.train.plus = X_train.copy() imputed.X.test.plus
■X_test.copy()
cols.with.missing = (col for col in X.train.columns
                     if X_train[col].isnull().any())
for col in cols.with.missing:
    imputed_X_train_plus[col * '.was.missing[1]]
    ■iraputed_X_train_plus[col].isnull() imput-
    ed_X_test_plus[col + '.was.missing[1]] =
    i»puted_X_test_plus[col].isnull()
* Imputation
my.imputer ■SimpleIoputer()
imputed.X.train.plus =
my.imputer.fit_transform(imputed.X.train.plus) imput-
ed.X.test.plus * my.imputer.transform(imputed.X.test.plus)
```

Hình 4.62: Đoạn code bổ sung dữ liệu thiếu bằng dữ liệu mưa dự đoán mới

Nhận xét: Việc xử lý các giá trị mưa mất mát cho phép chúng ta cải thiện mô hình AI dự báo mưa so với việc bỏ các cột đó.

4.6.3. Chuẩn hóa dữ liệu mưa

4.6.3.1. Công cụ thực hiện

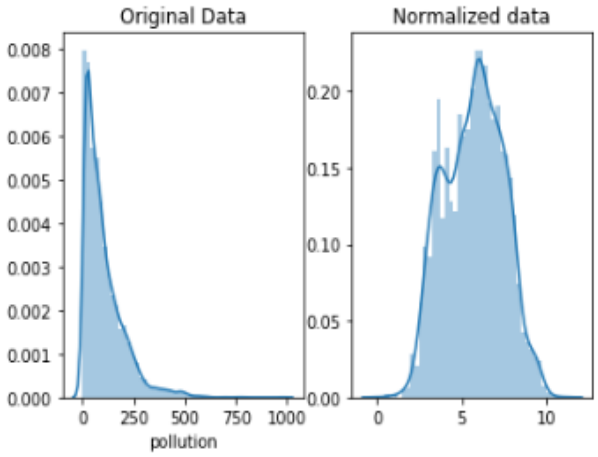
Sử dụng công cụ Pandas trong thư viện mã nguồn mở của Python để thực hiện chuẩn hóa dữ liệu lượng mưa trong mô hình AI dự báo mưa lớn diện rộng.

4.6.3.2. Kết quả triển khai

Kết quả chuẩn hóa dữ liệu lượng mưa như sau:

```
# lấy chỉ số của tất cả các giá trị thể hiện lượng mưa dương (Box-Cox chỉ nhận các giá trị dương)
index_of_positive_pledges = kickstarters_2017.pollution > 0
# lấy các giá trị dương (sử dụng chỉ số của chúng)
positive_pledges = kickstarters_2017.pollution.loc[index_of_positive_pledges]
# chuẩn hóa (w/ Box-Cox)
normalized_pledges = stats.boxcox(positive_pledges)[0]
# vẽ hình so sánh
fig, ax=plt.subplots(1,2)
sns.distplot(positive_pledges, ax=ax[0])
ax[0].set_title("Original Data")
sns.distplot(normalized_pledges, ax=ax[1])
ax[1].set_title("Normalized data")
```

Hình 4.63: Thuật toán chuẩn hóa dữ liệu



Hình 4.64: Kết quả chuẩn hóa dữ liệu

4.6.4. Trực quan hóa dữ liệu mưa

4.6.4.1. Công cụ thực hiện

- Sử dụng các công cụ biểu đồ cột (Bar/Column chart); Biểu đồ đường (Line Graph); Biểu đồ tròn, biểu đồ cột chồng (Pie chart, Stacked Column/bar chart); Biểu đồ phân tán (Scatter plot); Biểu đồ hộp và râu (Box- plot) để trực quan hóa dữ liệu mưa.

- Thông tin dữ liệu mưa lớn diện rộng cần trực quan là: 441 trận mưa lớn;

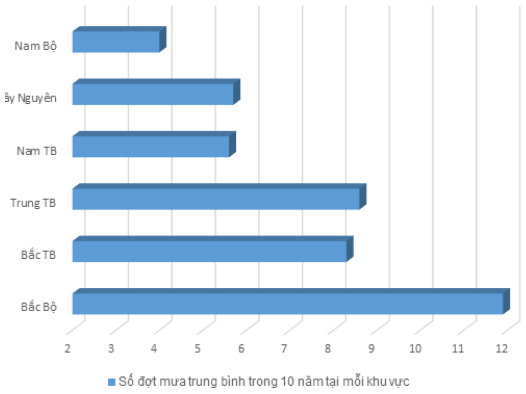
- Thời gian dữ liệu mưa lớn là: 10 năm, từ năm 2008 - 2017;

4.6.4.2. Kết quả triển khai

Kết quả triển khai trực quan hóa mưa lớn trung bình theo khu vực bằng biểu đồ cột như sau:

Bảng 4.10: Thống kê số lượng các trận mưa lớn trong 10 năm (2008-2017)

Khu vực	Số trận mưa/ tháng trong 10 năm									
	3	4	5	6	7	8	9	10	11	12
Bắc Bộ	2	6	12	12	31	19	22	9	4	2
Bắc TB	2	2	2	7	15	9	23	17	6	0
Trung TB	2	1	2	0	3	2	17	28	24	7
Nam TB	0	2	2	0	2	2	6	16	20	6
Tây Nguyên	0	2	5	8	9	3	12	10	7	1
Nam Bộ	0	1	2	9	7	2	11	6	2	0
Cộng	6	14	25	36	67	37	91	86	63	16



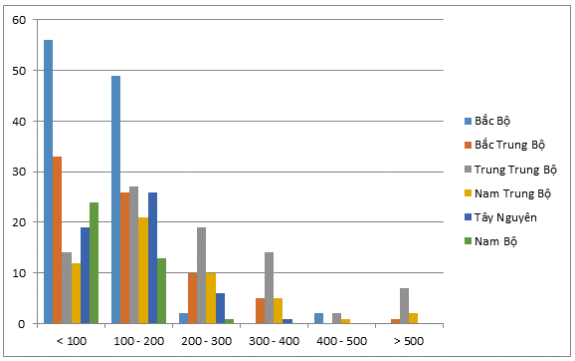
Hình 4.65: Trực quan hóa số đợt mưa lớn trung bình từ năm 2008-2017 tại các

	khu vực bằng biểu đồ cột

Kết quả triển khai trực quan hóa mưa lớn dựa vào việc thống kê tổng lượng mưa qua các đợt mưa lớn diện rộng trên từng lưu vực có thể chia lượng mưa trong các đợt mưa lớn theo các ngưỡng nhỏ hơn 100mm, 100-200mm, 200-300mm, 300-400mm, 400-500m và ngưỡng lớn hơn 500mm như sau:

Bảng 4.11: Tổng số đợt mưa lớn theo ngưỡng mưa trong 10 năm (2008-2017)

Khu vực	Ngưỡng lượng mưa					
	< 100	100 - 200	200 - 300	300 - 400	400 - 500	> 500
Bắc Bộ	56	49	2	0	2	0
Bắc TB	33	26	10	5	0	1
Trung TB	14	27	19	14	2	7
Nam TB	12	21	10	5	1	2
Tây Nguyên	19	26	6	1	0	0
Nam Bộ	24	13	1	0	0	0

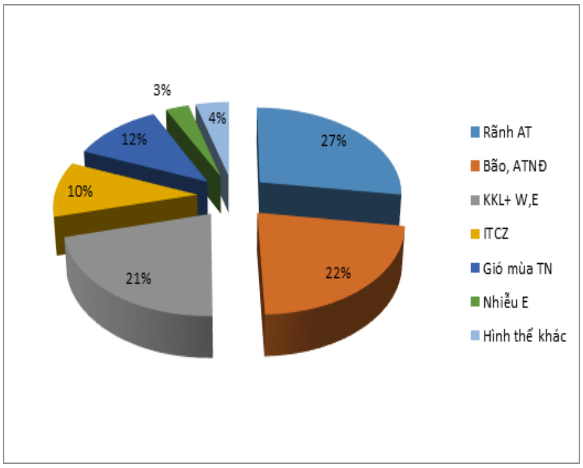


Hình 4.66: Trực quan hóa tổng số đợt có lượng mưa theo ngưỡng xảy ra trên khu vực dự báo bằng biểu đồ cột

Kết quả triển khai trực quan hóa đánh giá phân loại hình thế synop đặc trưng gây mưa lớn diện rộng thường đánh giá theo nguyên nhân và vai trò chủ đạo của các hệ thống synop như sau:

Bảng 4.12: Các hình thế synop gây mưa lớn diện rộng trong 10 năm (2008-2017)

Năm	Hình thế thời tiết gây mưa lớn						
	Rãnh AT	Bão, ATNĐ	KKL	ITCZ	Gió mùa TN	Nhiễu E	Hình thế khác
2008	11	4	10	1	1	0	1
2009	5	8	1	2	1	2	0
2010	5	2	8	2	0	0	1
2011	2	5	7	3	1	0	0
2012	5	4	3	3	0	0	2
2013	9	8	2	3	7	0	2
2014	8	4	6	3	5	3	2
2015	9	3	7	0	4	0	0
2016	4	6	4	4	3	0	1
2017	6	8	4	3	5	0	1



Hình 4.67: Trực quan hóa các hình thế synop gây mưa lớn diện rộng (2008-2017) bằng biểu đồ tỷ lệ

Nhận xét:

Kết quả triển khai trực quan hóa dữ liệu mưa lớn diện rộng trong 10 năm (2008 - 2017) bao gồm:

- Dữ liệu thống kê về số đợt mưa diện rộng;
- Dữ liệu thống kê về phân bổ các đợt mưa lớn theo tháng các khu vực;
- Dữ liệu thống kê về tổng lượng mưa theo đợt xảy ra tại các khu vực dự báo;
- Dữ liệu thống kê về các hình thế Synop chính gây mưa lớn diện rộng;
- Dữ liệu thống kê về số ngày kéo dài của các đợt mưa lớn riện rộng.

Dựa vào trực quan hóa các dữ liệu thống kê trên, có thể thấy những năm gần đây số đợt mưa lớn có xu hướng tăng dần, trong đó năm 2013 và 2014 là 2 năm có số đợt mưa lớn diện rộng nhiều nhất (31 đợt), cao hơn so với gía trị TBNN 7 đợt. Năm 2012 là năm có số đợt mưa lớn ít nhất (17 đợt), thấp hơn TBNN 7 đợt và thấp hơn so với năm có đợt mưa lớn cao nhất là 14 đợt. Nhiều thông tin quan trọng khác chúng ta có thể dễ dàng có được thông qua các biểu đồ trực quan.

4.6.5. Dự báo mưa lớn bằng mô hình AI với tập dữ liệu huấn luyện

4.6.5.1. Công cụ thực hiện

Sử dụng phương pháp thuật toán Gradient Boosting (GB) để huấn luyện dự báo mưa lớn diện rộng.

4.6.5.2. Kết quả triển khai

a) Dữ liệu đầu vào

Thời gian lấy dữ liệu từ 1/1/2014 đến 31/12/2019, cụ thể danh sách các trạm quan trắc đã trình bày tại mục 4.1.1, tần suất quan trắc 6h một lần.

b) Triển khai thuật toán Gradient Boosting

Bước 1: Tối ưu mô hình phân loại GBM sử dụng tối ưu bayes, ở đây do thời gian tối ưu rất lâu nên chọn một bộ tham số tối ưu.

Bước 2: Chia ngẫu nhiên dữ liệu huấn luyện thành các phần sao cho tỉ lệ có/không có mưa lớn diện rộng = 1 trong đó lớp số ít có thể được lặp lại, sau đó kết hợp các mô hình GBM tương ứng với các lần chia bằng phương pháp voting cho nhãn và lấy trung vị cho xác suất mưa lớn diện rộng.

c) Đánh giá chất lượng huấn luyện của mô hình bằng chỉ số ROC-AUC

Đường cong đo lường hiệu suất của bài toán phân lớp với nhiều thiết đặt ngưỡng khác nhau. ROC nghĩa là đường cong xác suất còn AUC thể hiện mức độ hay đo lường sự khả tách. Nó cho thấy mô hình có khả năng phân biệt giữa các lớp như thế nào? Nếu giá trị AUC lớn hơn nghĩa là mô hình có khả năng dự báo lớp 0 đúng lớp 0 hơn, lớp 1 đúng lớp 1 hơn. Điều đó có nghĩa là nếu giá trị AUC lớn mô

hình có khả năng phân biệt tốt giữa 2 trường hợp có xảy ra mưa lớn và không có mưa lớn. Đường cong ROC dùng 2 chỉ số TPR (true positive ratio) trên trục Ox và FPR (false positive ratio) trên trục Oy. Mô hình có giá trị AUC càng gần 1 thì càng tốt, điều đó chỉ ra rằng mô hình có khả năng phân tách tốt. Mô hình có giá trị AUC càng gần 0 thì càng tồi, thể hiện khả năng phân biệt kém. Khi giá trị AUC bằng 0.5 thì mô hình không có khả năng phân tách lớp.

d) Kết quả triển khai

Dự báo mưa lớn diện rộng được huấn luyện thông qua trình rain_model.py.

```
def train_model(data_file_path, save_model_file_path, nr_days, stations):
    if nr_days is None or len(nr_days) == 0:
        nr_days = [1, 2, 3]
    save_models = {}
    …….

    processingData.mean_col = mean_col
    processingData.std_col = std_col
    labels = pd.DataFrame()
    n_obs = 4
    invalidate_stations = []
```

Hình 4.68: Minh họa đoạn code huấn luyện mô hình AI dự báo mưa

Kết quả huấn luyện mô hình được gọi trả về thông qua api_call_rain_model.py sử dụng API.

```
def get_data_total_rain(stations, datetimeObs, data):
    response_data = {}
    data = data.sort_values(by=['TimeVN'], ascending=True)
    for station in stations:
        …….
        response = {
                'real_total': float(total),
                'real_data': mongo_data
            }
        response_data[station] = response
    return response_data
```

Hình 4.69: Minh họa đoạn code gọi API trả kết quả huấn luyện mô hình AI dự báo mưa

Sử dụng dữ liệu lượng mưa tại trong 36h trước 19h hàng ngày để dự báo tổng lượng mưa của ngày kế tiếp, kết quả cụ thể khi học thuật toán trên cho dữ liệu huấn luyện từ 2014 đến hết năm 2018, và kiểm tra đánh giá trên tập dữ liệu năm 2019 của một số trạm như sau:

Bảng 4.13: Kết quả huấn luyện dự báo mưa lớn diện rộng bằng GB

STT	Trạm	AUC	STT	Trạm	AUC
1.	Bắc Quang	0.7692537313432836	10.	Nam Đông	0.8365384615384616
2.	Tam Đảo	0.761904761904762	11.	Tam Kỳ	0.8925714285714287

STT	Trạm	AUC	STT	Trạm	AUC
3.	Quảng Hà	0.8810991268618387	12.	Trà My	0.8535100286532951
4.	Cửa Ông	0.7656054931335831	13.	Quảng Ngãi	0.9229340761374187
5.	Móng Cái	0.7701551566778218	14.	Ba Tơ	0.9126760563380282
6.	Hà Tĩnh	0.842292089249493	15.	Cát Tiên	0.7981733524355301
7.	Kỳ Anh	0.7683809523809523	16.	Đồng Phú	0.7396389524535978
8.	Huế	0.8016901408450704	17.	Thủ Dầu Một	0.6960049937578028
9.	A Lưới	0.8944761904761905	18.		

Với dự báo mưa lớn diện rộng (tổng lượng mưa trong ngày lớn hơn 50mm) trung bình AUC trên tất cả các trạm: 0.8090980442926845.

4.6.6. Trình diễn kết quả dự báo mưa lớn

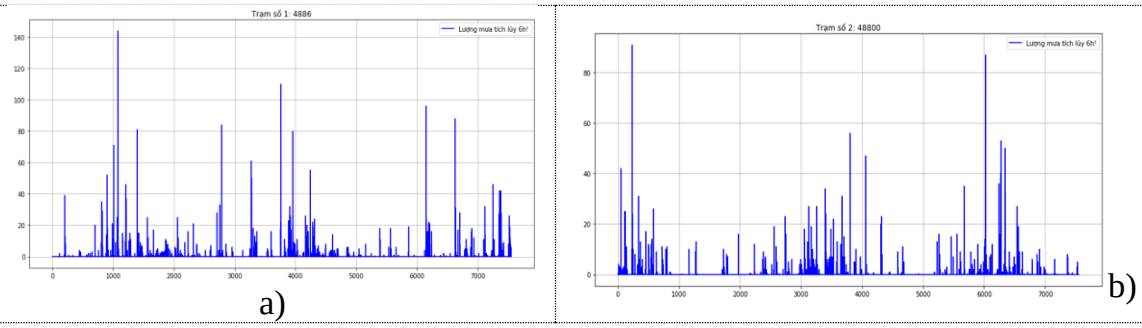
4.6.6.1. Công cụ thực hiện

Các phương pháp, công cụ được áp dụng để thực hiện trình diễn kết quả dự báo mưa lớn diện rộng gồm:

- Phương pháp trình diễn dữ liệu sử dụng bản đồ;
- Phương pháp sử dụng Box - plot;
- Phương pháp mô hình cây;
- Phương pháp trình diễn biểu đồ cột và đường;
- Công cụ trình diễn Trie;
- Công cụ trình diễn Succinct Tries.

4.6.6.2. Kết quả triển khai

Sử dụng các phương pháp, công cụ nêu trên để trình diễn dữ liệu quan trắc và kết quả dự báo mưa lớn diện rộng. Sử dụng thư viện Matplotlib trong Python để trình diễn dữ liệu quan trắc mưa lớn diện rộng theo dạng đồ thị, cách sử dụng tương tự như trong Matlab, kết quả như sau:



a) b)

Hình 4.70: Trình diễn dữ liệu quan trắc mưa tại trạm Kỳ Anh (a) và Mường Lay (b)

bằng phương pháp đồ thị

Hình 4.71: Trình diễn kết quả dự báo mưa trên bản đồ

4.6.7. Kết luận

Các kết quả triển khai cho thấy mô hình AI để dự báo mưa lớn diện rộng về hiệu năng tương đối tốt. Tuy nhiên, giá trị AUC của mô hình vẫn cần được cải thiện hơn nữa, kết quả dự báo vẫn còn chưa đoán được đúng xu thế của dữ liệu. Chính vì vậy, trong tương lai nhóm nghiên cứu sẽ tiếp tục cải tiến phương pháp để thu được kết quả dự báo tốt hơn nữa. Ngoài ra số tham số phụ thuộc vào số giá trị thời điểm trước cũng cần được điều chỉnh linh hoạt để có được kết quả dự báo phù hợp với thực tế.

4.7. Triển khai hệ thống AI hỗ trợ dự báo không khí lạnh khu vực Bắc Bộ

4.7.1. Dữ liệu cho hệ thống AI hỗ trợ dự báo không khí lạnh (KKL)

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo KKL sử dụng bộ dữ liệu quan trắc synop (nhiệt độ, tốc độ gió, hướng gió, lượng mưa, ...) của 43 trạm khí tượng, cụ thể tại bảng sau:

Bảng 4.14: Danh sách 43 trạm quan trắc sử dụng dự báo KKL

TT	Tên trạm	TT	Tên trạm	TT	Tên trạm	TT	Tên trạm
1	Mường Lay	12	Bãi Cháy	23	Huế	34	Phú Quý
2	Điện Biên	13	Phù Liễn	24	Hoàng Sa	35	Pleiku
3	Sơn La	14	Bạch Long Vĩ	25	Đà Nẵng	36	Buôn Ma Thuật
4	Hòa Bình	15	Hà Đông	26	Quảng Ngãi	37	Nhà Bè
5	Lào Cai	16	Nam Định	27	Hoài Nhơn	38	Vũng Tàu
6	Hà Giang	17	Hồi Xuân	28	Quy Nhơn	39	Côn Đảo
7	Tuyên Quang	18	Thanh Hóa	29	Tuy Hòa	40	Cần Thơ
8	Cao Bằng	19	Vinh	30	Nha Trang	41	Phú Quốc

TT	Tên trạm	TT	Tên trạm	TT	Tên trạm	TT	Tên trạm
9	Lạng Sơn	20	Hà Tĩnh	31	Trường Sa	42	Rạch Giá
10	Móng Cái	21	Kỳ Anh	32	Phan Rang	43	Cà Mau
11	Tiên Yên	22	Đồng Hới	33	Phan Thiết		

- Tần suất dữ liệu quan trắc: 8 obs/ ngày; 3 giờ/ lần.

- Số lượng đợt KKL là: 275 đợt; trung bình khoảng 27,5đợt/ năm.

- Thời gian dữ liệu KKL là: 10 năm, từ năm 2008 - 2018.

- Các dữ liệu quan trắc KKL được trích xuất từ CSDL MongoDB và lưu trữ trong file .csv.

4.7.2. Xử lý các dữ liệu không khí lạnh (KKL)

4.7.2.1. Công cụ thực hiện

- Sử dụng phương pháp Z-Score, biểu đồ Box-plot để phát hiện dữ liệu không khí lạnh (KKL) ngoại lai.

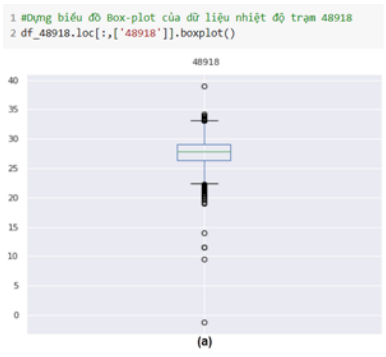
- Sử dụng 02 phương pháp xử lý dữ liệu thiếu đó là: (1) Bỏ qua các biến chứa giá trị thiếu, (2) Ước tính các giá trị bị thiếu.

4.7.2.2. Kết quả triển khai

Kết quả phát hiện dữ liệu KKL ngoại lai như sau:

TimeVN	48918	zscore	outlier
2015-02-01 16:00:00	-1.3	-14.283356	1
2015-02-10 19:00:00	14.0	-6.785862	1
2015-09-10 10:00:00	39.0	5.464945	1
2016-03-20 13:00:00	11.5	-8.010942	1
2016-04-30 22:00:00	11.5	-8.010942	1
2017-06-15 16:00:00	9.5	-8.991007	1

a)



b)

Hình 4.72: Phát hiện dữ liệu nhiệt độ ngoại lai tại trạm Côn Đảo bằng Z-Score (a) và box-plot (b)

Kết quả kiểm chứng và xử lý dữ liệu KKL ngoại lai như sau:

```
#Dữ liệu tại các thời điểm trong ngày
df_48918['2015-09-10']
```

TimeVN	48918
2015-09-10 01:00:00	27.4
2015-09-10 04:00:00	26.8
2015-09-10 07:00:00	27.8
2015-09-10 10:00:00	39.0
2015-09-10 13:00:00	30.7
2015-09-10 16:00:00	27.6
2015-09-10 19:00:00	27.6
2015-09-10 22:00:00	27.6

a)

```
#Vẽ biểu đồ nhiệt độ trong 5 ngày
df_48918['2015-09-10':'2015-09-15'].plot(marker='o',color='Red')
```


```
#Xử lý dữ liệu ngoại lai của trạm 48918
df_48918.loc['2015-02-01 16:00:00',['48918']] = None
df_48918.loc['2015-02-10 19:00:00',['48918']] = None
df_48918.loc['2016-03-20 13:00:00',['48918']] = None
df_48918.loc['2016-04-30 22:00:00',['48918']] = None
df_48918.loc['2017-06-15 16:00:00',['48918']] = None
df_48918.loc['2015-09-10 10:00:00',['48918']] = None
```

b)

Hình 4.73: Kiểm chứng (a) và xử lý (b) dữ liệu nhiệt độ ngoại lai tại trạm Côn Đảo

Nhận xét: Việc triển khai công cụ phát hiện, xử lý dữ liệu KKL ngoại lai là khâu quan trọng để nâng cao chất lượng dữ liệu đầu vào của mô hình AI dự báo không khí lạnh.

4.7.3. Chuẩn hóa dữ liệu không khí lạnh

4.7.3.1. Công cụ và dữ liệu thực hiện

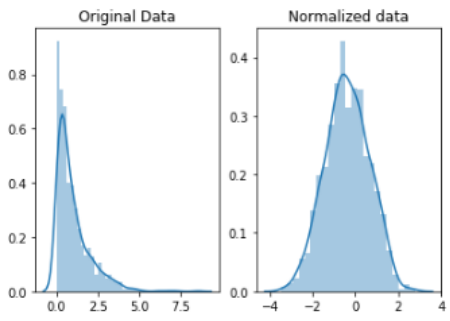
Sử dụng công cụ Pandas trong thư viện mã nguồn mở của Python để thực hiện chuẩn hóa dữ liệu nhiệt độ trong mô hình AI dự báo KKL.

4.7.3.2. Kết quả triển khai

Kết quả chuẩn hóa dữ liệu lượng KKL như sau:

```
# dành cho biến đổi Box-Cox
from scipy import stats
# cho phép co dãn min_max
from mlxtend.preprocessing import
minmax_scaling
# các module vẽ hình
import seaborn as sns
import matplotlib.pyplot as plt
# đọc dữ liệu
kickstarters_2017 =
pd.read_csv("KKL.csv")
# đặt giá trị sinh cho thử nghiệm
np.random.seed(0)
```

a)



b)

Hình 4.74: Thuật toán (a) và kết quả (b) chuẩn hóa dữ liệu nhiệt độ

4.7.4. Trực quan hóa dữ liệu không khí lạnh (KKL)

4.7.4.1. Công cụ thực hiện

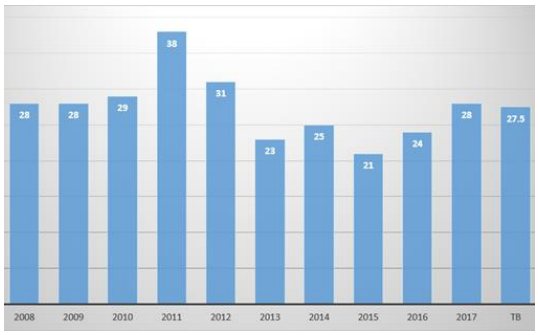
Công cụ thực hiện: Sử dụng các công cụ biểu đồ cột (Bar/Column chart); Biểu đồ đường (Line Graph); Biểu đồ tròn, biểu đồ cột chồng (Pie chart, Stacked Column/bar chart); Biểu đồ phân tán (Scatter plot); Biểu đồ hộp và râu (Box- plot) để trực quan hóa dữ liệu KKL.

4.7.4.2. Kết quả triển khai

Kết quả triển khai trực quan hóa số đợt KKL trung bình theo khu vực bằng biểu đồ cột như sau:

Bảng 4.15: Thống kê số lượng các đợt KKL trong 10 năm (2008-2017)

Năm	Số đợt KKL/ tháng trong năm										Cộng
	T1	T2	T3	T4	T5	T6	T9	T10	T11	T12	
2008	4	6	2	1	4			1	5	5	28
2009	6	4	2	1	2			1	5	5	28
2010	5	2	4	5		1		3	4	5	29
2011	9	2	8	2	4		1	4	3	5	38
2012	5	5	3	2	1		2	2	5	6	31
2013	5	3	2	3		1	1	1	2	5	23
2014	4	4	3	1	1			1	4	7	25
2015	2	2	4	3	1			2	2	5	21
2016	4	4	3	1	1			2	4	5	24
2017	2	2	6	3	3			2	4	6	28
TB	4.6	3.4	3.7	2.2	1.7	0.2	0.4	1.9	3.8	5.4	27.5



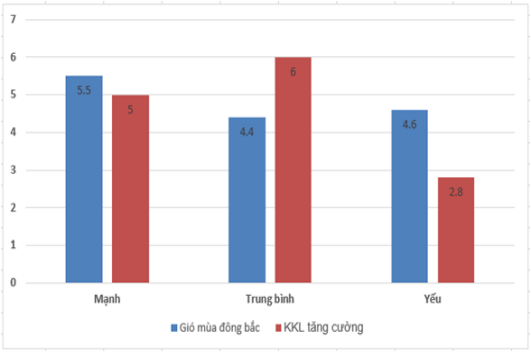
Hình 4.75: Trực quan hóa số đợt KKL theo năm (2008-2017) bằng biểu đồ cột

Kết quả triển khai trực quan hóa theo cường độ KKL thông qua tốc độ gió đông bắc tại vịnh Bắc Bộ bằng biểu đồ cột như sau:

- KKL yếu: chỉ gây ra gió cường độ cấp 5/cấp 6, đồng thời làm thay đổi rõ rệt về thời tiết ở một số khu vực (chuyển đầy mây, diện mưa tăng lên đột ngột và nhiệt độ trung bình ngày giảm từ 3-5 độ trở lên).
- KKL trung bình: gây ra gió mạnh lên cấp 6, giật trên cấp 6.
- KKL mạnh: gây ra gió mạnh cấp 7 trở lên.

Bảng 4.16: Phân bố các đợt KKL tăng cường giai đoạn 2008-2017

Năm	Gió mùa Đông Bắc				KKL tăng cường				Tổng cộng
	Mạnh	TB	Yếu	Cộng	Mạnh	TB	Yếu	Cộng	
2008	5	2	5	*12*	6	6	4	*16*	**28**
2009	5	4	7	*16*	7	4	1	*12*	**28**
2010	8	6	2	*16*	5	3	5	*13*	**29**
2011	5	3	10	*18*	10	7	3	*20*	**38**
2012	10	6	3	*19*	5	4	3	*12*	**31**
2013	3	3	3	*9*	4	11	4	*19*	**28**
2014	3	5	2	*10*	6	8	1	*15*	**25**
2015	4	4	8	*16*	1	3	4	*8*	**24**
2016	8	5	3	*16*	4	4	0	*8*	**24**
2017	4	6	3	*13*	2	10	3	*15*	**28**
TB	**5,5**	**4,4**	**4,6**	***14,5***	**5,0**	**6,0**	**2,8**	***13,8***	**28,3**



Hình 4.76: Trực quan hóa số đợt KKL theo năm (2008-2017) bằng biểu đồ cột

Nhận xét: Triển khai trực quan hóa dữ liệu thống kê về số đợt KKL theo năm, theo cường độ bằng biểu đồ cột. Dựa vào trực quan hóa các dữ liệu thống kê trên, có thể thấy trong 10 năm trở lại đây, số đợt gió mùa Đông Bắc có cường độ mạnh xảy ra nhiều hơn số đợt có cường độ trung bình và yếu. Ngược lại, số đợt KKL tăng cường có cường độ trung bình lại nhiều hơn số đợt KKL mạnh và đợt KKL yếu.

4.7.5. Dự báo không không khí lạnh bằng mô hình AI với tập dữ liệu huấn luyện

4.7.5.1. Công cụ thực hiện

Sử dụng thuật toán GB để huấn luyện dự báo không khí lạnh (tương tự như huấn luyện mưa lớn diện rộng).

4.7.5.2. Kết quả triển khai huấn luyện dự báo KKL

Thời gian lấy dữ liệu từ 1/1/2014 đến 31/12/2019, tần suất dữ liệu 3h/lần. Các trường dữ liệu sử dụng và thực hiện mô phỏng dữ liệu KKL như sau.

Bảng 4.17: Các trường dữ liệu nhiệt độ T

TT	Tên trường	Kiểu dữ liệu	Kô tả
1	StationID	Nvarchar	Mã trạm quan trắc(Station)
2	DtDate	ISODate	Ngày quan trắc
3	DtObs	Nvarchar	Giờ quan trắc: 7h, 13h, 19h, 1h
4	TT	Int	Nhiệt độ tại thời điểm đo
5	Tx	Real	Tx là nhiệt độ min day
6	Tm	Real	Tx là nhiệt độ max day
7	RRR6	Real	RRR6 là nhiệt độ trong 6h qua
8	Lat	Real	Kinh độ
9	Lon	Real	Vĩ độ

Bảng 4.18: Tiêu chí xác định KKL

TT	Nhiệt độ trung bình	Kết luận	Nhãn
1	Nhiệt độ Trung bình ngày <15	Rét đậm	0
2	Nhiệt độ Trung bình ngày <13	Rét hại	1
3	Nhiệt độ trung bình > 15	BT	2

Dự báo không khí lạnh được huấn luyện thông qua temp_model.py

```
def weighted_median(dff, wei):
```

```
    def wei_med(scores, w):
        scores = np.array(scores)
        sort_idx = np.argsort(scores)
        scores = scores[sort_idx]
        w = w[sort_idx]
……..
        model = lgb.LGBMClassifier()
        model.set_params(**params)
        model.fit(train_x, train_y)
        return  metrics.roc_auc_score(test_y,  model.predict_proba(test_x)[:,
1])
```

Hình 4.77: Minh họa đoạn code huấn luyện mô hình AI dự báo KKL

Kết quả huấn luyện mô hình được gọi trả về thông qua api_call_temp_model.py sử dụng API.

```
def get_data_and_mean(stations, datetimeObs, data):
    response_data = {}
    data = data.sort_values(by=['TimeVN'], ascending=True)
    for station in stations:
        start_date = datetimeObs + timedelta(days=-1)
……..
            response = {
                    'real_mean': -999,
                    'real_data': mongo_data
                }
        response_data[station] = response
    return response_data
```

Hình 4.78: Một đoạn code gọi api trả kết quả mô hình AI dự báo KKL

Sử dụng dữ liệu nhiệt độ trong 36h trước 19h hàng ngày để dự báo nhiệt độ trung bình của ngày kế tiếp. Kết quả khi học thuật toán trên cho dữ liệu huấn luyện từ 2014 đến hết năm 2018, và kiểm tra đánh giá trên tập dữ liệu năm 2019 như sau:

Bảng 4.19: Kết quả huấn luyện dự báo KKL bằng GB

STT	Trạm	AUC	STT	Trạm	AUC
1	Mường Lay	0.9958217270194986	9	Lạng Sơn	0.96132744547036
2	Điện Biên	0.9971988795518206	10	Móng Cái	0.9787063720781299
3	Sơn La	0.9832657200811359	11	Tiên Yên	0.9719405003380662
4	Hòa Bình	0.9873949579831933	12	Bãi Cháy	0.9842154131847725
5	Lào Cai	0.9954481792717087	13	Bạch Long Vỹ	0.9923822714681441
6	Hà Giang	0.9932394366197184	14	Phủ Liễn	0.980037140204271
7	Tuyên Quang	0.9856442577030813	15	Hà Đông	0.9767873723305479
8	Cao Bằng	0.9812804782129936	16		

Với dự báo rét đậm với nhiệt độ trung bình trong ngày dưới 15 độ C, giá trị AUC trung bình trên tất cả các trạm là 0.9798546099808678.

4.7.6. Trình diễn kết quả dự báo không khí lạnh

4.7.6.1. Công cụ thực hiện

Sử dụng các phương pháp, công cụ để thực hiện trình diễn kết quả dự báo **KKL** bằng gồm:

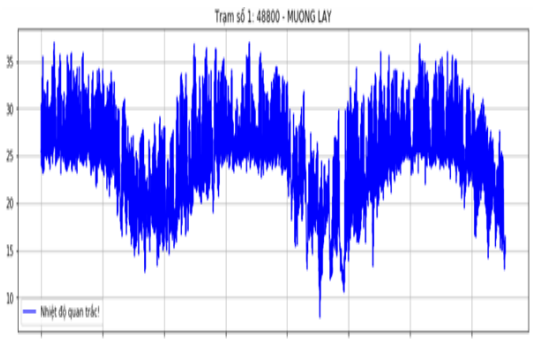
- Phương pháp mô hình cây và phương pháp trình diễn bản đồ.
- Công cụ trình diễn Trie và công cụ trình diễn Succinct Tries.

4.7.6.2. Kết quả triển khai

Sử dụng các phương pháp, công cụ nêu trên để trình diễn dữ liệu quan trắc và dự báo KKL, kết quả cụ thể như sau:

```
num_station = 16
fig, axs = plt.subplots(num_station, 1,
sharex=True,figsize=(15, 80))
# Remove horizontal space between axes
fig.subplots_adjust(hspace=0.2)
for i in range(num_station):
    axs[i].set_title('Trạm số ' + str(i+1)
+ ": " + name_station[i])
    axs[i].plot(data_temp.iloc[:,i+2],
color='blue', label='Nhiệt độ quan trắc!')
    axs[i].grid(True)
    axs[i].legend(loc="best")

#fig.suptitle('BIỂU ĐỒ SO SÁNH NHIỆT ĐỘ
QUAN TRẮC VÀ DỰ ĐOÁN CỦA MÔ HÌNH', font-
size=16)
plt.show()
```

Hình 4.79: Thuật toán trình diễn so sánh kết quả dự báo nhiệt độ và quan trắc



Hình 4.80: Trình diễn dữ liệu quan trắc nhiệt độ tại trạm Mường Lay bằng phương pháp biểu đồ đường

Nhận xét: Trình diễn kết quả dự báo KKL cho phép mô tả dữ liệu dưới dạng các hình ảnh trực quan như bảng, biểu đồ, đồ thị, ... Việc trực quan hóa dữ liệu hỗ trợ quan sát sự thay đổi dữ liệu, trên cơ sở đó lựa chọn được phương pháp học máy hiệu quả.

Hình 4.81: Trình diễn kết quả dự báo KKL bằng bản đồ

4.7.7. Kết luận

Việc sử dụng AI cho bài toán dự báo không khí lạnh cho kết quả rất khả quan, hầu như giá trị AUC của các trạm đều lớn hơn 90%. Vì vậy, cần tiếp tục nghiên cứu dự báo nhiệt độ trong khoảng thời gian dài ngày hơn nữa.

4.8. Triển khai hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng

4.8.1. Dữ liệu cho hệ thống AI hỗ trợ dự báo lũ

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo lũ trên hệ thống sông Hồng sử dụng dữ liệu quan trắc của 3 trạm thủy văn là Vụ Quang, Hà Nội, Hưng Yên, cụ thể như sau:

- **Về dữ liệu trạm Vụ Quang và trạm Hà Nội**: Thời gian dữ liệu là 10 năm (từ 01/01/2008 đến 31/12/2017). Tần suất dữ liệu quan trắc trong điều kiện thời tiết bình thường là 4 obs/ ngày vào thời điểm 1h, 7h, 13h, và 19h với bước dữ liệu là 6 giờ/ lần dữ liệu; trong điều kiện thời tiết nguy hiểm là lên 8 obs hoặc 12 obs/ ngày, với các bước dữ liệu là 2-3 giờ/ lần dữ liệu. Vì vậy, phải thực hiện chuẩn hóa dữ liệu về tần suất là 8 obs/ ngàytại các mốc thời gian là 1h,4h, 10h,13h,16h,19h,22h. Khi đó, sẽ có 2 trường hợp xảy ra: (i) Tần suất dữ liệu nhỏ hơn 8 obs/ ngày hoặc thiếu dữ liệu; trường hợp này sẽ coi như dữ liệu bị thiếu và áp dụng phương pháp nội suy dữ liệu và cửa sổ trượt để xử lý dữ liệu thiếu; (ii) Tần suất dữ liệu lớn hơn 8 obs/ ngày; trường hợp này sẽ phải tính toán lại giá trị mực nước dựa vào thời gian quan trắc là 1h,4h,7h,10h,13h,16h,19h và 22h.

- **Dữ liệu trạm Hưng Yên**: Thời gian dữ liệu từ 01/01/2008 đến 23/04/2015. Tần suất dữ liệu là 1 giờ/ lần.

Để đánh giá và so sánh thì tất cả dữ liệu thu thập được sẽ được chia thành 2 phần để huấn luyện và kiểm thử, Với dữ liệu trạm Vụ Quang và Hà Nội thì 80% dữ liệu để huấn luyện (từ 2008 đến 2015) và 20% (từ 2016-2017) còn lại được sử dụng để kiểm thử mô hình đã được huấn luyện.

4.8.2. Xử lý, chuẩn hóa, trực quan hóa và phân hạng, lưa chọn đặc trưng dữ liệu lũ

4.8.2.1. Công cụ thực hiện

- Về xử lý dữ liệu **lũ thiếu/ mất mát**: sử dụng ngôn ngữ R gồm: (i) Các gói bổ sung dữ liệu một chiều gồm: na.interp (forecast R); na.approx (zoo R-package); na.approx; na.aggregate; na.locf; (ii) Các gói bổ sung dữ liệu một nhiều chiều gồm: MI - Multiple Imputation (MI R-package); MICE - Multiple Imputation by Chained Equations (MICE R-package); (iii) Phương pháp bổ sung dựa trên học máy: And

miss Forest (missForest R-package): Dựa trên thuật toán rừng ngẫu nhiên để điền dữ liệu bị thiếu, đặc biệt trong trường hợp dữ liệu kiểu hỗn hợp (cả số và category); (iv) Phương pháp bổ sung dữ liệu sử dụng trực tiếp dữ liệu có sẵn: DTWUMI (Dynamic Time Warping-based Uncorrelated Multivariate Imputation).

- Về xử lý dữ liệu **lũ không chắc chắn:** sử dụng ngôn ngữ R và các hàm trong gói Keras.

- Về phát hiện và xử lý dữ liệu **lũ ngoại lai/ bất thường** (outlier): sử dụng công cụ Data mining của Microsoft add-ins trong Excel.

- Về **chuẩn hóa** dữ liệu lũ: sử dụng công cụ Excel của Microsoft và công cụ Pandas trong thư viện mã nguồn mở của Python.

- Về **trực quan hóa** dữ liệu lũ: Công cụ là cácgói của ngôn ngữ lập trình Python gồm: Numpy, Matplotlib, Jupyter Notebook, Pandas, Tensor Flow.Các phương pháp trực quan dữ liệu lũ gồm: biểu đồ đường (Line Graph), biểu đồ thanh, biểu đồ histogram, biểu đồ hộp, biểu đồ phân tán.

- Về công cụ **phân hạng đặc trưng** dữ liệu lũ: (i) Theo phương pháp lọc - Filter Methods: Test giả thuyết thống kê (Statistical Hypothesis Tests), Đa Test (Multiple Testing Problem), Phân hạng biến (Đặc trưng), Các bộ lọc đa biến (Multivariate Filters), Các phương pháp tìm kiếm đa biến (Multivariate Search Methods); (ii) Các phương pháp đóng gói - Wrapper Methods; (iii) Các phương pháp nhúng - Embedded Methods: Các bộ dự đoán thưa tuyến tính (Sparse Linear Predictor), Các phương pháp phi tuyến (Non-linear Methods).

- Về công cụ **lựa chọn đặc trưng** dữ liệu lũ: Sử dụng phương pháp lựa chọn đặc trưng văn bản, mô hình túi, Mô hình túi từ (Bag-of-Words), Mô hình Bag-of-ngrams, Mô hình TF-IDF, Biểu diễn cấu trúc (Structural representation), Biểu diễn ngữ nghĩa ẩn (Latend semantic representation), Biểu diễn sử dụng “embedding” (Embedding representation).

4.8.2.2. Kết quả triển khai xử lý dữ liệu lũ thiếu/ mất mát

Để so sánh hiệu năng giữa các mô hình, thực hiện xóa dữ liệu tại một số ngày trong tập dữ liệu lũ đầy đủ. Bảng dưới đây cho biết: Cột [,1]: vị trí dữ liệu thiếu; Cột [,2]: số dữ liệu bị thiếu.

```
D:/Project/Du_bao_lu_song_H
[[3]]
     [,1] [,2]
[1,]   13    4
[2,]   86    5
[3,]  115    1
[4,]  130    5
[5,]  184    1
[6,]  292    9
[[4]]
     [,1] [,2]
[1,]   57    6
[2,]  111    1
[3,]  125    1
[4,]  161    7
[5,]  279    6
```

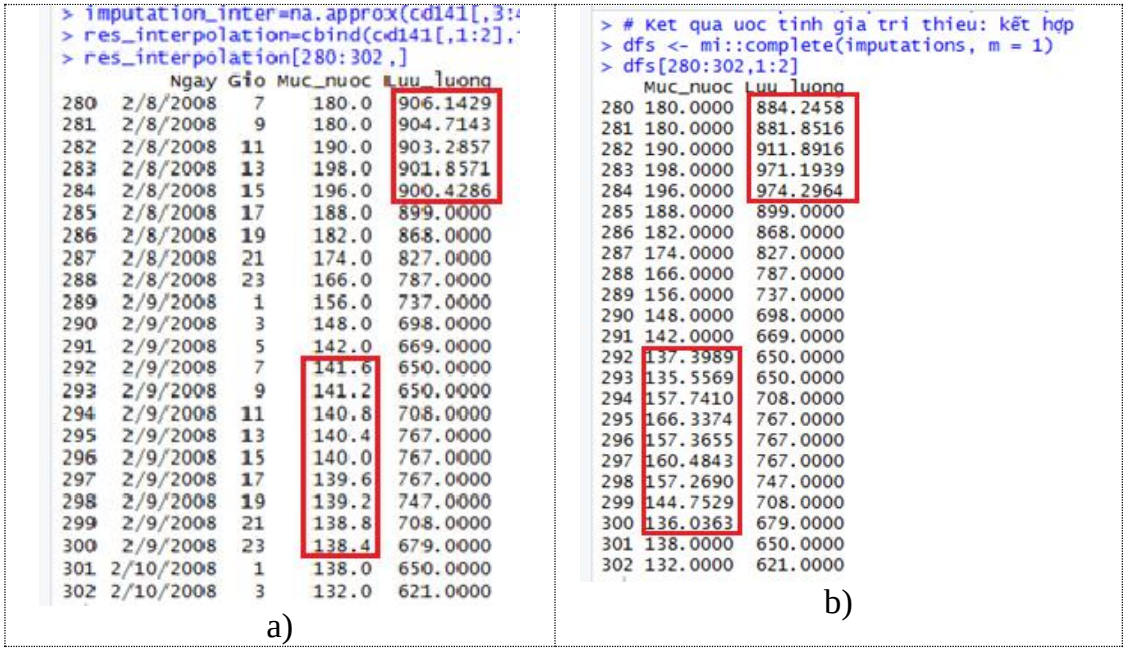
```
> cd141[280:302,]
         Ngay Gio Muc_nuoc Luu_luong
280  2/8/2008   7      180        NA
281  2/8/2008   9      180        NA
282  2/8/2008  11      190        NA
283  2/8/2008  13      198        NA
284  2/8/2008  15      196        NA
285  2/8/2008  17      188       899
286  2/8/2008  19      182       868
287  2/8/2008  21      174       827
288  2/8/2008  23      166       787
289  2/9/2008   1      156       737
290  2/9/2008   3      148       698
291  2/9/2008   5      142       669
292  2/9/2008   7       NA       650
293  2/9/2008   9       NA       650
294  2/9/2008  11       NA       708
295  2/9/2008  13       NA       767
296  2/9/2008  15       NA       767
297  2/9/2008  17       NA       767
298  2/9/2008  19       NA       747
299  2/9/2008  21       NA       708
300  2/9/2008  23       NA       679
>
```

Hình 4.82: Dữ liệu lũ mất mát/ thiếu

Trường hợp này, dữ liệu thiếu cả từng điểm và thiếu cả khoảng dữ liệu, đồng thời thiếu cả trên 2 biến (biến mực nước và lưu lượng nước). Do đó, sử dụng 4 phương pháp khác nhau: na.approx, MI, MICE và miss Forest thực hiện bổ sung dữ liệu lũ thiếu. Đoạn code thực hiện xử lý dữ liệu lũ với các phương pháp như sau:

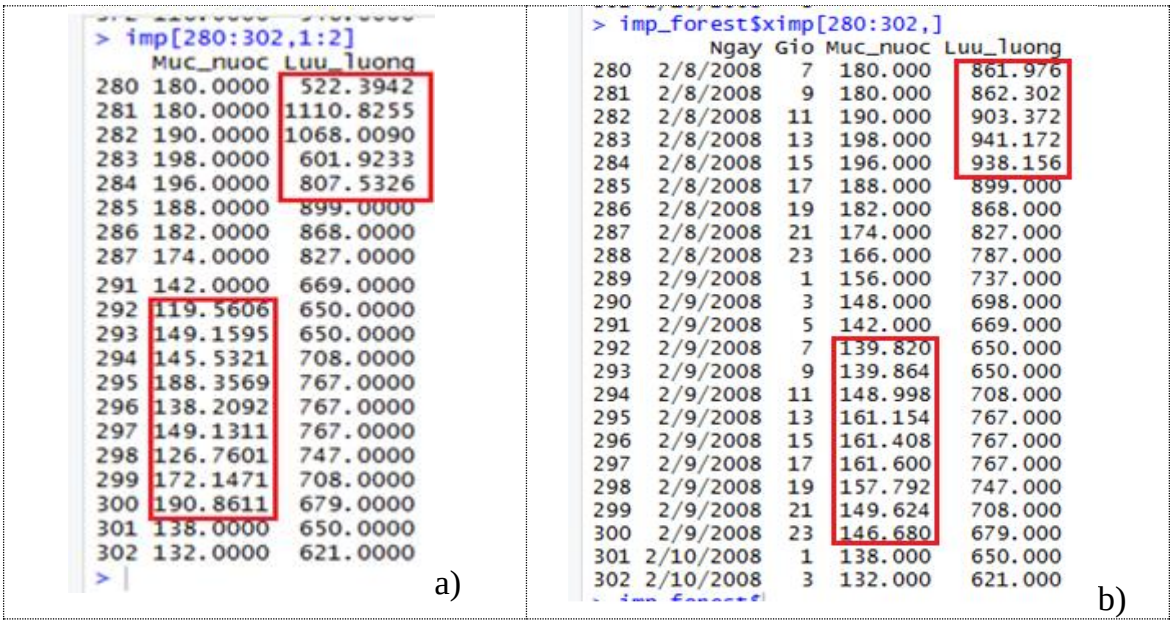
```
#--- Doc du lieu dau vao----
cd141=read.csv("D:\\Project\\Du_bao_lu_song_Hong\\Chuyende_Hong\\cd141.csv",h=T)
index_miss=Indexes_size_missing(cd141)

#--- Dien day du lieu su dung phuong phap noi suy-------------------------
imputation_inter=na.approx(cd141[,3:4])
res_interpolation=cbind(cd141[,1:2],imputation_inter)
res_interpolation[280:302,]

#---------------------Phuong phap mi-----------------------------------
# STEP 1: Convert to a missing_data.frame
mdf <- missing_data.frame(cd141[,3:4]) # warnings about missingness patterns
show(mdf)

# STEP 2: impute
imputations <- mi(mdf)

# Ket qua uoc tinh gia tri thieu: kết hợp cả các chuỗi thành kết quả cuối cùng
dfs <- mi::complete(imputations, m = 1)
dfs[280:302,1:2]

# --------------------phuong phap mice----------------------------
# Thuc hien dien day du lieu su dung phuong phap hoi quy tuyen tinh
tempData <- mice(cd141[,3:4],method="norm")
#imp_mice=tempData$imp
# Ket hop cac chuoi khac nhau thanh mot ket qua cuoi cung
imp=mice::complete(tempData)
imp[280:302,1:2]

#------------phuong phap missForest-----------------
imp_forest=missForest(cd141,ntree=500)
imp_forest$ximp[280:302,]
```

Kết quả xử lý dữ liệu lũ thiếu bằng phương pháp na.approx và MI như sau:



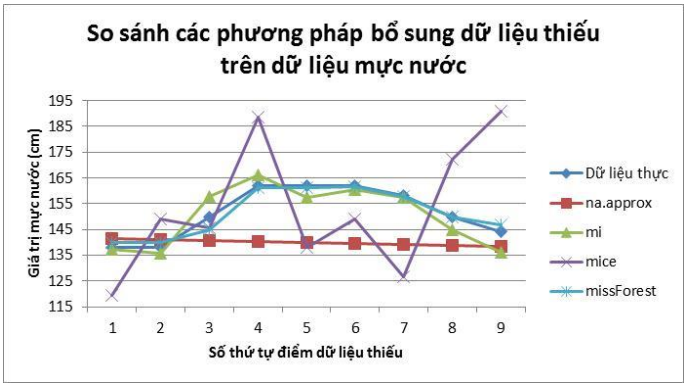
Hình 4.83: Xử lý giá trị dữ liệu lũ thiếu bằng phương pháp na.approx (a) và MI (b)

Kết quả xử lý dữ liệu lũ thiếu bằng phương pháp MICE, missForest như sau:



Hình 4.84: Xử lý giá trị dữ liệu lũ thiếu bằng phương pháp MICE (a) và missForest (b)

So sánh kết quả xử lý dữ liệu mực nước thiếu các phương pháp như sau:



Hình 4.85: So sánh kết quả xử lý dữ liệu mực nước thiếu của các phương pháp

Nhận xét: Kết quả triển khai cho thấy, không có phương pháp nào cho kết quả là hoàn hảo. Độ tương quan của 2 chuỗi dữ liệu mực nước và lưu lượng là rất cao (~1), do đó giá trị bổ sung của phương pháp missForest là tốt nhất, tiệm cận với giá trị thực đo, tiếp sau đó là giá trị bổ sung của phương pháp MI. Các phương pháp mi, mice, missForest là những phương pháp bổ sung dữ liệu thiếu nhiều chiều và chỉ cho kết quả tốt khi dữ liệu có độ tương quan cao.

4.8.2.3. Kết quả triển khai xử lý dữ liệu lũ không chắc chắn

Sử dụng dữ liệu lũ có tần số quan trắc là 12 obs/ ngày (2 giờ/ lần).

Sử dụng đoạn mã lệnh với các hàm trong gói keras để tính toán sự không chắc chắn sử dụng dữ liệu mực nước. Sử dụng 80% cho dữ liệu huấn luyện, 20% cho dữ liệu thử nghiệm. Cụ thể như sau:

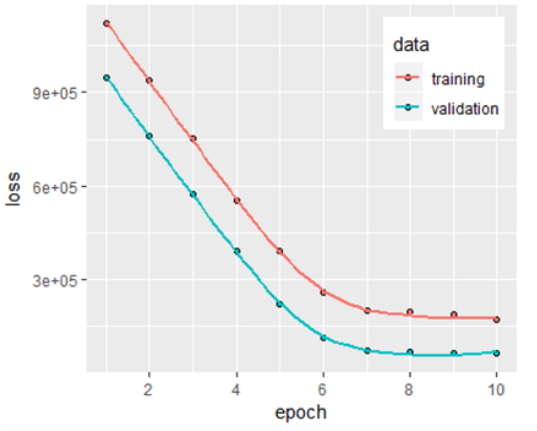
```
> hn_hy
      Ngay Gio Muc_nuoc Luu_luong
1   1/16/2008   1      158       747
2   1/16/2008   3      156       737
3   1/16/2008   5      154       728
4   1/16/2008   7      154       728
5   1/16/2008   9      152       718
6   1/16/2008  11      156       737
7   1/16/2008  13      162       767
8   1/16/2008  15      164       777
9   1/16/2008  17      166       787
10  1/16/2008  19      167       792
11  1/16/2008  21      166       787
12  1/16/2008  23      166       787
13  1/17/2008   1      166       787
14  1/17/2008   3      166       787
15  1/17/2008   5      166       787
16  1/17/2008   7      166       787
17  1/17/2008   9      166       787
18  1/17/2008  11      166       787
19  1/17/2008  13      166       787
20  1/17/2008  15      168       797
21  1/17/2008  17      170       807
22  1/17/2008  19      170       807
```

Hình 4.86: Dữ liệu H quan trắc 12 obs/ ngày

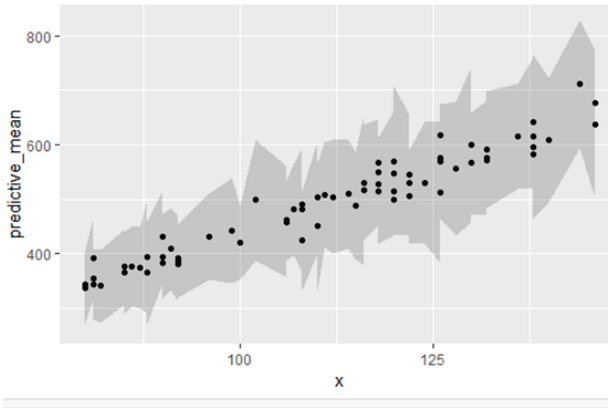
```
K <- keras::backend()
hn=read.csv("D: \\Du_bao_lu \\cd_clustering.csv",h=T)
n_samples <- 300#nrow(hn)
n_features <- 1
n_hidden1 <- 128
n_hidden2 <- 128
n_output <- 1
learning_rate <- 1e-6
num_epochs <- 10
batch_size <- n_samples / 10
dropout <- 0.5
l2 <- 0.1
X_train=matrix(hn[1:300,3],nrow=300,ncol=1)
y_train=hn[1:300,4]
X_test=matrix(hn[301:nrow(hn),3],nrow=72,ncol=1)
y_test=hn[301:nrow(hn),4]
fit <- lm(y_train ~ X_train)
summary(fit)
model <- keras_model_sequential()
model %>%
layer_dense(units = n_hidden1, activation = 'relu',
input_shape = c(n_features)) %>%
layer_dropout(rate = dropout) %>%
layer_activity_regularization(l1=0, l2=l2) %>%
layer_dense(units = n_hidden2, activation = 'relu')
layer_dropout(rate = dropout) %>%
layer_activity_regularization(l1=0, l2=l2) %>%
layer_dense(units = n_output, activation = 'linear')
```

Hình 4.87: Đoạn code tính toán độ không chắc chắn của dữ liệu H

Kết quả tính toán hàm lỗi và xử lý độ không chắc chắn của dữ liệu mực nước (H) như sau:



Hình 4.88: Trực quan hàm dữ liệu H lỗi

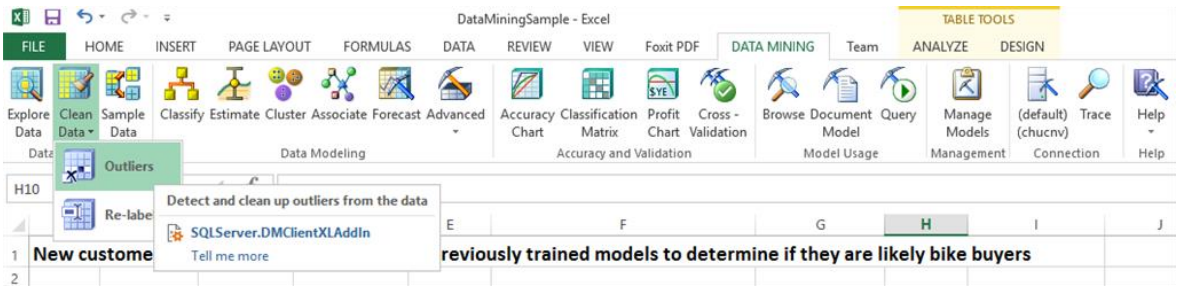


Hình 4.89: Độ không chắc chắn trên tập xác nhận dữ liệu mực nước để dự báo

Nhận xét: Kết quả dự báo lưu lượng nước (Q) dựa trên yếu tố mực nước đầu vào, trong đó có sử dụng phương pháp xử lý dữ liệu mực nước không chắc chắn. Phương pháp này cho kết quả dự báo lưu lượng Q một khoảng xung quanh giá trị quan trắc thực tế.Thực tế, dữ liệu quan trắc lũ (H, Q) được biểu diễn dưới dạng các con số đã bao hàm độ không chắc chắn trong đó. Do đó, cần có một công cụ để đánh giá độ không chắc chắn của một mô hình để xử lý dữ liệu không chắc trong mô hình AI dự báo lũ (H, Q).

4.8.2.4. Kết quả triển khai xử lý dữ liệu lũ bất thường/ ngoại lai

Sử dụng công cụ Data mining của Microsoft add-ins trong Excel để phát hiện và xử lý outlier. Cài Data Mining Add-ins - Excel, vào menu Clean Data => Outliers.

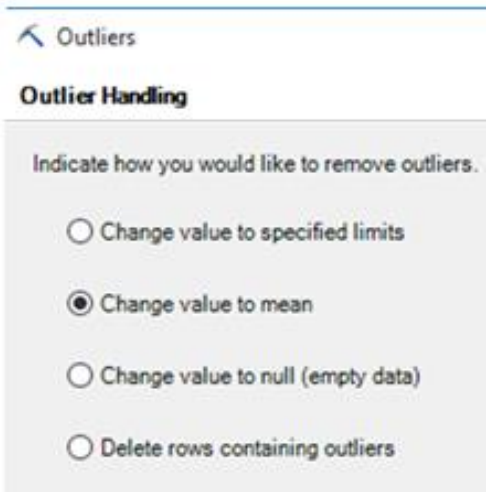


Hình 4.90: Giao diện công cụ xử lý dữ liệu bất thường

Chọn vùng dữ liệu, chọn cột cần phát hiện outlier trong tập dữ liệu quan trắc mực nước của trạm Sơn Tây, cụ thể giá trị H tăng đột ngột từ 264 cm đến 1123 cm rồi lại giảm đột về 264 cm. Khi xuất hiện các dữ liệu quan trắc mực nước H bất thường đó, sử dụng các phương thức xử lý trong công cụ Data mining như sau:

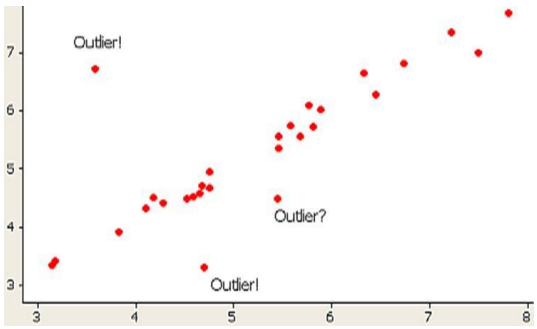
2013-02-1(	13	0	2	74162	331
2013-02-1	19	0	2	74162	554
2013-02-2	13	0	1	74162	282
2013-02-2	19	0	0	74162	259
2013-02-2(	13	0	0	74162	264
2013-03-0	17	0	1	74162	1123
2013-03-0	13	0	1	74162	271
2013-03-0(	13	0	2	74162	345
2012-01-3	7	0	0	74162	519
2012-02-0	13	0	1	74162	549
2012-02-1	19	0	1	74162	222
2012-02-1	1	0	0	74162	28
2012-02-1	19	0	0	74162	236
2012-03-0	7	0	0	74162	319
2012-03-0	7	0	2	74162	382
2012-03-0	19	0	0	74162	382

Hình 4.91: Dữ liệu H ngoại lai

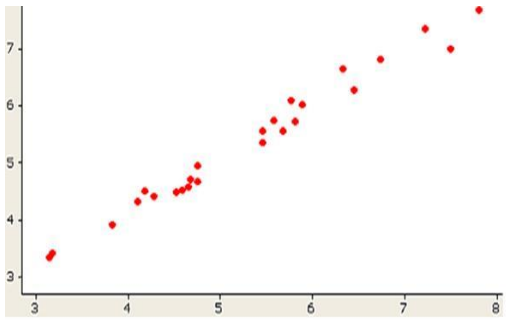


Hình 4.92: Xử lý dữ liệu H ngoại lai trong công cụ Data mining

Các dữ liệu quan trắc mực nước H bất thường được xem là các outliers và được thay thế bằng giá trị trung bình thu nhập của các quan trắc khác. Chọn thêm cột mới vào dữ liệu nguồn (worksheet đang làm việc), copy dữ liệu đã xử lý outlier ra sheet mới. Kết quả xử lý dữ liệu H bất thường như sau:



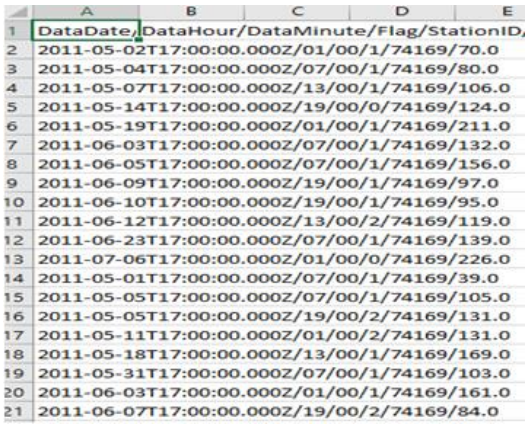
Hình 4.93: Xuất hiện dữ liệu H ngoại lai



Hình 4.94: Dữ liệu mực nước H sau khi xử lý ngoại lai

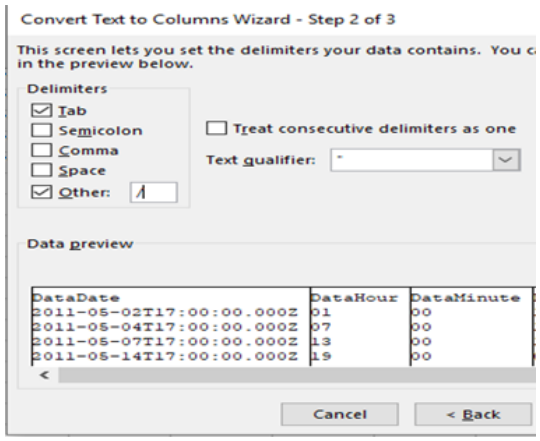
4.8.2.5. Kết quả triển khai chuẩn hóa dữ liệu lũ

Sử dụng phần mềm Excel để xử lý tách thông tin dữ liệu lũ dạng text thành các cột ngày, giờ, mực nước, lưu lượng nước liền kề nhau riêng biệt:



	A	B	C	D	E
1	DataDate/	DataHour/	DataMinute/	Flag/	StationID/
2	2011-05-02T17:00:00.000Z/	01/	00/	1/	74169/70.0
3	2011-05-04T17:00:00.000Z/	07/	00/	1/	74169/80.0
4	2011-05-07T17:00:00.000Z/	13/	00/	1/	74169/106.0
5	2011-05-14T17:00:00.000Z/	19/	00/	0/	74169/124.0
6	2011-05-19T17:00:00.000Z/	01/	00/	1/	74169/211.0
7	2011-06-03T17:00:00.000Z/	07/	00/	1/	74169/132.0
8	2011-06-05T17:00:00.000Z/	07/	00/	1/	74169/156.0
9	2011-06-09T17:00:00.000Z/	19/	00/	1/	74169/97.0
10	2011-06-10T17:00:00.000Z/	19/	00/	1/	74169/95.0
11	2011-06-12T17:00:00.000Z/	13/	00/	2/	74169/119.0
12	2011-06-23T17:00:00.000Z/	07/	00/	1/	74169/139.0
13	2011-07-06T17:00:00.000Z/	01/	00/	0/	74169/226.0
14	2011-05-01T17:00:00.000Z/	07/	00/	1/	74169/39.0
15	2011-05-05T17:00:00.000Z/	07/	00/	1/	74169/105.0
16	2011-05-05T17:00:00.000Z/	19/	00/	2/	74169/131.0
17	2011-05-11T17:00:00.000Z/	01/	00/	2/	74169/131.0
18	2011-05-18T17:00:00.000Z/	13/	00/	1/	74169/169.0
19	2011-05-31T17:00:00.000Z/	07/	00/	1/	74169/103.0
20	2011-06-03T17:00:00.000Z/	01/	00/	1/	74169/161.0
21	2011-06-07T17:00:00.000Z/	19/	00/	2/	74169/84.0

Hình 4.95: Thông tin dữ liệu lũ dạng text



Hình 4.96: Chuyển thông tin dữ liệu lũ dạng text (txt) thành dạng cột (xls)

Sử dụng thuật toán nhỏ trong thư viện pandas để xử lý vấn đề chuẩn hóa vì đúng định dạng thời gian dữ liệu theo ngày/ tháng/ năm:

```
[4]  X=pd.read_csv('74162Sontay.csv')
```

```
[5]  X.head()
```

	DataDate	Hour	Minute	Flag	ID	WaterLevel Value
0	2010-12-24T17:00:00.000Z	7	0	2	74162	470
1	2010-12-26T17:00:00.000Z	19	0	1	74162	373
2	2010-12-28T17:00:00.000Z	7	0	2	74162	402
3	2010-12-29T17:00:00.000Z	1	0	1	74162	414
4	2011-01-06T17:00:00.000Z	1	0	2	74162	425

Hình 4.97: Thông tin ngày tháng dữ liệu lũ trước khi chuẩn hóa

```
[16]  m.head()
```

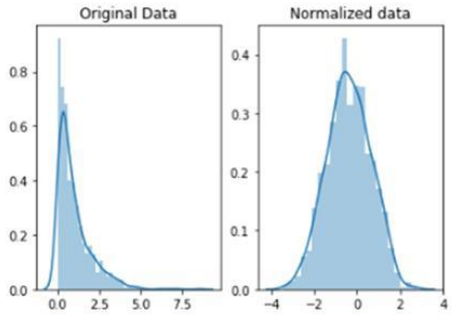
	DataDate	Hour	Minute	Flag	StationID	WaterLevel Value
0	2010-12-24	7	0	2	74162	470
1	2010-12-26	19	0	1	74162	373
2	2010-12-28	7	0	2	74162	402
3	2010-12-29	1	0	1	74162	414
4	2011-01-06	1	0	2	74162	425

Hình 4.98: Thông tin ngày tháng dữ liệu lũ sau khi chuẩn hóa

Dữ liệu lũ gồm nhiều đặc tính (cột), và mỗi đặc tính thì lại có các đơn vị và độ lớn nhỏ khác nhau. Điều này tác động tới tính hiệu quả của nhiều thuật toán, ví dụ thời gian thực hiện, quá trình hội tụ, hay thậm chí ảnh hưởng cả tới độ chính xác của thuật toán. Chính vì vậy, cần tiến hành điều chỉnh dữ liệu để các đặc tính cùng có chung một tỉ lệ (data scaling) và thường để các đặc tính có giá trị trong khoảng [0, 1]. Sử dụng thuật toán nhỏ trong thư viện pandas để điều chỉnh tỉ lệ dữ liệu trên dữ liệu mực nước như sau:

```
normalized_data = stats.boxcox(original_data)
# v? hình đ? so sánh
fig, ax=plt.subplots(1,2)
sns.distplot(original_data, ax=ax[0])
ax[0].set_title("Original Data")
sns.distplot(normalized_data[0], ax=ax[1])
ax[1].set_title("Normalized data"
```

Hình 4.99: Thuật toán điều chỉnh tỉ lệ dữ liệu trên dữ liệu mực nước (H)



Hình 4.100: Trực quan hoá kết quả chuẩn hóa dữ liệu lũ (H)

Nhận xét: Quá trình điều chỉnh và chuẩn hóa (normalization) dữ liệu lũ sẽ cho kết quả đầu ra hầu hết các giá trị nằm trong đoạn dữ liệu rất nhỏ. Lưu ý rằng, rất nhiều tập dữ liệu có các điểm kì dị.

4.8.2.6. Kết quả triển khai trực quan hóa dữ liệu lũ

Sử dụng thư viện nguồn mở pandas trong ngôn ngữ Python đọc file dữ liệu và trực quan hóa dữ liệu lũ ở dạng biểu đồ.

```
[4] X=pd.read_csv('74162Sontay.csv')
[5] X.head()
```

	DataDate	Hour	Minute	Flag	Station ID	WaterLevel Value
0	2010-12-24T17:00:00.000Z	7	0	2	74162	470
1	2010-12-26T17:00:00.000Z	19	0	1	74162	373
2	2010-12-28T17:00:00.000Z	7	0	2	74162	402
3	2010-12-29T17:00:00.000Z	1	0	1	74162	414
4	2011-01-06T17:00:00.000Z	1	0	2	74162	425

a)

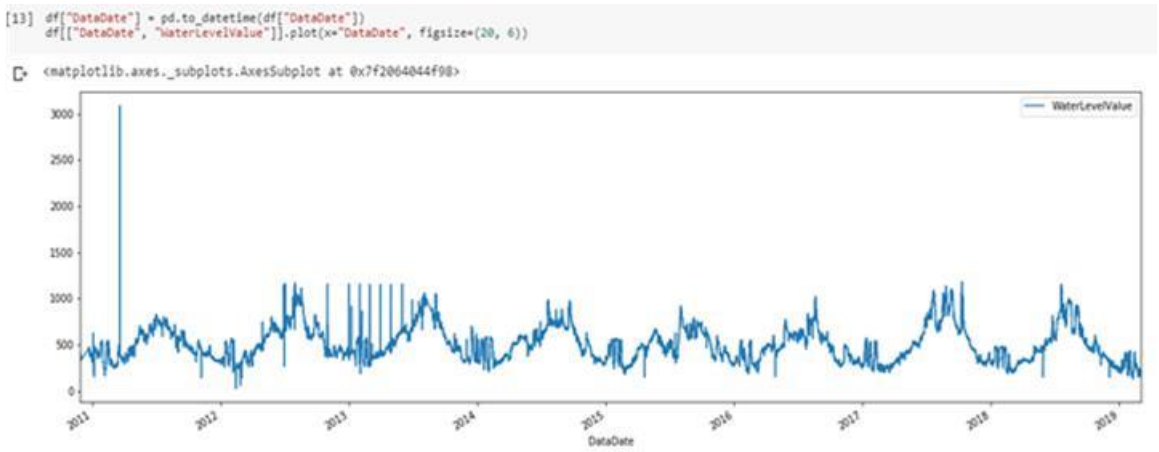
```
[6] import numpy as np
    import pandas as pd
    import seaborn as sns
[4] df=pd.read_csv('./processing74162Sontay.csv')
12] df.head()
```

	Unnamed: 0	DataDate	Hour	Minute	Flag	Station ID	WaterLevel Value
0	0	2010-12-24	7	0	2	74162	470
1	1	2010-12-26	19	0	1	74162	373
2	2	2010-12-28	7	0	2	74162	402
3	3	2010-12-29	1	0	1	74162	414
4	4	2011-01-06	1	0	2	74162	425

b)

Hình 4.101: Thuật toán đọc file dữ liệu (a) và trực quan dữ liệu lũ (b)

Kết quả trực quan hóa dữ liệu lũ như sau:



Hình 4.102: Kết quả trực quan hoá dữ liệu lũ (H) tại trạm Sơn Tây (2011-2019)

Nhận xét: Việc trực quan hóa dữ liệu lũ hỗ trợ cho chúng ta có thể quan sát sự thay đổi dữ liệu, chẳng hạn các dữ liệu bất thường, dị biệt trong quá trình thu thập dữ liệu, thông qua đó chọn được phương pháp học máy hiệu quả.

4.8.2.7. Kết quả triển khai phân hạng đặc trưng dữ liệu lũ

Sử dụng các thư viện skedm trong phương pháp Non-linear Methods để đọc, tính toán dữ liệu lũ cần phân hạng.

```python
from . import utilities
from scipy import stats as stats
from sklearn import neighbors
import numpy as np
from sklearn import metrics as skmetrics
```

a)

```python
def dist_calc(self,Xtest):
    """Calculates the distance from
    the testing set to the training set.
    Parameters
    ----------
    Xtest : 2D array
          Test features (nsamples, nfeatures).
    d,i = self.knn.kneighbors(Xtest)
    self.dist = d
    self.ind = i
    self.Xtest = Xtest
```

b)

Hình 4.103: Các thuật toán đọc (a) và tính toán để phân hạng dữ liệu lũ (b)

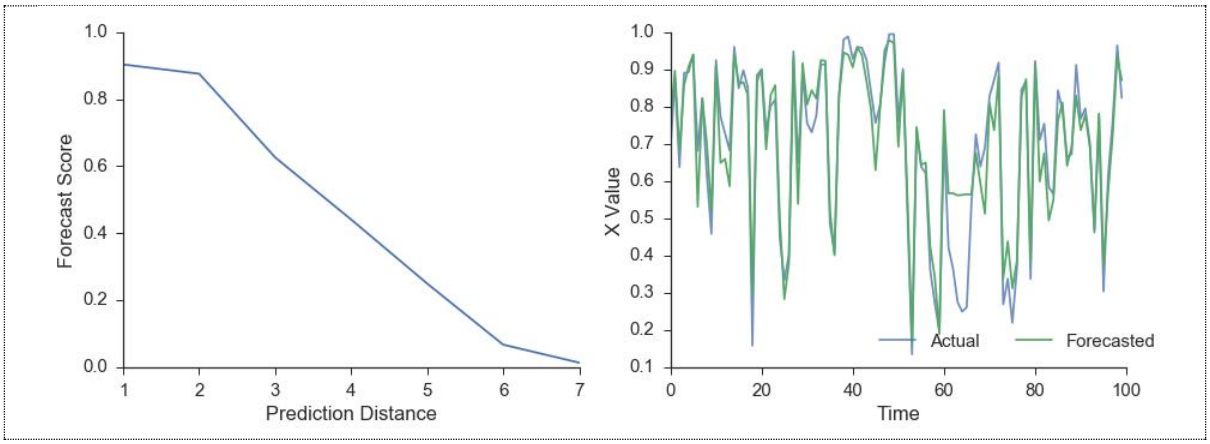
Sử dụng thuật toán K-nearest neighbors trong ML để dự đoán hạng đặc trưng của dữ liệu lũ:

```python
def dist_stats(self, nn_list):
    """Returns the mean and std of the
    """

    nn_list = np.array(nn_list)
    d = self.dist[:,nn_list-1]

    mean = np.mean(d,axis=0)
    std = np.std(d,axis=0)

    return mean, std
```

```python
def predict_individual(self,Xtest,nn_list):
    """Make a prediction for each neighbor.
    Parameters
    Xtest : 2d array
        Contains the test features.
    nn_list : 1d array of ints
        Neighbors to be tested.
    #calculate distances first
    self.dist_calc(Xtest)
    ypred = []
    for nn in nn_list:
        neigh_ind = self.ind[:,nn-1]# subtract 1
        ypred.append(self.ytrain[neigh_ind])
    self.ypred = ypred
    return ypred
```

Hình 4.104: Các thuật toán để phân hạng đặc trưng dữ liệu lũ

Kết quả phân hạng đặc trưng dữ liệu lũ như sau:



Hình 4.105: Kết quả phân hạng đặc trưng dữ liệu lũ

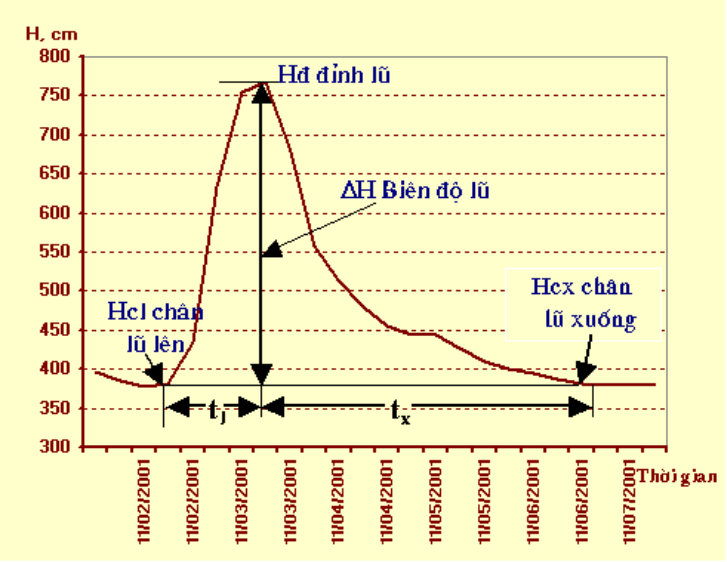
4.8.2.8. Kết quả triển khai lựa chọn đặc trưng dữ liệu lũ

Một cách tổng quát, sự đồng quan hệ giữa các đặc trưng được biết tới như sự dư thừa đặc trưng, trong khi sự đồng quan hệ giữa các lớp được xem như sự liên quan giữa các đặc trưng. Như vậy, toàn bộ tập đặc trưng có thể chia thành bốn phần: (1) các đặc trưng không liên quan, (2) các đặc trưng dư thừa, (3) các đặc trưng ít liên quan, và (4) các đặc trưng liên quan. Một thuật toán lựa chọn đặc trưng tốt nên lựa chọn các đặc trưng không dư thừa và liên quan.

Kết quả lựa chọn các đặc trưng chính dữ liệu của một trận lũ sau:

- Chân lũ lên là mực nước (Hcl) hay lưu lượng (Qcl) khi lũ bắt đầu lên.
- Đỉnh lũ là mực nước (Hđ) hay lưu lượng nước (Qđ) cao nhất trong một trận lũ.
- Thời gian lũ lên (tl) là khoảng thời gian từ khi lũ bắt đầu lên đến đỉnh lũ.
- Thời gian lũ xuống (tx) là khoảng thời gian đỉnh lũ đến khi hết lũ.
- Thời gian của một trận lũ (t) là khoảng thời gian từ khi lũ bắt đầu lên đến khi hết lũ (t = tl + tx).
- Biên độ mực nước lũ được xử lý dữ liệu là chênh lệch mực nước giữa mực nước đỉnh với mực nước khi lũ bắt đầu lên (DH). Biên độ lũ trên các sông miền núi có thể đạt 10-20 m, cá biệt, có nơi đạt trên 25 m (Lai Châu), ở vùng đồng bằng thường từ 3-8 m.

Kết quả lựa chọn đặc trưng dữ liệu lũ như sau:



Hình 4.106: Kết quả lựa chọn đặc trưng dữ liệu lũ

Nhận xét: Các đặc trưng dữ liệu lũ gồm: giá trị mực nước, lưu lượng tại thời điểm xuất hiện lũ và lũ đạt đỉnh; thời gian đạt đỉnh lũ, thời gian lũ; biên độ mực nước lũ.

4.8.3. Dự báo lũ bằng mô hình AI với tập dữ liệu huấn luyện

4.8.3.1. Công cụ thực hiện

- Các mô hình dự báo cho các thành phần **tuyến tính** theo chuỗi thời gian gồm: MA; ARMA; ARIMA (Mô hình trượt trung bình tích hợp tự hồi quy); SARIMA;

- Các mô hình dự báo cho các thành phần **phi tuyến** theo chuỗi thời gian gồm: LASSO; RF (Rừng ngẫu nhiên - Random Forest); SVR (Hồi quy vector hỗ trợ - Support Vector Regression); LR, KL, kNN (Thuật toán hàng xóm K gần nhất - K Nearest Neighbors).

Sử dụng mô hình lai để triển khai huấn luyện dự báo lũ, cụ thể:

- Sử dụng mô hình ARIMA để mô hình hóa thành phần lũ tuyến tính.

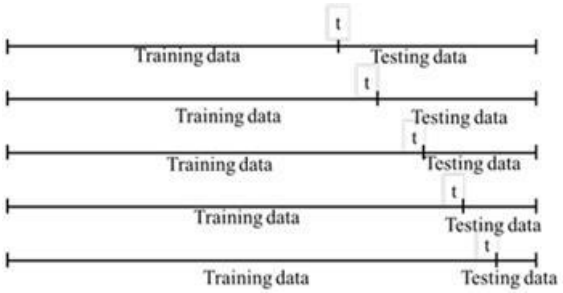
- Sử dụng SVM, RF, KNN hoặc LSTM để mô hình hóa các thành phần lũ phi tuyến.

4.8.3.2. Kết quả triển khai

Để đánh giá và so sánh thì tất cả dữ liệu thu thập được sẽ được chia thành 2 phần để huấn luyện và kiểm thử. Với dữ liệu trạm Vụ Quang và Hà Nội thì 80% dữ liệu để huấn luyện (từ 2008 đến 2015) và 20% (từ 2016-2017) còn lại được sử dụng để kiểm thử mô hình đã được huấn luyện. Triển khai huấn luyện dự báo lũ trên hệ thống sông Hồng với ARIMA, KNN, SVR, RF, LSTM và các mô hình lai được đề xuất. Thực hiện phát triển mô hình dự báo nhiều bước dựa trên phương pháp dự đoán 1 bước. Có nghĩa là để dự báo giá trị $\hat{y}(t)$ tại thời điểm t thì p các giá trị trong quá khư y(t-1), y(t-2)…y(t-p) được sử dụng. Ở bước dự báo tiếp theo thì $\hat{y}(t)$ sẽ được dùng như một giá trị đầu vào để dự báo giá trị $\hat{y}(t+1)$.

Tất cả mô hình ML và mô hình lai đều được huấn luyện một lần bằng tập dữ liệu huấn luyện và sau đó được áp dụng để dự báo tại các thời điểm t khác nhau trên tập dữ liệu kiểm thử. Tuy nhiên đối với ARIMA để có kết quả tốt hơn thì tất cả các

giá trị lịch sử cho đến thời điểm t-1 đều được sử dụng để xây dựng mô hình, nên mô hình này được sử dụng để dự báo T giá trị từ thời điểm t. Vì thế, với mỗi dự báo tại một thời điểm khác nhau thì cần phải huấn luyện lại mô hình ARIMA. Quá trình này được mô tả ở hình 4.108



Hình 4.107: Huấn luyện mô hình ARIMA tại các thời điểm khác nhau

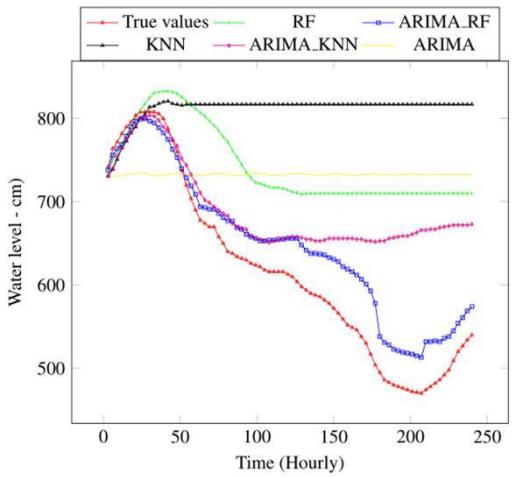
Để áp dụng phương pháp lai thì bước đầu tiên là xử lý các thành phần tuyến tính, phi tuyến của dữ liệu theo chuỗi thời gian và tham số thứ tự p. Trong bước tiếp theo cho mô hình 1 thì dữ liệu gốc và thành phần phi tuyến sẽ được tổng hợp như trong công thức dưới đây để tạo ra dữ liệu mới sử dụng cho phương pháp lai:

$$\hat{y}_{hybrid1}(t) = f(y(t-1), \ldots, y(t-p), \epsilon(t-1), \epsilon(t-2)) \quad (4.8.1)$$

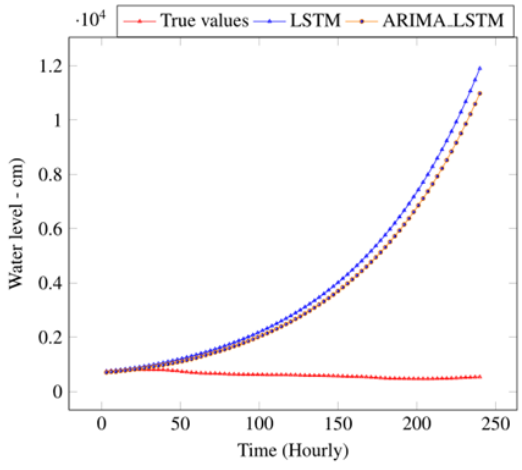
Với mô hình 2, thành phần dữ liệu phi tuyến được xử lý trực tiếp bằng thuật toán ML:

$$\hat{y}_{\epsilon}(t) = f(\epsilon(t-1), \ldots, \epsilon(t-p)) \quad (4.8.2)$$

Kết quả huấn luyện dự báo mực nước thời hạn 10 ngày trong điều kiện thời tiết bình thường (với dữ liệu tần suất 3h) sử dụng các phương pháp dự báo khác nhau trên tập dữ liệu tại trạm Hà Nội như sau:



a)

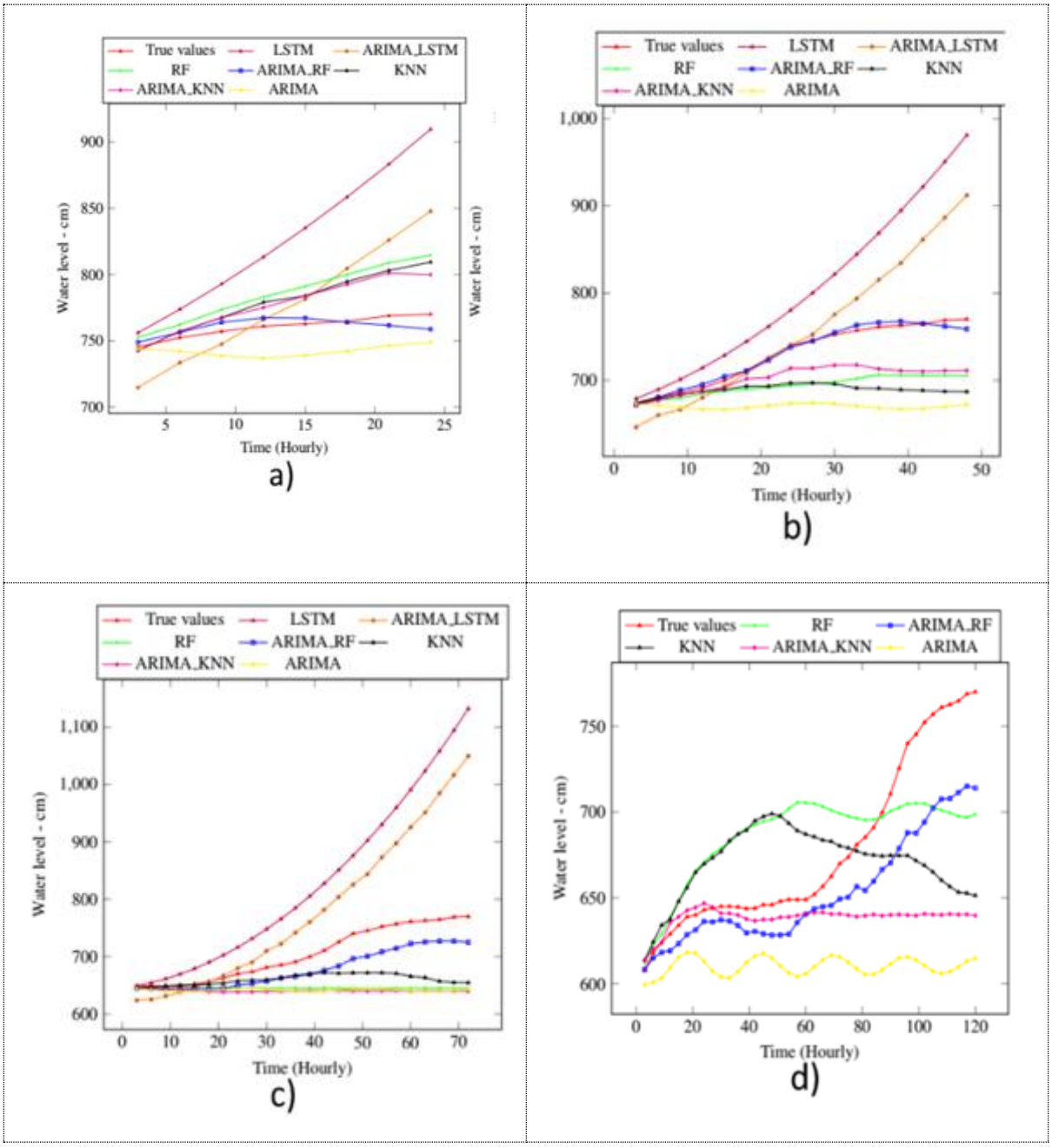


b)

Hình 4.108: So sánh dự báo H 10 ngày trên tập dữ liệu tại trạm Hà Nội

a): sử dụng phương pháp dự báo RF, ARIMA.RF, KNN, ARIMA.KNN, ARIMA; b): sử dụng phương pháp dự báo LSTM, ARIMA. LSTM.

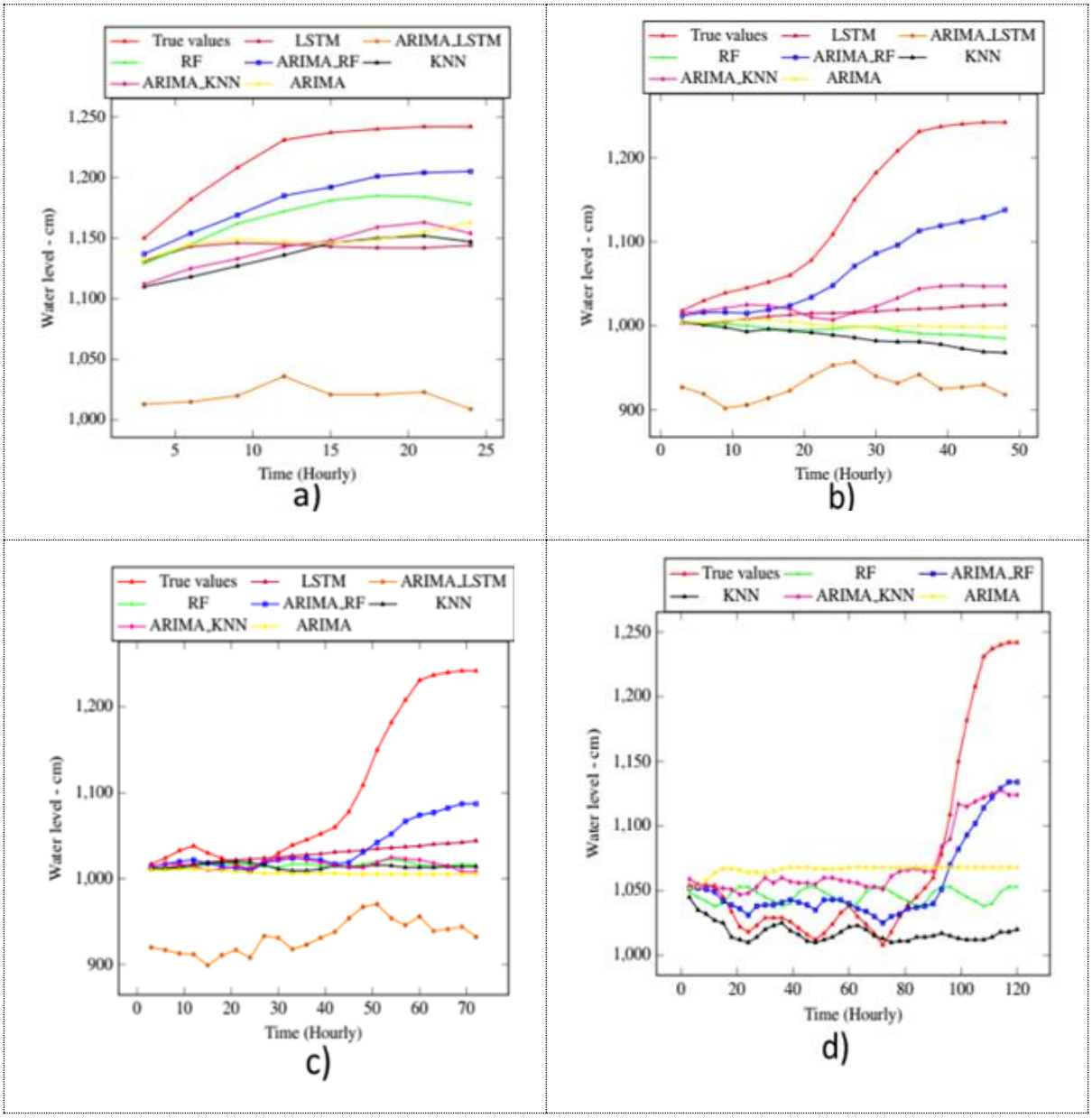
Kết quả huấn luyện dự báo mực nước trong điều kiện thời tiết nguy hiểm (với dữ liệu tần suất 3h) sử dụng các phương pháp dự báo khác nhau trên tập dữ liệu tại trạm Hà Nội như sau:



Hình 4.109: So sánh kết quả dự báo với các phương pháp khác nhau trên dữ liệu 24/7/2017 tại trạm Hà Nội

a) dự báo 24h trước khi đạt đỉnh, b) 72h trước khi đạt đỉnh, c)72h trước khi đỉnh, d) 5 ngày trước khi đỉnh.

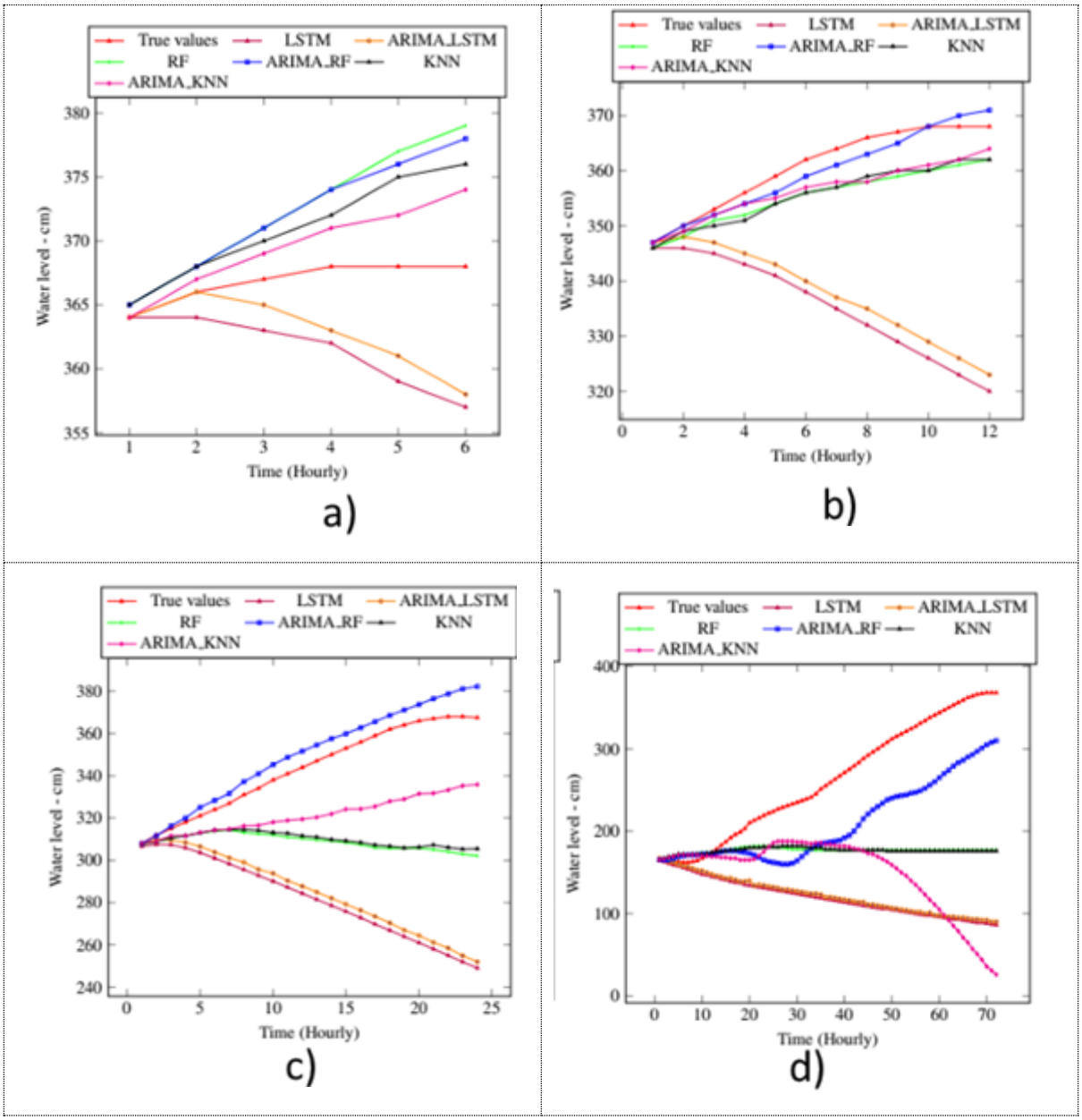
Kết quả huấn luyện dự báo mực nước trong điều kiện thời tiết nguy hiểm (với dữ liệu tần suất 3h) sử dụng các phương pháp dự báo khác nhau trên tập dữ liệu tại trạm Vụ Quang như sau:



Hình 4.110: So sánh kết quả dự báo với các phương pháp khác nhau trên dữ liệu 21/8/2016 tại trạm Vụ Quang

a) dự báo 24h trước khi đạt đỉnh, b) 72h trước khi đạt đỉnh, c)72h trước khi đỉnh, d) 5 ngày trước khi đỉnh

Kết quả huấn luyện dự báo mực nước trong điều kiện thời tiết nguy hiểm (với dữ liệu tần suất 3h) sử dụng các phương pháp dự báo khác nhau trên tập dữ liệu tại trạm Hưng Yên như sau:



Hình 4.111: So sánh kết quả dự báo với các phương pháp khác nhau trên dữ liệu 21/8/2016 tại trạm Hưng Yên

a) dự báo 24h trước khi đạt đỉnh, b) 72h trước khi đạt đỉnh, c)72h trước khi đỉnh, d) 5 ngày trước khi đỉnh.

Nhận xét:

Dự báo chính xác dữ liệu mực nước có cấu trúc chuỗi thời gian là một nhiệm vụ quan trọng trong việc cảnh báo lũ tuy nhiên cũng đầy thách thức. ARIMA, KNN, RF, SVM, và LSTM là những phương pháp dự báo hiệu quả và phổ biến rộng rãi đã được thử nghiệm với dữ liệu thủy văn. ARIMA có thể mô hình hóa thành phần tuyến tính tốt trong khi các mô hình học máy (ML) lại phù hợp với thành phần phi tuyến. Trong thực tế thì dữ liệu thủy văn có cấu trúc chuỗi thời gian thường bao gồm cả thành phần tuyến tính và phi tuyến tương quan lẫn nhau.

Việc triển khai đã thực hiện theo 2 phương pháp là mô hình lai (mô hình 1) là kết hợp từ dữ liệu gốc và phần dư sau khi áp dụng ARIMA để xây dựng mô hình dự đoán được gọi là ARIMA_KNN. ARIMA_RF, ARIMA_SVR và ARIMA_LSTM.

Và phương pháp lai thứ 2 là thực hiện dự đoán trên dữ liệu gốc và dữ liệu còn lại sau khi thực hiện ARIMA sau đó tổng hợp kết quả dự báo lại (mô hình 2). Các phương pháp này đã được thử nghiệm trên 3 bộ dữ liệu lớn thu thập tại 3 trạm trên sông Hồng (Vụ Quang, Hà Nội, Hưng Yên) và so sánh với từng mô hình riêng lẻ. Kết quả cho thấy phương pháp kết hợp ARIMA_RF và ARIMA_KNN của mô hình 1 vượt trội và đáng tin hơn các mô hình lai khác cũng như các phương pháp truyền thống. Mô hình ARIMA_RF và ARIMA_KNN là các mô hình cần phát triển thử nghiệm để dự báo mực nước trên các trạm thủy văn của các hệ thống sông trên phạm vi cả nước.

4.8.4. Tối ưu cấu hình, tham số và xác định độ tin cậy của hệ thống AI dự báo lũ

4.8.4.1. Công cụ thực hiện

a) Công cụ tối ưu cấu hình, tham số: Sử dụng ngôn ngữ R, các thư viện keras, tensorflow trên python để tối ưu các tham số phục vụ cho bài toán dự báo lũ tại trạm Vụ Quang, Hà Nội và Hưng Yên. Với mô hình ML, sử dụng phương pháp kiểm chứng chéo 5 lần để tìm ra giá trị tham số tối ưu của mỗi mô hình.Với mô hình ARIMA, sử dụng hàm auto.arima() từ package dự báo R.

b) Công cụ xác định độ tin cậy của mô hình AI dự báo lũ: Xây dựng mô hình học sâu DL (Deep learning), sử dụng giải thuật LSTM để dự báo lũ. Mô hình DL sử dụng ngôn ngữ lập trình Python và sử dụng các thư viện cải thiện tốc độ tính toán của Google Colab là PyTorch, TensorFlow, Keras... (GPU chạy code Python cho mục đích nghiên cứu).

c) Phương pháp để xác định độ tin cậy dự đoán của hệ thống AI dự đoán lũ là Accuracy score. Các chỉ số đánh giá độ tin cậy dự đoán của mô hình dự báo **lũ** sử dụng hồi quy Regression là:

- Sim: Chỉ số xác định % tương tự giữa giá trị dự báo và giá trị thực đo.
- MAE (Mean Absolute Error): Sai số tuyệt đối trung bình giữa giá trị dự báo và giá trị thực đo, khi MAE = 0, giá trị của mô hình hoàn toàn trùng khớp với giá trị quan trắc, mô hình được xem là lý tưởng.

- RMSE (Root mean square Error): Sai số bình phương trung bình quân phương, là giá trị trung bình của bình phương sai số giữa giá trị dự báo và giá trị thực đo tương ứng.

- R score: Hệ số tương quan giữa 2 biến dự báo và thực đo. Chỉ số này nói lên chất lượng của mô hình dự báo.

- FSD: Độ lệch chuẩn, FSD tiến tới 0 thì phương pháp này là hoàn hảo.

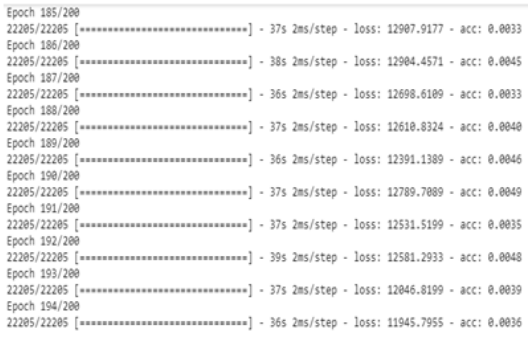
- NSE: Hệ số hiệu quả mô hình NashTHER để đánh giá khả năng dự báo của mô hình thủy văn. NSE càng cao thì giá trị dự báo càng sát giá trị quan trắc.

4.8.4.2. Kết quả triển khai tối ưu cấu hình, tham số cho mô hình sử dụng LSTM

Để có thể áp dụng được các phương pháp học máy cho bài toán dự báo mực nước, thực hiện chuyển dữ liệu 1 chiều mực nước thành dữ liệu nhiều chiều. Điều này có nghĩa là để dự đoán cho thời điểm t, sẽ dựa vào dữ liệu của p thời điểm trước đó (x(t-1), x(t-2),…,x(t-p)); trong đó, sử dụng giá trị p=5. Sau khi chuyển đổi dữ liệu thành công, thực hiện tối ưu hóa tham số cho phương pháp Long Short Term Memory (LSTM). Kết quả như sau:

```
#model.compile(optimizer='adam', loss='mse')


#Stacked LSTM (Nhiều lớp LSTM)
model = Sequential()
model.add(LSTM(50, activation='relu',
               return_sequences=True,
               batch_input_shape=(None,1,4)))
model.add(LSTM(50,
               return_sequences=False,
               activation='relu'))
model.add(LSTM(units=num_dimensions,
                                # return_sequences=True,
                                input_shape=(timestep, num_dimer
                                dropout=drop_out,
                                recurrent_dropout=drop_out
                                ))
model.add(Dense(1))
model.compile(optimizer='adam',
              loss='mse',
              metrics=['accuracy'])
```

Hình 4.112: Thực hiện tối ưu hóa tham số cho mô hình LSTM



```
Epoch 185/200
22205/22205 [==============================] - 37s 2ms/step - loss: 12907.9177 - acc: 0.0033
Epoch 186/200
22205/22205 [==============================] - 38s 2ms/step - loss: 12904.4571 - acc: 0.0045
Epoch 187/200
22205/22205 [==============================] - 36s 2ms/step - loss: 12698.6109 - acc: 0.0033
Epoch 188/200
22205/22205 [==============================] - 37s 2ms/step - loss: 12610.8324 - acc: 0.0040
Epoch 189/200
22205/22205 [==============================] - 36s 2ms/step - loss: 12391.1389 - acc: 0.0046
Epoch 190/200
22205/22205 [==============================] - 37s 2ms/step - loss: 12789.7089 - acc: 0.0049
Epoch 191/200
22205/22205 [==============================] - 37s 2ms/step - loss: 12531.5199 - acc: 0.0035
Epoch 192/200
22205/22205 [==============================] - 39s 2ms/step - loss: 12581.2933 - acc: 0.0048
Epoch 193/200
22205/22205 [==============================] - 37s 2ms/step - loss: 12046.8199 - acc: 0.0039
Epoch 194/200
22205/22205 [==============================] - 36s 2ms/step - loss: 11945.7955 - acc: 0.0036
```

Hình 4.113: Kết quả từng bước chạy tối ưu hóa tham số

Sau khi thực hiện tối ưu hóa tham số, bộ tham số tối ưu thu được là (Batchsize=3, dropout=0.3). Sau đó, sử dụng bộ tham số này để thực hiện dự đoán mực nước tại trạm Hà Nội. Kết quả dự đoán mực nước trước 48h (dữ liệu 8 obs/ ngày) như sau:

Bảng 4.20: Kết quả dự báo mực nước tại trạm Hà Nội thời hạn đến 24h sau khi tối ưu hóa tham số LSTM

True value	Kết quả dự báo		Sai số tuyệt đối	
	LSTM	LSTM-optimize	LSTM	LSTM-optimize

788	797.3	791.6	9.3	5.7
774	790.5	784.9	16.5	5.6
760	784.9	776.9	24.9	8.0
738	775.1	770.3	37.1	4.8
720	770.9	758.8	50.9	12.0
704	761.9	748.3	57.9	13.7
690	756.5	738.5	66.5	18.0
678	747.7	726.4	69.7	21.3
674	742.5	723.4	68.5	19.1
670	739.0	722.8	69.0	16.2
670	735.9	730.4	65.9	5.5
658	732.8	726.6	74.8	6.2
650	728.6	724.0	78.6	4.6
640	725.9	720.7	85.9	5.3
638	724.7	716.9	86.7	7.9
634	725.3	711.2	91.3	14.1
Sai số tuyệt đối trung bình			59.6	**10.5**

Nhận xét: Sau khi thực hiện tối ưu hóa các tham số, kết quả cho thấy mô hình dự báo đã chính xác hơn rất nhiều, sai số tuyệt đối trung bình giảm từ 60cm xuống 10.5 cm, trong khi đó sai số cho phép tại trạm Hà Nội là 25cm cho dự báo trước 24h. Các kết quả này cho thấy độ chính xác của dự báo tương đối cao, khả năng ứng dụng của việc tối ưu hóa tham số của các thuật toán học máy và học sâu để giải quyết các bài toán dự báo với dữ liệu thủy văn là rất khả quan trong thực tế.

4.8.4.3. Kết quả tối ưu cấu hình, tham số cho các mô hình lai (ML và ARIMA)

Với các mô hình lai sử dụng để huấn luyện dự báo thì kết quả dự báo cuối cùng là sự tổng hợp của dự báo tuyến tính bằng ARIMA và dự báo phi tuyến bằng phương pháp ML.Việc xác định được các tham số của các mô hình này đóng vai trò quan trọng. Những tham số này ảnh hưởng đến độ chính xác của dự báo đáng kể. Khi xây dựng dự báo bằng phương pháp ML, áp dụng phương pháp kiểm chứng chéo 5 lần để tìm ra giá trị tham số tối ưu của mỗi mô hình.Đối với ARIMA, áp dụng hàm auto.arima() từ package dự báo R để tìm ra tham số p,d và q. Đồng thời xét các thành phần theo mùa của tất cả tập dữ liệu khi huấn luyện mô hình ARIMA.Kết quả thiết lập tham số của các mô hình như sau:

Bảng 4.21: Các tham số được lựa chọn đối với các mô hình dự báo khác nhau

Dataset name		**Method**	
Station	ARIMA (p.d.q)	KNN	RF
Vu Quang	(5.1.1)	K=6	mtry=4. nlree= 100
Hanoi	(5,1,5)	K=7	mtry=4,ntree= 1000
HungYen	(5,1,3)	K=5	nitry=4, ntrce=300
		ARIMA-KNN	ARIMA.RF

Dataset name		Method	
Vu Quang		K=6	mtry=6. nlree=200
Hanoi		K=5	mtry=7,ntree=1000
HungYen		K=6	mtry=6, ntree= 100
		SVM	
Vu Quang		c=50, gamma= 0.001, kernel= rbf	
Hanoi		c=50, gamma= 0.001, kernel=rht'	
HungYen		c=50, gammas 0.01. kemel=rbf	
		ARIMA.SVM	
Vu Quang		c=50, gamma= 0.001. kemel=rbf	
Hanoi		c=5, gamma=0.001, kernel=rbf	
HungYen		c=50, gamma= 0.001, kemel=rbf	
Parameters of LSTM & AR1MA.LSTM for all experiments			
	Library: keras	Metrics: Accuracy	Loss function: MSE
	Epochs=200	Validation-	Batch size=3
	Optimizer function: Adam		

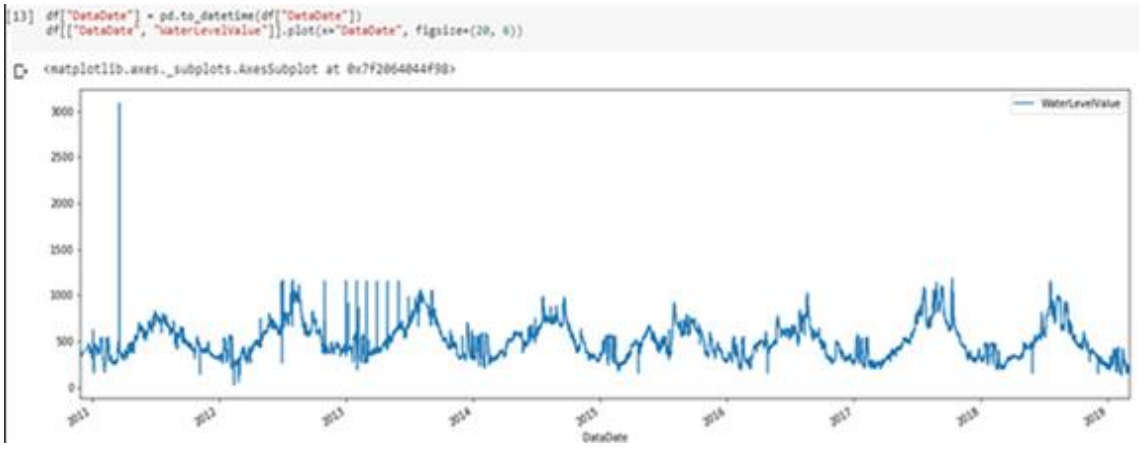
4.8.4.4. Kết quả triển khai xác định độ tin cậy của hệ thống AI dự báo lũ

Xây dựng các hàm xác định độ tin cậy của dự đoán như sau:

```
#Xây dựng hàm tính lỗi của mô hình dự báo
def mse(actual, predicted):
    """ Mean Squared Error """
    error = np.mean(abs(actual -  predicted))
    return round(error,3)
def rmse(actual, predicted):
    """ Root Mean Squared Error """
    error = np.sqrt(np.mean(np.square(actual -  predicted)))
    return round(error,3)
def nrmse(actual, predicted):
    """ Normalized Root Mean Squared Error """
    error = rmse(actual, predicted) / (actual.max() - actual.min())
    return round(error,3)
```

a) Kết quả xác định độ tin cậy dự báo của mô hình LSTM

Thực hiện huấn luyện mô hình và sử dụng bộ dữ liệu kiểm thử để đánh giá độ tin cậy của các dự đoán, với các dạng LSTM khác nhau cho kết quả khác nhau. Kết quả xác định độ tin cậy và dự đoán lũ tại trạm Sơn Tây với mô hình Vanilla LSTM như sau: MSE: 0.942; RMSE: 1.21; NRMSE: 0.073.



Hình 4.114: Độ tin cậy kết quả dự báo lũ tại trạm Sơn Tây với mạng Vanilla LSTM

b) Kết quả xác định độ tin cậy của các mô hình lai (ML và ARIMA)

Thực hiện huấn luyện mô hình và sử dụng bộ dữ liệu kiểm thử để đánh giá độ tin cậy của các dự đoán với các mô hình khác nhau cho kết quả khác nhau. Kết quả xác định độ tin cậy và dự đoán lũ tại Vụ Quang và trạm Hà Nội với mô hình lai và các thời hạn dự báo 12h, 24h, 48h, 72h và 5 ngày như sau:

Bảng 4.22: Độ tin cậy dự báo H tại trạm Vụ Quang và trạm Hà Nội của mô hình lai 1

Method	Size	Vu Quang - Model 1						Hanoi - Model 1					
		Sim	MAE	RMSE	FSD	R	NSE	Sim	MAE	RMSE	FSD	R	NSE
ARIMA	12h	0.67	10.0	11.63	1.15	0.92	-4.4	0.65	15.1	17.30	1.01	0.61	-3.08
LSTM		0.64	11.7	13.36	0.95	0.87	-5.4	0.63	17.0	20.10	0.62	**0.85**	-8.13
ARIMA_LSTM		0.39	42.1	43.08	**0.60**	0.72	-117	0.68	13.1	14.75	0.64	0.76	-3.16
RF		0.69	9.4	10.92	1.04	0.85	-2.4	0.72	11.3	13.16	0.70	0.84	-1.72
ARIMA_RF		**0.78**	**6.0**	**7.01**	0.64	**0.93**	**-0.2**	**0.81**	**7.0**	**8.26**	**0.37**	0.84	**-0.10**
KNN		0.71	8.7	9.95	0.88	0.84	-2.0	0.64	15.4	17.18	0.71	0.81	-4.10
ARIMA_KNN		0.71	7.1	7.96	0.74	0.91	-1.5	0.68	12.0	13.33	0.61	0.83	-2.13
SVM		0.62	12.8	14.76	0.95	0.92	-7.2	0.65	39.9	45.71	0.70	**0.85**	-28.1
ARIMA_SVM		0.61	20.6	25.50	0.84	0.93	-43	0.59	66.7	80.17	0.87	0.84	-72.3
ARIMA	24h	0.70	16.8	19.97	1.11	0.74	-3.2	0.69	26.3	30.63	1.06	0.63	-4.12
LSTM		0.67	18.9	21.81	0.88	0.66	-4.4	0.62	38.8	47.25	0.75	0.78	-27.2
ARIMA_LSTM		0.52	39.9	41.83	**0.48**	0.41	-43	0.70	25.0	30.95	0.76	0.71	-8.36
RF		0.71	15.7	18.89	0.93	0.57	-2.8	0.74	20.2	24.21	0.68	0.72	-3.38
ARIMA_RF		**0.77**	10.9	13.27	0.66	0.73	**-1.1**	**0.82**	**12.5**	**15.16**	**0.46**	0.82	**-0.34**
KNN		0.71	14.5	17.47	0.96	0.68	-2.6	0.71	23.7	27.50	0.81	0.74	-4.40
ARIMA_KNN		0.76	**10.5**	**12.30**	0.61	**0.87**	-1.2	0.75	18.2	20.98	0.61	**0.83**	-1.67
SVM		0.65	19.9	23.59	1.05	0.68	-5.8	0.68	58.5	66.25	0.80	0.80	-19.1
ARIMA_SVM		0.59	57.6	76.60	0.93	0.67	-90	0.57	113.1	132.92	0.84	0.78	-148
ARIMA	48h	0.74	29.7	36.93	1.42	0.59	-2.0	0.71	44.2	50.74	1.17	0.47	-2.88
LSTM		0.72	32.5	39.80	0.88	0.65	-3.2	0.58	92.9	113.27	0.74	0.70	-30.7
ARIMA_LSTM		0.65	42.3	48.31	0.73	0.44	-8.5	0.64	70.7	89.71	0.69	0.71	-16.8
RF		0.75	27.5	34.29	1.16	0.40	-1.5	0.74	35.9	42.42	0.65	0.68	-3.44
ARIMA_RF		0.80	18.1	22.21	0.70	0.74	-0.8	**0.82**	**20.0**	**23.62**	**0.42**	0.76	**-0.69**
KNN		0.75	27.9	35.61	1.05	0.53	-2.0	0.72	39.3	45.14	0.86	0.65	-3.12
ARIMA_KNN		**0.81**	**18.0**	**22.03**	**0.69**	**0.88**	**-0.6**	0.78	26.7	30.44	0.57	**0.81**	-0.86
SVM		0.71	35.7	43.78	1.11	0.61	-3.5	0.68	76.5	86.43	0.75	0.67	-28.3
ARIMA_SVM		0.55	115.2	144.50	0.97	0.59	-161	0.57	138.0	151.45	0.73	0.64	-53.2
ARIMA	72h	0.76	42.4	53.03	1.55	0.44	-1.7	0.72	55.3	62.61	1.25	0.35	-2.31
LSTM		0.73	45.2	55.50	0.94	0.62	-2.6	0.55	159.8	197.02	0.82	0.70	-42.1
ARIMA_LSTM		0.70	49.1	57.66	0.80	0.48	-3.3	0.60	130.3	165.92	0.76	0.72	-28.6
RF		0.76	40.2	51.00	1.26	0.36	-1.2	0.75	45.9	53.49	0.60	0.63	-2.57
ARIMA_RF		**0.82**	**25.5**	**31.47**	**0.72**	0.78	**-0.8**	**0.83**	**25.1**	**29.06**	**0.40**	0.78	**-0.21**
KNN		0.77	40.2	50.35	1.11	0.49	-1.2	0.73	49.6	56.79	0.86	0.56	-1.98
ARIMA_KNN		**0.82**	25.7	32.41	0.78	**0.86**	-1.3	0.80	30.8	34.72	0.50	**0.82**	-0.36
SVM		0.72	49.3	60.75	1.17	0.58	-3.9	0.70	80.9	89.82	0.72	0.60	-7.47
ARIMA_SVM		0.54	147.8	176.32	0.98	0.58	-105	0.59	140.9	152.66	0.59	0.57	-17.9
ARIMA	5 days	0.74	62.1	74.51	1.66	0.43	-2.1	0.73	70.5	79.36	1.40	0.28	-2.51
LSTM		0.71	68.2	82.20	0.78	0.66	-4.2	0.48	356.9	454.77	1.15	0.67	-130
ARIMA_LSTM		0.73	58.1	68.99	0.80	0.50	-1.9	0.52	309.6	402.00	1.11	0.67	-101
RF		0.75	58.4	70.43	1.35	0.34	-1.5	0.76	56.9	64.73	0.54	0.62	-2.26
ARIMA_RF		0.80	**37.41**	**44.38**	**0.77**	**0.79**	**-1.0**	**0.82**	**30.8**	**34.82**	**0.34**	**0.78**	**-0.32**
KNN		0.76	54.5	65.42	1.18	0.40	-1.4	0.73	64.4	73.17	0.94	0.56	-2.54
ARIMA_KNN		**0.81**	37.43	45.59	0.82	0.78	-1.9	**0.82**	34.7	39.15	0.54	0.77	-0.45
SVM		0.72	64.9	76.99	1.08	0.54	-3.4	0.72	81.8	91.08	0.69	0.59	-4.19
ARIMA_SVM		0.51	182.3	206.60	0.91	0.57	-79	0.61	138.4	149.74	0.53	0.58	-12.4

Bảng 4.23: Xác định độ tin cậy dự báo H tại trạm Hưng Yên của mô hình lai 1

Method	Sim	MAE	RMSE	FSD	R	NSE	Sim	MAE	RMSE	FSD	R	NSE
			6h						48h			
ARIMA	0.78	12.95	14.66	0.6	0.93	-0.71	0.81	24.2	29.34	0.93	0.6	-0.38
LSTM	0.85	8.92	10.16	0.3	0.97	0.42	0.79	28.45	36.06	0.52	0.52	-0.98
ARIMA_LSTM	0.82	9.87	11.54	0.31	0.93	0.32	0.8	27.96	35.64	0.48	0.5	-0.91
RF	0.77	13.16	15.18	0.46	0.92	-2.35	0.76	36.05	45.18	**0.41**	0.36	-2.14
ARIMA_RF	**0.87**	**6.43**	**7.37**	**0.28**	**0.98**	**0.44**	0.78	30.92	38.19	0.42	0.48	-1.4
KNN	0.8	11.41	13.1	0.37	0.93	-1.79	0.79	29.27	36.7	0.45	0.32	-1.1
ARIMA_KNN	0.82	9.79	11.28	0.39	0.97	-0.25	**0.82**	**23.15**	**29.01**	0.48	**0.67**	**-0.35**
SVM	0.79	16.18	19.65	0.52	0.86	-2.12	0.76	36.75	44.19	0.49	0.38	-3.06
ARIMA_SVM	0.73	25.34	27.74	0.41	0.93	-6.08	0.6	77.71	86.65	0.56	0.46	-15.62
			12h						72h			
ARIMA	0.79	17.25	19.92	0.36	0.86	-1.07	0.8	28.9	34.87	1.1	0.48	-0.62
LSTM	0.83	13.66	16.24	**0.33**	0.86	0.05	0.79	33.7	41.51	0.66	0.45	-1.22
ARIMA_LSTM	0.82	14.23	16.75	**0.33**	0.85	0.04	0.79	33.04	40.94	0.61	0.43	-1.14
RF	0.77	20.95	25.37	0.35	0.81	-1.63	0.74	42.52	52.43	0.39	0.33	-3.17
ARIMA_RF	**0.86**	**10.42**	**12.84**	0.41	0.88	**0.36**	0.76	38.95	48.15	**0.35**	0.32	-2.44
KNN	0.81	16.86	20.66	0.34	0.81	-0.91	0.8	30.59	37.91	0.46	0.32	-0.92
ARIMA_KNN	0.82	14.52	17.16	0.58	**0.9**	-0.39	**0.82**	**28.05**	**34.72**	0.4	**0.64**	**-0.61**
SVM	0.78	24.22	27.65	0.44	0.83	-2.54	0.76	38.53	46.53	0.5	0.25	-2.22
ARIMA_SVM	0.7	36.92	42.9	0.5	0.79	-12.52	0.59	85.69	93.73	0.47	0.44	-17.58
			24h						5 days			
ARIMA	**0.82**	19.81	24.16	0.6	**0.75**	-0.62	0.80	**33.78**	**40.35**	1.35	0.31	-0.90
LSTM	**0.82**	21.48	27.22	0.42	0.67	-0.58	0.79	40.49	48.30	0.84	0.38	-1.31
ARIMA_LSTM	**0.82**	19.94	25.29	**0.41**	0.66	-0.44	0.79	39.47	47.33	0.80	0.37	-1.21
RF	0.79	26.11	33.26	0.52	0.6	-1.22	0.75	49.86	60.77	**0.34**	0.24	-2.97
ARIMA_RF	**0.82**	19.92	25.42	0.46	0.59	-0.27	0.75	47.87	58.74	0.36	0.35	-3.09
KNN	0.79	24.78	31.94	0.47	0.56	-0.99	0.80	36.17	44.50	0.47	0.29	**-0.85**
ARIMA_KNN	0.81	**19.78**	**23.98**	0.67	**0.75**	**-0.26**	**0.81**	37.09	45.00	0.43	**0.53**	-1.20
SVM	0.77	32.14	38.01	0.54	0.67	-3.48	0.77	42.95	51.50	0.58	0.20	-1.67
ARIMA_SVM	0.66	58.15	67.13	0.68	0.64	-11	0.56	113.02	122.27	0.45	0.32	-17.11

Bảng 4.24: Độ tin cậy dự báo H tại trạm Vụ Quang của mô hình 2

Method	Sim	MEA	RMSE	FSD	R	NSE	Sim	MEA	RMSE	FSD	R	NSE
			12h						72h			
ARIMA	0.67	10.04	11.63	1.15	0.92	-4.38	0.756	42.35	53.03	1.55	0.44	-1.65
ARIMA_LSTM2	0.671	10.26	11.835	1.18	0.89	-4.45	0.758	42.212	52.868	1.557	0.454	-1.65
ARIMA_RF2	0.666	10.31	11.86	1.14	0.78	-4.48	0.758	42.191	52.832	1.552	0.422	-1.64
ARIMA_KNN2	0.669	10.37	11.889	1.13	0.84	-4.71	0.758	42.212	52.853	1.561	0.44	-1.64
ARIMA_SVM2	0.665	10.5	12.004	1.16	0.81	-4.91	0.757	42.26	52.887	1.568	0.45	-1.65
			24h						5 days			
ARIMA	0.7	16.82	19.97	1.11	0.74	-3.18	0.74	62.12	74.51	1.66	0.43	-2.06
ARIMA_LSTM2	0.7	16.6	19.77	1.11	0.74	-3.17	0.737	62.7	75.277	1.667	0.426	-2.05
ARIMA_RF2	0.7	16.61	19.75	1.11	0.73	-3.14	0.737	62.67	75.255	1.664	0.421	-2.04
ARIMA_KNN2	0.7	16.64	19.75	1.11	0.75	-3.19	0.737	62.69	75.263	1.669	0.42	-2.04
ARIMA_SVM2	0.69	16.77	19.88	1.13	0.72	-3.25	0.737	62.71	75.273	1.675	0.427	-2.04
			48h									
ARIMA	0.74	29.69	36.93	1.42	0.59	-2.04						
ARIMA_LSTM2	0.75	29.37	36.71	1.44	0.56	-2.04						
ARIMA_RF2	0.75	29.36	36.68	1.42	0.57	-2.04						
ARIMA_KNN2	0.75	29.37	36.7	1.44	0.56	-2.04						
ARIMA_SVM2	0.75	29.45	36.76	1.45	0.56	-2.05						

4.8.5. Giải thích dự đoán và ra quyết định thống kê của hệ thống AI dự báo lũ

4.8.5.1. Kết quả giải thích dự đoán của hệ thống AI dự báo lũ

- Theo kết quả dự báo H trung bình theo các thuật toán khác nhau trên tập dữ liệu quan trắc tại trạm Vụ Quang và Hà Nội với phương pháp lai **mô hình 1**, có thể giải thích dự đoán của các mô hình dự báo lũ như sau:

+ Kết quả cho thấy rằng ARIMA_RF đạt MAE và RMSE thấp nhấp nhất, FSD và NSE và Sim cao nhất trong hầu hết các giai đoạn dự báo.

+ Xếp thứ 2 là **ARIMA_KNN** cho thấy các chỉ số tốt và kết quả tốt nhất với dữ liệu trạm Vụ Quang dự báo 48h; điều này chứng tỏ rằng ARIMA_RF và ARIMA_KNN chính xác hơn các phương pháp khác và có thể xử lý triệt để các thành phần tuyến tính và phi tuyến tách biệt.

+ Kết quả của ARIMA_SVM,SVM, ARIMA_LSTM,LSTM cho thấy ARIMA_SVM và ARIMA_LSTM không cải thiện so với SVM và LSTM tương ứng; điều đó cho thấy, khi kết hợp thành phần phi tuyến với dữ liệu gốc thì cả SVM và LSTM **không xử lý tốt** thành phần phi tuyến của dữ liệu.

- Theo kết quả dự báo H trung bình theo các thuật toán khác nhau trên tập dữ liệu quan trắc tại trạm Hưng Yên với phương pháp lai **mô hình 1** (Bảng 4.23), có thể giải thích dự đoán của các mô hình dự báo lũ như sau: Từ kết quả cho thấy **ARIMA_RF và ARIMA_KNN** không đạt kết quả tốt với khoảng dự báo như ở trạm Hà Nội và Vụ Quang nhưng vẫn có hiệu năng cải thiện nếu so sánh với các mô hình học máy như RF, KNN. Các phương pháp ARIMA_SVM và ARIMA_LSTM không xử lý tốt thành phần phi tuyến của dữ liệu, thậm chí còn kém hơn cả các phương pháp học máy SVM và LSTM.

Theo kết quả dự báo H trung bình theo các thuật toán khác nhau trên tập dữ liệu quan trắc tại trạm Vụ Quang với phương pháp lai **mô hình 2** (Bảng 4.24), có thể giải thích dự đoán của các mô hình dự báo lũ như sau: phương pháp này không hiệu quả hơn ARIMA. Nhìn chung, phương pháp lai sử dụng mô hình 1 vượt trội hơn đáng kể so với phương pháp lai sử dụng mô hình 2.

4.8.5.2. Kết quả ra quyết định dự báo lũ trong điều kiện thời tiết bình thường

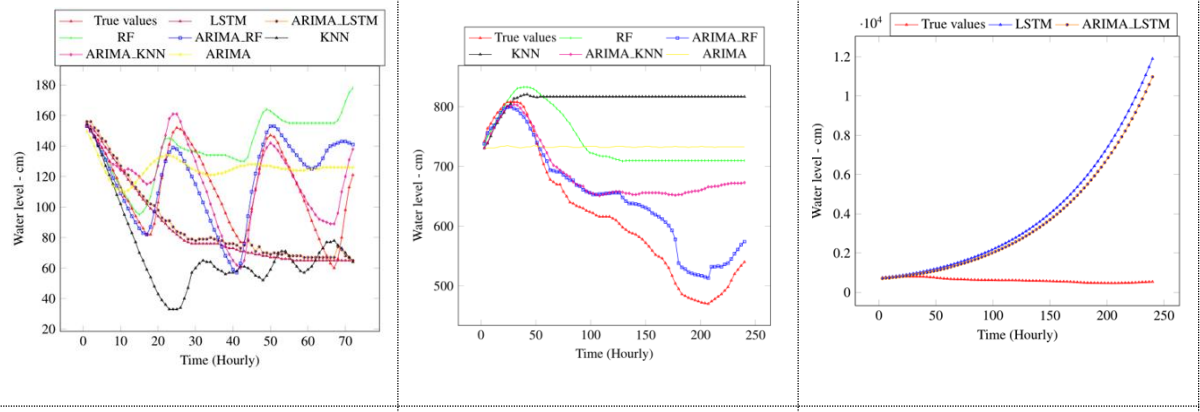
So sánh giá trị quan trắc với giá trị dự báo mực nước (H) tại các trạm quan trắc (Hà Nội, Vụ Quang và Hưng Yên) để phân tích, đánh giá độ tin cậy chính xác và các quyết định thống kê của các mô hình **dự báo lũ** trong điều kiện thời tiết bình thường với các thời hạn dự báo khác nhau (24h, 48h, 72h và 5 ngày) tại như sau:

- ARIMA_RF cho kết quả tốt nhất trong ngày đầu tiên, kết quả gần giống với thực tế tuy nhiên khi dự báo cho ngày thứ 2 và thứ 3 thì kết quả kém hơn ngày đầu nhưng sai số dự báo nằm trong ngưỡng quy định.Ngược lại ARIMA KNN dự báo không tốt ngày đầu tiên nhưng lại tương đối tốt với 2 ngày tiếp. Phương pháp RF, KNN, LSTM,ARIMA tạo ra giá trị dự đoán khởi đầu tốt nhưng nhanh chóng không nắm bắt được xu hướng của dữ liệu thực.

- ARIMA không phù hợp với dự báo mực nước trên sông Hồng mặc dù ban đầu cho kết quả tốt (hình 4.115a).

- ARIMA_RF và ARIMA_KNN thể hiện tính ưu việt cho dự báo mực nước với bộ dữ liệu Hà Nội trong 5 ngày (hình 4.115b).

- LSTM và ARIMA_LSTM có hiệu suất không tốt khi dự báo 10 ngày (hình 4.115c).



a) b) c)

Hình 4.115: So sánh giá trị dự báo H với giá trị quan trắc tại trạm với các phương pháp khác nhau

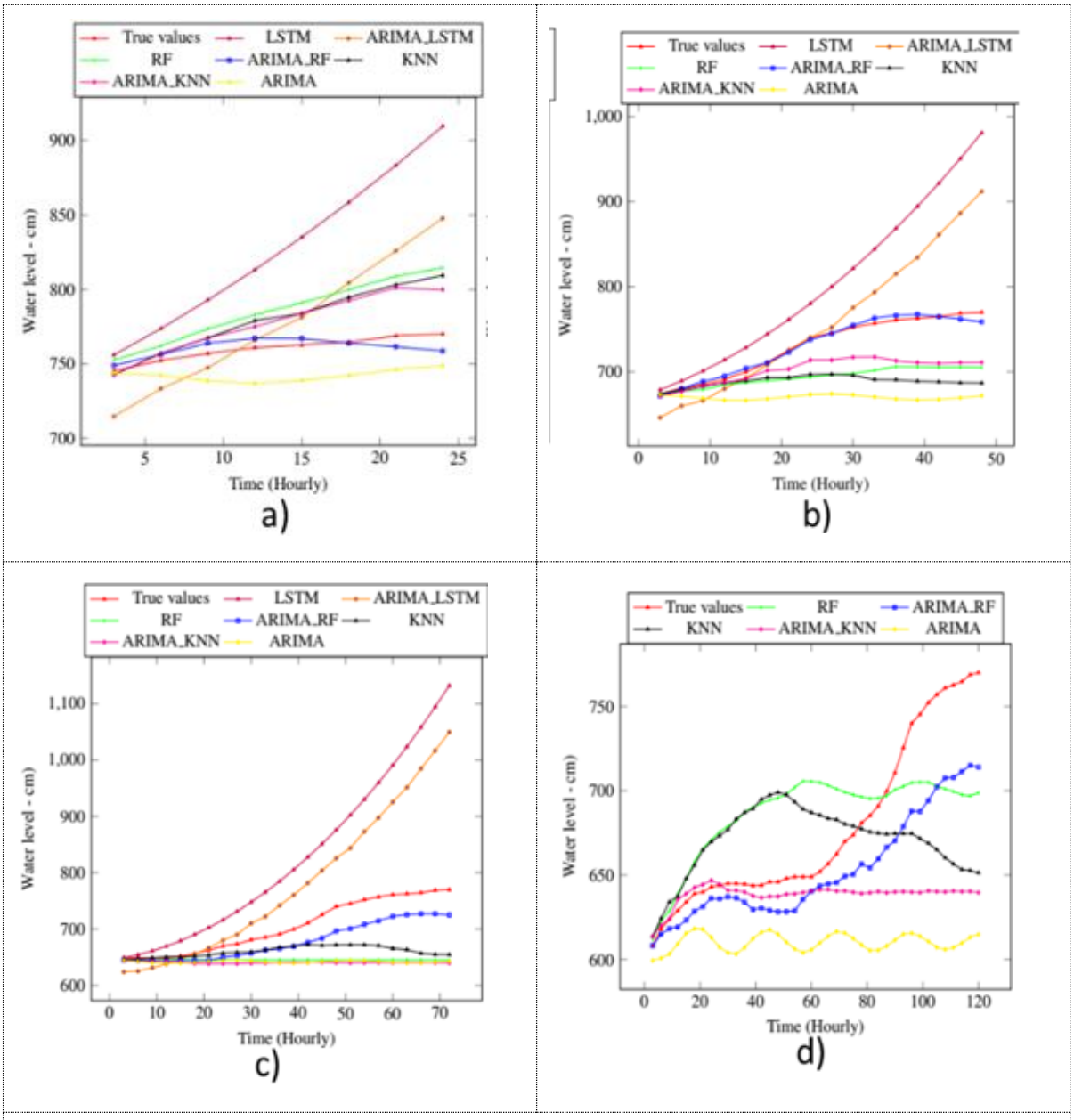
a) Dự báo 72 giờ với giá trị quan trắc (1 giờ/ lần) của trạm Hưng Yên; b) Dự báo 10 ngày với giá trị quan trắc (3 giờ/ lần) của trạm Hà Nội; c) Dự báo 10 ngày sử dụng các phương pháp khác nhau trên tập dữ liệu tại trạm Hà Nội.

*4.8.5.3. Kết quả quyết định dự báo **lũ** trong điều kiện thời tiết nguy hiểm*

So sánh giá trị quan trắc với giá trị dự báomực nước (H) tại các trạm quan trắc (Hà Nội, Vụ Quang và Hưng Yên) để phân tích, đánh giá độ tin cậy chính xác và các quyết định thống kê của các mô hình **dự báo lũ** với các thời hạn dự báo khác nhau (24h, 48h, 72h và 5 ngày) trong điều kiện thời tiết nguy hiểm như sau:

- ARIMA_RF cho kết quả tốt nhất dự báo 1 ngày 2 ngày tương đối gần với giá trị đo thực tế, với dự báo 3 ngày, 5 ngày thì ARMIA_RF có thể nắm bắt xu

hướng của dữ liệu gần với thực tế tuy nhiên các lỗi dự báo vẫn khá cao. Các phương pháp khác không đạt hiệu quả tốt (hình 4.116 a, b,c, d).



Hình 4.116: So sánh kết quả dự báo đưa ra bởi các phương pháp khác nhau trên dữ liệu ngày 24/7/2017 tại trạm Hà Nội

a) dự báo 24h trước khi đạt đỉnh, b) 48h trước khi đạt đỉnh; c) 72h trước khi đạt đỉnh; d) 5 ngày trước khi đạt đỉnh (với dữ liệu 3h/ lần)

Kết luận: Trong tương lai, đề tài sẽ áp dụng ARIMA_RF và ARIMA_KNN để dự báo mực nước tại các trạm thuộc các hệ thống sông trên phạm vi cả nước.

4.8.6. Trình diễn kết quả dự báo lũ

4.8.6.1. Công cụ thực hiện

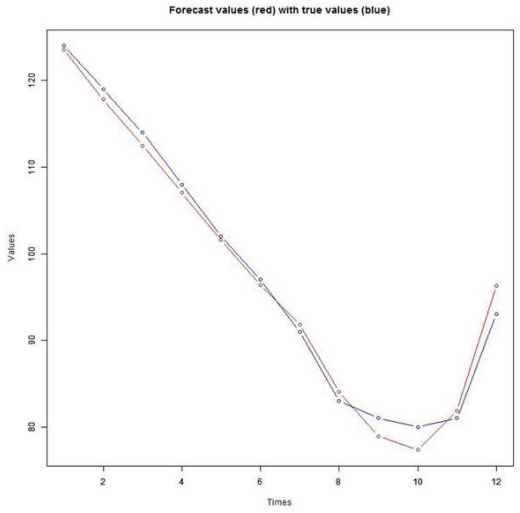
Các phương pháp trình diễn kết quả dự báo lũ gồm: biểu đồ đường (Line Graph), biểu đồ thanh, biểu đồ histogram, biểu đồ hộp, biểu đồ phân tán.

4.8.6.2. Kết quả triển khai

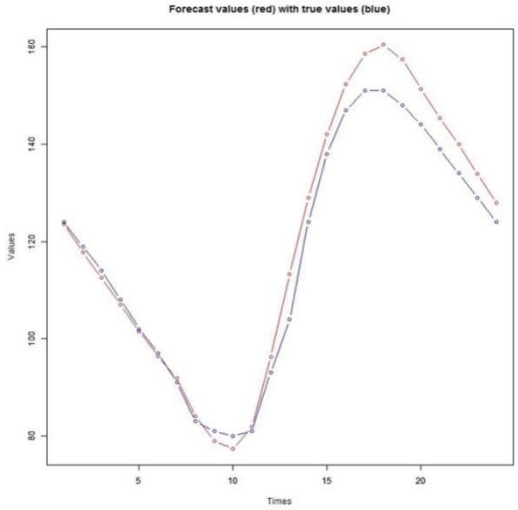
Kết quả trình diễn kết quả dự báo mực nước bằng mô hình FFNN với các thời hạn dự báo tại trạm Hưng Yên cụ thể như sau:

Bảng 4.25: RMSE và C giữa giá trị dự báo và quan trắc tại trạm Hưng Yên

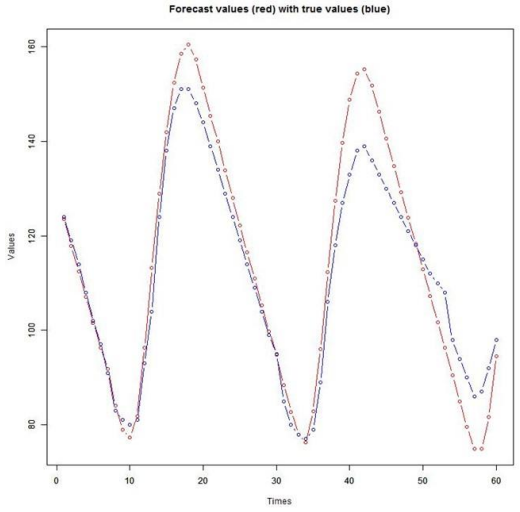
Thời hạn dự báo	Corelation	RMSE
12h	0.9946	1.5858
24h	0.9950	4.9634
36h	0.9947	4.4038
48h	0.9884	7.0686
60h	0.9776	7.3734



a)



b)



c)

Hình 4.117: Trình diễn kết quả dự báo H thời hạn 24h (a), 24h (b), 5 ngày (c) bằng FFNN

Nhận xét: Kết quả dự báo mực nước tại trạm Hưng Yên bằng mô hình FFNN có RMSE thấp hệ số tương quan C giữa giá trị dự báo và giá trị thực cao, cho thấy độ tin cậy dự báo của mô hình FFNN cao, khả năng ứng dụng của mô hình để dự báo lũ theo chuỗi thời gian là rất khả quan.

4.8.7. Kết luận

Dự báo chính xác dữ liệu mực nước có cấu trúc chuỗi thời gian là một nhiệm vụ quan trọng trong việc cảnh báo lũ. ARIMA, KNN, RF, SVM, và LSTM là những phương pháp dự báo hiệu quả và phổ biến rộng rãi đã được thử nghiệm với dữ liệu thủy văn. ARIMA có thể mô hình hóa thành phần tuyến tính tốt trong khi các mô hình học máy (ML) lại phù hợp với thành phần phi tuyến. Trong thực tế thì dữ liệu thủy văn có cấu trúc chuỗi thời gian thường bao gồm cả thành phần tuyến tính và phi tuyến tương quan lẫn nhau.

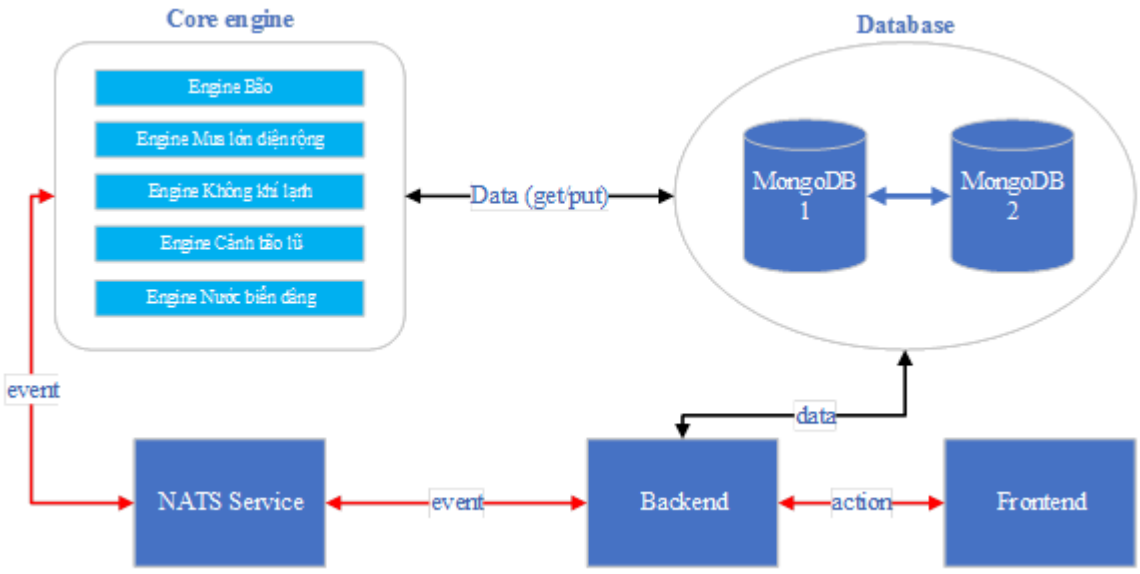
Vì vậy, đề tài đã đề xuất mô hình lai (mô hình 1) là kết hợp từ dữ liệu gốc và phần dư sau khi áp dụng ARIMA để xây dựng mô hình dự đoán được gọi là ARIMA_KNN. ARIMA_RF, ARIMA_SVR và ARIMA_LSTM. Phương pháp lai thứ 2 là thực hiện dự đoán trên dữ liệu gốc và dữ liệu còn lại sau khi thực hiện ARIMA sau đó tổng hợp kết quả dự báo lại (mô hình 2). Phương pháp này đã được thử nghiệm trên 3 bộ dữ liệu lớn thu thập tại 3 trạm tại sông Hồng và so sánh với từng mô hình riêng lẻ. Kết quả cho thấy phương pháp kết hợp ARIMA_RF và ARIMA_KNN của mô hình 1 vượt trội và đáng tin hơn các mô hình lai khác cũng như các phương pháp truyền thống. Phương pháp ARIMA_RF và ARIMA_KNN cần nghiên cứu phát triển để dự báo mực nước trên các trạm thủy văn khác của sông Hồng cũng như các con sông khác.

4.9. Thiết lập và triển khai framework tích hợp các module AI dự báo KTTV

4.9.1. Mô hình, kiến trúc và nguyên lý hoạt động của framework AI KTTV

4.9.1.1. Mô hình của hệ thống framework AI KTTV

Mô hình hệ thống framework tích hợp tích hợp các module AI hỗ trợ dự báo KTTV (gọi tắt là framework AI KTTV) cụ thể như sau:



Hình 4.118: Sơ đồ mô hình hệ thống Framework AI hỗ trợ dự báo KTTV

Mô hình hệ thống Framework tích hợp các module AI hỗ trợ dự báo KTTV gồm: khối cơ sở dữ liệu Bigdata MongoDB; khối Core Engine; khối Back-end API; khối hiển thị Front-end.

4.9.1.2. Nguyên lý hoạt động của Framework AI hỗ trợ dự báo KTTV

Khi có hành động yêu cầu từ phía giao diện người dùng Front-end sẽ được gọi tới các API tương ứng tại Back-end. Tại Back-end sẽ phát sinh ra sự kiện tương ứng với yêu cầu cụ thể, các sự kiện này được gửi tới NATS Service. NATS là một hệ thống nhắn tin mã nguồn mở PubSub đơn giản, an toàn và hiệu suất cao cho các ứng dụng đám mây, IoTmessaging, và kiến trúc microservices. Sự kiện được NATS chuyển hệ thống Core Engine, tại đây sự kiện sẽ được nhận dạng tương ứng với loại Engine thời tiết nguy hiểm cụ thể. Sau khi nhận dạng, hệ thống tiến hành lấy dữ liệu học máy tương ứng từ cơ sở dữ liệu làm đầu vào quá trình học máy với các tham số hiệu chỉnh tự động hoặc các tham số hiệu chỉnh từ các dự báo viên. Kết quả học máy được lưu trữ vào cơ sở dữ liệu và trình bày trên giao diện hiển thị phía Front-end cho người dùng.

4.9.1.3. Thiết kế chức năng của Framework AI hỗ trợ dự báo KTTV

Hệ thống Framework tích hợp các module AI dự báo KTTV bao gồm các chức năng cơ bản sau:

Bảng 4.26: Các chức năng của Framework tích hợp các module AI dự báo KTTV

STT	Tên chức năng	Mô tả chức năng
1)	Quản trị nhóm người dùng và người dùng hệ thống	Quản trị nhóm người dùng và người dùng hệ thống trên giao diện phía front-end được gọi tới các API tương ứng phía Back-end
2)	Tạo lập và chuẩn bị dữ liệu	Kết nối vào cơ sở dữ liệu và trích xuất các dữ liệu tương ứng với từng loại mô hình thời tiết nguy hiểm
3)	Nhận dạng các hiện tượng thời tiết nguy hiểm	Nhận dạng loại hình thời tiết nguy hiểm trên cơ sở dữ liệu học máy đầu
4)	Quản trị tham số hệ thống	Quản trị các tham số về loại mô hình AI và về trường dự báo KTTV
5)	Quản trị thời gian dự báo	Quản trị về thời hạn dự báo của mô hình AI dự báo KTTV
6)	Quản trị huấn luyện dự báo	Quản trị các thông tin về huấn luyện mô hình AI dự báo KTTV
7)	Quản trị mô hình dự báo	Quản trị các thông tin của mô hình AI dự báo KTTV

4.9.1.4. Thiết kế quy trình nghiệp vụ của Framework AI hỗ trợ dự báo KTTV

Quy trình nghiệp vụ của hệ thống Framework tích hợp các module AI dự báo KTTV cụ thể như sau:

Bảng 4.27: Quy trình nghiệp vụ của Framework tích hợp các module AI dự báo KTTV

STT	Tên quy trình nghiệp vụ	Mô tả
1	Quản trị người dùng và hệ thống	Thực hiện phân quyền người dùng trên hệ thống và quản trị thông tin dữ liệu, bảo mật an toàn hệ thống.
2	Cập nhật, lưu trữ dữ liệu vào CSDL	Thực hiện việc cập nhật và lưu trữ toàn bộ các loại dữ liệu cần sử dụng của hệ thống vào CSDL
3	Huấn luyện và hiệu chỉnh tham số đầu của mô hình AI	Thực hiện toàn bộ việc huấn luyện và hiệu chỉnh các tham số đầu vào của mô hình AI dự báo các hiện tượng thời tiết nguy hiểm
4	Hiển thị các kết quả dự báo	Thực hiện hiển thị các kết quả nhận dạng, dự báo KTTV của mô hình AI trên giao diện đồ họa
5	Lưu trữ kết quả nhận dạng, dự báo	Thực hiện lưu trữ toàn bộ các kết quả nhận dạng, dự báo hiện tượng thời tiết nguy hiểm của mô hình AI vào CSDL

STT	Tên quy trình nghiệp vụ	Mô tả
6	Thực thi lại quá trình học máy	Thực hiện lại quá trình học máy để tạo mẫu nhận dạng, dự báo các hiện tượng thời tiết nguy hiểm
7	Hiệu chỉnh lại các tham số đầu của mô hình AI	Thực hiện việc hiệu chỉnh lại các tham số đầu vào của mô hình AI nhận dạng, dự báo các hiện tượng thời tiết nguy hiểm
8	Quy trình tự động điều chỉnh các tham số đầu vào của mô hình AI	Thực hiện việc tự động điều chỉnh các tham số đầu vào của mô hình AI theo hiệu chỉnh của dự báo viên với các hiện tượng thời tiết nguy hiểm
9	Quy trình lưu log hoạt động của hệ thống	Thực hiện lưu log hoạt động của hệ thống, luồng dữ liệu, thao tác người dùng
10	Quy trình cung cấp đầu API	Cung cấp các đầu API cho bên thứ 3 để đưa dữ liệu đầu vào, hiệu chỉnh tham số và thực thi mô hình AI nhận dạng, báo các hiện tượng thời tiết nguy hiểm

4.9.1.5. Thiết kế kiến trúc của Framework AI hỗ trợ dự báo KTTV

a) Kiến trúc ứng dụng của Framework AI hỗ trợ dự báo KTTV

Kiến trúc ứng dụng của hệ thống Framework tích hợp các module AI dự báo KTTV bao gồm 03 lớp:

- Lớp giao tiếp người dùng: Cung cấp giao diện cho phép người dùng khai thác và hiển thị kết quả học máy và dự báo các hiện tượng KTTV nguy hiểm. Ngoài ra còn cung cấp các công cụ để hiệu chỉnh các tham số đầu vào của học máy.

- Lớp tính toán xử lý dữ liệu: Xử lý dữ liệu được người dùng nhập vào hoặc dữ liệu được người dùng tìm kiếm từ lớp lưu trữ để trả về các kết quả phù hợp phục vụ cho mục đích lưu trữ hoặc hiển thị.

- Lớp lưu trữ: Lưu trữ dữ liệu đã được xử lý bởi lớp tính toán xử lý dữ liệu.

b) Kiến trúc dữ liệu của Framework AI hỗ trợ dự báo KTTV

Kiến trúc dữ liệu của hệ thống Framework tích hợp các module AI dự báo KTTV bao gồm:

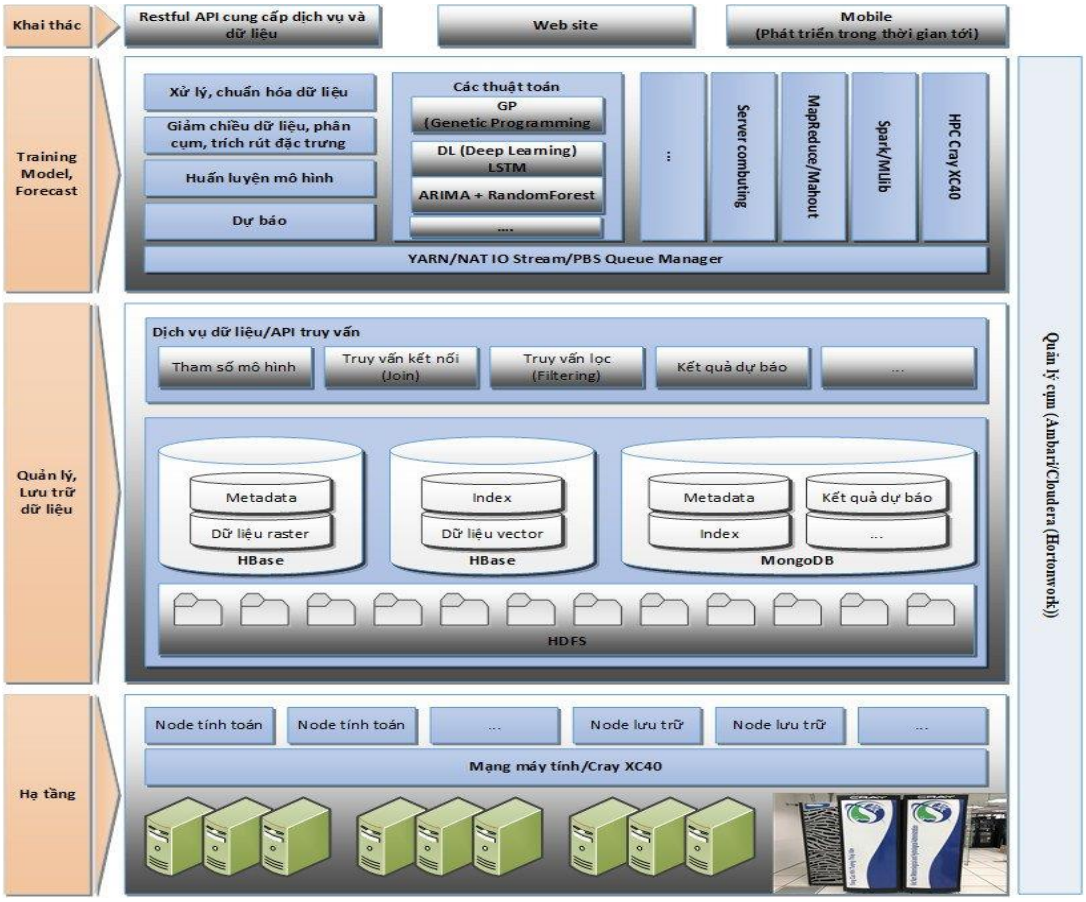
- Thiết kế danh mục dữ liệu chính và các bảng mã: bao gồm các danh mục: người dùng, nhóm người dụng, nhóm API, trường dự báo, trung tâm dự báo, danh trạm quan trắc, sông, ….;

- Các dữ liệu hoạt động chính (dữ liệu gốc): Dữ liệu P0, PP, Rain 1h, Rain 6h, nước biển dâng, nhiệt độ, mực nước, gió;

- Các dữ liệu thứ cấp: Các dữ liệu được xử lý, nội suy ra từ các dữ liệu gốc.

c) Kiến trúc cơ sở hạ tầng CNTT của Framework AI hỗ trợ dự báo KTTV

Kiến trúc cơ sở hạ tầng CNTT của hệ thống Framework tích hợp các module AI dự báo KTTV (hình 4.100) bao gồm các khối sau: Khối hạ tầng kỹ thuật (máy chủ, kết nối mạng, ….); khối quản lý và lưu trữ dữ liệu (thiết bị và công cụ lưu trữ, …); khối huấn luyện dự báo (hệ thống máy tính hiệu năng cao, các công cụ huấn luyện dự báo, ….) ; Khối khai thác hệ thống.



Hình 4.119: Kiến trúc cơ sở hạ tầng CNTT của hệ thống Framework AI KTTV

4.9.1.6. Thiết kế biểu đồ hoạt động, tuần tự và usecase của Framework AI hỗ trợ dự báo KTTV

Các biểu đồ hoạt động, biểu đồ tuần tự và biểu đồ Usecase của Framework gồm: (i) Biểu đồ hoạt động chung; (ii) Biểu đồ tạo lập dữ liệu, kết nối vào kho dữ liệu KTTV dùng chung để lấy dữ liệu làm đầu vào cho các mô-đun nhận dạng; (iii) Biểu đồ nhận dạng và dự báo các hiện tượng KTTV nguy hiểm; (iv) Biểu đồ lưu trữ kết quả của các mô-đun nhận dạng vào kho dữ liệu KTTV dùng chung; (v) Biểu đồ đánh giá kết quả nhận dạng; (vi) Biểu đồ tự động học máy, điều chỉnh các tham số; (vii) Biểu đồ quản trị người dùng hệ thống; (vii) Biểu đồ Usecase.

Chi tiết các hình vẽ biểu đồ tại Sản phẩm chính số 6: Báo cáo thiết lập hệ thống Framework tích hợp và kết nối được với các mô-đun nhận dạng, hỗ trợ dự báo, cảnh báo các hiện tượng KTTV nguy hiểm.

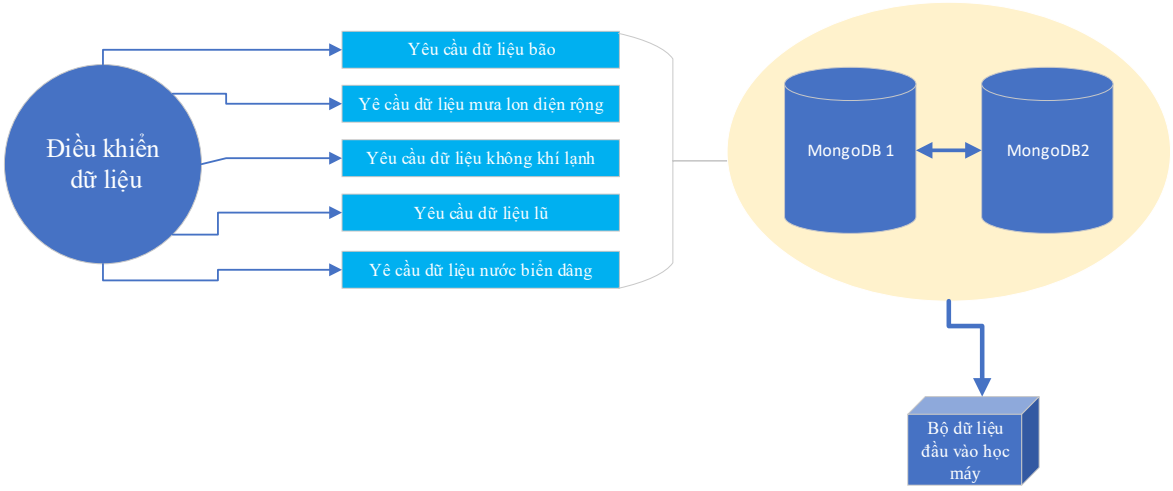
4.9.2. Thiết lập và triển khai module phân hệ dữ liệu nguồn

4.9.2.1. Thiết lập các module phân hệ dữ liệu nguồn

a) Mô tả chức năng

Module này chịu trách nhiệm kết nối vào cơ sở dữ liệu MongoDB trích xuất dữ liệu tương ứng với từng loại mô hình để làm đầu vào cho các module huấn luyện học máy cho mô hình AI hỗ trợ dự báo với từng loại thời tiết cụ thể bao gồm: dữ liệu để huấn luyện mô hình AI hỗ trợ dự báo bão, dữ liệu để huấn luyện mô hình AI hỗ trợ dự báo mưa lớn diện rộng, dữ liệu để huấn luyện mô hình AI hỗ trợ dự báo không khí lạnh, dữ liệu để huấn luyện mô hình AI hỗ trợ dự báo lũ, dữ liệu để huấn luyện mô hình AI hỗ trợ dự báo nước biển dâng do bão.

b) Nguyên lý hoạt động



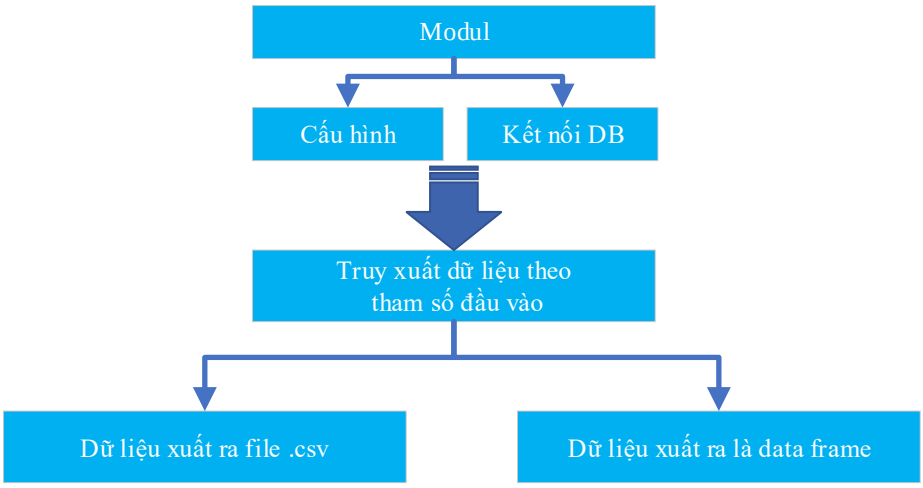
Hình 4.120: Sơ đồ nguyên lý của module quản lý và khai thác dữ liệu nguồn

Khối điều khiển dữ liệu nhận lệnh từng loại dữ liệu đầu vào học máy tương ứng các loại hình hiện tượng thời tiết tương ứng; thực hiện kết nối vào CSDL MongoDB và truy xuất các dữ liệu theo các tham số yêu cầu đầu vào để khai thác các tệp dữ liệu cần thiết cho các module AI huấn luyện dự báo.

4.9.2.2. Triển khai module tạo lập, lưu trữ, khai thác dữ liệu nguồn

Sử dụng ngôn ngữ lập trình Python 3.6 và công cụ lập trình Visual Studio 2017 để triển khai module tạo lập, lưu trữ, khai thác dữ liệu nguồn.

Sơ đồ và các triển khai của module truy xuất dữ liệu từ MongoDB làm đầu vào cho các mô hình AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm như sau:



Hình 4.121: Các bước thực hiện của module truy xuất dữ liệu từ MongoDB

Bước 1: Đọc file cấu hình, kết nối vào cơ sở dữ liệu MongoDB. Toàn bộ các tham số cấu hình của hệ thống nằm tại file cấu hình có tên dev_config.json. Khi chương trình khởi hoạt sẽ đọc toàn bộ cấu hình từ file cấu hình để thực hiện kết nối vào MongoDB. Các hàm thuật toán thực hiện kết nối tới CSDL MongoDB với các tham số cấu hình bao gồm: địa chỉ máy chủ, user, password và xác thực. Cấu hình tham số tương ứng với các loại dữ liệu thời tiết nguy hiểm.

Bước 2: Thực hiện truy xuất dữ liệu từ MongoDB. các hàm lấy dữ liệu tương ứng các loại hình thời tiết nguy hiểm theo các tham số đầu vào để lấy dữ liệu chuẩn bị cho các Module học máy theo các định dạng dữ liệu khác nhau.

4.9.3. Triển khai module quản trị hệ thống Framework

4.9.3.1. Triển khai module quản trị nhóm người dùng hệ thống

Triển khai module nhóm người dùng hệ thống Framework như sau:

- GET/groups/: lấy toàn bộ danh sách các nhóm người dùng của hệ thống.
- GET/groups/{code}: hiển thị chi tiết 1 nhóm người dùng với tham số mã của nhóm.
- POST/groups/: Phương thức khởi tạo 1 nhóm người dùng mới trên hệ thống.
- PUT /groups/{code}: phương thức sửa thông tin 1 nhóm theo tham số đầu vào là mã nhóm.

4.9.3.2. Triển khai module quản trị người dùng hệ thống

Triển khai module quản trị người dùng hệ thống gồm:

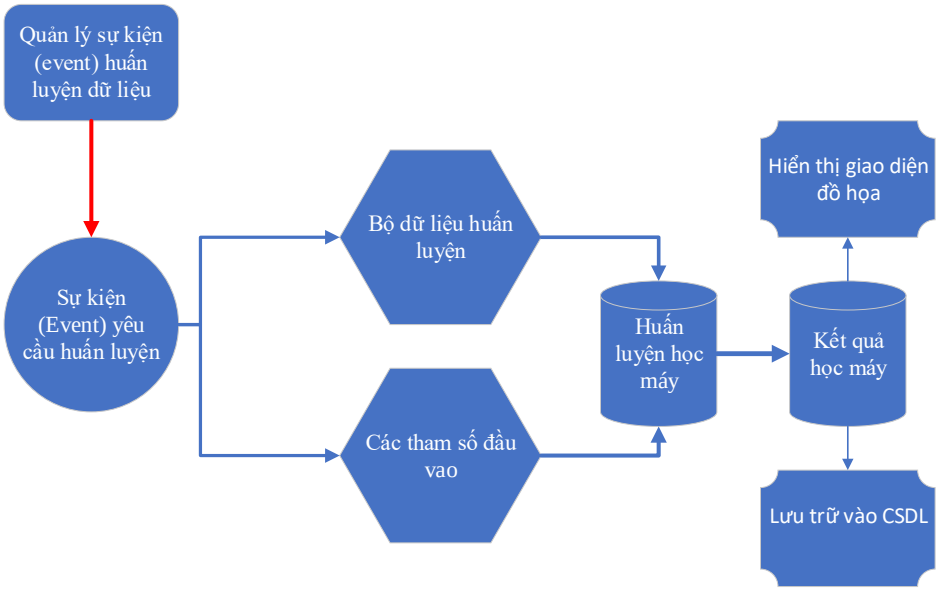
- POST /users/login: xử lý đăng nhập của người dùng. Khi người dùng đăng nhập thành công hệ thống sẽ tự sinh 1 chuỗi token.

- GET /users/: hiển thị toàn bộ thông tin chi tiết của người dùng.
- POST /users/: thực hiện khởi tạo 1 người dùng mới trong hệ thống với các thông tin yêu cầu.
- GET /users/{user_name}: lấy toàn bộ thông tin người dùng cụ thể theo tham số tên người dùng.
- PUT /users/{user_name}: cho phép sửa đổi thông tin của 1 người dùng cụ thể.

4.9.4. Triển khai các module tích hợp các mô hình AI nhận dạng và huấn luyện dự báo KTTV

4.9.4.1. Sơ đồ cấu trúc mô hình AI nhận dạng và huấn luyện dự báo

Sơ đồ cấu trúc của các module nhận dạng và huấn luyện dự báo các hiện tượng thời tiết nguy hiểm gồm: Khối quản lý dữ liệu sự kiện; khối huấn luyện nhận dạng, dự báo; khối hiển thị và lưu trữ kết quả nhận dạng, dự báo.



Hình 4.122: Sơ đồ cấu trúc của các module nhận dạng, hỗ trợ dự báo KTTV

4.9.4.2. Nguyên lý hoạt động của các module nhận dạng và huấn luyện dự báo KTTV

Với mỗi loại hình thời tiết nguy hiểm khác nhau, cụ thể ở đây là 5 loại thời tiết nguy hiểm bao gồm: bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão sẽ có những bộ dữ liệu học máy đầu vào tương ứng được truy xuất từ cơ sở dữ liệu MongoDB. Ứng mỗi loại hình thời tiết nguy hiểm, các module nhận dạng sẽ tự động thực thi mô hình tương ứng, thuật toán học máy từ các bộ dữ liệu đầu vào để nhận dạng được chính xác hiện tượng thời tiết nguy hiểm. Tiếp theo, các module hiệu chỉnh tự động các tham số đầu của hệ thống vào hoặc hiệu chỉnh thủ công các

tham số đầu vào theo nhận định của dự báo viên về hiện tượng thời tiết nguy hiểm cụ thể. Sau đó, các kết quả huấn luyện nhận dạng, dự báo sẽ được hiển thị, trình diễn trên giao diện đồ họa. Kết quả huấn luyện dự báo bởi các mô hình học máy cũng được lưu trữ vào Big data.

4.9.4.3. Triển khai các module quản trị tham số mô hình AI

Triển khai module quản trị tham số của mô hình được gọi thông qua API:

- GET /parameters/: lấy thông tin toàn bộ các tham số.
- GET /parameters/{code}: hiển thị thông tin chi tiết của 1 tham số cụ thể với đầu vào là mã tham số mô hình.
- POST /parameters/: thêm 1 tham số mới.
- PUT /parameters/{code}: Sửa thông tin 1 tham số cụ thể với mã tham số mô hình tương ứng.

4.9.4.4. Triển khai module quản trị loại mô hình AI

Triển khai module quản trị loại mô hình AI được gọi thông qua API:

- GET /modeltypes/: lấy toàn bộ thông tin các loại mô hình.
- GET /modeltypes/{code}: hiển thị thông tin 1 loại mô hình cụ thể theo mã.
- POST /modeltypes/: thêm 1 loại mô hình mới.
- PUT /modeltypes/{code}: sửa thông tin 1 loại mô hình cụ thể theo mã.

4.9.4.5. Triển khai module quản trị mô hình

Triển khai module quản trị mô hình AI được gọi thông qua API:

- GET /modelapis/: hiển thị thông tin toàn bộ các mô hình.
- GET /modeapis/{code}: lấy thông tin 1 mô hình cụ thể theo mã.
- POST /modelapis/: tạo mới 1 mô hình.
- PUT /modelapis/{code}: sửa thông tin 1 mô hình cụ thể theo mã.

4.9.4.6. Triển khai module quản trị trường dự báo của mô hình AI

Triển khai module quản trị trường dự báo của mô hình AI được gọi thông qua API:

- GET/fieldforecasts/: lấy thông tin toàn bộ loại trường dự báo trong hệ thống.
- GET/fieldforecasts/{code}: lấy thông tin 1 trường dự báo cụ theo mã.
- POST/fieldforecasts/: tạo mới 1 trường dự báo.

- PUT/fieldforecasts/{code}: sửa thông tin 1 trường dự báo theo mã.

4.9.4.7. Triển khai module quản trị thời gian dự báo

Triển khai module quản trị thời gian dự báo của mô hình AI được gọi thông qua API:

- GET/datetimeforecasts/: lấy toàn bộ thông tin thời gian dự báo trong hệ thống.
- GET/datetimeforecast/{code}: lấy thông tin 1 thời gian dự báo cụ thể theo mã.
- POST/datetimeforecasts/: tạo 1 thời gian dự báo mới.
- PUT/datetimeforecasts/{code}: chỉnh sửa thông tin 1 thời gian dự báo cụ thể theo mã.

4.9.4.8. Triển khai module quản trị huấn luyện mô hình AI

Triển khai module quản trị huấn luyện dự báo của mô hình AI được gọi thông qua API:

- GET /modeltrains/: hiển thị thông tin toàn bộ các huấn luyện mô hình trong hệ thống.
- POST /modeltrains/: tạo mới 1 huấn luyện mô hình.

4.10. Kết chương 4

Các nội dung Chương 4 trên đã trình bày các nội dung nghiên cứu về xây dựng và triển khai hệ thống AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm gồm: (i) Xây dựng mô hình hệ thống học máy/ AI để hỗ trợ dự báo KTTV; (ii) Triển khai các hệ thống AI để hỗ trợ dự báo bão, mưa lớn diện rộng, không khí lạnh khu vực Bắc Bộ; hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng; hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ; (iii) Thiết lập và triển khai Framework tích hợp các module AI hỗ trợ dự báo KTTV.

Chương tiếp theo sẽ trình bày kết quả chuyển giao, thử nghiệm và đánh giá hệ thống AI hỗ trợ dự báo bão khu vực Bắc Bộ; hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng; hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ tại Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc Bộ và Bắc Trung Bộ.

5. CHƯƠNG 5: CHUYỂN GIAO, ĐÀO TẠO, THỬ NGHIỆM VÀ ĐÁNH GIÁ

5.1. Xây dựng quy trình vận hành và đào tạo vận hành hệ thống

5.1.1. Xây dựng quy trình vận hành hệ thống AI hỗ trợ dự báo KTTV

Tài liệu quy trình vận hành hệ thống hỗ trợ dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão bằng mô hình AI bao gồm:

a. Mô tả chung của hệ thống AI hỗ trợ nhận dạng và dự báo các hiện tượng KTTV nguy hiểm.

b. Quy trình quản trị Website gồm các bước: Bước 1: Đăng ký người dùng; Bước 2: Phân nhóm người dùng; Bước 3: Phân quyền truy cập dữ liệu; Bước 4. Kiểm soát hoạt động tức thời của người dùng.

c. Quy trình vận hành hệ thống AI hỗ trợ dự báo KTTV gồm:

- Quy trình thiết lập tham số và huấn luyện mô hình AI gồm các bước: Bước 1. Xác định trường dự báo; Bước 2. Xác định loại mô hình; Bước 3. Xác định tham số mô hình; Bước 4. Xác định thời hạn dự báo; Bước 5. Thực thi huấn luyện mô hình.

- Quy trình dự báo một số hiện tượng KTTV nguy hiểm gồm: Quy trình vận hành module dự báo bão; Quy trình vận hành module dự báo mưa lớn diện rộng; Quy trình vận hành module dự báo không khí lạnh; Quy trình vận hành module dự báo lũ; Quy trình vận hành module dự báo nước biển dâng do bão.

d. Quy trình vận hành hệ thống hạ tầng gồm: Quy trình kiểm tra tình trạng máy chủ; Quy trình kiểm tra các kết nối mạng; kiểm tra Big data; Kiểm tra dịch vụ Website.

Tài liệu quy trình vận hành hệ thống hỗ trợ dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão bằng mô hình AI tại Sản phẩm 06.

5.1.2. Chuyển giao và đào tạo vận hành hệ thống

Sau khi hoàn thành nghiên cứu đề xuất phương pháp, mô hình và xây dựng quy trình vận hành hệ thống AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm; đơn vị chủ trì thực hiện đề tài đã có công văn gửi Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực: Đông Bắc, Đồng bằng Bắc Bộ, Bắc Trung Bộ mời tham gia thực hiện "Chuyển giao, đào tạo, vận hành thử nghiệm sản phẩm và đánh giá khả năng ứng dụng của mô hình nhận dạng hình thế và hỗ trợ dự báo, cảnh báo

bão, mưa lớn diện rộng, không khí lạnh, lũ và nước biển dâng do bão sử dụng trí tuệ nhân tạo" (Công văn gửi các đơn vị tại Phụ lục 2 kèm theo).

Báo cáo kết quả tổ chức chuyển giao, đào tạo vận hành hệ thống AI hỗ trợ dự báo, cảnh báo một số các hiện tượng KTTV nguy hiểm, cụ thể như sau:

- **Về mục đích chuyển giao, đào tạo**: Chuyển giao và đào tạo vận hành hệ thống AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm để các cán bộ dự báo và kỹ thuật của Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực: Đông Bắc, Đồng bằng Bắc Bộ, Bắc Trung Bộ phối hợp triển khai thử nghiệm, đánh giá khả năng ứng dụng của mô hình AI nhận dạng hình thế và hỗ trợ dự báo, cảnh báo bão, nước biển dâng do bão, mưa lớn diện rộng, không khí lạnh, lũ trong hoạt động nghiệp vụ thực tế tại đơn vị dự báo.

- **Về nội dung và địa điểm đào tạo**: (i) Chuyển giao, đào tạo, hướng dẫn sử dụng và vận hành hệ thống AI hỗ trợ dự báo, cảnh báo **bão, mưa lớn diện rộng, không khí lạnh** khu vực Bắc Bộ tại Trung tâm Dự báo KTTV quốc gia và Đài KTTV khu vực Đồng bằng Bắc Bộ; (ii) Chuyển giao, đào tạo, hướng dẫn sử dụng và vận hành hệ thống AI hỗ trợ dự báo, cảnh báo **lũ** khu vực Bắc Bộ tại Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực Đồng bằng Bắc Bộ;(iii) Chuyển giao, đào tạo, hướng dẫn sử dụng và vận hành hệ thống AI hỗ trợ dự báo, cảnh báo **nước biển dâng do bão** khu vực ven biển Bắc Bộ và Bắc Trung Bộ tại Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc Bộ và Bắc Trung Bộ.

- **Về giảng viên gồm:** Chủ nhiệm đề tài và các thành viên, chuyên gia tham gia thực hiện đề tài.

- **Về học viên và thành phần tham dự**: Là các dự báo viên, quan trắc viên, cán bộ kỹ thuật của Trung tâm Dự báo KTTV quốc gia và các Đài KTTV khu vực tham gia thử nghiệm hệ thống.

- **Về thời gian đào tạo**: Tháng 9 và tháng 10 năm 2020.

Báo cáo kết quả đào tạo vận hành hệ thống hỗ trợ dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão bằng mô hình AI tại Sản phẩm số 06.

Sau khi hoàn thành nội dung chuyển giao, đào tạo vận hành hệ thống AI hỗ trợ dự báo một số các hiện tượng KTTV nguy hiểm, đơn vị chủ trì đã phối hợp với các đơn vị tham gia thử nghiệm và đánh giá hệ thống, kết quả cụ thể như sau:

5.2. Thử nghiệm mô hình AI hỗ trợ dự báo bão khu vực Bắc Bộ

5.2.1. Thông tin chung về thử nghiệm mô hình AI dự báo bão

5.2.1.1. Về nội dung, và khối lượng thực hiện

- Nội dung thực hiện: Thử nghiệm và đánh giá mô hình hệ thống AI hỗ trợ dự báo **bão** khu vực Bắc Bộ.

- Về khối lượng thử nghiệm: cho **10 cơn bão** đổ bộ vào khu vực Bắc Bộ trong giai đoạn 2008 - 2018.

5.2.1.2. Về thành phần, thời gian và địa điểm thực hiện

- Về thành phần tham gia: Các dự báo viên, kỹ thuật viên thuộc Trung tâm Dự báo KTTV quốc gia, Đài KTTV khu vực Đồng bằng Bắc Bộ và đơn vị chủ trì.

- Về thời gian thực hiện: 30 ngày, trong khoảng từ 18/9- 18/10/2020.

- Địa điểm thực hiện: Tại các phòng dự báo nghiệp vụ thuộc Trung tâm Dự báo KTTV quốc gia và Đài KTTV khu vực Đồng bằng Bắc Bộ.

5.2.1.3. Về công cụ mô hình thử nghiệm

Đơn vị chủ trì cung cấp các công cụ phục vụ thử nghiệm bao gồm:

- Website hỗ trợ dữ báo **bão** tại địa chỉ: http://ai.thoitietnguyhiem.gov.vn/;

- Tài khoản (username/pass): dubao/dubao123;

- Tài liệu hướng dẫn sử dụng đã được đơn vị chủ trì bàn giao cho các đơn vị tham gia thử nghiệm tại đợt chuyển giao, đào tạo sử dụng hệ thống AI hỗ trợ dự báo một số hiện tượng thời tiếp nguy hiểm.

5.2.1.4. Về dữ liệu đầu vào cho mô hình thử nghiệm

- Dữ liệu của 103 cơn bão (2010 - 2019): Tại thời điểm 00 và dữ liệu dự báo;

- Lat, Lon, Vmax, Pressure, bán kính gió ảnh hưởng R30KT, R50KT, R64KT;

- Dữ liệu dự báo của 6 trung tâm: JMA - Nhật Bản, NMC - Trung Quốc, JTWC - Hải quân Mỹ, KMA - Hàn Quốc, HKO - Hồng Kông, CWB - Đài Loan.

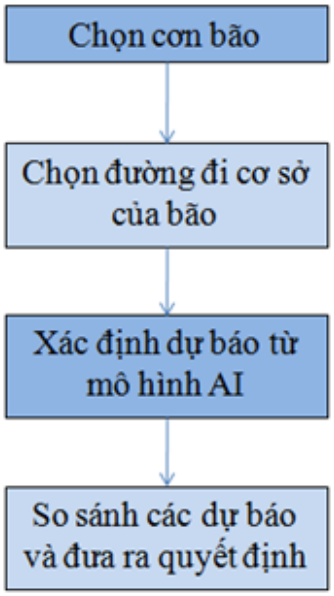
5.2.1.5. Về phương pháp, quy trình vận hành thử nghiệm hệ thống AI hỗ trợ dự báo bão

- Phương pháp dự báo bão bằng mô hình AI gồm 3 bước: Bước 1: Phân vùng dự báo; Bước 2: Xác định trung tâm dự báo tốt nhất; Bước 3: Tham khảo ý kiến chuyên gia.

- Quy trình vận hành thử nghiệm hệ thống AI hỗ trợ dự báo bão (hình 5.1) gồm 4 bước:

  + Bước 1. Chọn cơn bão;
  + Bước 2. Chọn đường đi cơ sở của bão;
  + Bước 3. Xác định kết quả dự báo từ mô hình AI;
  + Bước 4. So sánh các dự báo và đưa ra quyết định.

Chi tiết về phương pháp và quy trình dự báo bão được trình bày tại Báo cáo công việc 183 và 184.



Hình 5.1: Quy trình hỗ trợ dự báo bão bằng mô hình AI

5.2.1.6. Về so sánh, tham khảo kết quả dự báo

Kết quả dự báo bằng mô hình AI được so sánh với các kết quả dự báo của một số Trung tâm dự báo trong và ngoài nước cung cấp thông tin bão trên menu gồm:

- Đường mầu xanh dương nét đứt là kết quả dự báo của mô hình AI;
- Đường mầu xanh tím than là kết quả dự báo của HKO: Cơ quan Khí tượng Hồng Kông;
- Đường mầu đỏ là kết quả dự báo của JTWC: Trung tâm Cảnh báo bão liên hợp của Mỹ;
- Đường mầu xanh lá cây là kết quả dự báo của JMA: Cơ quan Khí tượng Nhật Bản;
- Đường mầu ghi xám là kết quả dự báo của NMC: Trung tâm Khí tượng quốc gia Trung Quốc;
- Đường mầu nâu là kết quả dự báo của CWB: Cơ quan thời tiết Đài Loan;
- Đường mầu vàng là kết quả dự báo của KMA: Cơ quan Khí tượng Hàn Quốc;
- Đường mầu đen là kết quả dự báo của PAGASA: Cơ quan Khí tượng Philippines.

Kết quả thử nghiệm và đánh giá mô hình AI hỗ trợ dự báo **bão** cụ thể như sau:

5.2.2. Kết quả thử nghiệm tại Trung tâm Dự báo KTTV quốc gia

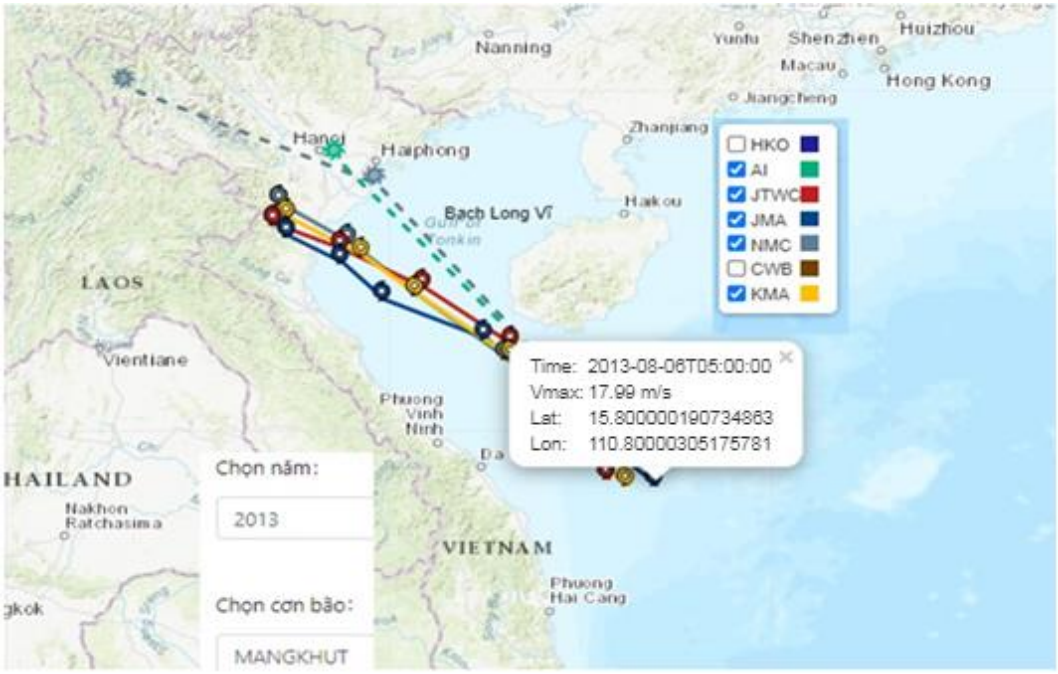
Thực hiện kế hoạch thử nghiệm đánh giá hệ thống, từ 18/9 đến18/10/2010, đơn vị chủ trì thực hiện đề tài đã phối hợp với Trung tâm Dự báo KTTV quốc gia tiến hành thử nghiệm hệ thống AI dự báo **bão khu vực Bắc Bộ**. Khối lượng thử nghiệm: cho **05 cơn bão** đổ bộ vào khu vực Bắc Bộ trong giai đoạn 2008 – 2017 (Bảng 5.1).

Bảng 5.1: Danh sách các cơn bão thử nghiệm hệ thống AI dự báo bão tại Trung tâm Dự báo KTTV quốc gia

STT	Năm	Tháng	Loại	Tên bão và ATNĐ			Cường độ		Số ngày tồn tại	Phạm vi hoạt động	
				Mã QT	Tên QT	Mã VN	Pmin	Vmax		Nơi phát sinh	Nơi kết thúc
1	2013	8	TS	1310	MANGKHUT	Bão số 6	905	110	10	12.2-166.3	23.7-107.3
2	2013	11	TYP	1330	HAIYAN	Bão số 13	895	125	8	6.1-153.3	22.8-108.6
3	2014	9	TYP	1415	KALMAEGI	Bão số 3	960	75	5	13.7-142.2	22.3-104.7
4	2015	6	TS	1508	KUJIRA	Bão số 1	985	45	3	15.7-111.9	21.3-106.4
5	2016	10	TYP	1621	SARIKA	Bão số 7	935	95	6	13.2-130.2	21.6-108.2

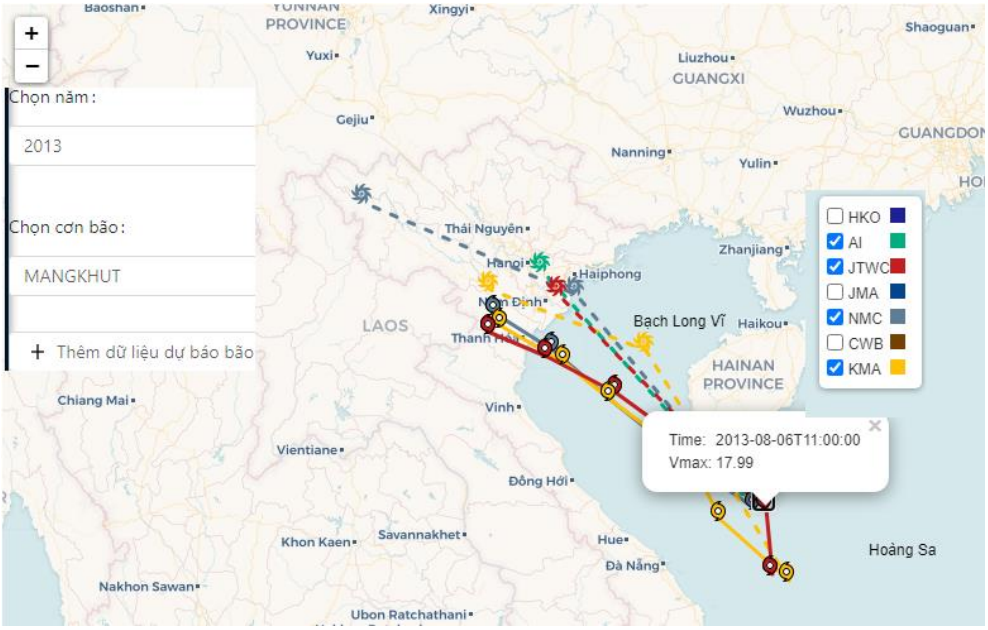
5.2.2.1. Kết quả thử nghiệm cho cơn bão MANGKHUT

MANGKHUT (cơn bão số 6) đổ bộ vào khu vực Thanh Hóa, Việt Nam vào tháng 8/2013, thời gian tồn tại trong 02 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo trong khu vực khác như sau:



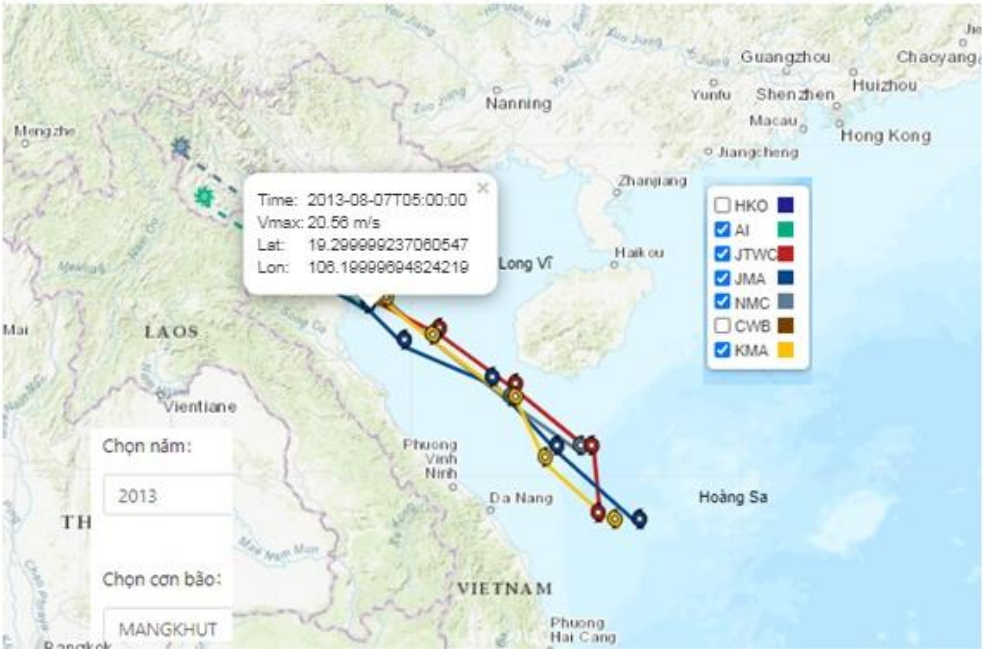
Hình 5.2: Kết quả dự báo cơn bão MANGKHUT lúc 05h00 ngày 06/8/2013

Kết quả dự báo cơn bão MANGKHUT lúc 05h00 ngày 06/8/2013 là thời điểm bão xuất hiện trên khu vực quần đảo Hoàng Sa - Việt Nam. Vận tốc gió max là 17.99 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.



Hình 5.3: Kết quả dự báo cơn bão MANGKHUT lúc 11h00 ngày 06/8/2013

Kết quả dự báo cơn bão MANGKHUT **lúc 11h00** ngày 06/8/2013 là thời điểm bão bão đi qua khu vực đảo Hải Nam. Vận tốc gió max là: 17.99 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

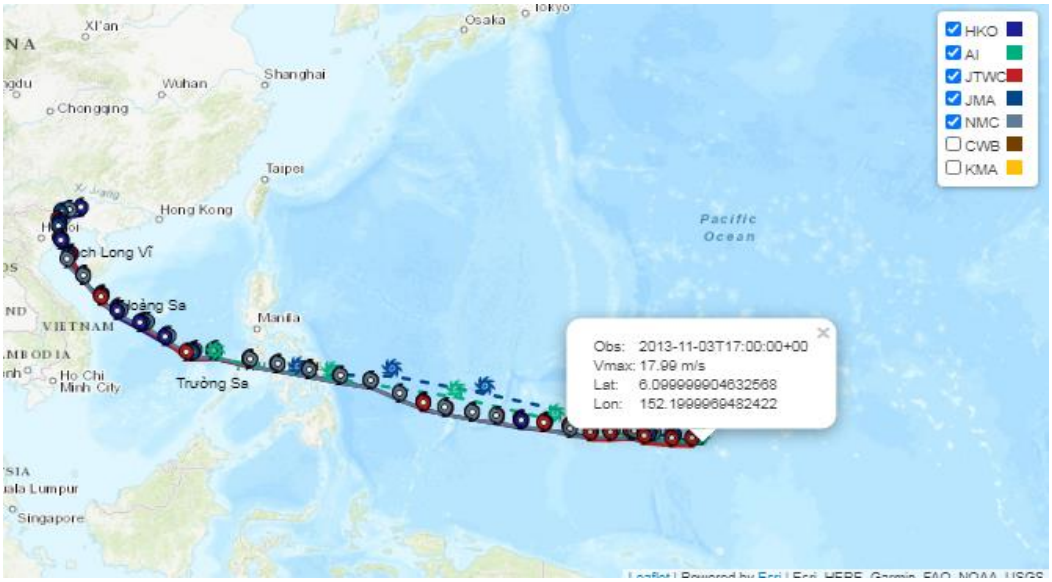


Hình 5.4: Kết quả dự báo cơn bão MANGKHUT lúc 05h00 ngày 07/8/2013

Kết quả dự báo cơn bão MANGKHUT **lúc 05h00** ngày 07/8/2013 là thời điểm bão sắp đổ bộ vào bờ khu vực Thanh Hóa. Vị trí tâm bão là Lat: 19.299; Lon" 106.199; Vận tốc gió max là: 20.56 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

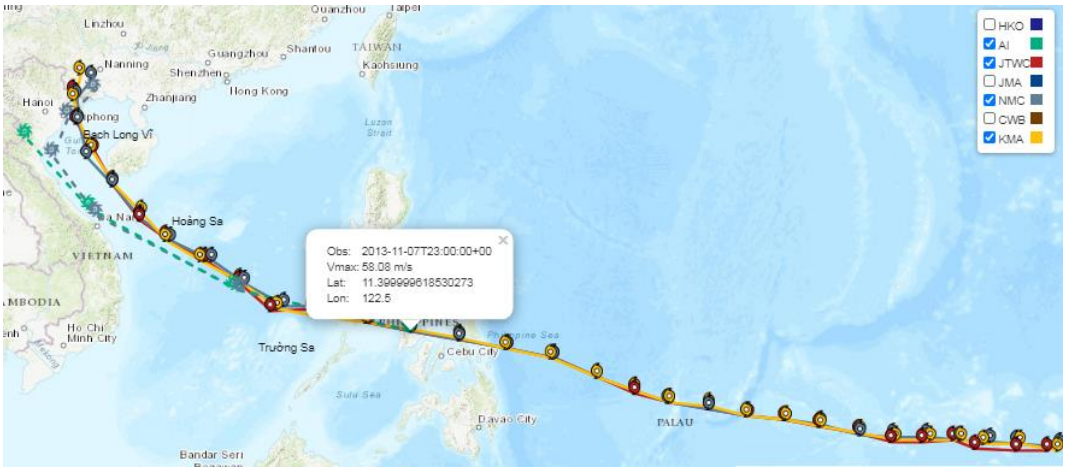
5.2.2.2. Kết quả thử nghiệm cho cơn bão HAIYAN

HAIYAN (cơn bão số 13) đổ bộ vào khu vực Quảng Ninh, Việt Nam vào tháng 11/2013, thời gian tồn tại trong 08 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo khác như sau:



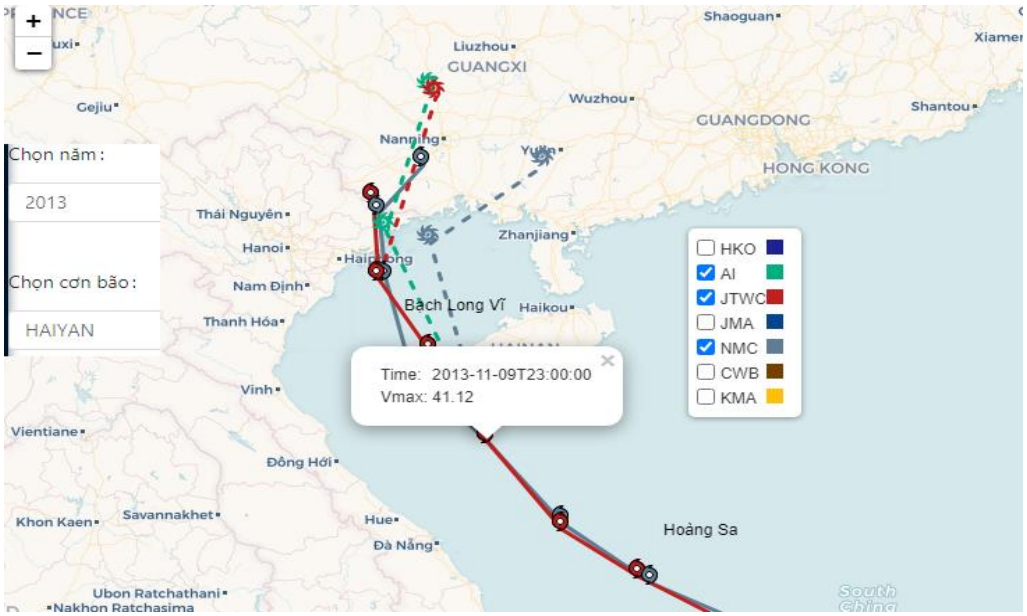
Hình 5.5: Kết quả dự báo cơn bão HAIYAN lúc 17h00 ngày 03/11/2013

Kết quả dự báo cơn bão HAIYAN lúc **17h00** ngày 03/11/2013 là thời điểm bão xuất hiện. Vị trí tâm bão là Lat: 6.099; Lon" 152.199; Vận tốc gió max là 17.99 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.



Hình 5.6: Kết quả dự báo cơn bão HAIYAN lúc 23h00 ngày 07/11/2013

Kết quả dự báo cơn bão HAIYAN lúc **23h00** ngày 07/11/2013 là thời điểm bão đổ bộ vào Philippines. Vận tốc gió max là: 58.08 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

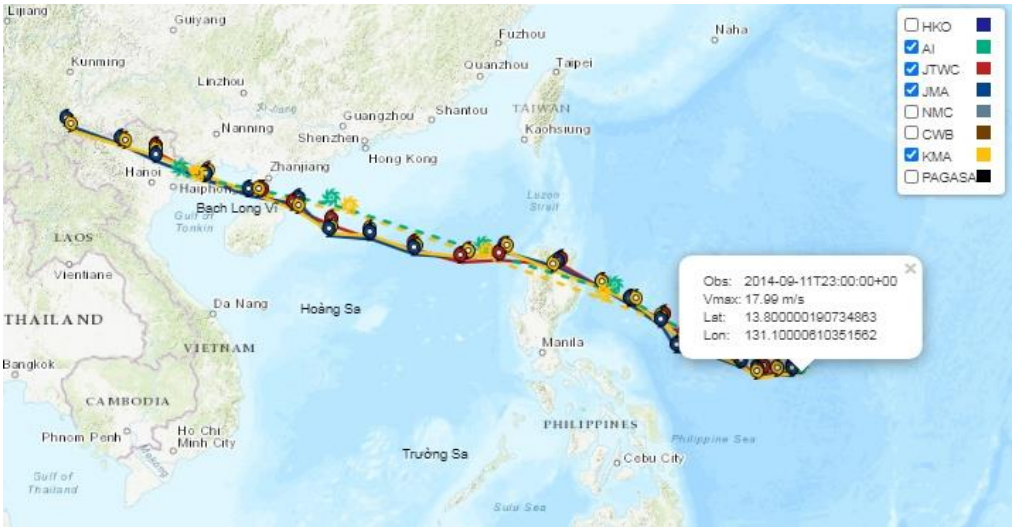


Hình 5.7: Kết quả dự báo cơn bão HAIYAN lúc 23h00 ngày 09/11/2013

Kết quả dự báo cơn bão HAIYAN lúc **23h00** ngày 0911/2013 là thời điểm bão chuẩn bị đổ bộ vào khu vực Quảng Ninh. Vận tốc gió max là: 41.12 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

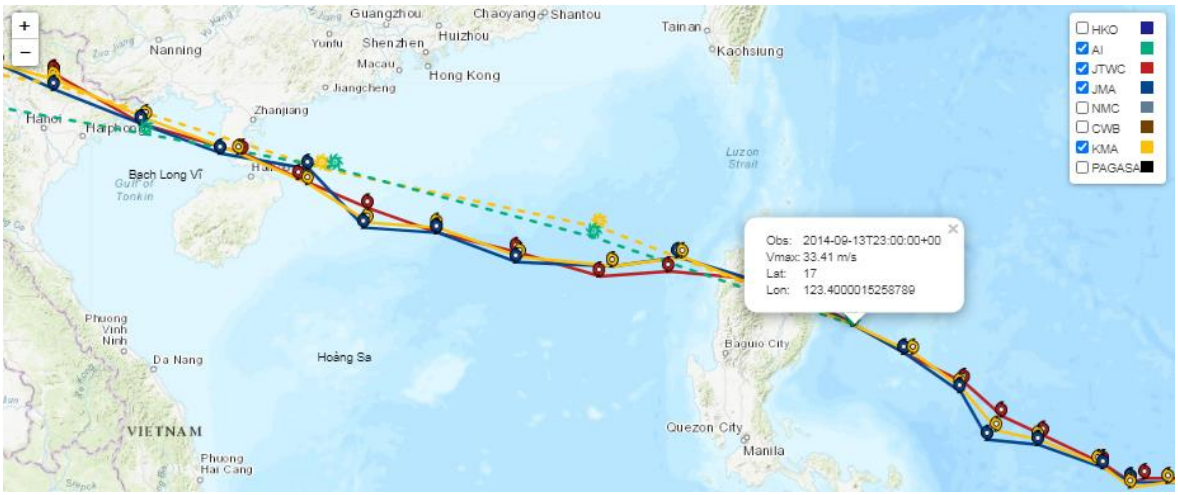
5.2.2.3. Kết quả thử nghiệm cho cơn bão KALMAEGI

KALMAEGI (bão số 9) đổ bộ vào khu vực Quảng Ninh, Việt Nam vào tháng 9/2014. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo như sau:



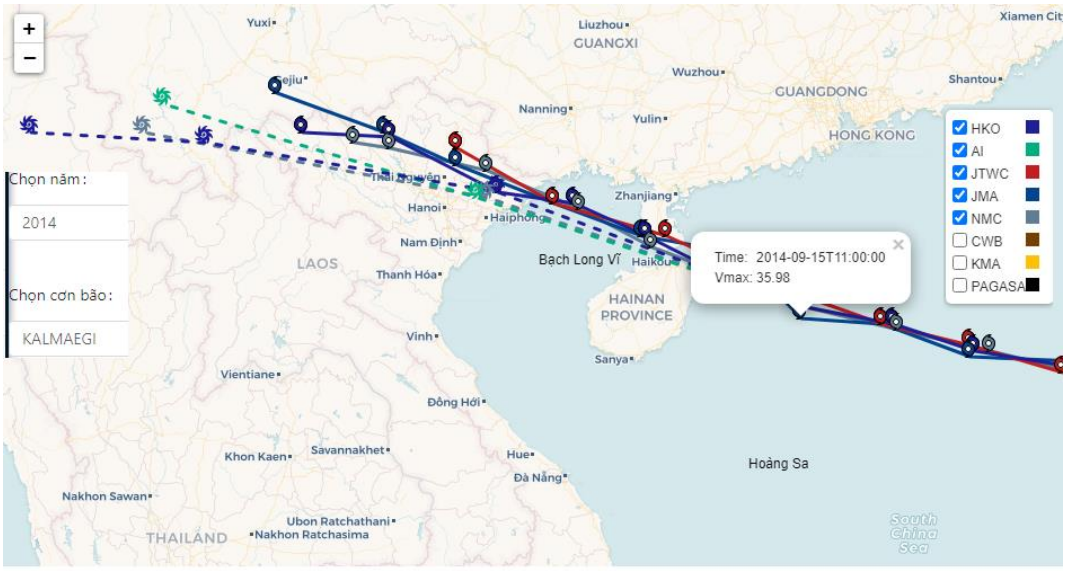
Hình 5.8: Kết quả dự báo cơn bão KALMAEGI lúc 23h00 ngày 11/9/2014

Kết quả dự báo cơn bão **KALMAEGI** lúc 23h00 ngày 11/9/2014 là thời điểm bão xuất hiện. Vị trí tâm bão Lat: 13.8; Lon" 130.1; vận tốc gió max là: 17.99m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.



Hình 5.9: Kết quả dự báo cơn bão KALMAEGI lúc 23h00 ngày 13/9/2014

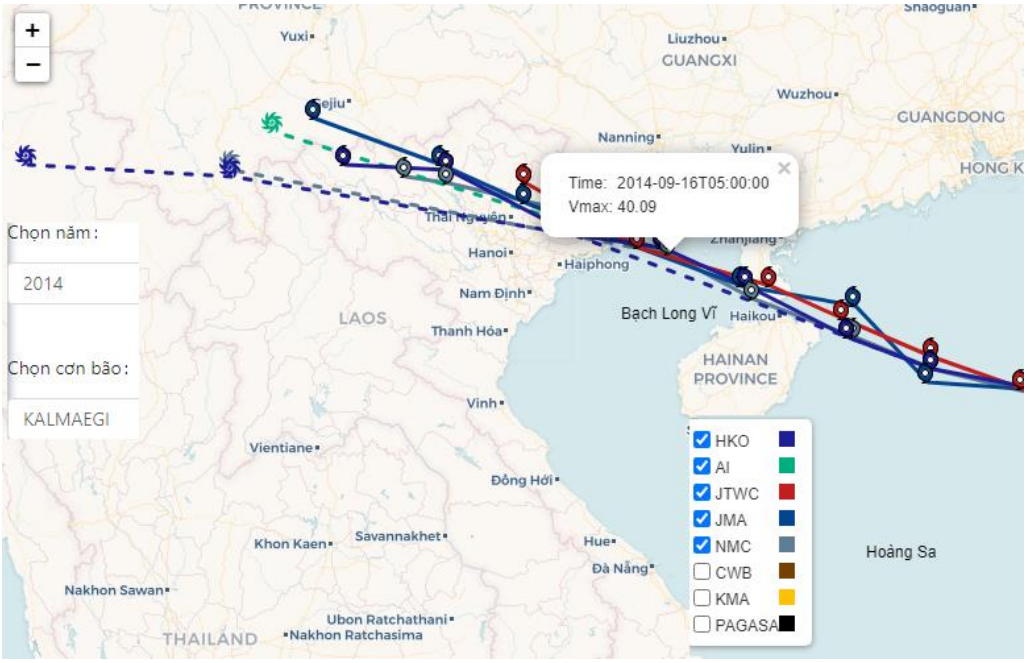
Kết quả dự báo cơn bão **KALMAEGI lúc 23h00** ngày 13/9/2014 là thời điểm bão đi qua đảo Luzon - Philippines. Vị trí tâm bão là Lat: 17; Lon" 123.400; Vận tốc gió max là: 33.41m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực, thế giới.



Hình 5.10: Kết quả dự báo cơn bão KALMAEGI lúc 11h00 ngày 15/9/2014

Kết quả dự báo cơn bão **KALMAEGI lúc 11h00** ngày 15/9/2014 là thời điểm chuẩn bị đổ bộ vào khu vực đảo Hải Nam. Vận tốc gió max là: 35.98 m/s. Dự báo

quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

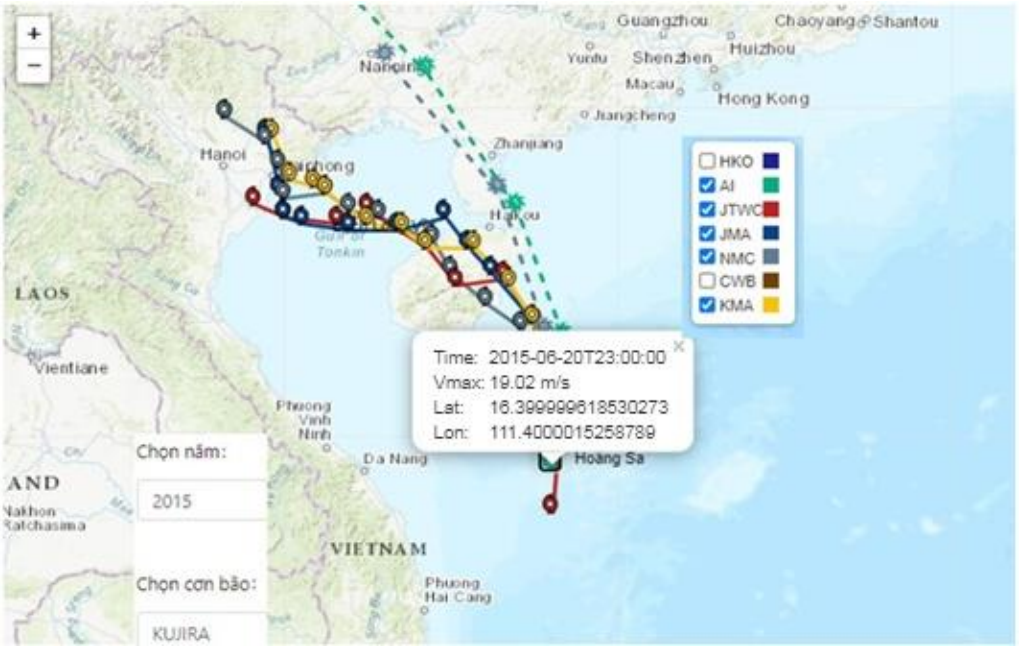


Hình 5.11: Kết quả dự báo cơn bão KALMAEGI lúc 05h00 ngày 16/9/2014

Kết quả dự báo cơn bão **KALMAEGI lúc 05h00** ngày 16/9/2014 là thời điểm bão chuẩn bị đổ bộ vào khu vực Quảng Ninh.Vận tốc gió max là: 40.09 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

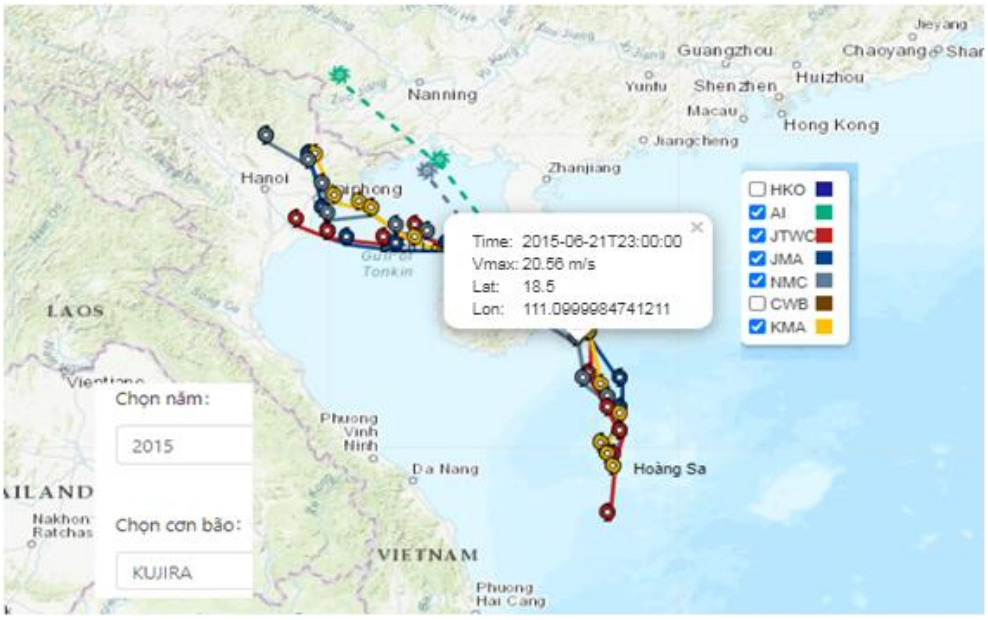
5.2.2.4. Kết quả thử nghiệm cho cơn bão KUJIRA

KUJIRA (bão số 4) đổ bộ vào khu vực Quảng Ninh, Việt Nam vào tháng 6/2015, thời gian tồn tại trong 3 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo khác như sau:



Hình 5.12: Kết quả dự báo cơn bão KUJIRA lúc 23h00 ngày 20/6/2015

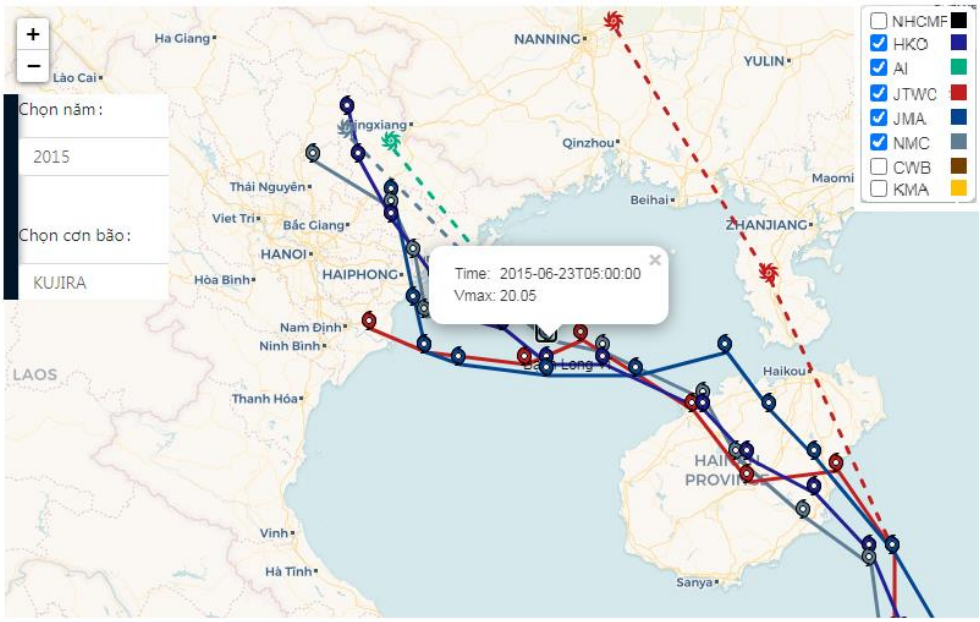
Kết quả dự báo cơn bão **KUJIRA lúc 23h00** ngày 20/6/2015 là thời điểm bão xuất hiện trên khu vực quần đảo Hoàng Sa - Việt Nam. Vị trí tâm bão là Lat: 16.399; Lon" 111.400; Vận tốc gió max là: 19.02m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.



Hình 5.13: Kết quả dự báo cơn bão KUJIRA lúc 23h00 ngày 21/6/2015

Kết quả dự báo cơn bão **KUJIRA lúc 23h00** ngày 21/6/2015 là thời điểm bão chuẩn bị đổ bộ vào khu vực đảo Hải Nam. Vị trí tâm bão là Lat: 18.5; Lon" 110.099; Vận tốc gió max là: 20.56m/s. Dự báo quỹ đạo bão của mô hình AI

(đường màu xanh dương nét đứt) bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.

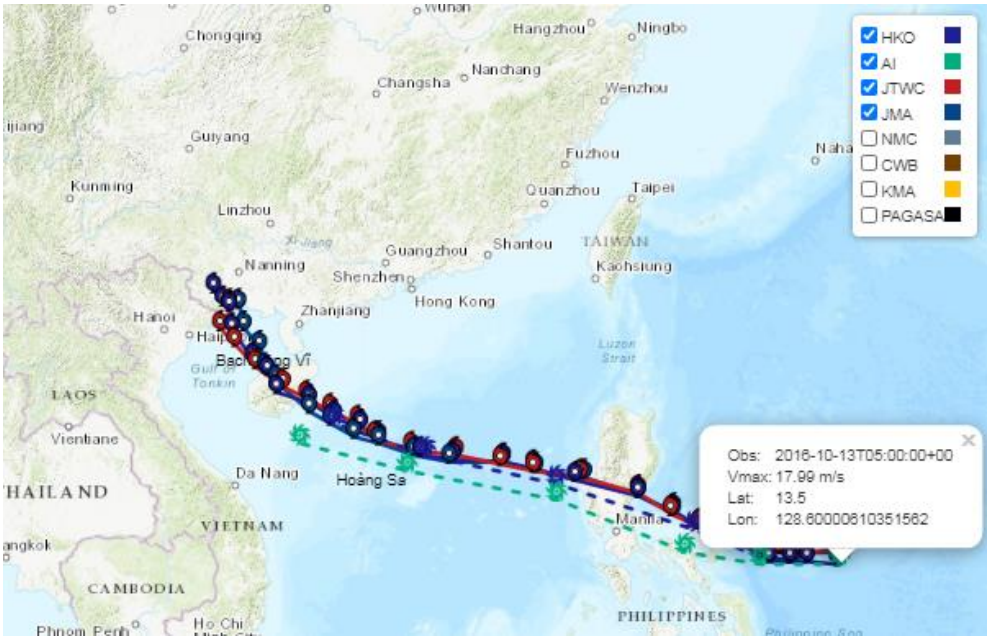


Hình 5.14: Kết quả dự báo cơn bão KUJIRA lúc 05h00 ngày 23/6/2015

Kết quả dự báo cơn bão **KUJIRA lúc 05h00** ngày 23/6/2015 là thời điểm bão chuẩn bị đổ bộ vào khu vực khu vực Quảng Ninh. Vận tốc gió max là: 20.05 m/s. Dự báo quỹ đạo bão của mô hình AI bám sát kết quả dự báo của các Trung tâm dự báo.

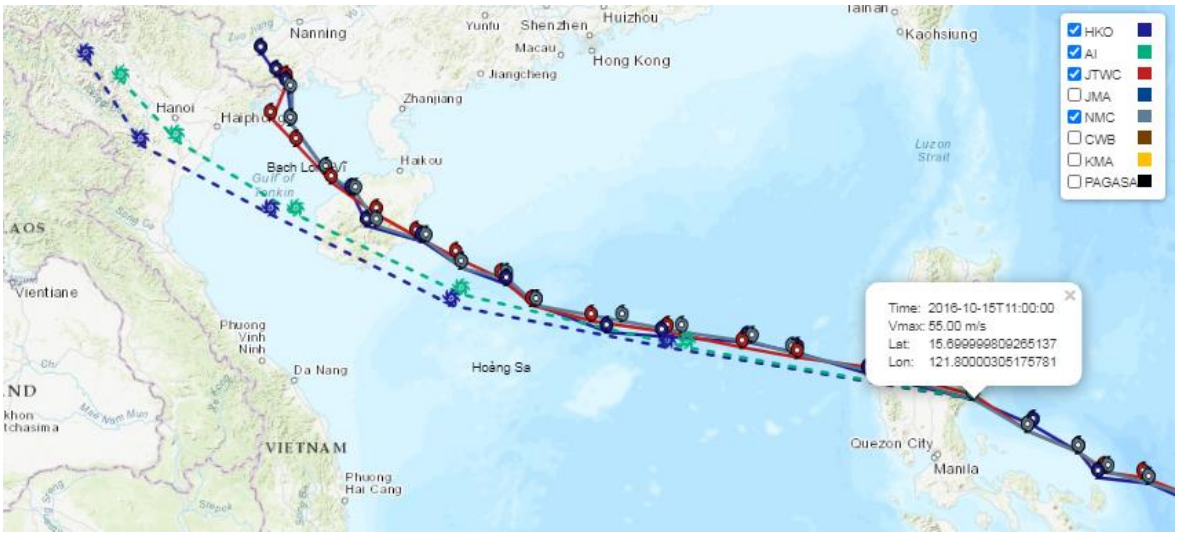
5.2.2.5. Kết quả thử nghiệm cho cơn bão SARIKA

SARIKA (cơn bão số 7) đổ bộ vào khu vực Quảng Ninh, Việt Nam vào tháng 8/2016. Kết quả dự báo của mô hình AI và các Trung tâm dự báo khác như sau:



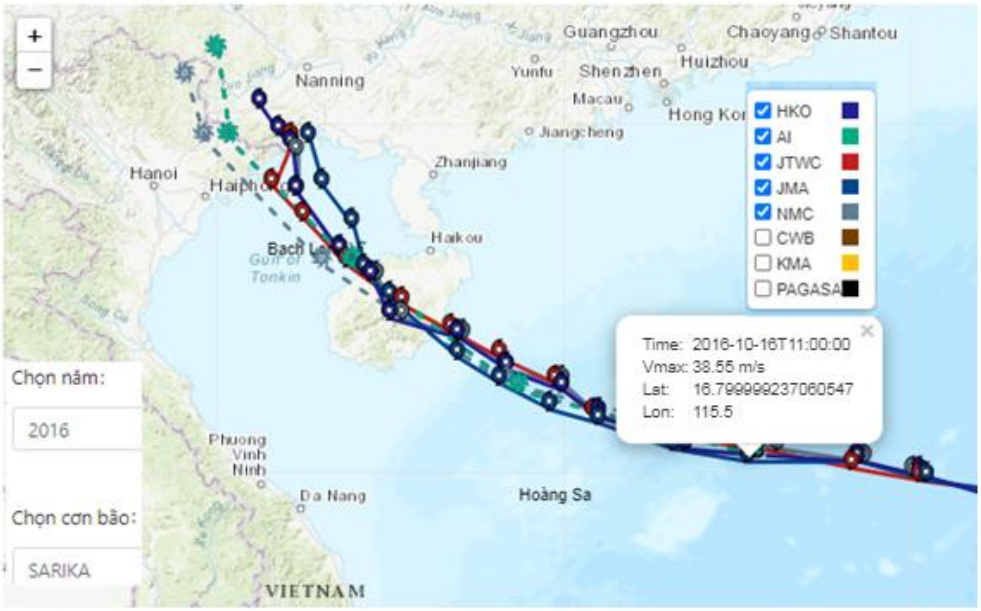
Hình 5.15: Kết quả dự báo cơn bão SARIKA lúc 05h00 ngày 13/10/2016

Kết quả dự báo cơn bão SARIKA lúc 05h00 ngày 13/10/2016 là thời điểm bão xuất hiện. Vị trí tâm bão là Lat: 13.5; Lon" 128.600; Vận tốc gió max là: 17.99m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.



Hình 5.16: Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 15/10/2016

Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 15/10/2016 là thời điểm bão đổ bộ vào Philippines. Vị trí tâm bão là Lat: 15.699; Lon" 121.800; Vận tốc gió max là: 55.00 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo.



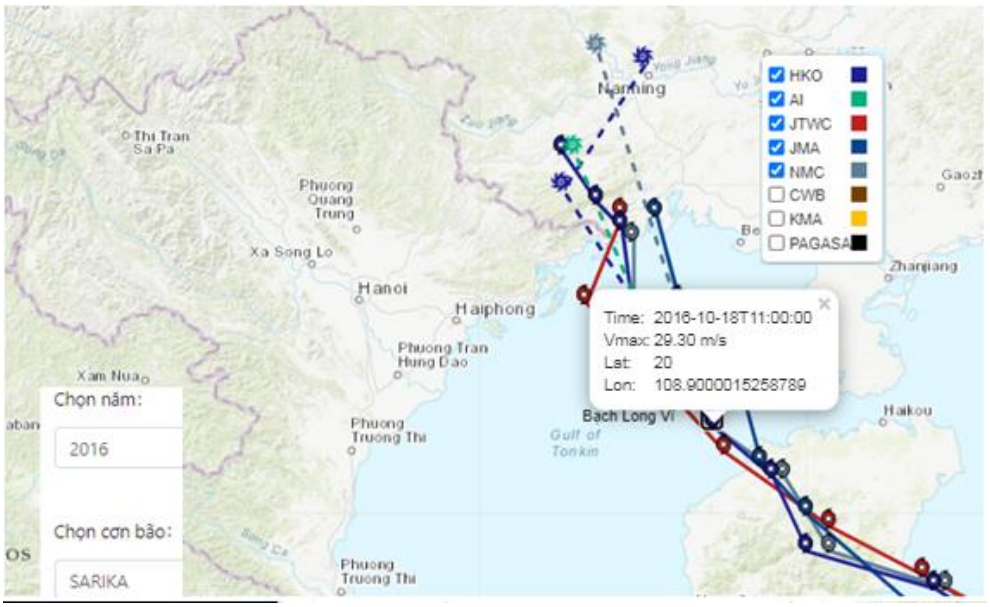
Hình 5.17: Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 16/10/2016

Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 16/10/2016 là thời điểm bão trên vùng biển thuộc quần đảo Hoàng Sa - Việt Nam. Vị trí tâm bão là Lat: 16.799;

Lon” 115.5; Vận tốc gió max là: 38.55m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo.

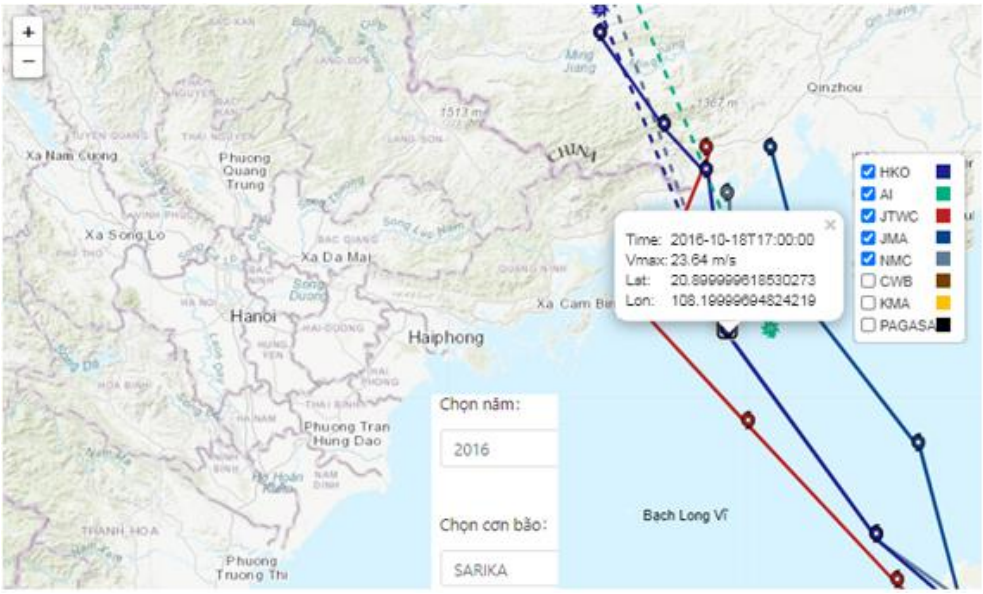
Hình 5.18: Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 17/10/2016

Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 17/10/2016 là thời điểm bão chuẩn bị đổ bộ vào đảo Hải Nam. Vận tốc gió max là: 38.55m/s. Dự báo quỹ đạo bão của mô hình AI bám sát kết quả dự báo của các Trung tâm dự báo.



Hình 5.19: Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 18/10/2016

Kết quả dự báo cơn bão SARIKA lúc 11h00 ngày 18/10/2016 là thời điểm bão trên khu vực đảo Bạch Long Vĩ - Việt Nam. Vị trí tâm bão là Lat: 20; Lon” 108.900; Vận tốc gió max là: 29.30m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo.



Hình 5.20: Kết quả dự báo cơn bão SARIKA lúc 17h00 ngày 18/10/2016

Kết quả dự báo cơn bão SARIKA lúc 17h00 ngày 18/10/2016 là thời điểm bão chuẩn bị đổ bộ vào bờ khu vực tỉnh Quảng Ninh . Vị trí tâm bão là Lat: 20.899; Lon" 108.199; Vận tốc gió max là:23.64m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bám sát kết quả dự báo của các Trung tâm dự báo.

5.2.3. Kết quả thử nghiệm tại Đài KTTV khu vực Đồng bằng Bắc Bộ

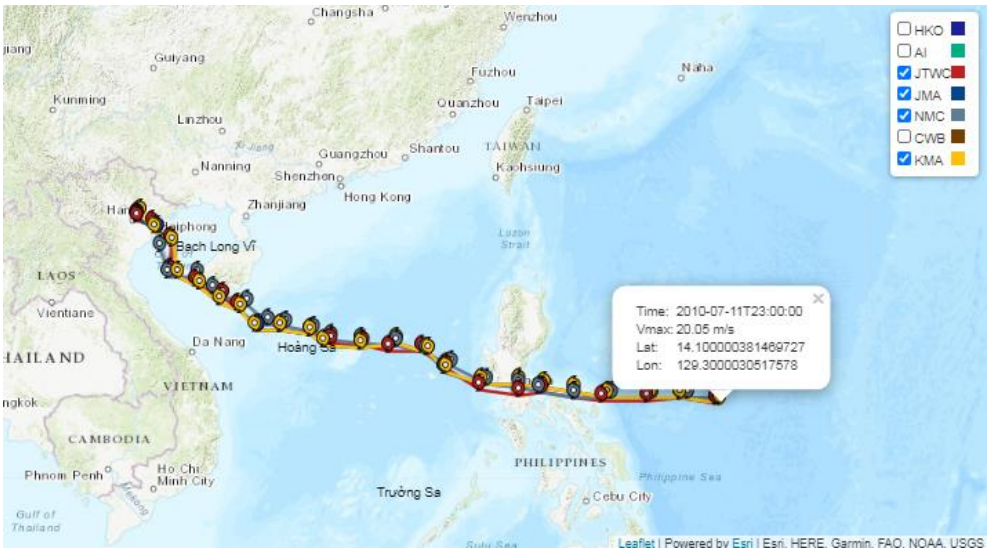
Từ 18/9-18-10/2010, đơn vị chủ trì thực hiện đề tài đã phối hợp với Đài KTTV khu vực Đồng bằng Bắc Bộ tiến hành thử nghiệm hệ thống AI hỗ trợ dự báo **bão khu vực Bắc Bộ**. Khối lượng thử nghiệm: cho **05 cơn bão** đổ bộ vào khu vực Hải Phòng, Nam Định, Thái Bình, Thanh Hóa trong giai đoạn 2008 - 2018 (bảng 5.2).

Bảng 5.2: Danh sách các cơn bão thử nghiệm hệ thống AI dự báo bão tại Đài KTTV khu vực Đồng bằng Bắc Bộ

STT	Năm	Tháng	Loại	Tên bão và ATNĐ			Cường độ		Số ngày tồn tại	Phạm vi hoạt động	
				Số hiệu QT	Tên QT	Số hiệu VN	Pmin	Vmax		Nơi phát sinh	Nơi kết thúc
1	2010	7	*TYP*	1002	CONSON	Bão số 1	970	70	6	14.2-131.8	21.2-105.9
2	2012	10	*TYP*	1223	SON TINH	Bão số 8	945	85	5	8.3-128.1	21.3-107.1
3	2013	6	*STS*	1306	RUMBIA	Bão số 3	980	55	4	9.9-128.6	23.9-108.3
4	2016	8	TS	1608	DIANMU	Bão số 3	980	45	2	20.6-112.6	20.8-105.9
5	2018	8	*STS*	1816	BEBINCA	Bão số 4	985	50	4	20.6-112.5	19.6-105.9

*5.2.3.1. Kết quả thử nghiệm cho cơn bão **CONSON***

CONSON (cơn bão số 1) đổ bộ vào khu vực Hải Phòng - Quảng Ninh, Việt Nam vào tháng 7/2010, thời gian tồn tại trong 06 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo như sau:



Hình 5.21: Kết quả dự báo cơn bão CONSON lúc 23h00 ngày 11/7/2010

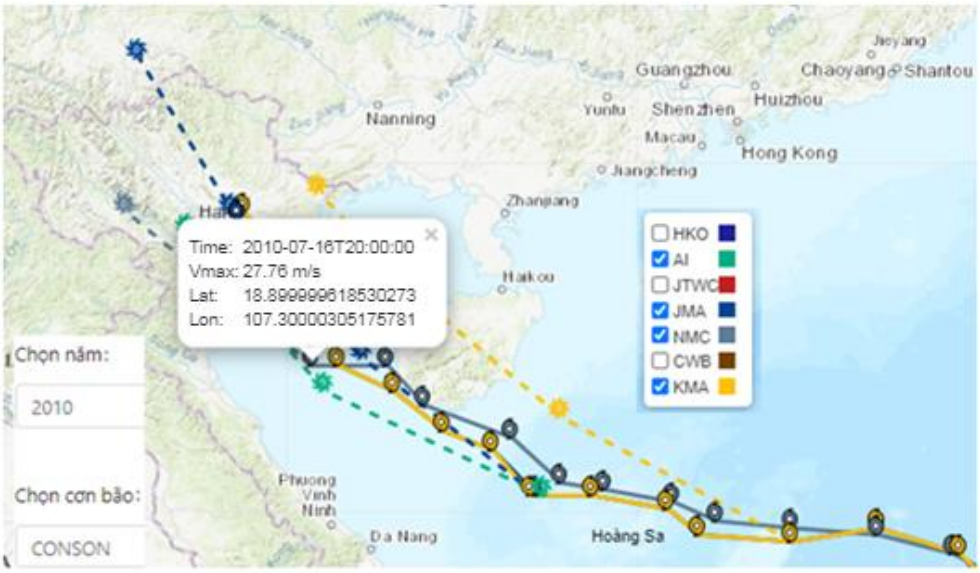
Kết quả dự báo cơn bão **CONSON lúc 23h00** ngày 11/7/2010 là thời điểm bão xuất hiện. Vị trí tâm bão là Lat: 14.100; Lon" 129.300; vận tốc gió max là: 20.05m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo trong khu vực và thế giới.

Hình 5.22: Kết quả dự báo cơn bão CONSON lúc 05h00 ngày 13/7/2010

Kết quả dự báo cơn bão CONSON lúc 05h00 ngày 13/7/2010 là thời điểm bão đổ bộ vào Philippines. Vị trí tâm bão là Lat: 14.5; Lon" 122.099; Vận tốc gió max là: 29.81 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.

Hình 5.23: Kết quả dự báo cơn bão CONSON lúc 17h00 ngày 14/7/2010

Kết quả dự báo cơn bão CONSON lúc **17h00** ngày 14/7/2010 là thời điểm bão trên khu vực quần đảo Hoàng Sa - Việt Nam. Vị trí tâm bão là Lat: 16.200; Lon: 115; Vận tốc gió max là: 25.70m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương nét đứt) bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.

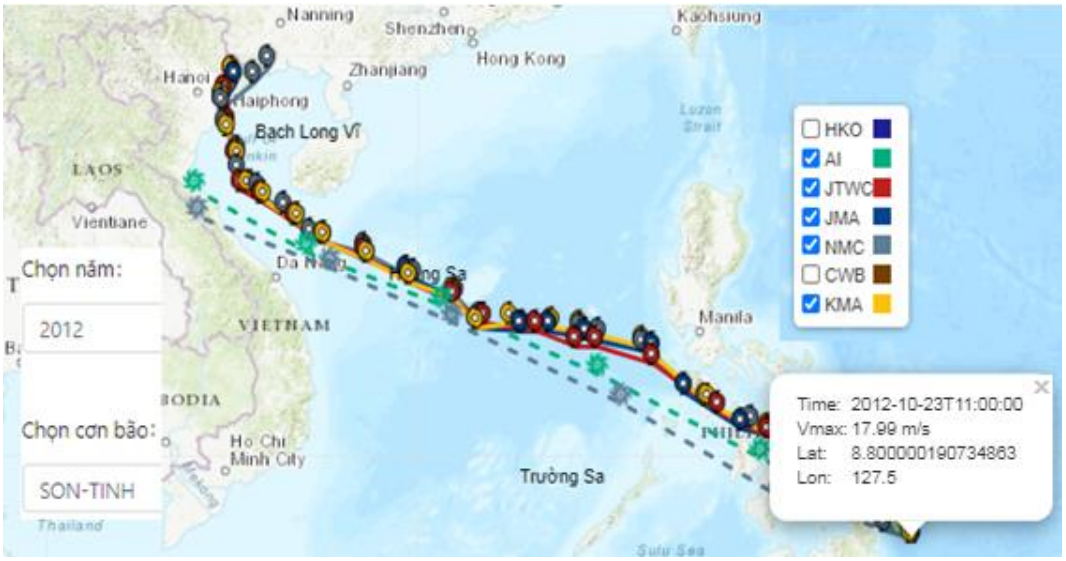


Hình 5.24: Kết quả dự báo cơn bão CONSON lúc 20h00 ngày 16/7/2010

Kết quả dự báo cơn bão CONSON lúc **lúc 20h00** ngày 16/7/2010 là thời điểm bão chuẩn bị đổ bộ vào bờ khu vực Quảng Ninh. Vị trí tâm bão: Lat 18.899; Lon 107.300; vận tốc gió max: 27.76m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.

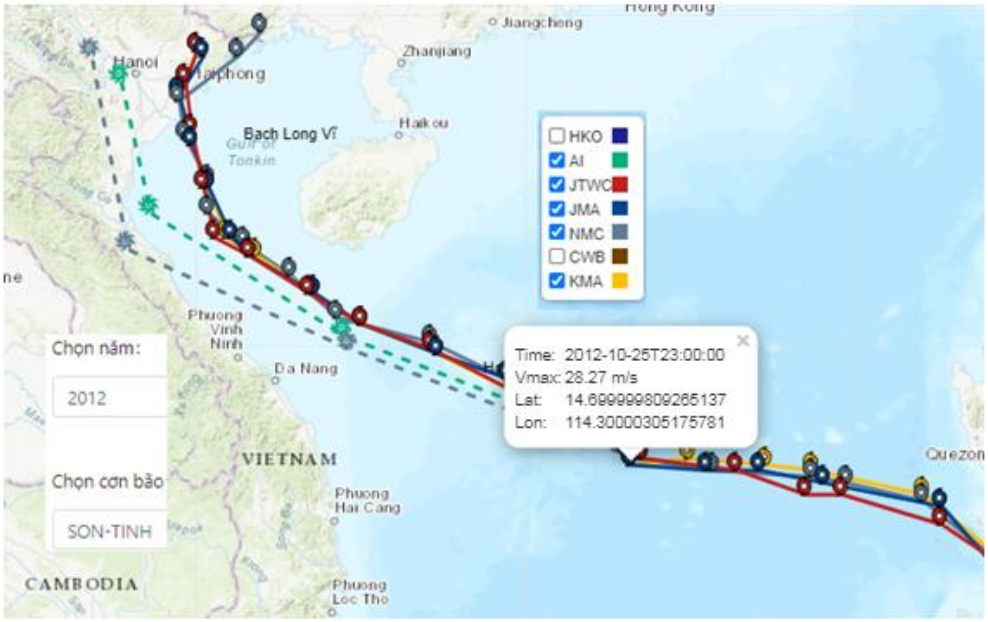
5.2.3.2. Kết quả thử nghiệm cho cơn bão SON TINH

SON TINH (cơn bão số 8) đổ bộ vào khu vực Thái Bình - Hải Phòng vào tháng 10/2012. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo khác như sau:



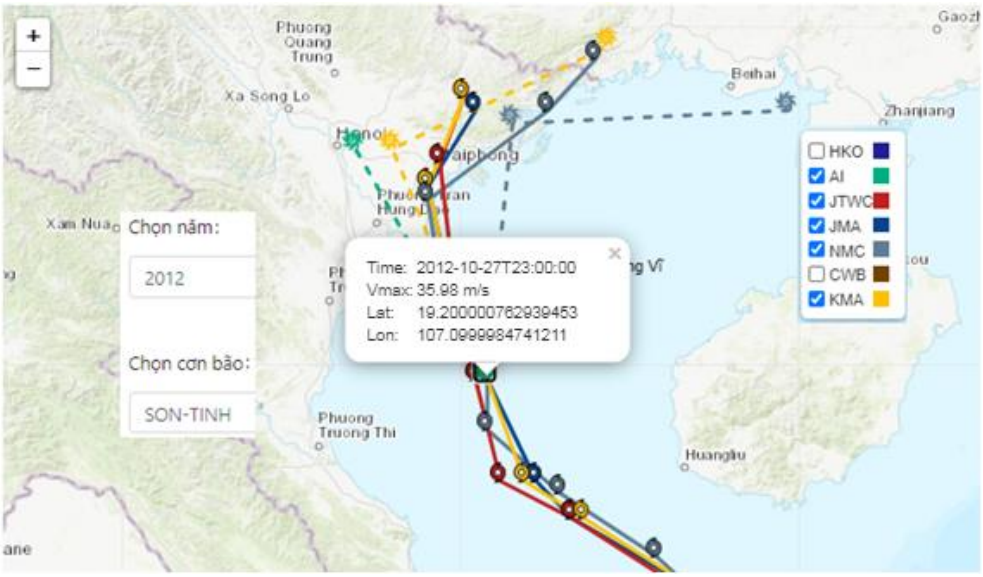
Hình 5.25: Kết quả dự báo cơn bão SON TINH lúc 11h00 ngày 23/10/2012

Kết quả dự báo cơn bão **SON TINH** l**úc 11h00** ngày 23/10/2012 là thời điểm xuất hiện bão trên vùng biển Philippines. Vị trí tâm bão là Lat: 8.800; Lon: 127.500; Vận tốc gió max là: 17.99m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.



Hình 5.26: Kết quả dự báo cơn bão SON TINH lúc 23h00 ngày 25/10/2012

Kết quả dự báo cơn bão **SON TINH** l**úc 23h00** ngày 25/10/2012 là thời điểm bão trên khu vực quần đảo Hoàng Sa - Việt Nam. Vị trí tâm bão là Lat: 14.699; Lon" 114.300; Vận tốc gió max là: 28.27m/s.



Hình 5.27: Kết quả dự báo cơn bão SON TINH lúc 23h00 ngày 27/10/2012

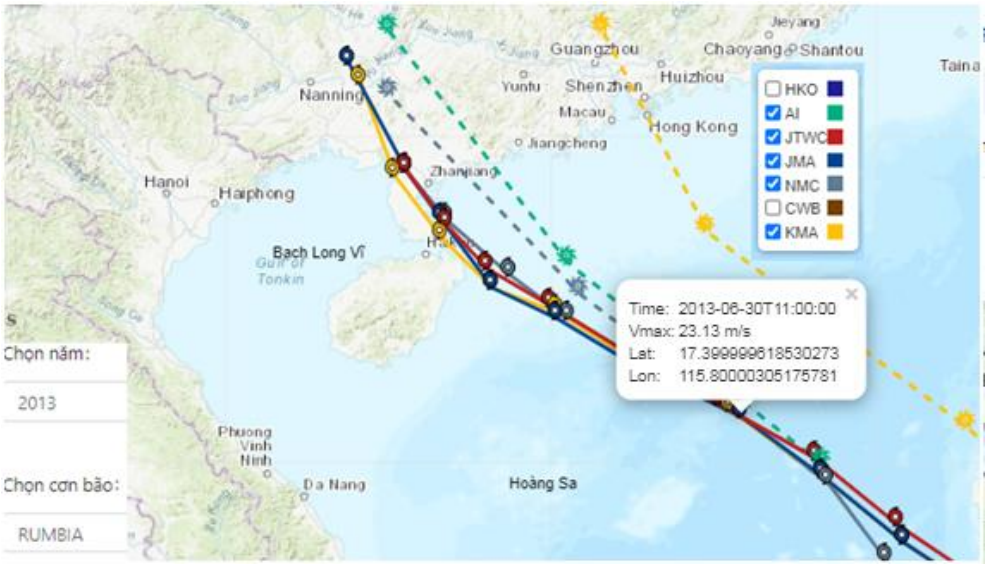
Kết quả dự báo cơn bão **SON TINH** l**úc 23h00** ngày 27/10/2012 là thời điểm bão chuẩn bị đổ bộ vào bờ trên khu vực Thái Bình - Hải Phòng. Vị trí tâm bão là Lat: 19.200; Lon" 107.099; Vận tốc gió max là: 35.98m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.

5.2.3.3. Kết quả thử nghiệm cho cơn bão RUMBIA

RUMBIA (cơn bão số 3) xuất hiện vào tháng 6/2013, thời gian tồn tại trong 04 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo khác như sau:

Hình 5.28: Kết quả dự báo cơn bão RUMBIA lúc 05h00 ngày 28/6/2013

Kết quả dự báo cơn bão **RUMBIA** l**úc 05h00** ngày 28/6/2013 là thời điểm bão chuẩn bị đổ bộ vào Philippines. Vị trí tâm bão là Lat: 10.300; Lon: 127.900; Vận tốc gió max: 17.99m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.

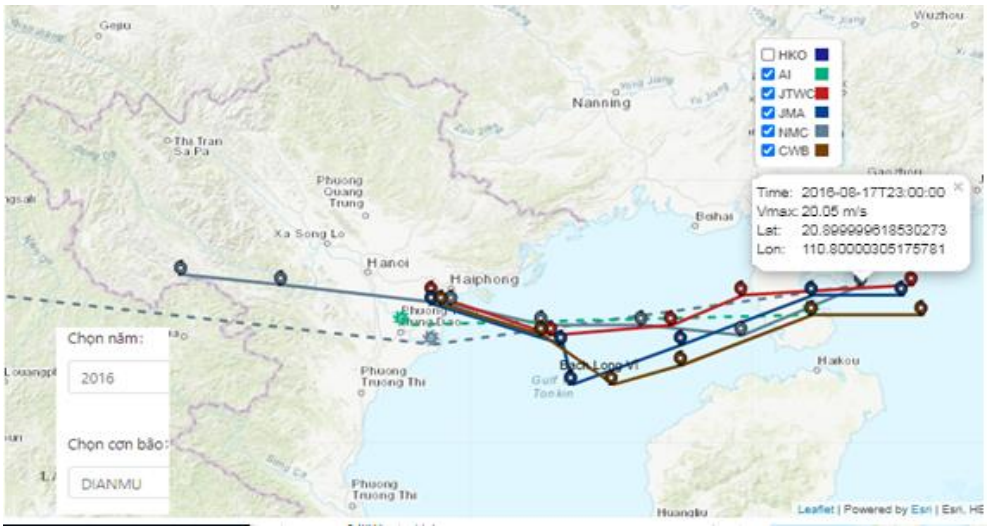


Hình 5.29: Kết quả dự báo cơn bão RUMBIA lúc 11h00 ngày 30/6/2013

Kết quả dự báo cơn bão **RUMBIA** l**úc 11h00** ngày 30/6/2013 là thời điểm bãotrên khu vực quần đảo Hoàng Sa - Việt Nam. Vị trí tâm bão là Lat: 17.399; Lon" 115.800; vận tốc gió max là: 23.13 m/s. Dự báo quỹ đạo bão của mô hình AI bị lệch tương đối lớn so với kết quả dự báo của các Trung tâm dự báo.

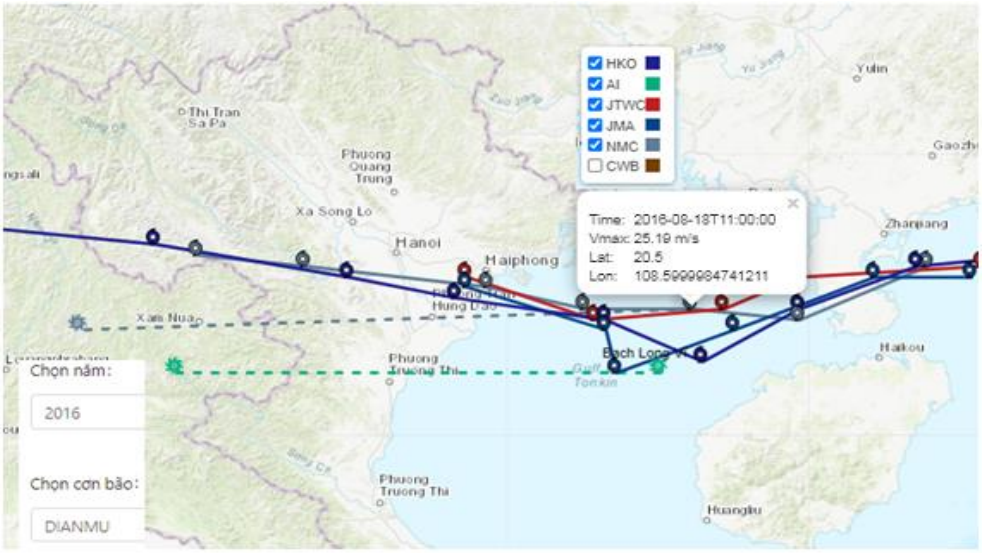
5.2.3.4. Kết quả thử nghiệm cho cơn bão DIANMU

DIANMU (cơn bão số 3) đổ bộ vào khu vực Hải Phòng, Thái Bình, Việt Nam vào tháng 8/2016, thời gian tồn tại trong 02 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo khác như sau:



Hình 5.30: Kết quả dự báo cơn bão DIANMU lúc 23h00 ngày 17/8/2016

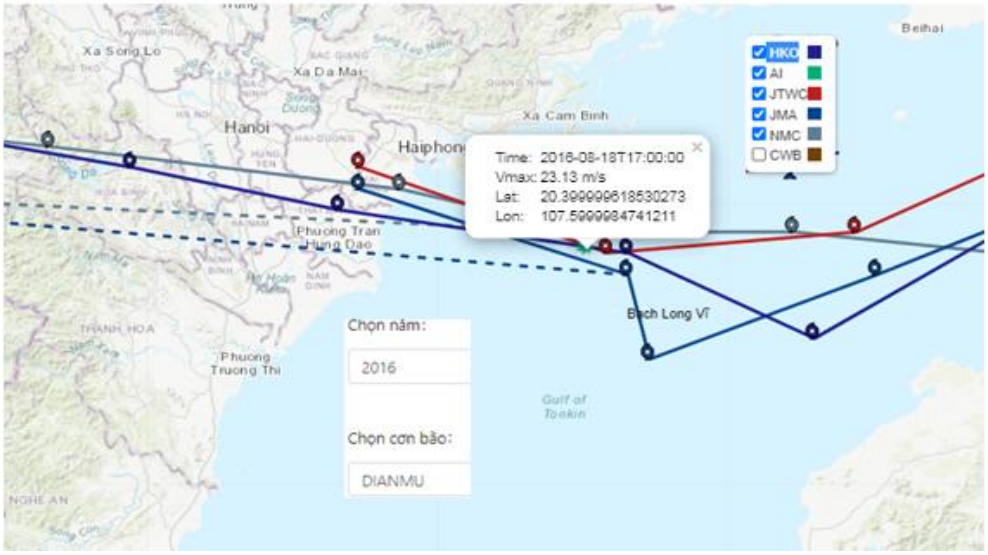
Kết quả dự báo cơn bão **DIANMU** l**úc 23h00** ngày 17/8/2016 là thời điểm xuất hiện bão trên vùng biển Dongshan - Trung Quốc. Vị trí tâm bão là Lat: 20.899; Lon" 110.800; Vận tốc gió max là: 20.50m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.



Hình 5.31: Kết quả dự báo cơn bão DIANMU lúc 11h00 ngày 18/8/2016

Kết quả dự báo cơn bão **DIANMU** l**úc 11h00** ngày 18/8/2016 là thời điểm bão trên khu vực đảo Bạch Long Vĩ - Việt Nam. Vị trí tâm bão là Lat: 20.5; Lon" 108.599; Vận tốc gió max là: 25.19 m/s. Dự báo quỹ đạo bão của mô hình AI

(đường màu xanh dương) bị lệch tương đối lớn so với các kết quả dự báo của các Trung tâm dự báo.

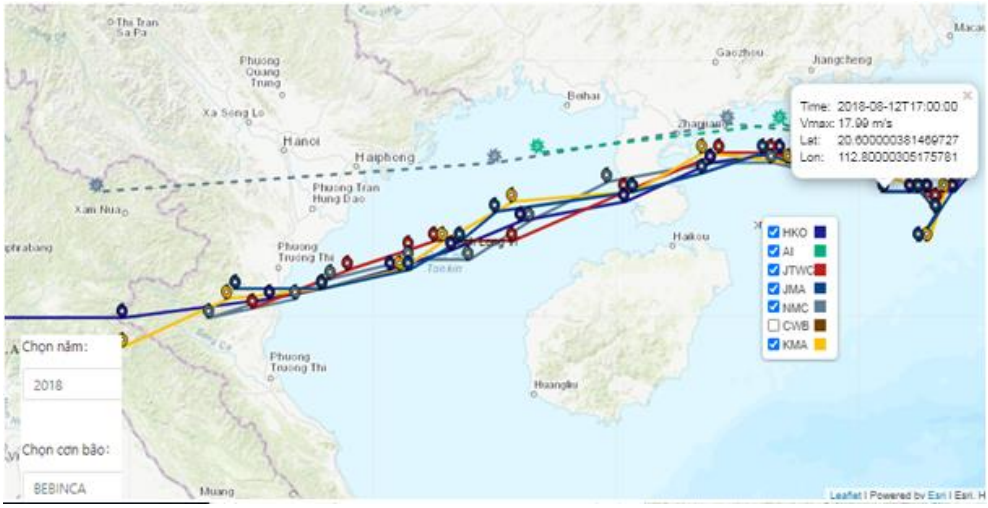


Hình 5.32: Kết quả dự báo cơn bão DIANMU lúc 17h00 ngày 18/8/2016

Kết quả dự báo cơn bão **DIANMU l**úc **17h00** ngày 18/8/2016 là thời điểm bão chuẩn bị đổ bộ vào khu vực Hải Phòng, Thái Bình. Vị trí tâm bão là Lat: 20.399; Lon" 107.599; Vận tốc gió max là: 23.13m/s.

5.2.3.5. Kết quả thử nghiệm cho cơn bão BEBINCA

BEBINCA (cơn bão số 4) đổ bộ vào khu vực Nam Định, Thái Bình, Việt Nam vào tháng 8/2018, thời gian tồn tại trong 04 ngày. Kết quả dự báo của mô hình AI, và các Trung tâm dự báo khác như sau:



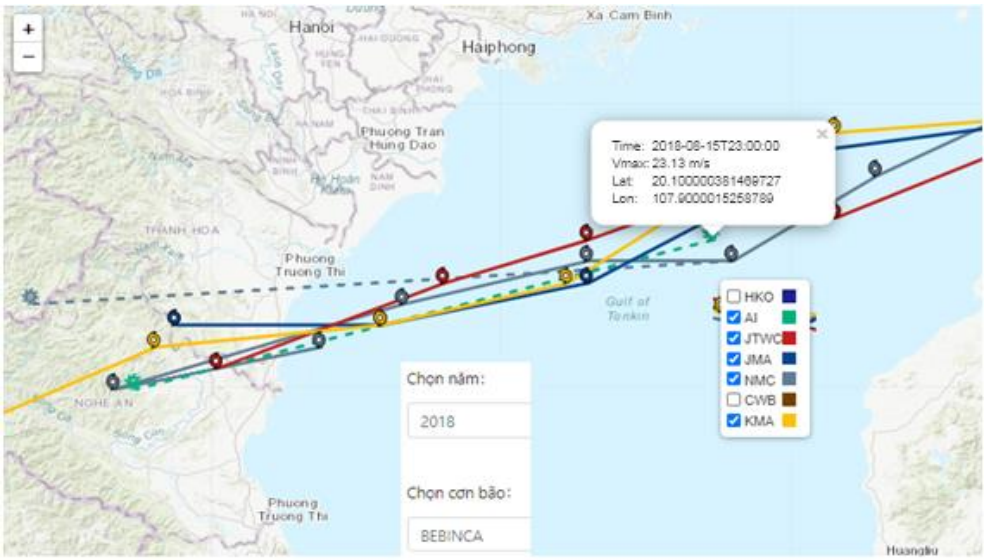
Hình 5.33: Kết quả dự báo cơn bão BEBINCA lúc 17h00 ngày 12/8/2018

Kết quả dự báo cơn bão **BEBINCAlúc 17h00** ngày 12/8/2018 là thời điểm xuất hiện bão trên vùng biển Trung Quốc. Vị trí tâm bão là Lat: 20.600; Lon" 112.800; Vận tốc gió max là: 17.99m/s. Dự báo quỹ đạo bão của mô hình AI

(đường màu xanh dương) bị lệch tương đối lớn so với các kết quả dự báo của các Trung tâm dự báo.

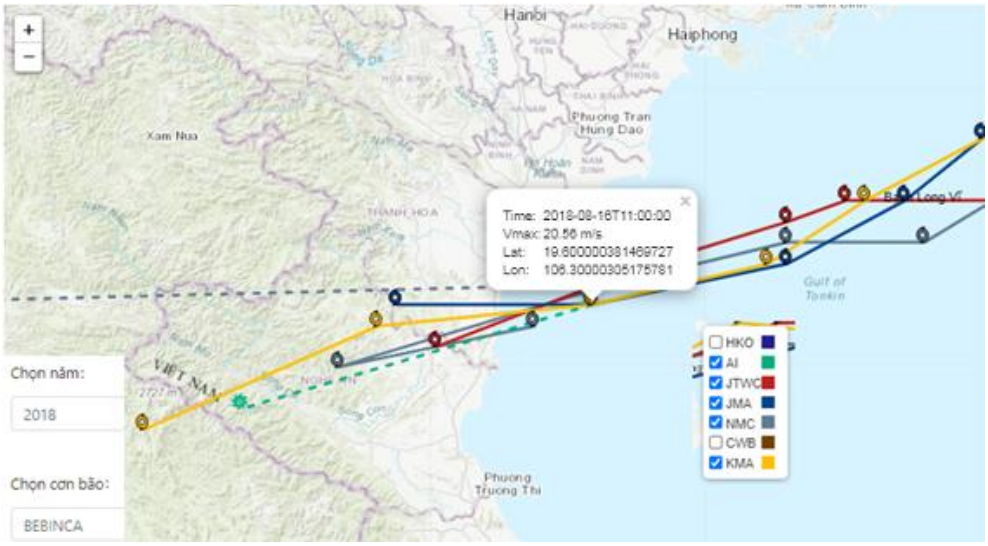
Hình 5.34: Kết quả dự báo cơn bão BEBINCA lúc 05h00 ngày 15/8/2018

Kết quả dự báo cơn bão **BEBINCA** l**úc 05h00** ngày 15/8/2018 là thời điểm bão đổ bộ vào khu vực tỉnh Quảng Đông - Trung Quốc. Vị trí tâm bão là Lat: 20.799; Lon" 110.400; vận tốc gió max là: 20.56m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.



Hình 5.35: Kết quả dự báo cơn bão BEBINCA lúc 23h00 ngày 15/8/2018

Kết quả dự báo cơn bão **BEBINCA** l**úc 23h00** ngày 15/8/2018 là thời điểm bão trên khu vực đảo Bạch Long Vĩ - Việt Nam. Vị trí tâm bão là Lat: 20.100; Lon" 107.900; vận tốc gió max là: 23.13m/s.Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.



Hình 5.36: Kết quả dự báo cơn bão BEBINCA lúc 11h00 ngày 16/8/2018

Kết quả dự báo cơn bão **BEBINCAlúc 11h00** ngày 16/8/2018 là thời điểm bão chuẩn bị đổ bộ vào khu vực Nam Định - Thái Bình. Vị trí tâm bão là Lat: 19.600; Lon" 106.300; Vận tốc gió max là: 20.56 m/s. Dự báo quỹ đạo bão của mô hình AI (đường màu xanh dương) bám sát kết quả dự báo của các Trung tâm dự báo.

5.3. Thử nghiệm mô hình AI hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ, Bắc Trung Bộ

5.3.1. Thông tin chung về thử nghiệm mô hình AI dự báo nước biển dâng do bão

5.3.1.1. Về nội dung, và khối lượng thực hiện

- Nội dung thực hiện: Thử nghiệm và đánh giá mô hình hệ thống AI hỗ trợ dự báo **nước biển dâng do bão** khu vực ven biển Bắc Bộ và Bắc Trung Bộ.

- Về khối lượng thử nghiệm: cho **20 cơn bão** đổ bộ vào khu vực ven biển Bắc Bộ và Bắc Trung Bộ trong giai đoạn 2008 - 2017.

5.3.1.2. Về thành phần, thời gian và địa điểm thực hiện

- Về thành phần tham gia: Các dự báo viên, kỹ thuật viên thuộc Trung tâm Dự báo KTTV quốc gia, Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc Bộ, Bắc Trung Bộ và đơn vị chủ trì.

- Về thời gian thực hiện: 30 ngày, trong khoảng từ 18/9- 18/10/2020.

- Địa điểm thực hiện: Tại các phòng dự báo nghiệp vụ thuộc Trung tâm Dự báo KTTV quốc gia và các Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc Bộ, Bắc Trung Bộ.

5.3.1.3. Về công cụ mô hình thử nghiệm

Đơn vị chủ trì cung cấp các công cụ phục vụ thử nghiệm bao gồm:

- Website hỗ trợ dữ báo nước biển dâng do bão tại địa chỉ: http://ai.thoitietnguyhiem.gov.vn/
- Tài khoản (username/pass): dubao/dubao123;
- Tài liệu hướng dẫn sử dụng đã được đơn vị chủ trì bàn giao cho các đơn vị tham gia thử nghiệm tại đợt chuyển giao, đào tạo sử dụng hệ thống AI hỗ trợ dự báo một số hiện tượng thời tiếp nguy hiểm.

5.3.1.4. Về dữ liệu đầu vào cho mô hình thử nghiệm

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo nước biển dâng do bão khu vực Bắc Bộ và Bắc Trung Bộ cụ thể như sau:

- Dữ liệu quan trắc của 02 trạm khí tượng hải văn trên vùng biển Quảng Ninh đến Nghệ An.
- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.
- Các yếu tố hải văn gồm: độ cao sóng, chu kỳ sóng, nước biển dâng do bão, mực nước thủy triều.
- Các trường dữ liệu sử dụng cho mô hình dự báo nước biển dâng do bão sau khi được trích chọn các đặc trưng được mô tả như ở dưới bảng sau:

Bảng 5.3: Mô tả các đặc trưng dữ liệu nước biển dâng do bão

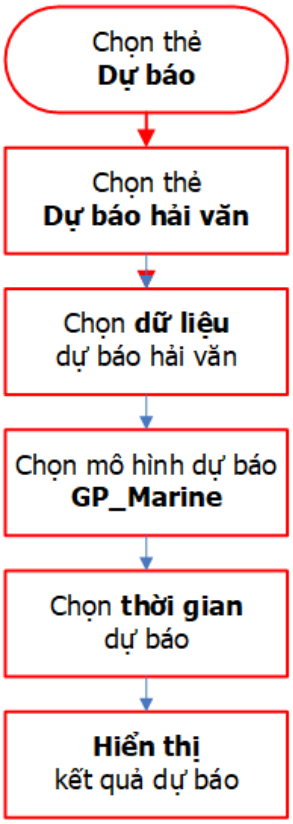
STT	Tên trường	Ý nghĩa	Kiểu giá trị
1	Date (dd/mm/yyyy), Time (hh.mm.ss)	Ngày Giờ thu thập dữ liệu	String
2	water_level	Mực nước dâng	Float
3	wind_spe	Tốc độ gió	Float
4	wind_dir	Hướng gió	Float
5	latitude	Kinh độ của bão	Float
6	longitude	Vĩ độ của bão	Float
7	central_press	Áp lực tâm bão	Float
8	max_win_spe	Vận tốc gió tối đa	Float
9	sea_surface_level	Mực nước dâng trên bề mặt	Float
10	storm_surge	mực nước dâng do bão	Float

5.3.1.5. Về Quy trình vận hành thử nghiệm hệ thống AI hỗ trợ dự báo nước biển dâng do bão

Quy trình vận hành thử nghiệm hệ thống AI hỗ trợ dự báo nước biển dâng do bão (hình 5.37) gồm 6 bước:

- Bước 1. Chọn thẻ dự báo;
- Bước 2. Chọn thẻ dự báo hải văn;
- Bước 3. Chọn dữ liệu cho mô hình AI dự báo nước biển dâng do bão;
- Bước 4. Chọn mô hình AI dự báo;
- Bước 5. Chọn thời gian dự báo;
- Bước 6: Hiển thị kết quả dự báo nước biển dâng do bão

Chi tiết về quy trình dự báo nước biển dâng do bão được trình bày tại Báo cáo công việc 188, 189, 190, 191, 192.



Hình 5.37: Quy trình hỗ trợ dự báo nước biển dâng do bão bằng AI

Kết quả thử nghiệm mô hình AI hỗ trợ dự báo **nước biển dâng do bão** khu vực ven biển Bắc Bộ và Bắc Trung Bộ cụ thể như sau:

5.3.2. Kết quả thử nghiệm Trung tâm Dự báo KTTV quốc gia

Đơn vị chủ trì đã phối hợp với Trung tâm Dự báo KTTV quốc gia thử nghiệm cho 05 đợt nước biển dâng do **05 cơn bão** đổ bộ vào khu vực ven biển Bắc Bộ và Bắc Trung Bộ trong giai đoạn 2008 - 2017.

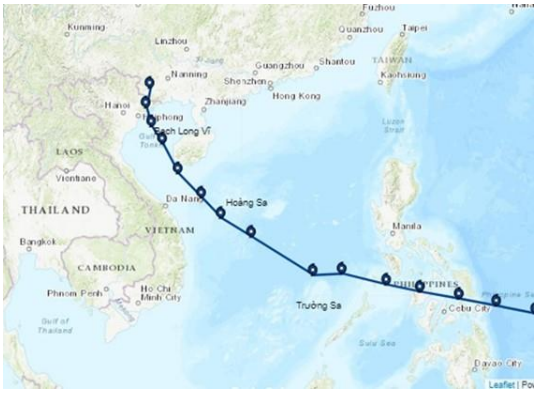
Bảng 5.4: Danh sách các cơn bão phục vụ thử nghiệm dự báo nước biển dâng tại Trung tâm Dự báo KTTV quốc gia

STT	Năm	Tên cơn bão ảnh hưởng	Trạm quan trắc
1.	2013	Haiyan	Hòn Ngư
2.	2015	Kujira	Hòn Ngư
3.	2016	Mirinae	Hòn Ngư
4.	2017	Doksuri	Hòn Ngư
5.	2018	Sơn Tinh	Hòn Ngư

5.3.2.1. Đợt 1: Thử nghiệm dự báo nước biển dâng do bão Haiyan năm 2013

a) Thông tin về cơn bão Haiyan

Thông tin cơ bản về cơn bão Haiyan, cơn bão số 13 ảnh hưởng đến Việt Nam năm 2013: (i) Áp suất thấp nhất Pmin = 895 hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax= 125 mph; (iii) số ngày tồn tại: 8 ngày; (iv) vị trí phát sinh: vĩ độ 6.1 - kinh độ 153.3; vị trí kết thúc: vĩ độ 22.8 - kinh độ 108.6.



Hình 5.38: Đường đi của bão Haiyan 2013

b) Kết quả dự báo nước biển dâng do bão Haiyan của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão bằng mô hình AI tại Hòn Ngư vào lúc **10h0**0 ngày 10/11/2013 như sau:

- Nước biển dâng quan trắc: 0.45 m
- Nước biển dâng dự báo: 0.49 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.39: Kết quả dự báo nước biển dâng do bão Haiyan lúc 10h00, 10/11/2013

Kết quả quan trắc và dự báo nước biển dâng do bão bằng mô hình AI tại Hòn Ngư vào lúc 23h00 ngày 10/11/2013 như sau:

- Nước biển dâng quan trắc: 0.608192 m;
- Nước biển dâng dự báo: 0.674697261989391 m.

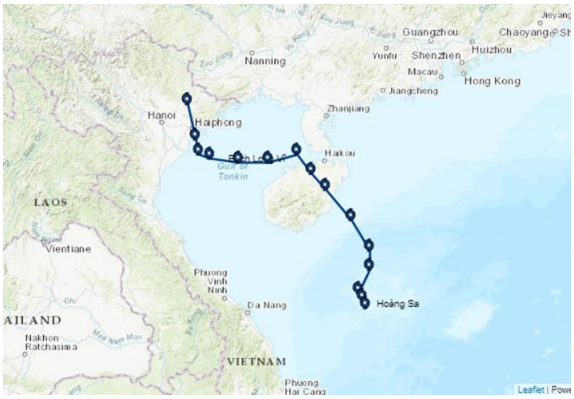
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.40: Kết quả dự báo nước biển dâng do bão Haiyan lúc 23h00, 10/11/2013

5.3.2.2. Đợt 2: Thử nghiệm dự báo nước biển dâng do bão Kujira năm 2015

a) Thông tin về cơn bão Kujira:

Thông tin cơ bản về cơn bão Kujira, là cơn bão số 01 ảnh hưởng đến Việt Nam năm 2015: (i) Áp suất thấp nhất:Pmin = 985 hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax =45 mph; (iii) số ngày tồn tại: 3 ngày; (iv) vị trí phát sinh: vĩ độ 15.7 - kinh độ 111.9; vị trí kết thúc: vĩ độ 21.3 - kinh độ 106.4.



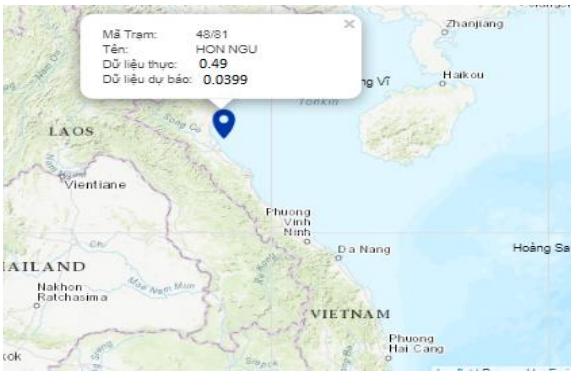
Hình 5.41: Đường đi của bão Kujira

b) Kết quả dự báo nước biển dâng do bão Kujira của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão Kujirabằng mô hình AI tại Hòn Ngư vào lúc **13h00** ngày 20/6/2015 như sau:

- Nước biển dâng quan trắc: 0.49 m;
- Nước biển dâng dự báo: 0.04 m.

Kết quả dự báo của mô hình AI có độ sai lệch nhất định so với dữ liệu quan trắc thực tế.

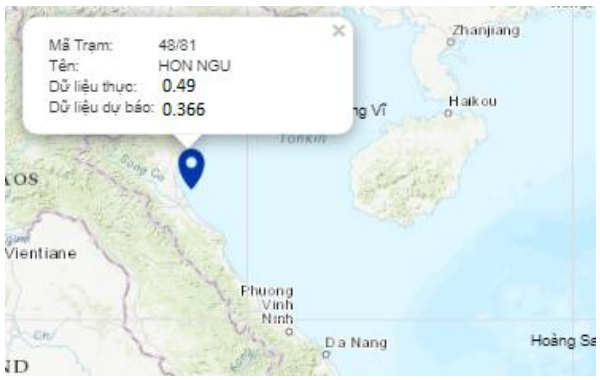


Hình 5.42: Kết quả dự báo nước biển dâng do bão Kujira lúc 13h00, 20/6/2015

Kết quả quan trắc và dự báo nước biển dâng do bão Kujira bằng mô hình AI tại Hòn Ngư vào lúc **15h00** ngày 20/6/2015 như sau:

- Nước biển dâng quan trắc: 0.49 m;
- Nước biển dâng dự báo: 0.37 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

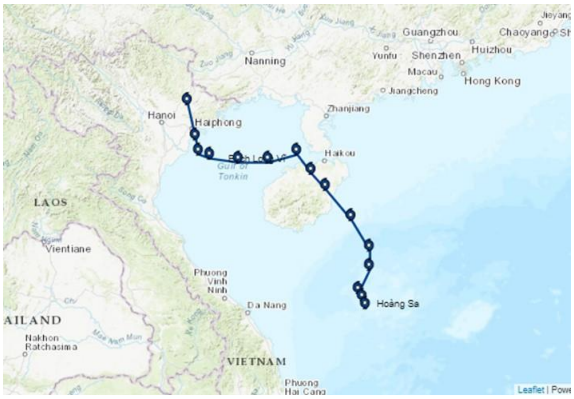


Hình 5.43: Kết quả dự báo nước biển dâng do bão Kujira lúc 15h00, 20/6/2015

5.3.2.3. Đợt 3: Thử nghiệm dự báo nước biển dâng do bão Mirinae (2016)

a) Thông tin về cơn bão Mirinae:

Thông tin cơ bản về cơn bão Mirinae, cơn bão số **01** ảnh hưởng đến Việt Nam năm **2016**: (i) Áp suất thấp nhất: Pmin = **985** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **60** mph; (iv) số ngày tồn tại: **3** ngày; (v) vị trí phát sinh: vĩ độ 17.9 - kinh độ 113.1; Vị trí kết thúc: vĩ độ 20.8 – kinh độ 105.5



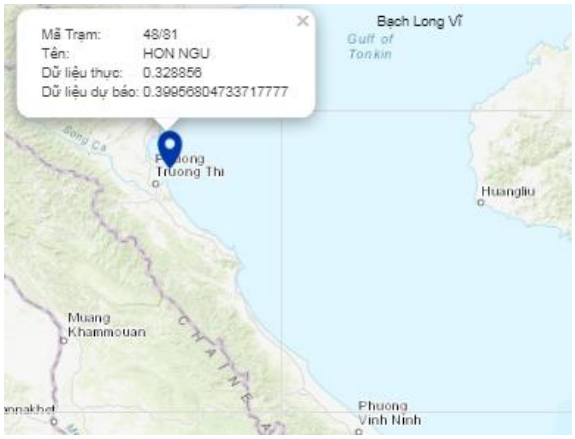
Hình 5.44: Đường đi của bão Mirinae

b) Kết quả dự báo nước biển dâng do bão Mirinae của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão Mirinae bằng mô hình AI tại Hòn Ngư vào lúc **16h00** ngày 27/7/2016 như sau:

- Nước biển dâng quan trắc: 0.33 m;
- Nước biển dâng dự báo: 0.39 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.



Hình 5.45: Kết quả dự báo nước biển dâng do bão Mirinae lúc 16h00, 27/7/2016

Kết quả quan trắc và dự báo nước biển dâng do bão Mirinae bằng mô hình AI tại Hòn Ngư vào lúc **20h00** ngày 27/7/2016 như sau:

- Nước biển dâng quan trắc: 0.216828 m;
- Nước biển dâng dự báo: 0.316962638528935 m.

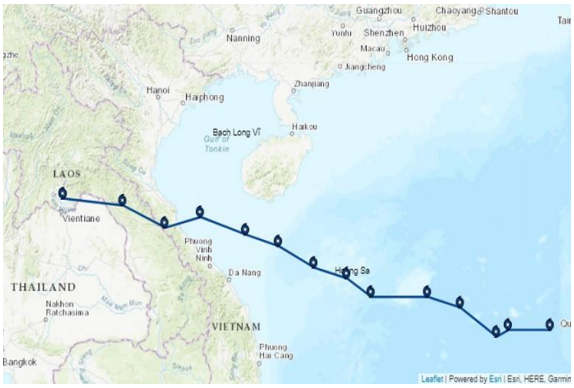
Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.

Hình 5.46: Kết quả dự báo nước biển dâng do bão Mirinae lúc 20h00, 27/7/2016

5.3.2.4. Đợt 4: Thử nghiệm dự báo nước biển dâng do bão Doksuri (2017)

a) Thông tin về cơn bão Doksuri:

Thông tin cơ bản về cơn bão Doksuri, là cơn bão số **10** ảnh hưởng đến Việt Nam năm **2017**: (i) Áp suất thấp nhất: Pmin = **955** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **80** mph; (iv) số ngày tồn tại: **4** ngày; (v) vị trí phát sinh: vĩ độ 14.7- kinh độ 119.6; Vị trí kết thúc: vĩ độ 18.5 - kinh độ 101.2



Hình 5.47: Đường đi của bão Doksuri

b) Kết quả dự báo nước biển dâng do bão Doksuri của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão Doksuri bằng mô hình AI tại Hòn Ngư vào lúc **00h00** ngày 15/09/2017 như sau:

- Nước biển dâng quan trắc: 1.046 m;
- Nước biển dâng dự báo: 0.88 m.

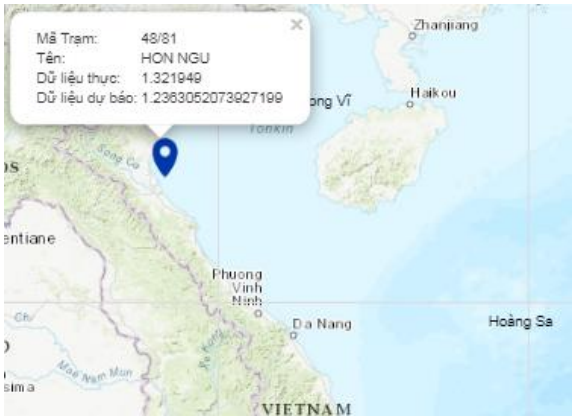
Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.

Hình 5.48: Kết quả dự báo nước biển dâng do bão Doksuri lúc 00h00, 15/09/2017

Kết quả quan trắc và dự báo nước biển dâng do bão Doksuri bằng mô hình AI tại Hòn Ngư vào lúc **04h00** ngày 15/09/2017 như sau:

- Nước biển dâng quan trắc: 1.32 m;
- Nước biển dâng dự báo: 1.24 m.

Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.

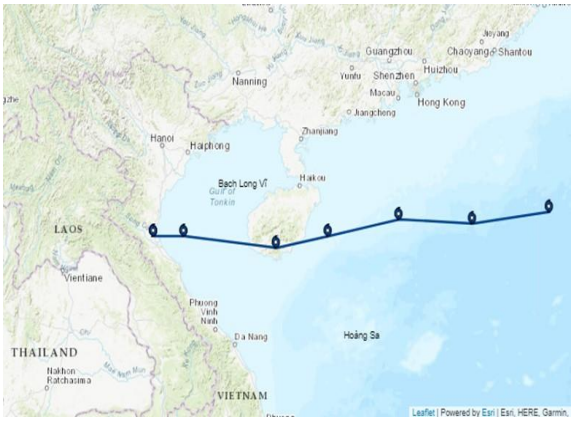


Hình 5.49: Kết quả dự báo nước biển dâng do bão Doksuri lúc 04h00, 15/09/2017

*5.3.2.5. Đợt 5: Thử nghiệm dự báo nước biển dâng do bão **Sơn Tinh** (2018)*

a) Thông tin về cơn bão Sơn Tinh:

Thông tin cơ bản về cơn bão Sơn Tinh, cơn bão số 3 ảnh hưởng đến Việt Nam năm 2018: (i) Áp suất thấp nhất: Pmin = 990 hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = 45 mph; (iv) số ngày tồn tại: 2 ngày; (v) vị trí phát sinh: vĩ độ 19.0 - kinh độ 120.6; Vị trí kết thúc: vĩ độ 19.1 – kinh độ 104.6.



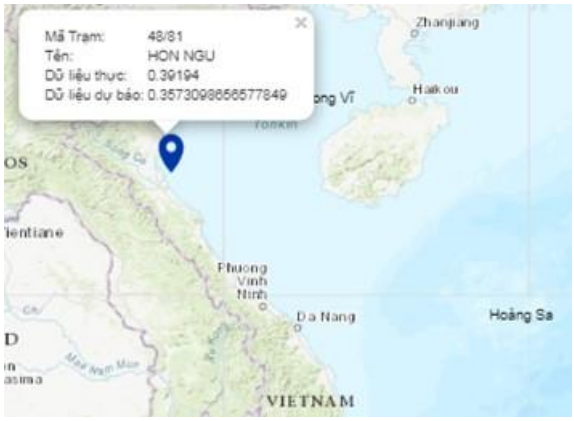
Hình 5.50: Đường đi của bão Sơn Tinh

*b) Kết quả dự báo nước biển dâng do bão **Sơn Tinh** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Sơn Tinh** bằng mô hình AI tại Hòn Ngư vào lúc **13h00** ngày 16/07/2018 như sau:

- Nước biển dâng quan trắc: **0.39** m;
- Nước biển dâng dự báo: **0.36** m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

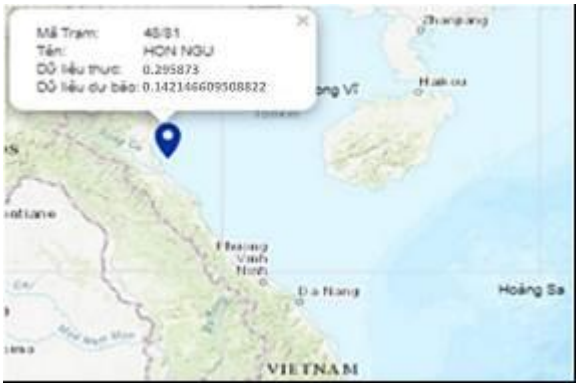


Hình 5.51: Kết quả dự báo nước biển dâng do bão Sơn Tinh lúc 13h00, 16/07/2018

Kết quả quan trắc và dự báo nước biển dâng do bão **Sơn Tinh**bằng mô hình AI tại Hòn Ngư vào lúc **18h00** ngày 16/07/2018 như sau:

- Nước biển dâng quan trắc: 0.295873 m;
- Nước biển dâng dự báo: 0.142146609508822 m.

Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.



Hình 5.52: Kết quả dự báo nước biển dâng do bão Sơn Tinh lúc 18h00, 16/07/2018

5.3.3. Kết quả thử nghiệm Đài KTTV khu vực Đồng bằng Bắc Bộ

Đơn vị chủ trì đã phối hợp với Đài KTTV khu vực Đồng bằng Bắc Bộ thử nghiệm cho 05 đợt nước biển dâng do **05 cơn bão** đổ bộ vào khu vực ven biển Bắc Bộ và Bắc Trung Bộ trong giai đoạn 2008 - 2017.

Bảng 5.5: Danh sách các cơn bão phục vụ thử nghiệm dự báo nước biển dâng tại Đài KTTV khu vực Đồng bằng Bắc Bộ

STT	Năm	Tên cơn bão ảnh hưởng	Trạm quan trắc
1.	2012	Sơn Tinh	Hòn Ngư
2.	2014	Rammasun	Hòn Dấu
3.	2015	Mujigae	Hòn Dấu
4.	2016	Dianmu	Hòn Ngư
5.	2018	Bebinca	Hòn Ngư

*5.3.3.1. Đợt 1: Thử nghiệm dự báo nước biển dâng do bão **Sơn Tinh (2012)***

a) Thông tin về cơn bão Sơn Tinh:

Thông tin về cơn bão **Sơn Tinh**, cơn bão số 8 ảnh hưởng đến Việt Nam năm **2012**: (i) Áp suất thấp nhất: Pmin = **945** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax -**85** mph; (iv) số ngày tồn tại: **5** ngày; (v) vị trí phát sinh: vĩ độ 8.3 - kinh độ 128.1; Vị trí kết thúc: vĩ độ 21.3 - kinh độ 107.1.

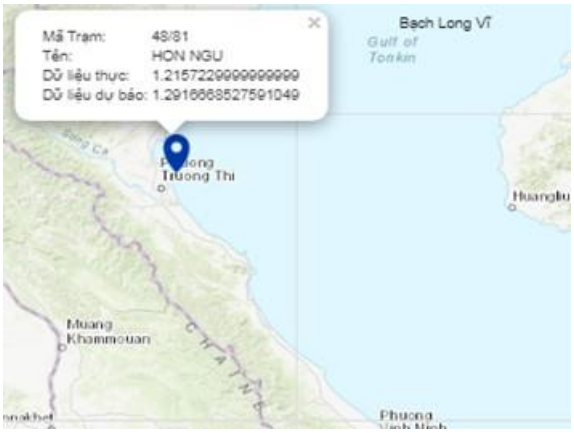
Hình 5.53: Đường đi của bão Sơn Tinh (2012)

*b) Kết quả dự báo nước biển dâng do bão **Sơn Tinh** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Sơn Tinh** bằng mô hình AI tại Hòn Ngư vào lúc **23h00** ngày **27/10/2012** như sau:

- Nước biển dâng quan trắc: **1.22** m;
- Nước biển dâng dự báo: **1.29** m.

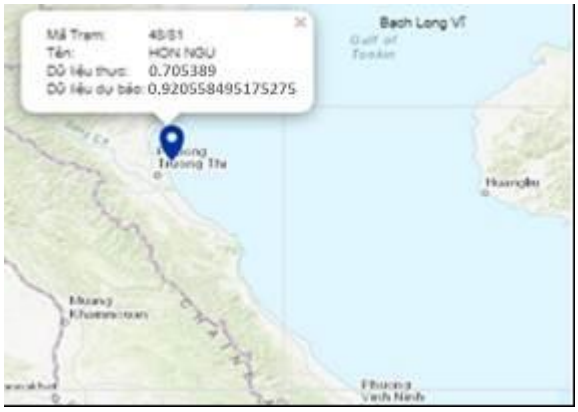
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.



Hình 5.54: Kết quả dự báo nước biển dâng do bão Sơn Tinh lúc 23h00, 27/10/2012

Kết quả quan trắc và dự báo nước biển dâng do bão **Sơn Tinh** bằng mô hình AI tại Hòn Ngư vào lúc **9h00** ngày **28/10/2012** như sau:

- Nước biển dâng quan trắc: **0.705389** m;
- Nước biển dâng dự báo: **0.920558495175275** m.
- *Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.*

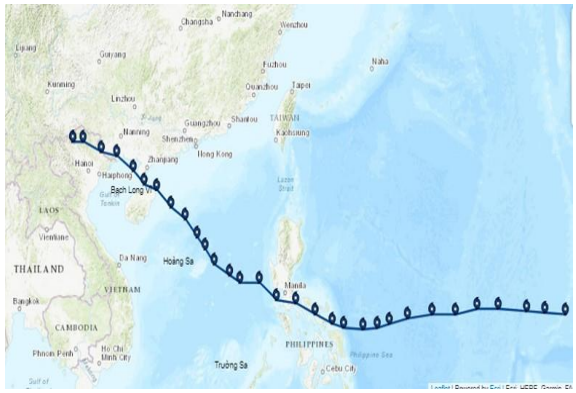


Hình 5.55: Kết quả dự báo nước biển dâng do bão Sơn Tinh lúc 09h00, 28/10/2012

*5.3.3.2. Đợt 2: Thử nghiệm dự báo nước biển dâng do bão **Rammasun** (2014)*

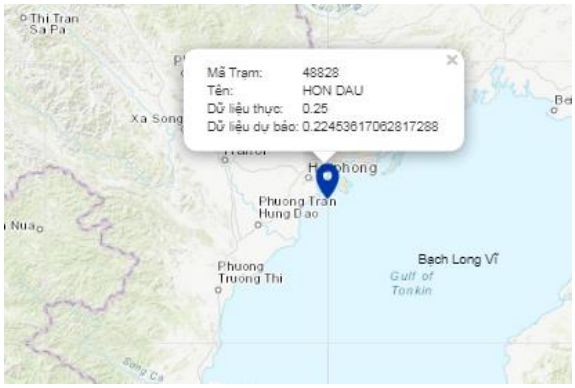
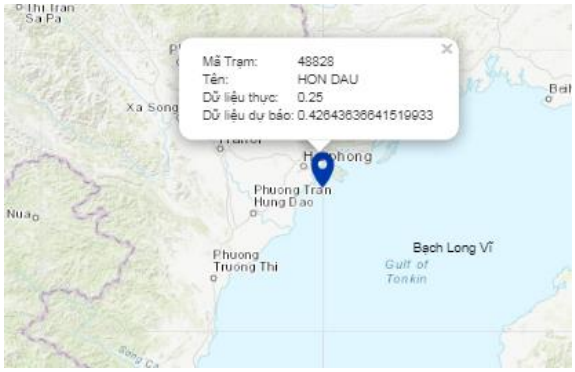
*a) Thông tin về cơn bão **Rammasun**:*

Thông tin về cơn bão **Rammasun**, cơn bão số **2** ảnh hưởng đến Việt Nam năm **2014**: (i) Áp suất thấp nhất: Pmin = **940** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax =**90** mph; (iv) số ngày tồn tại: **8** ngày; (v) vị trí phát sinh: vĩ độ 13.7-kinh độ 145.2; Vị trí kết thúc: vĩ độ 22.1- kinh độ 107.0.

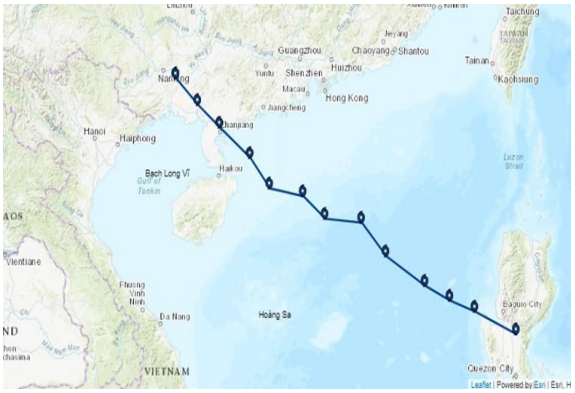


Hình 5.56: Đường đi của bão Rammasun

*b) Kết quả dự báo nước biển dâng do bão **Rammasun** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Rammasun** bằng mô hình AI tại Hòn Dấu vào lúc **07h00** ngày **18/7/2014** như sau: - Nước biển dâng quan trắc: **0.25** m; - Nước biển dâng dự báo: **0.22** m. *Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.*	 Hình 5.57: Kết quả dự báo nước biển dâng do bão Rammasun lúc 7h00, 18/7/2014
Kết quả quan trắc và dự báo nước biển dâng do bão **Rammasun** bằng mô hình AI tại Hòn Dấu vào lúc **08h00** ngày **18/7/2014** như sau: - Nước biển dâng quan trắc: **0.25** m; - Nước biển dâng dự báo: **0.42** m. *Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.*	 Hình 5.58: Kết quả dự báo nước biển dâng do bão Rammasun lúc 8h00, 18/7/2014

*5.3.3.3. Đợt 3: Thử nghiệm dự báo nước biển dâng do **bão Mujigae** (2015)*

*a) Thông tin về cơn bão **Mujigae**:* Thông tin về cơn bão **Mujigae**, cơn bão số **4** ảnh hưởng đến Việt Nam năm **2015**: (i) Áp suất thấp nhất: Pmin = **960** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **75** mph; (iv) số ngày tồn tại: **4** ngày; (v) vị trí phát sinh: vĩ độ 15.2 - kinh độ 122.7; Vị trí kết thúc: vĩ độ 23.2 - kinh độ 108.3.	 Hình 5.59: Đường đi của bão Mujigae

*b) Kết quả dự báo nước biển dâng do bão **Mujigae** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Mujigae** bằng mô hình AI tại Hòn Dấu vào lúc **19h00** ngày **05/10/2015** như sau:

- Nước biển dâng quan trắc: **0.38** m;
- Nước biển dâng dự báo: **0.31** m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.60: Kết quả dự báo nước biển dâng do bão Mujigaelúc 19h00, 05/10/2015

Kết quả quan trắc và dự báo nước biển dâng do bão **Mujigae** bằng mô hình AI tại Hòn Dấu vào lúc **21h00** ngày **05/10/2015** như sau:

- Nước biển dâng quan trắc: **0.38** m;
- Nước biển dâng dự báo: **0.3899402** m.

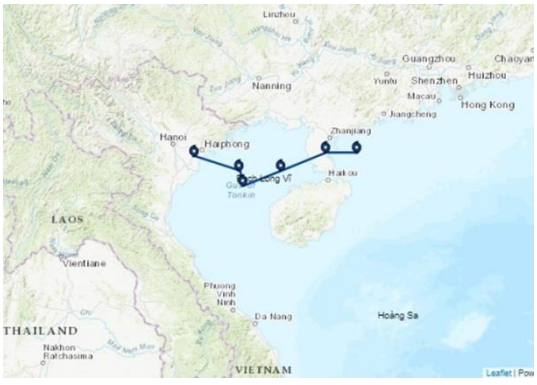
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.61: Kết quả dự báo nước biển dâng do bão Mujigaelúc 21h00, 05/10/2015

*5.3.3.4. Đợt 4: Thử nghiệm dự báo nước biển dâng do bão **Dianmu** (2016)*

*a) Thông tin về cơn bão **Dianmu**:*

Thông tin về cơn bão **Dianmu**, cơn bão số **3** ảnh hưởng đến Việt Nam năm **2016**: (i) Áp suất thấp nhất: Pmin = **980** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax =**45** mph; (iv) số ngày tồn tại: **2** ngày; (v) vị trí phát sinh: vĩ độ 20.6 – kinh độ 112.6; Vị trí kết thúc: vĩ độ 20.8 – kinh độ 105.9;



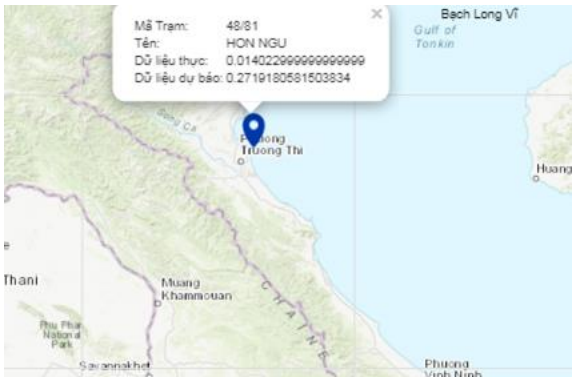
Hình 5.62: Đường đi của bão Dianmu

*b) Kết quả dự báo nước biển dâng do bão **Dianmu** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Dianmu** bằng mô hình AI tại Hòn Ngư vào lúc **16h00 ngày 18/08/2016** như sau:

- Nước biển dâng quan trắc: **0.014** m;
- Nước biển dâng dự báo: **0.27** m.

Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.



Hình 5.63: Kết quả dự báo nước biển dâng do bão Dianmulúc 16h00, 18/08/2016

Kết quả quan trắc và dự báo nước biển dâng do bão **Dianmu** bằng mô hình AI tại Hòn Ngư vào lúc **23h00 ngày 18/08/2016** như sau:

- Nước biển dâng quan trắc: **0.254205** m;
- Nước biển dâng dự báo: **0.272872433877755** m.

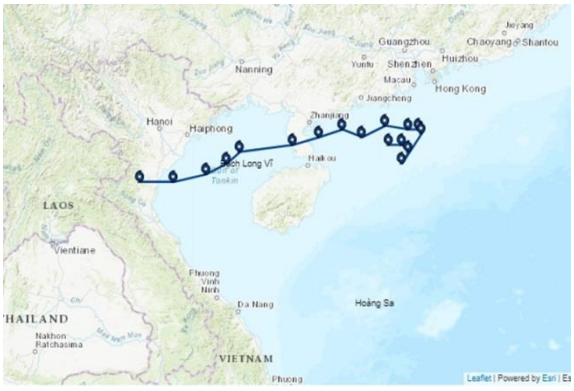
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.64: Kết quả dự báo nước biển dâng do bão Dianmulúc 22h00, 18/08/2016

*5.3.3.5. Đợt 5: Thử nghiệm dự báo nước biển dâng do bão **Bebinca** (2018)*

*a) Thông tin về cơn bão **Bebinca***

Thông tin về cơn bão **Bebinca**, cơn bão số **3** ảnh hưởng đến Việt Nam năm **2018**: (i) Áp suất thấp nhất: Pmin = **985** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **50** mph; (iv) số ngày tồn tại: **4** ngày; (v) vị trí phát sinh: vĩ độ 20.6 – kinh độ 112.5; Vị trí kết thúc: vĩ độ 19.6 – kinh độ 105.9;



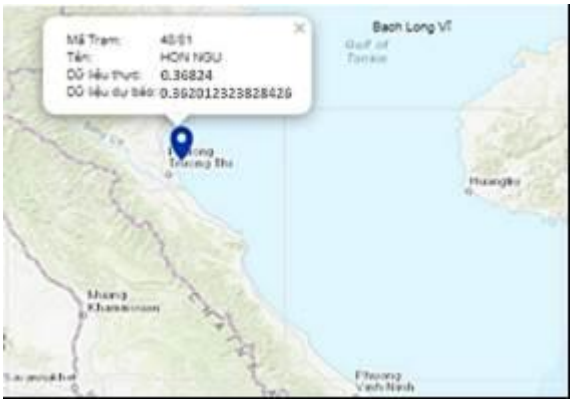
Hình 5.65: Đường đi của bão Bebinca

*b) Kết quả dự báo nước biển dâng do bão **Bebinca** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Bebinca**bằng mô hình AI tại Hòn Ngư vào lúc **19h00 ngày 16/08/2018** như sau:

- Nước biển dâng quan trắc: **0.36824** m;
- Nước biển dâng dự báo: **0.362012323828426** m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

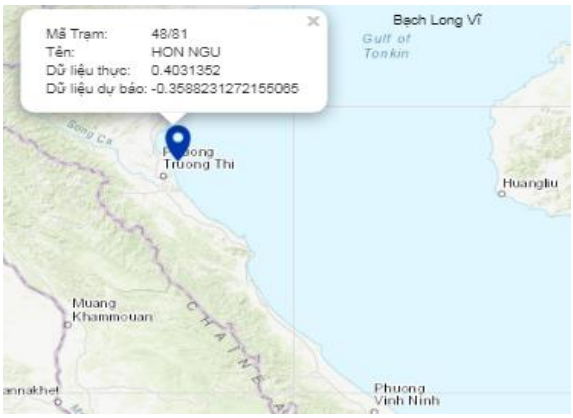


Hình 5.66: Kết quả dự báo nước biển dâng do bão Bebinca lúc 19h00, 16/08/2018

Kết quả quan trắc và dự báo nước biển dâng do bão **Bebinca** bằng mô hình AI tại Hòn Ngư vào lúc **22h00 ngày 16/08/2018** như sau:

- Nước biển dâng quan trắc: **0.4** m;
- Nước biển dâng dự báo: **0.36** m.

Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.



Hình 5.67: Kết quả dự báo nước biển dâng do bão Bebinca lúc 22h00, 16/08/2018

5.3.4. Kết quả thử nghiệm KTTV khu vực Đông Bắc

Đơn vị chủ trì đã phối hợp với Đài KTTV khu vực Đông Bắc thử nghiệm cho 05 đợt nước biển dâng do **05 cơn bão** đổ bộ vào khu vực ven biển Bắc Bộ và Bắc Trung Bộ trong giai đoạn 2008 - 2017.

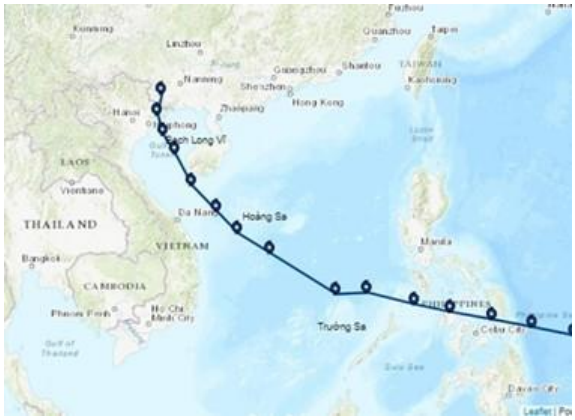
Bảng 5.6: Danh sách các cơn bão phục vụ thử nghiệm dự báo nước biển dâng tại Đài KTTV khu vực Đông Bắc

STT	Năm	Tên cơn bão ảnh hưởng	Trạm quan trắc
1.	2013	Haiyan	Hòn Dấu
2.	2015	Kujira	Hòn Dấu
3.	2016	Haima	Hòn Dấu
4.	2017	Hato	Hòn Dấu
5.	2018	Doksuri	Hòn Dấu

*5.3.4.1. Đợt 1: Thử nghiệm dự báo nước biển dâng do bão **Haiyan** (2013)*

*a) Thông tin về cơn bão **Haiyan***

Thông tin cơ bản về cơn bão Haiyan, là cơn bão số 13 ảnh hưởng đến Việt Nam năm 2013 như sau: (i) Áp suất thấp nhất Pmin = 895 hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = 125 mph; (iii) số ngày tồn tại: 8 ngày; (iv) vị trí phát sinh: vĩ độ 6.1 - kinh độ 153.3; vị trí kết thúc: vĩ độ 22.8 - kinh độ 108.6.



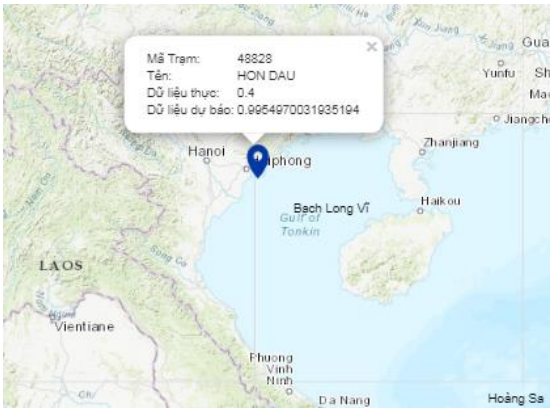
Hình 5.68: Đường đi của bão Haiyan 2013

b) Kết quả dự báo nước biển dâng do bão Haiyan của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão bằng mô hình AI tại Hòn Dấu vào lúc **5h00** ngày 10/11/2013 như sau:

- Nước biển dâng quan trắc: 0.40 m
- Nước biển dâng dự báo: 0.99 m.

Kết quả dự báo của mô hình AI có độ sai lệch nhất định so với dữ liệu quan trắc thực tế.



Hình 5.69: Kết quả dự báo nước biển dâng do bão Haiyan lúc 5h00, 10/11/2013

Kết quả quan trắc và dự báo nước biển dâng do bão bằng mô hình AI tại Hòn Dấu vào lúc **10h00** ngày 10/11/2013 như sau:

- Nước biển dâng quan trắc: 0.40 m
- Nước biển dâng dự báo: 0.4341861 m.

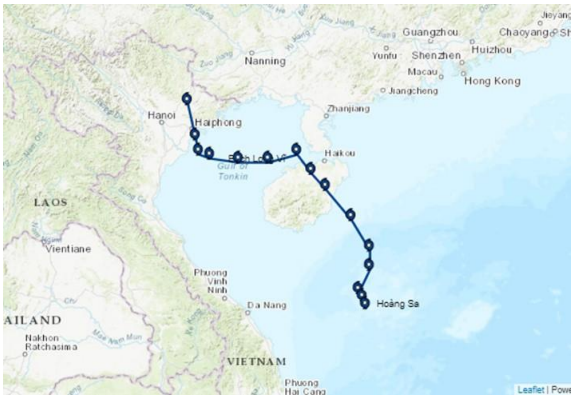
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.70: Kết quả dự báo nước biển dâng do bão Haiyan lúc 10h00, 10/11/2013

5.3.4.2. Đợt 2: Thử nghiệm dự báo nước biển dâng do bão Kujira (2015)

a) Thông tin về cơn bão Kujira:

Thông tin cơ bản về cơn bão Kujira, là cơn bão số 01 ảnh hưởng đến Việt Nam năm 2015 như sau: (i) Áp suất thấp nhất:Pmin = 985 hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = 45 mph; (iii) số ngày tồn tại: 3 ngày; (iv) vị trí phát sinh: vĩ độ 15.7- kinh độ 111.9; vị trí kết thúc: vĩ độ 21.3 - kinh độ 106.4.



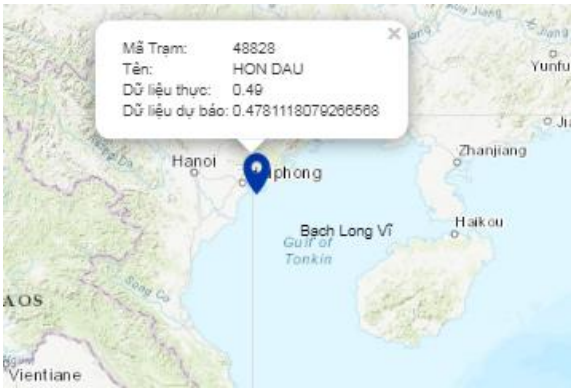
Hình 5.71: Đường đi của bão Kujira

b) Kết quả dự báo nước biển dâng do bão Kujira của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão Kujira bằng mô hình AI tại Hòn Dấu vào lúc **19h00** ngày 20/6/2015 như sau:

- Nước biển dâng quan trắc: 0.49 m;
- Nước biển dâng dự báo: 0.48 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.



Hình 5.72: Kết quả dự báo nước biển dâng do bão Kujira lúc 19h00, 20/6/2015

Kết quả quan trắc và dự báo nước biển dâng do bão Kujira bằng mô hình AI tại Hòn Dấu vào lúc **21h00** ngày 20/6/2015 như sau:

- Nước biển dâng quan trắc: 0.49 m;
- Nước biển dâng dự báo: 0.47998023 m.

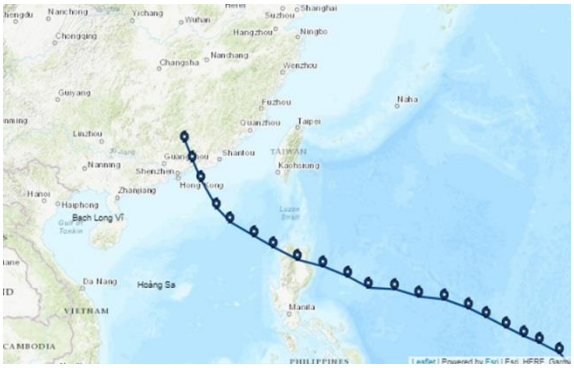
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.73: Kết quả dự báo nước biển dâng do bão Kujira lúc 21h00, 20/6/2015

*5.3.4.3. Đợt 3: Thử nghiệm dự báo nước biển dâng do **bão Haima** (2016)*

*a) Thông tin về cơn bão **Haima**:*

Thông tin cơ bản về cơn bão **Haima**, là cơn bão số 8 ảnh hưởng đến Việt Nam năm 2016 như sau: (i) Áp suất thấp nhất:Pmin = **900** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **115** mph; (iii) số ngày tồn tại: **7 ngày**; (iv) vị trí phát sinh: vĩ độ 7.9 - kinh độ 144.3; vị trí kết thúc:vĩ độ 25.6 - kinh độ 115.2.



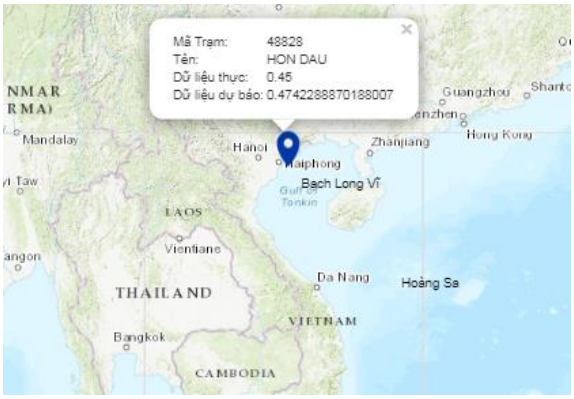
Hình 5.74: Đường đi của bão Haima

*b) Kết quả dự báo nước biển dâng do bão **Haima**của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Haima** bằng mô hình AI tại Hòn Dấu vào lúc **10h00** ngày 20/10/2016 như sau:

- Nước biển dâng quan trắc: 0.45 m;
- Nước biển dâng dự báo: 0.47 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

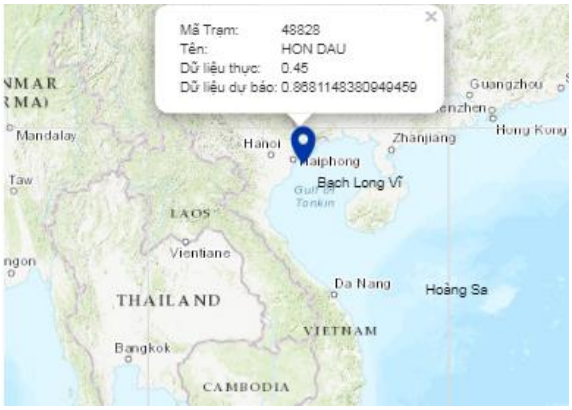


Hình 5.75: Kết quả dự báo nước biển dâng do bão Haima lúc 10h00, 20/10/2016

Kết quả quan trắc và dự báo nước biển dâng do bão **Haima** bằng mô hình AI tại Hòn Dấu vào lúc **12h00** ngày 20/10/2016 như sau:

- Nước biển dâng quan trắc: 0.45 m;
- Nước biển dâng dự báo: 0.87 m.

Kết quả dự báo của mô hình AI có độ sai lệch nhất định so với dữ liệu quan trắc thực tế.

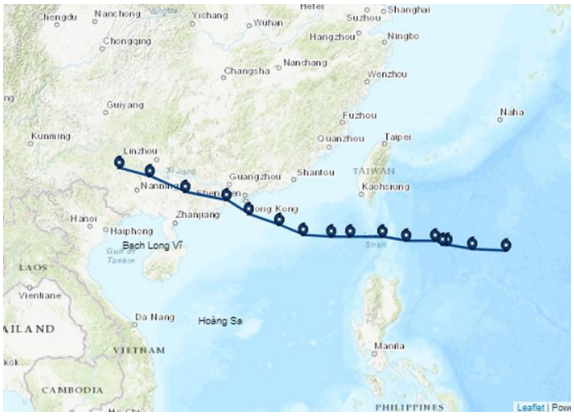


Hình 5.76: Kết quả dự báo nước biển dâng do bão Haima lúc 12h00, 20/10/2016

*5.3.4.4. Đợt 4: Thử nghiệm dự báo nước biển dâng do bão **Hato** (2017)*

*a) Thông tin về cơn bão **Hato***

Thông tin cơ bản về cơn bão **Haima**, là cơn bão số **6** ảnh hưởng đến Việt Nam năm **2017** như sau: (i) Áp suất thấp nhất:Pmin = **960** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **80** mph; (iii) số ngày tồn tại: **4 ngày**; (iv) vị trí phát sinh: vĩ độ 19.3 - kinh độ 129.0; vị trí kết thúc: vĩ độ 23.2 - kinh độ 107.5;



Hình 5.77: Đường đi của bão Hato

*b) Kết quả dự báo nước biển dâng do bão **Hato** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Hato** bằng mô hình AI tại Hòn Dấu vào lúc **19h00** ngày 21/8/2017 như sau:

- Nước biển dâng quan trắc: 1.21 m;
- Nước biển dâng dự báo: 1.47 m.

Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.

Hình 5.78: Kết quả dự báo nước biển dâng do bão Hato lúc 19h00, 21/8/2017

Kết quả quan trắc và dự báo nước biển dâng do bão **Hato** bằng mô hình AI tại Hòn Dấu vào lúc **21h00** ngày 21/8/2017 như sau:

- Nước biển dâng quan trắc: 1.21 m;
- Nước biển dâng dự báo: 1.1593637 m.

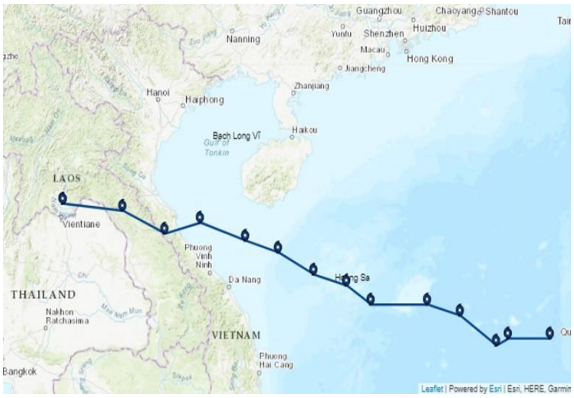
Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

Hình 5.79: Kết quả dự báo nước biển dâng do bão Hato lúc 21h00, 21/8/2017

5.3.4.5. Đợt 5: Thử nghiệm dự báo nước biển dâng do bão Doksuri(2018)

*a) Thông tin về cơn bão **Doksuri***

Thông tin cơ bản về cơn bão **Doksuri,** là cơn bão số **10** ảnh hưởng đến Việt Nam năm **2017**: (i) Áp suất thấp nhất:Pmin = **955** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **80** mph; (iii) số ngày tồn tại: **4 ngày**; (iv) vị trí phát sinh: vĩ độ 14.7 – kinh độ 119.6; vị trí kết thúc: vĩ độ 18.5 – kinh độ 101.2;



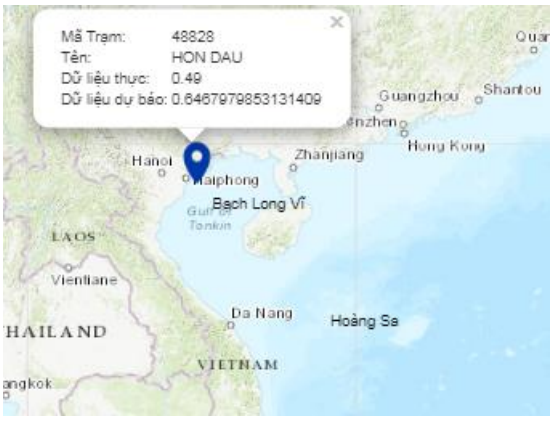
Hình 5.80: Đường đi của bão Doksuri

b) Kết quả dự báo nước biển dâng do bão Doksuri của mô hình AI

Kết quả quan trắc và dự báo nước biển dâng do bão **Doksuri** bằng mô hình AI tại Hòn Dấu vào lúc **19h00** ngày 15/9/2017 như sau:

- Nước biển dâng quan trắc: 0.49 m;
- Nước biển dâng dự báo: 0.65 m.

Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.

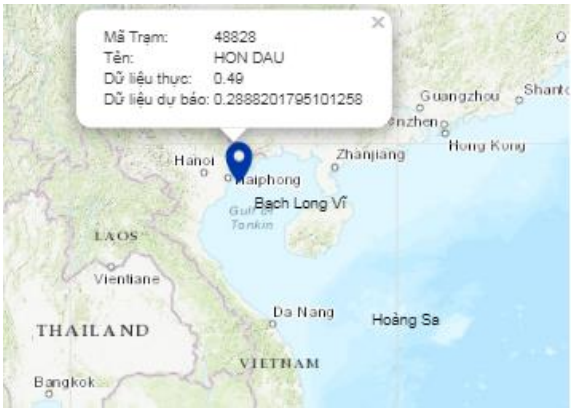


Hình 5.81: Kết quả dự báo nước biển dâng do bão Doksuri lúc 19h00, 15/9/2017

Kết quả quan trắc và dự báo nước biển dâng do bão **Doksuri** bằng mô hình AI tại Hòn Dấu vào lúc **07h00** ngày 16/9/2017 như sau:

- Nước biển dâng quan trắc: 0.49 m;
- Nước biển dâng dự báo: 0.29 m.

Kết quả dự báo của mô hình AI có độ sai khác nhất định so với dữ liệu quan trắc thực tế.



Hình 5.82: Kết quả dự báo nước biển dâng do bão Doksuri lúc 07h00, 16/9/2017

5.3.5. Kết quả thử nghiệm Đài KTTV khu vực Bắc Trung Bộ

Đơn vị chủ trì đã phối hợp với Đài KTTV khu vực Bắc Trung Bộ thử nghiệm cho 05 đợt nước biển dâng do **05 cơn bão** đổ bộ vào khu vực ven biển Bắc Bộ và Bắc Trung Bộ trong giai đoạn 2008 - 2017.

Bảng 5.7: Danh sách các cơn bão phục vụ thử nghiệm dự báo nước biển dâng tại Đài KTTV khu vực Bắc Trung Bộ

STT	Năm	Tên cơn bão ảnh hưởng	Trạm quan trắc
1.	2013	Mangkhut	Hòn Ngư
2.	2013	Wutip	Hòn Ngư
3.	2016	Dianmu	Hòn Ngư
4.	2016	Sarika	Hòn Ngư
5.	2018	Bebinca	Hòn Dấu

*5.3.5.1. Đợt 1: Thử nghiệm dự báo nước biển dâng do bão **Mangkhut** (2013)*

*a) Thông tin về cơn bão **Mangkhut***

Cơn bão **Mangkhut,** là cơn bão số **06** ảnh hưởng đến Việt Nam năm **2013**: (i) Áp suất thấp nhất: Pmin = **994** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **45** mph; (iii) số ngày tồn tại: **2 ngày**; (iv) vị trí phát sinh: vĩ độ 13.2 - kinh độ 114.6; vị trí kết thúc: vĩ độ 20.1 - kinh độ 105.2.

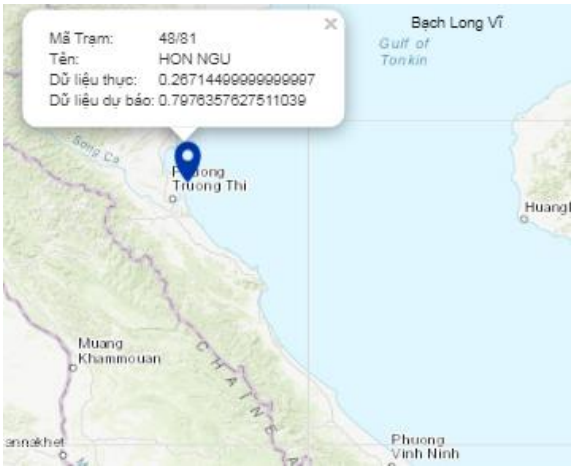
Hình 5.83: Đường đi của bão Mangkhut

*b) Kết quả dự báo nước biển dâng do bão **Mangkhut** của mô hình AI*

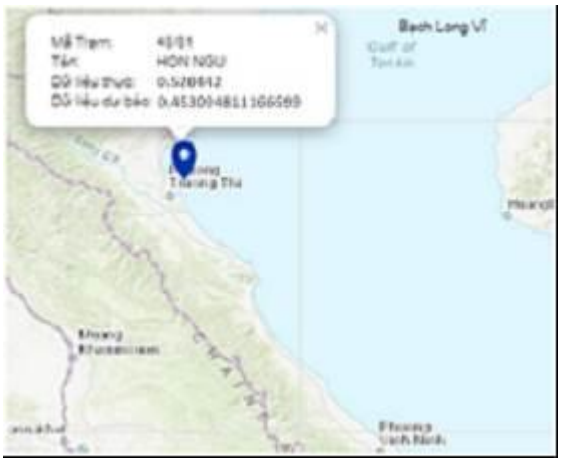
Kết quả quan trắc và dự báo nước biển dâng do bão **Mangkhut** bằng mô hình AI tại Hòn Dấu vào lúc **4h00 ngày 07/08/2013** như sau:

- Nước biển dâng quan trắc: 0.27 m;
- Nước biển dâng dự báo: 0.79 m.

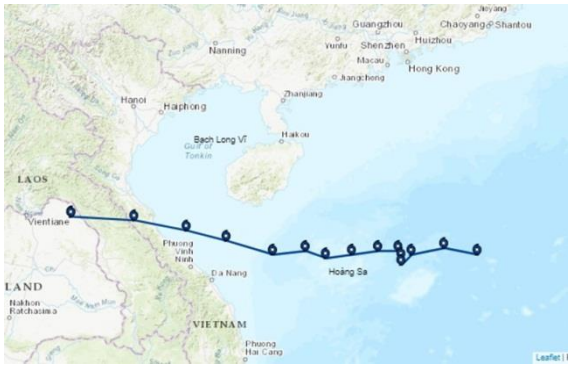
Kết quả dự báo của mô hình AI có độ sai khác nhất định so với dữ liệu quan trắc thực tế.



Hình 5.84: Kết quả dự báo nước biển dâng

	do bão Mangkhut lúc 4h00, 07/8/2013
Kết quả quan trắc và dự báo nước biển dâng do bão **Mangkhut** bằng mô hình AI tại Hòn Dấu vào lúc **16h00 ngày 07/08/2013** như sau: - Nước biển dâng quan trắc: 0.520442 m; - Nước biển dâng dự báo: 0.453094811166599 m. *Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.*	 Hình 5.85: Kết quả dự báo nước biển dâng do bão Mangkhut lúc 16h00, 07/8/2013

*5.3.5.2. Đợt 2: Thử nghiệm dự báo nước biển dâng do bão **Wutip** (2013)*

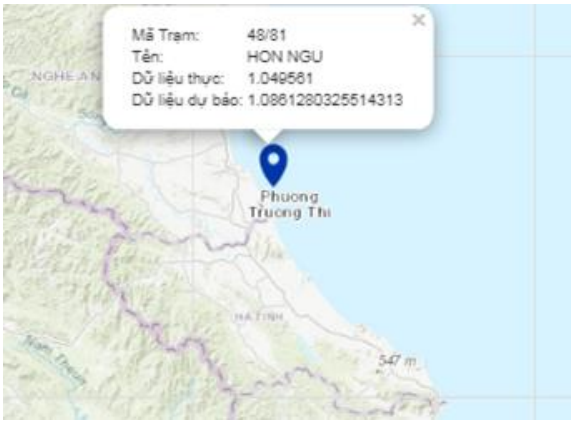
*a) Thông tin về cơn bão **Wutip*** Thông tin cơ bản về cơn bão **Wutip,** là cơn bão số **10** ảnh hưởng đến Việt Nam năm **2013**: (i) Áp suất thấp nhất: Pmin = **960** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **75** mph; (iii) số ngày tồn tại: **5 ngày**; (iv) vị trí phát sinh: vĩ độ 15.8 - kinh độ 117.6; vị trí kết thúc: vĩ độ 18.0 – kinh độ 101.8;	 Hình 5.86: Đường đi của bão Wutip

*b) Kết quả dự báo nước biển dâng do bão **Wutip** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Wutip** bằng mô hình AI tại Hòn Dấu vào lúc **01h00 ngày 30/09/2013** như sau:

- Nước biển dâng quan trắc: 1.04 m;
- Nước biển dâng dự báo: 1.09 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

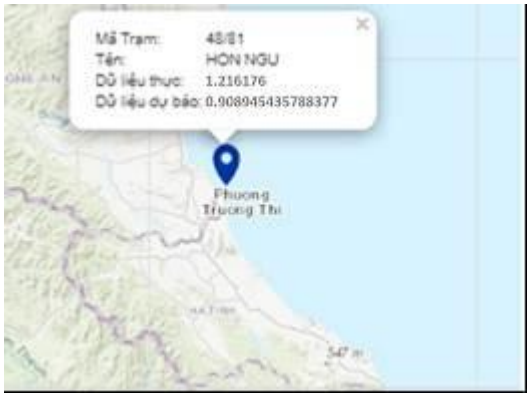


Hình 5.87: Kết quả dự báo nước biển dâng do bão Wutip lúc 01h00, 30/09/2013

Kết quả quan trắc và dự báo nước biển dâng do bão **Wutip** bằng mô hình AI tại Hòn Dấu vào lúc **11h00 ngày 30/09/2013** như sau:

- Nước biển dâng quan trắc: 1.216176 m;
- Nước biển dâng dự báo: 0.908945435788377 m.

Kết quả dự báo của mô hình AI tương đối phù hợp với dữ liệu quan trắc thực tế.

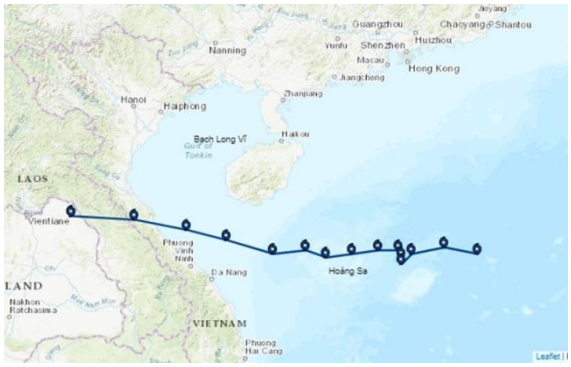


Hình 5.88: Kết quả dự báo nước biển dâng do bão Wutip lúc 11h00, 30/09/2013

*5.3.5.3. Đợt 3: Thử nghiệm dự báo nước biển dâng do bão **Dianmu** (2016)*

*a) Thông tin về cơn bão **Dianmu***

Thông tin cơ bản về cơn bão **Dianmu,** là cơn bão số **3** ảnh hưởng đến Việt Nam năm **2016**: (i) Áp suất thấp nhất: Pmin = **980** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **45** mph; (iii) số ngày tồn tại: **2 ngày**; (iv) vị trí phát sinh: vĩ độ 20.6 – kinh độ 112.6; vị trí kết thúc: vĩ độ 20.8 – kinh độ 105.9.



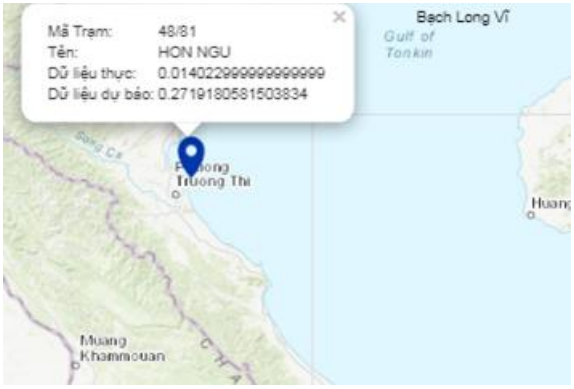
Hình 5.89: Đường đi của bão Dianmu

*b) Kết quả dự báo nước biển dâng do bão **Dianmu** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Dianmu** bằng mô hình AI tại Hòn Ngư vào lúc **16h00 ngày 18/08/2016** như sau:

- Nước biển dâng quan trắc: 0.14 m;
- Nước biển dâng dự báo: 0.27 m.

Kết quả dự báo của mô hình AI phù hợp dữ liệu quan trắc thực tế.

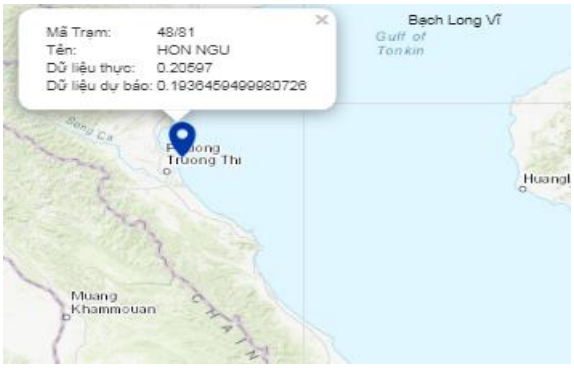


Hình 5.90: Kết quả dự báo nước biển dâng do bão Dianmu lúc 16h00, 18/08/2016

Kết quả quan trắc và dự báo nước biển dâng do bão **Dianmu** bằng mô hình AI tại Hòn Ngư vào lúc **22h00 ngày 18/08/2016** như sau:

- Nước biển dâng quan trắc: 0.14 m;
- Nước biển dâng dự báo: 0.27 m.

Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.

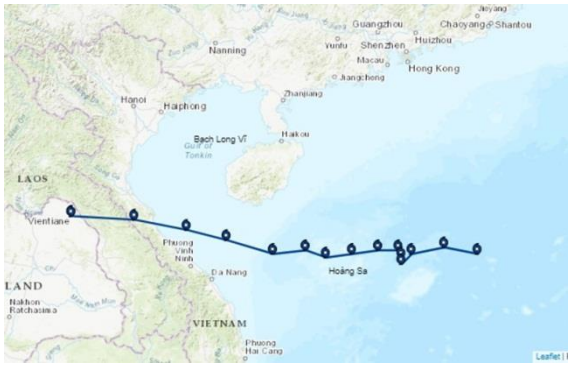


Hình 5.91: Kết quả dự báo nước biển dâng do bão Dianmu lúc 22h00, 18/08/2016

*5.3.5.4. Đợt 4: Thử nghiệm dự báo nước biển dâng do bão **Sarika** (2016)*

*a) Thông tin về cơn bão **Sarika***

Thông tin cơ bản về cơn bão **Sarika,** là cơn bão số **7** ảnh hưởng đến Việt Nam năm **2016**: (i) Áp suất thấp nhất: Pmin = **935** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax = **95** mph; (iii) số ngày tồn tại: **6 ngày**; (iv) vị trí phát sinh: vĩ độ 13.2 – kinh độ 130.2; vị trí kết thúc: vĩ độ 21.6 – kinh độ 108.2.



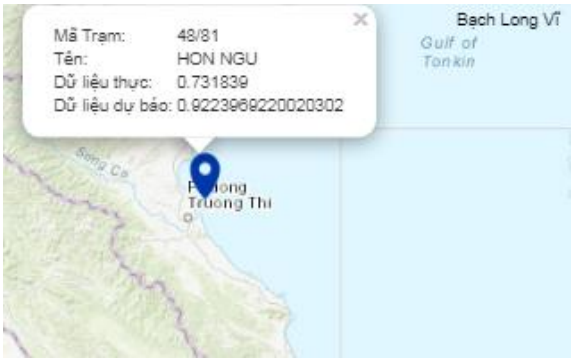
Hình 5.92: Đường đi của bão Sarika

*b) Kết quả dự báo nước biển dâng do bão **Sarika** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Sarika** bằng mô hình AI tại Hòn Ngư vào lúc **13h00 ngày 14/10/2016** như sau:

- Nước biển dâng quan trắc: 0.73 m;
- Nước biển dâng dự báo: 0.92 m.

Kết quả dự báo của mô hình AI phù hợp dữ liệu quan trắc thực tế.

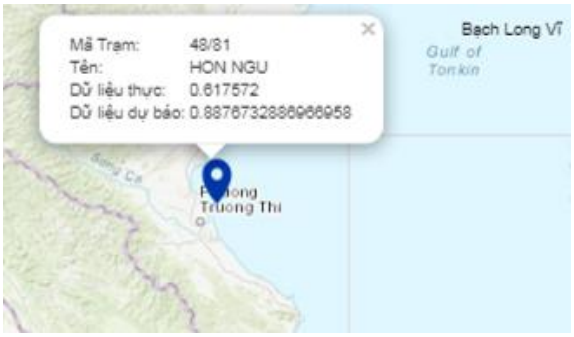


Hình 5.93: Kết quả dự báo nước biển dâng do bão Sarika lúc 13h00, 14/10/2016

Kết quả quan trắc và dự báo nước biển dâng do bão **Sarika** bằng mô hình AI tại Hòn Ngư vào lúc **17h00 ngày 14/10/2016** như sau:

- Nước biển dâng quan trắc: 0.61 m;
- Nước biển dâng dự báo: 0.88 m.

Kết quả dự báo của mô hình AI phù hợp dữ liệu quan trắc thực tế.

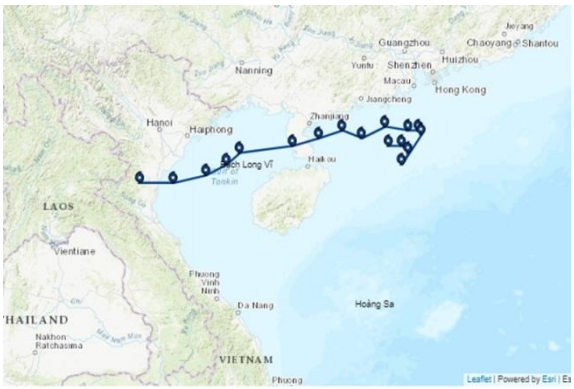


Hình 5.94: Kết quả dự báo nước biển dâng do bão Sarika lúc 17h00, 14/10/2016

5.3.5.5. Đợt 5: Thử nghiệm dự báo nước biển dâng do bão Bebinca (2018)

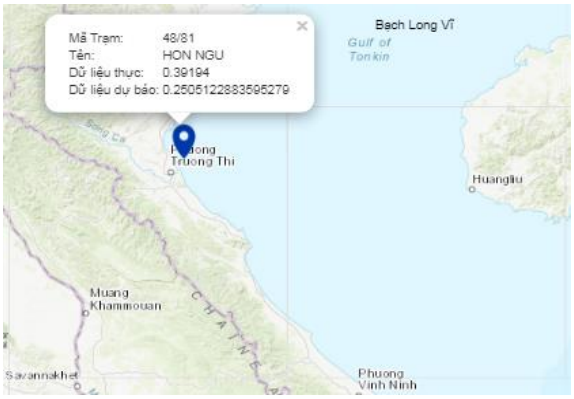
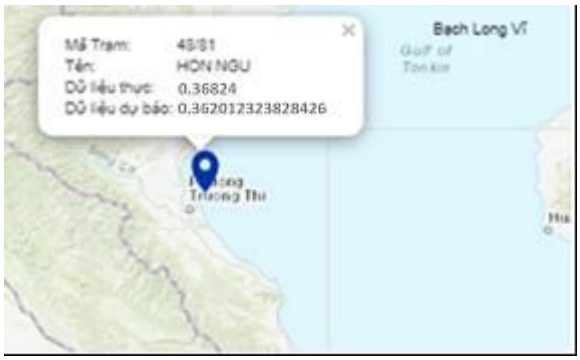
*a) Thông tin về cơn bão **Bebinca***

Thông tin về cơn bão **Bebinca**, cơn bão số **3** ảnh hưởng đến Việt Nam năm **2018**: (i) Áp suất thấp nhất: Pmin = **985** hPa; (ii) Tốc độ gió lớn nhất duy trì trong 10 phút: Vmax =**50** mph; (iv) số ngày tồn tại: **4** ngày; (v) vị trí phát sinh: vĩ độ 20.6 – kinh độ 112.5; Vị trí kết thúc: vĩ độ 19.6 – kinh độ 105.9;



Hình 5.95: Đường đi của bão Bebinca

*b) Kết quả dự báo nước biển dâng do bão **Bebinca** của mô hình AI*

Kết quả quan trắc và dự báo nước biển dâng do bão **Bebinca** bằng mô hình AI tại Hòn Ngư vào lúc **16h00 ngày 16/08/2018** như sau: - Nước biển dâng quan trắc: **0.40** m; - Nước biển dâng dự báo: **0.25** m. *Kết quả dự báo của mô hình AI tương đối phù hợp so với dữ liệu quan trắc thực tế.*	 Hình 5.96: Kết quả dự báo nước biển dâng do bão Bebinca lúc 16h00, 16/08/2018
Kết quả quan trắc và dự báo nước biển dâng do bão **Bebinca** bằng mô hình AI tại Hòn Ngư vào lúc **19h00 ngày 16/08/2018** như sau: - Nước biển dâng quan trắc: **0.36824** m; - Nước biển dâng dự báo: 0.362012323828426 m. *Kết quả dự báo của mô hình AI bám sát dữ liệu quan trắc thực tế.*	 Hình 5.97: Kết quả dự báo nước biển dâng do bão Bebinca lúc 19h00, 16/08/2018

5.4. Thử nghiệm mô hình AI hỗ trợ dự báo lũ trên hệ thống sông Hồng

5.4.1. Thông tin chung về thử nghiệm mô hình AI dự báo lũ

5.4.1.1. Về nội dung, và khối lượng thực hiện

- Nội dung thực hiện: Thử nghiệm và đánh giá mô hình hệ thống AI hỗ trợ dự báo **lũ** trên hệ thống sông Hồng.

- Về khối lượng thử nghiệm: cho **10 đợt lũ** trên hệ thống sông Hồng trong giai đoạn 2008 - 2019.

5.4.1.2. Về thành phần, thời gian và địa điểm thực hiện

- Về thành phần tham gia: Các dự báo viên, kỹ thuật viên thuộc Trung tâm Dự báo KTTV quốc gia, Đài KTTV khu vực Đồng bằng Bắc Bộ và đơn vị chủ trì.

- Về thời gian thực hiện: 30 ngày, trong khoảng từ 18/9- 18/10/2020.

- Địa điểm thực hiện: Tại các phòng dự báo nghiệp vụ thuộc Trung tâm Dự báo KTTV quốc gia và các Đài KTTV khu vực Đồng bằng Bắc Bộ.

5.4.1.3. Về công cụ mô hình thử nghiệm

Đơn vị chủ trì cung cấp các công cụ phục vụ thử nghiệm bao gồm:

- Website hỗ trợ dữ báo lũ tại địa chỉ: http://ai.thoitietnguyhiem.gov.vn/
- Tài khoản (username/pass): dubao/dubao123;
- Tài liệu hướng dẫn sử dụng đã được đơn vị chủ trì bàn giao cho các đơn vị tham gia thử nghiệm tại đợt chuyển giao, đào tạo sử dụng hệ thống AI hỗ trợ dự báo một số hiện tượng thời tiếp nguy hiểm.

5.4.1.4. Về dữ liệu đầu vào cho mô hình thử nghiệm

Dữ liệu triển khai để phục vụ hệ thống AI hỗ trợ dự báo lũ trên hệ thống sông Hồng cụ thể như sau:

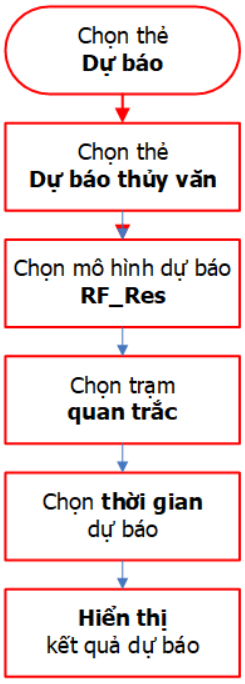
- Dữ liệu quan trắc của 3 trạm thủy văn là Vụ Quang, Hà Nội, Hưng Yên;
- Thời gian số liệu thu thập: trong thời đoạn 10 năm, từ năm 2008 - 2017.

*5.4.1.5. Về Quy trình vận hành thử nghiệm hệ thống AI hỗ trợ dự báo **lũ***

Quy trình vận hành thử nghiệm hệ thống AI hỗ trợ dự báo **lũ** (hình 5.98) gồm 6 bước:

- Bước 1. Chọn thẻ dự báo;
- Bước 2. Chọn thẻ dự báothủy văn;
- Bước 3. Chọn mô hình AI dự báo lũ (RF-RES);
- Bước 4. Chọn trạm quan trắc để lấy dữ liệu cho mô hình AI dự báo lũ;
- Bước 5. Chọn thời gian dự báo;
- Bước 6: Hiển thị kết quả dự báo lũ.

Chi tiết về quy trình dự báo nước biển dâng do bão được trình bày tại Báo cáo công việc 186, 187.



Hình 5.98: Quy trình hỗ trợ dự báo lũ bằng mô hình AI

Thực hiện thử nghiệm dự báo lũ theo khoảng thời gian thiết lập như sau:

- t = 3: Số ngày muốn dự báo;

- nump = t*8 # số lần dự báo trong ngày;
- Bước dự báo: 3 giờ/lần;
- Mô hình AI: Kết hợp ARIMA và Random Forest.

Kết quả thử nghiệm và đánh giá mô hình AI hỗ trợ dự báo **lũ** cụ thể như sau:

5.4.2. Kết quả thử nghiệm Trung tâm Dự báo KTTV quốc gia

Kết quả dự báo cho các đợt lũ như sau.

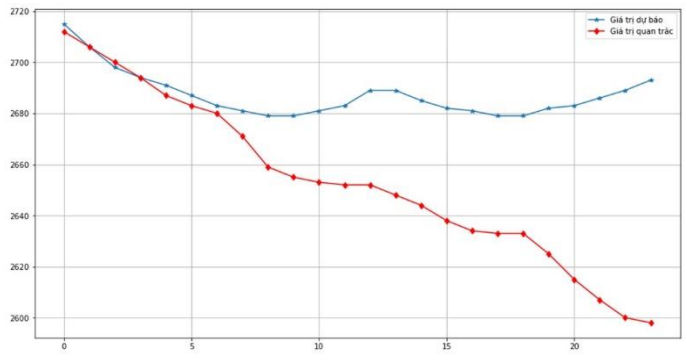
5.4.2.1. Đợt lũ 1: từ 11/01/2019 đến 13/01/2019 trạm Yên Bái 74129

Bảng 5.8: So sánh kết quả dự báo mực nước và thực đo đợt lũ 1 tại trạm Yên Bái

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	11/01/2019 1:00	2712	2715	12	12/01/2019 13:00	2652	2689
1	11/01/2019 4:00	2706	2706	13	12/01/2019 16:00	2648	2689
2	11/01/2019 7:00	2700	2698	14	12/01/2019 19:00	2644	2685
3	11/01/2019 10:00	2694	2694	15	12/01/2019 22:00	2638	2682
4	11/01/2019 13:00	2687	2691	16	13/01/2019 1:00	2634	2681
5	11/01/2019 16:00	2683	2687	17	13/01/2019 4:00	2633	2679
6	11/01/2019 19:00	2680	2683	18	13/01/2019 7:00	2633	2679
7	11/01/2019 22:00	2671	2681	19	13/01/2019 10:00	2625	2682
8	12/01/2019 1:00	2659	2679	20	13/01/2019 13:00	2615	2683
9	12/01/2019 4:00	2655	2679	21	13/01/2019 16:00	2607	2686
10	12/01/2019 7:00	2653	2681	22	13/01/2019 19:00	2600	2689
11	12/01/2019 10:00	2652	2683	23	13/01/2019 22:00	2598	2693

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.7966735308436067
2) MAE: 34.125
3) MSE: 1960.125
4) RMSE: 44.27329895094785
5) FB: 0.012717186927043975
6) FSD: 1.145568082892166
7) FA2: 1.0
8) Cor: 0.5505136044583602



Hình 5.99: Kết quả dự báo đợt lũ 1 - trạm Yên Bái

5.4.2.2. Đợt lũ 2: từ 03/8/2019 đến 06/8/2019 trạm Yên Bái 74129

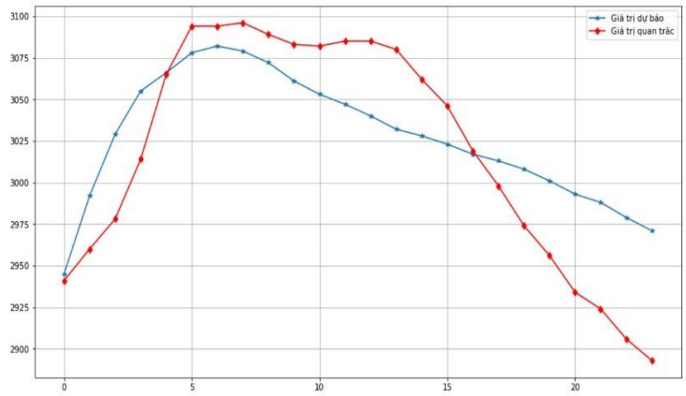
Bảng 5.9: So sánh kết quả dự báo mực nước và thực đo đợt lũ 2 tại trạm Yên Bái

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2019-08-03 10:00:00	2941	2945	12	2019-08-04 22:00:00	3085	3040

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
1	2019-08-03 13:00:00	2960	2992	13	2019-08-05 1:00:00	3080	3032
2	2019-08-03 16:00:00	2978	3029	14	2019-08-05 4:00:00	3062	3028
3	2019-08-03 19:00:00	3014	3055	15	2019-08-05 7:00:00	3046	3023
4	2019-08-03 22:00:00	3065	3066	16	2019-08-05 10:00:00	3019	3017
5	2019-08-04 1:00:00	3094	3078	17	2019-08-05 13:00:00	2998	3013
6	2019-08-04 4:00:00	3094	3082	18	2019-08-05 16:00:00	2974	3008
7	2019-08-04 7:00:00	3096	3079	19	2019-08-05 19:00:00	2956	3001
8	2019-08-04 10:00:00	3089	3072	20	2019-08-05 22:00:00	2934	2993
9	2019-08-04 13:00:00	3083	3061	21	2019-08-06 1:00:00	2924	2988
10	2019-08-04 16:00:00	3082	3053	22	2019-08-06 4:00:00	2906	2979
11	2019-08-04 19:00:00	3085	3047	23	2019-08-06 7:00:00	2893	2971

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.8761896339312804
2) MAE: 53.92857142857143
3) MSE: 3728.107142857142
4) RMSE: 61.0582274788348
5) FB: 0.01478632112165341
6) FSD: 0.269820405685545
7) FA2: 1.0
8) Cor: 0.9529845174625102



Hình 5.100: Kết quả dự báo đợt lũ 2 - trạm Yên Bái

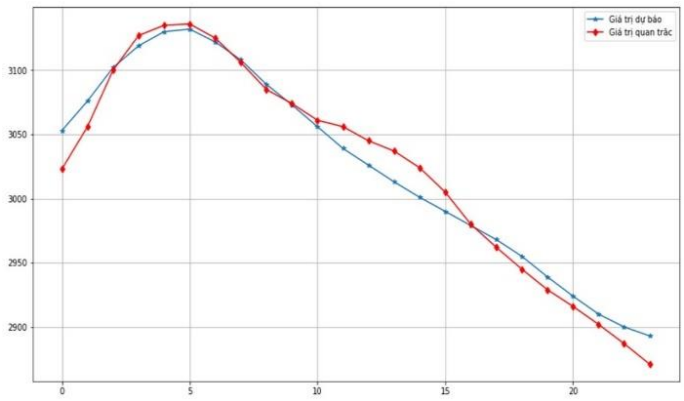
5.4.2.1. Đợt lũ 3: từ 30/8/2019 đến 02/9/2019 trạm Yên Bái 74129

Bảng 5.10: So sánh kết quả dự báo mực nước và thực đo đợt lũ 3 tại trạm Yên Bái

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2019-08-30 10:00:00	3023	3053	12	2019-08-31 22:00:00	3045	3026
1	2019-08-30 13:00:00	3056	3076	13	2019-09-01 1:00:00	3037	3013
2	2019-08-30 16:00:00	3100	3102	14	2019-09-01 4:00:00	3024	3001
3	2019-08-30 19:00:00	3127	3119	15	2019-09-01 7:00:00	3005	2990
4	2019-08-30 22:00:00	3135	3130	16	2019-09-01 10:00:00	2980	2979
5	2019-08-31 1:00:00	3136	3132	17	2019-09-01 13:00:00	2962	2968
6	2019-08-31 4:00:00	3125	3122	18	2019-09-01 16:00:00	2945	2955
7	2019-08-31 7:00:00	3106	3108	19	2019-09-01 19:00:00	2929	2939
8	2019-08-31 10:00:00	3085	3089	20	2019-09-01 22:00:00	2916	2924
9	2019-08-31 13:00:00	3074	3073	21	2019-09-02 1:00:00	2902	2910
10	2019-08-31 16:00:00	3061	3056	22	2019-09-02 4:00:00	2887	2900
11	2019-08-31 19:00:00	3056	3039	23	2019-09-02 7:00:00	2871	2893

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.9615749537041328
2) MAE: 10.833333333333334
3) MSE: 185.91666666666666
4) RMSE: 13.63512620648106
5) FB: 0.000137756226581391
6) FSD: 0.06020353733108702
7) FA2: 1.0
8) Cor: 0.9868833134598719



Hình 5.101: Kết quả dự báo đợt lũ 3 trạm Yên Bái

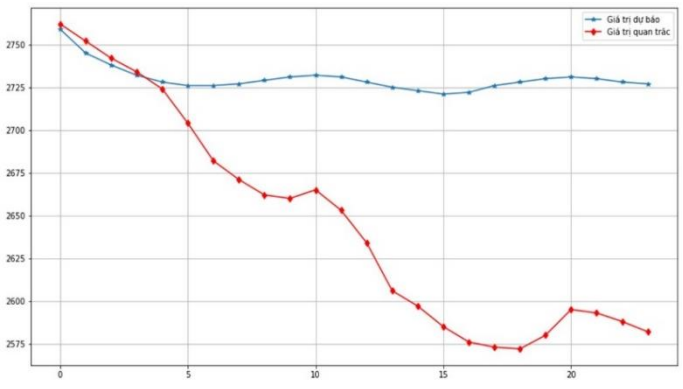
5.4.2.2. Đợt lũ 4: từ 14/7/2020 đến 16/7/2020 tại trạm Yên Bái 74129

Bảng 5.11: So sánh kết quả dự báo mực nước và thực đo đợt lũ 4 tại trạm Yên Bái

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2020-07-14 1:00:00	2762	2759	12	2020-07-15 13:00:00	2634	2728
1	2020-07-14 4:00:00	2752	2745	13	2020-07-15 16:00:00	2606	2725
2	2020-07-14 7:00:00	2742	2738	14	2020-07-15 19:00:00	2597	2723
3	2020-07-14 10:00:00	2734	2732	15	2020-07-15 22:00:00	2585	2721
4	2020-07-14 13:00:00	2724	2728	16	2020-07-16 1:00:00	2576	2722
5	2020-07-14 16:00:00	2704	2726	17	2020-07-16 4:00:00	2573	2726
6	2020-07-14 19:00:00	2682	2726	18	2020-07-16 7:00:00	2572	2728
7	2020-07-14 22:00:00	2671	2727	19	2020-07-16 10:00:00	2580	2730
8	2020-07-15 1:00:00	2662	2729	20	2020-07-16 13:00:00	2595	2731
9	2020-07-15 4:00:00	2660	2731	21	2020-07-16 16:00:00	2593	2730
10	2020-07-15 7:00:00	2665	2732	22	2020-07-16 19:00:00	2588	2728
11	2020-07-15 10:00:00	2653	2731	23	2020-07-16 22:00:00	2582	2727

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.7206451573801235
2) MAE: 85.95833333333333
3) MSE: 10523.875
4) RMSE: 102.58593958238137
5) FB: 0.03148471108010696
6) FSD: 1.5604878621597262
7) FA2: 1.0
8) Cor: 0.6595956554896106



Hình 5.102: Kết quả dự báo đợt lũ 4 trạm Yên Bái

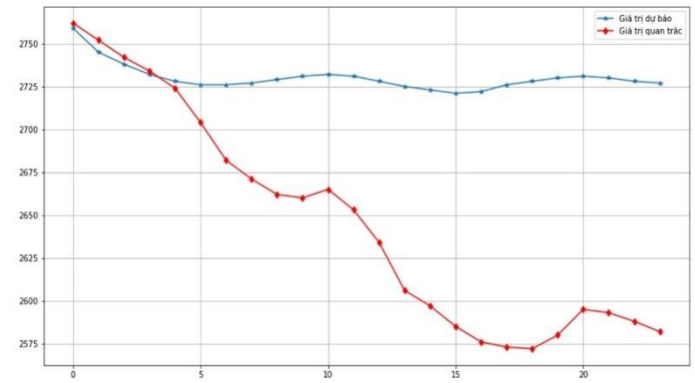
5.4.2.3. Đợt lũ 5: từ 06/8/2020 đến 06/8/2020 tại trạm Yên Bái 74129

Bảng 5.12: So sánh kết quả dự báo mực nước và thực đo đợt lũ 5 tại trạm Yên Bái

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2020-08-06 22:00:00	2972	2968	12	2020-08-08 10:00:00	2897	2939
1	2020-08-07 1:00:00	3010	2988	13	2020-08-08 13:00:00	2878	2933
2	2020-08-07 4:00:00	3010	2992	14	2020-08-08 16:00:00	2864	2928
3	2020-08-07 7:00:00	2998	2995	15	2020-08-08 19:00:00	2856	2921
4	2020-08-07 10:00:00	2986	2993	16	2020-08-08 22:00:00	2847	2917
5	2020-08-07 13:00:00	2972	2990	17	2020-08-09 1:00:00	2839	2918
6	2020-08-07 16:00:00	2954	2980	18	2020-08-09 4:00:00	2826	2916
7	2020-08-07 19:00:00	2935	2971	19	2020-08-09 7:00:00	2810	2916
8	2020-08-07 22:00:00	2921	2964	20	2020-08-09 10:00:00	2795	2920
9	2020-08-08 1:00:00	2919	2956	21	2020-08-09 13:00:00	2773	2920
10	2020-08-08 4:00:00	2922	2951	22	2020-08-09 16:00:00	2755	2918
11	2020-08-08 7:00:00	2915	2944	23	2020-08-09 19:00:00	2737	2916

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.7206451573801235
2) MAE: 85.95833333333333
3) MSE: 10523.875
4) RMSE: 102.58593958238137
5) FB: 0.03148471108010696
6) FSD: 1.5604878621597262
7) FA2: 1.0
8) Cor: 0.6595956554896106



Hình 5.103: Kết quả dự báo đợt lũ 5 trạm Yên Bái

5.4.3. Kết quả thử nghiệm Đài KTTV khu vực Đồng bằng Bắc Bộ

5.4.3.1. Đợt lũ 1: từ 31/5/2019 đến 03/6/2019 tại trạm Vụ Quang 74155

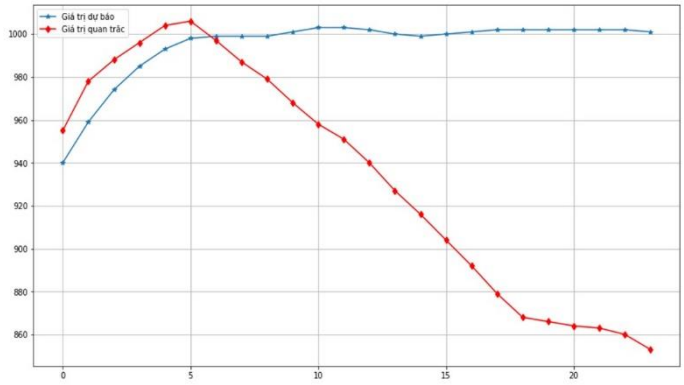
Bảng 5.13: So sánh kết quả dự báo mực nước và thực đo đợt lũ 1 tại trạm Vụ Quang

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2019-05-31 7:00:00	955	940	12	2019-06-01 19:00:00	940	1002
1	2019-05-31 10:00:00	978	959	13	2019-06-01 22:00:00	927	1000
2	2019-05-31 13:00:00	988	974	14	2019-06-02 1:00:00	916	999
3	2019-05-31 16:00:00	996	985	15	2019-06-02 4:00:00	904	1000
4	2019-05-31 19:00:00	1004	993	16	2019-06-02 7:00:00	892	1001
5	2019-05-31 22:00:00	1006	998	17	2019-06-02 10:00:00	879	1002
6	2019-06-01 1:00:00	997	999	18	2019-06-02 13:00:00	868	1002

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
7	2019-06-01 4:00:00	987	999	19	2019-06-02 16:00:00	866	1002
8	2019-06-01 7:00:00	979	999	20	2019-06-02 19:00:00	864	1002
9	2019-06-01 10:00:00	968	1001	21	2019-06-02 22:00:00	863	1002
10	2019-06-01 13:00:00	958	1003	22	2019-06-03 1:00:00	860	1002
11	2019-06-01 16:00:00	951	1003	23	2019-06-03 4:00:00	853	1001

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.7334125244784695
2) MAE: 67.70833333333333
3) MSE: 7359.458333333333
4) RMSE: 85.78728538270303
5) FB: 0.06350098342230968
6) FSD: 1.1057539761831916
7) FA2: 1.0
8) Cor: -0.3639529098054367



Hình 5.104: Kết quả dự báo đợt lũ 1 trạm Vụ Quang

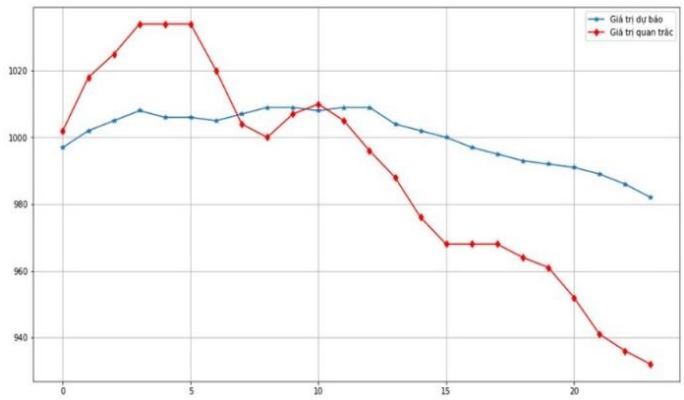
5.4.3.2. Đợt lũ 2: từ 04/8/2019 đến 07/8/2019 tại trạm Vụ Quang 74155

Bảng 5.14: So sánh kết quả dự báo mực nước và thực đo đợt lũ 2 tại trạm Vụ Quang

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2019-08-04 16:00:00	1002	997	12	2019-08-06 4:00:00	996	1009
1	2019-08-04 19:00:00	1018	1002	13	2019-08-06 7:00:00	988	1004
2	2019-08-04 22:00:00	1025	1005	14	2019-08-06 10:00:00	976	1002
3	2019-08-05 1:00:00	1034	1008	15	2019-08-06 13:00:00	968	1000
4	2019-08-05 4:00:00	1034	1006	16	2019-08-06 16:00:00	968	997
5	2019-08-05 7:00:00	1034	1006	17	2019-08-06 19:00:00	968	995
6	2019-08-05 10:00:00	1020	1005	18	2019-08-06 22:00:00	964	993
7	2019-08-05 13:00:00	1004	1007	19	2019-08-07 1:00:00	961	992
8	2019-08-05 16:00:00	1000	1009	20	2019-08-07 4:00:00	952	991
9	2019-08-05 19:00:00	1007	1009	21	2019-08-07 7:00:00	941	989
10	2019-08-05 22:00:00	1010	1008	22	2019-08-07 10:00:00	936	986
11	2019-08-06 1:00:00	1005	1009	23	2019-08-07 13:00:00	932	982

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.8281501187958522
2) MAE: 22.833333333333332
3) MSE: 732.75
4) RMSE: 27.06935536727832
5) FB: 0.011224190643715786
6) FSD: 1.1874445021778723
7) FA2: 1.0
8) Cor: 0.8650525977590188



Hình 5.105: Kết quả dự báo đợt lũ 2 trạm Vụ Quang

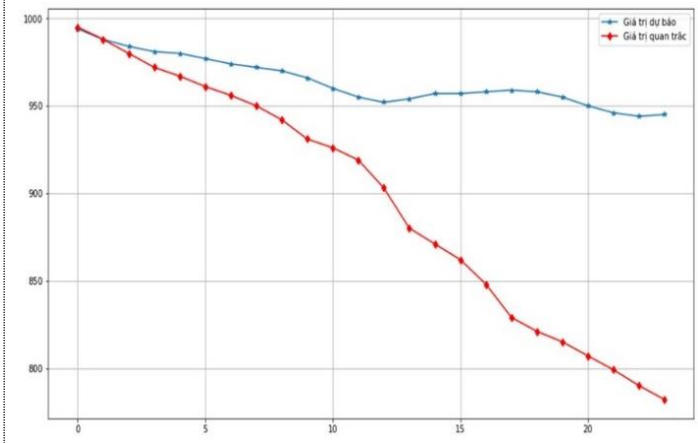
5.4.3.3. Đợt lũ 3: từ 11/9/2019 đến 14/9/2019 tại trạm Vụ Quang 74155

Bảng 5.15: So sánh kết quả dự báo mực nước và thực đo đợt lũ 3 tại trạm Vụ Quang

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2019-09-11 4:00:00	995	994	12	2019-09-12 16:00:00	903	952
1	2019-09-11 7:00:00	988	988	13	2019-09-12 19:00:00	880	954
2	2019-09-11 10:00:00	980	984	14	2019-09-12 22:00:00	871	957
3	2019-09-11 13:00:00	972	981	15	2019-09-13 1:00:00	862	957
4	2019-09-11 16:00:00	967	980	16	2019-09-13 4:00:00	848	958
5	2019-09-11 19:00:00	961	977	17	2019-09-13 7:00:00	829	959
6	2019-09-11 22:00:00	956	974	18	2019-09-13 10:00:00	821	958
7	2019-09-12 1:00:00	950	972	19	2019-09-13 13:00:00	815	955
8	2019-09-12 4:00:00	942	970	20	2019-09-13 16:00:00	807	950
9	2019-09-12 7:00:00	931	966	21	2019-09-13 19:00:00	799	946
10	2019-09-12 10:00:00	926	960	22	2019-09-13 22:00:00	790	944
11	2019-09-12 13:00:00	919	955	23	2019-09-14 1:00:00	782	945

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.7869253654961977
2) MAE: 68.5
3) MSE: 7908.416666666667
4) RMSE: 88.9292790180302
5) FB: 0.07358279184405105
6) FSD: 1.326036450887718
7) FA2: 1.0
8) Cor: 0.8961700612091738



Hình 5.106: Kết quả dự báo đợt lũ 3 trạm Vụ Quang

5.4.3.4. Đợt lũ 4: từ 22/7/2020 đến 25/7/2020 tại trạm Vụ Quang 74155

Bảng 5.16: So sánh kết quả dự báo mực nước và thực đo đợt lũ 4 tại trạm Vụ Quang

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2020-07-22 10:00:00	1013	995	12	2020-07-23 22:00:00	936	1110
1	2020-07-22 13:00:00	1045	1018	13	2020-07-24 1:00:00	932	1111
2	2020-07-22 16:00:00	1072	1038	14	2020-07-24 4:00:00	932	1112
3	2020-07-22 19:00:00	1088	1056	15	2020-07-24 7:00:00	932	1112
4	2020-07-22 22:00:00	1087	1068	16	2020-07-24 10:00:00	925	1111
5	2020-07-23 1:00:00	1076	1078	17	2020-07-24 13:00:00	916	1112
6	2020-07-23 4:00:00	1053	1086	18	2020-07-24 16:00:00	903	1111
7	2020-07-23 7:00:00	1029	1088	19	2020-07-24 19:00:00	898	1110
8	2020-07-23 10:00:00	1006	1091	20	2020-07-24 22:00:00	890	1109
9	2020-07-23 13:00:00	980	1098	21	2020-07-25 1:00:00	881	1107
10	2020-07-23 16:00:00	957	1103	22	2020-07-25 4:00:00	871	1106
11	2020-07-23 19:00:00	940	1108	23	2020-07-25 7:00:00	861	1107

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.6693013679514795

2) MAE: 132.58333333333334

3) MSE: 24248.833333333332

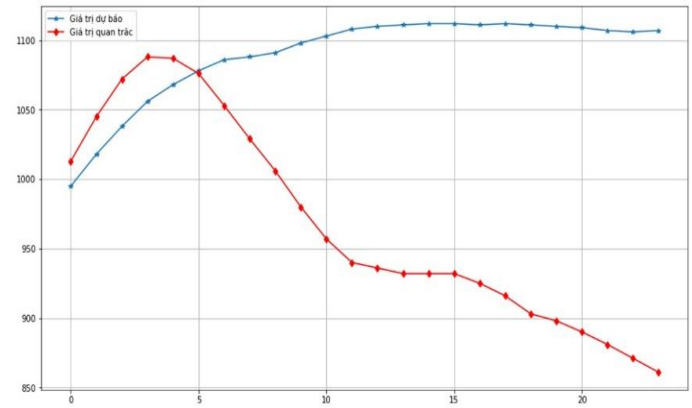
4) RMSE: 155.7203690380078

5) FB: 0.11837627613028683

6) FSD: 0.7824661895871681

7) FA2: 1.0

8) Cor: -0.6977260866179591



Hình 5.107: Kết quả dự báo đợt lũ 4 trạm Vụ Quang

5.4.3.5. Đợt lũ 5: từ 07/8/2020 đến 10/8/2020 tại trạm Vụ Quang 74155

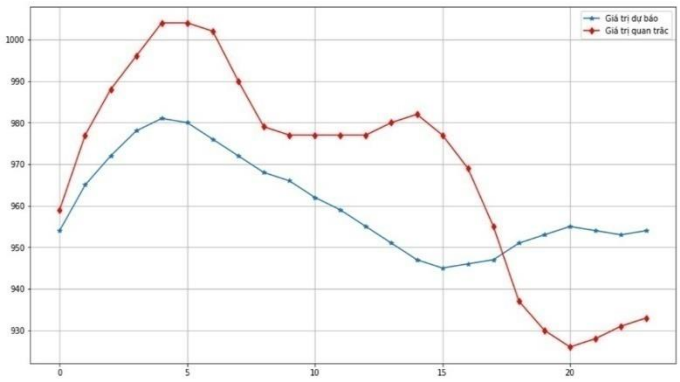
Bảng 5.17: So sánh kết quả dự báo mực nước và thực đo đợt lũ 5 tại trạm Vụ Quang

STT	Thời điểm dự báo	Giá trị (cm)		STT	Thời điểm dự báo	Giá trị (cm)	
		Quan trắc	Dự báo			Quan trắc	Dự báo
0	2020-08-07 13:00:00	959	954	12	2020-08-09 1:00:00	977	955
1	2020-08-07 16:00:00	977	965	13	2020-08-09 4:00:00	980	951
2	2020-08-07 19:00:00	988	972	14	2020-08-09 7:00:00	982	947
3	2020-08-07 22:00:00	996	978	15	2020-08-09 10:00:00	977	945
4	2020-08-08 1:00:00	1004	981	16	2020-08-09 13:00:00	969	946
5	2020-08-08 4:00:00	1004	980	17	2020-08-09 16:00:00	955	947
6	2020-08-08 7:00:00	1002	976	18	2020-08-09 19:00:00	937	951
7	2020-08-08 10:00:00	990	972	19	2020-08-09 22:00:00	930	953
8	2020-08-08 13:00:00	979	968	20	2020-08-10 1:00:00	926	955
9	2020-08-08 16:00:00	977	966	21	2020-08-10 4:00:00	928	954

10	2020-08-08 19:00:00	977	962	22	2020-08-10 7:00:00	931	953
11	2020-08-08 22:00:00	977	959	23	2020-08-10 10:00:00	933	954

Tổng hợp sai số của giá trị dự báo và giá trị quan trắc:

1) Simi: 0.800281296192917
2) MAE: 20.041666666666668
3) MSE: 457.4583333333333
4) RMSE: 21.3882756044832
5) FB: 0.0091146677034062
6) FSD: 0.760467931760200
7) FA2: 1.0
8) Cor: 0.6586778148054887

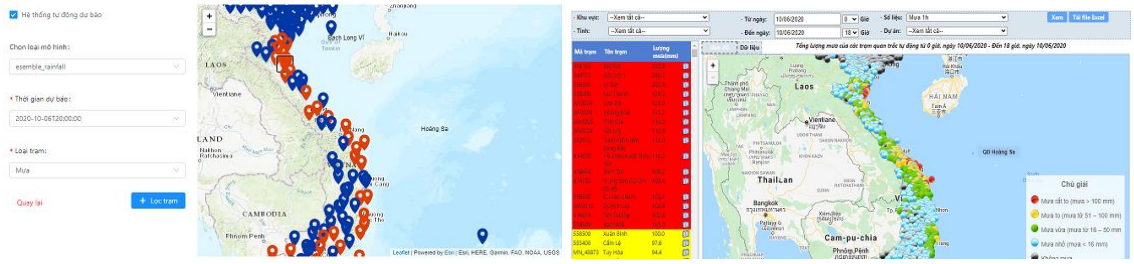


Hình 5.108: Kết quả dự báo đợt lũ 5 trạm Vụ Quang

5.5. Kết quả tự thử nghiệm mô hình AI hỗ trợ dự báo của đơn vị chủ trì

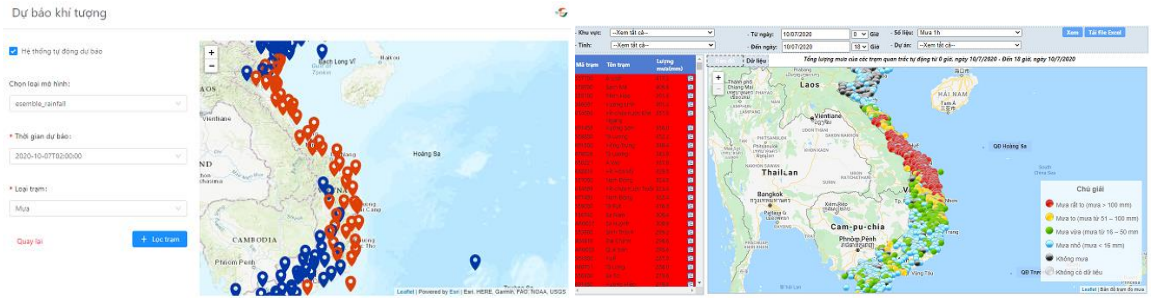
5.5.1. Kết quả thử nghiệm dự báo mưa lớn diện rộng

Nội dung nghiên cứu của đề tài không yêu cầu kết quả thử nghiệm mô hình AI hỗ trợ dự báo mưa lớn diện rộng. Tuy nhiên, đơn vị chủ trì thực hiện đề tài vẫn tiến hành chạy mô hình AI dự báo mưa lớn diện rộng hỗ trợ cho hoạt động nghiệp vụ thực tế. Kết quả dự báo mưa lớn diện rộng của mô hình AI là khả quan.



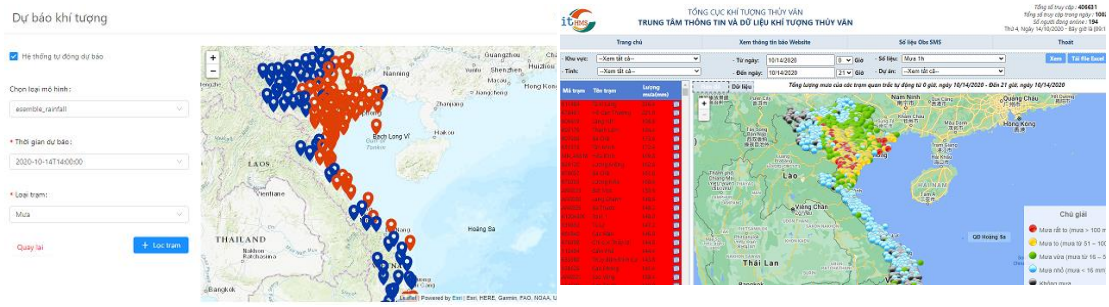
Hình 5.109: Kết quả dự báo mưa lớn diện rộng ngày 06/10/2020

Trong đợt mưa lớn diện rộng tháng 10/2020, mô hình AI cho kết quả dự báo bám sát với dữ liệu thực tế. Các điểm/ khu vực được mô hình AI dự báo có mưa lớn (thời đoạn dự báo 24h, điểm màu đỏ ở hình 5.109a) đã quan trắc được lượng mưa trên 50mm và hiển thị trên website của trung tâm Thông tin và dữ liêu Khí tượng thủy văn (các điểm đỏ và vàng ở hình 5.109b). Kết quả dự báo ngày 06/10/2020 cho thấy ba vùng mưa lớn diện rộng ở Thanh Hóa-Nghệ An, Huế-Đà Nẵng, Nha Trang.



Hình 5.110: Kết quả dự báo mưa lớn diện rộng ngày 07/10/2020

Vào ngày 07/10/2020 mưa lớn mở rộng trên toàn khu vực Trung Trung Bộ. Mô hình AI cho kết quả dự báo bám sát với dữ liệu quan trắc thực tế. Đặc biệt, mô hình Ai dự báo được điểm mưa lớn tương đối riêng lẻ tại khu vực thành phố Hồ Chí Minh phù hợp với dữ liệu quan trắc.



Hình 5.111: Kết quả dự báo mưa lớn diện rộng ngày 14/10/2020

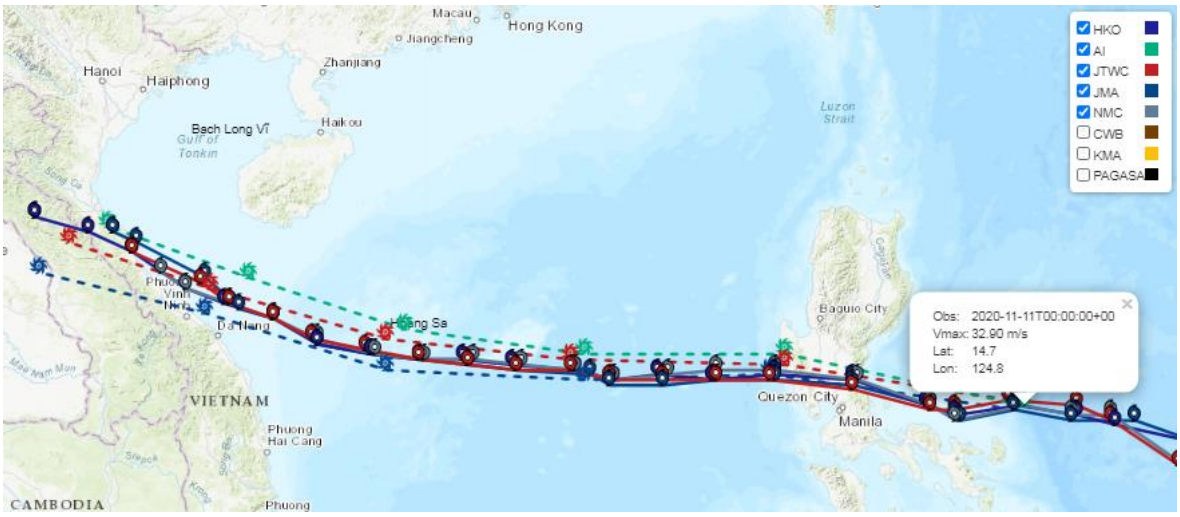
Ngày 14/10/2020, mưa lớn dịch chuyển mạnh lên phía bắc. Các điểm được mô hình AI dự báo có mưa lớn tập trung dày tại vùng đồng bằng Bắc Bộ, phù hợp với dữ liệu quan trắc hiển thị. Việc mô hình AI dự báo được vùng mưa lớn dịch chuyển này là kết quả rất khả quan trong nghiên cứu và hỗ trợ nghiệp vụ dự báo.

5.5.2. Kết quả thử nghiệm dự báo bão

Đơn vị chủ trì thực hiện đề tài cũng tiến hành thử nghiệm dự báo bão sử dụng mô hình AI với dữ liệu quan trắc và dự báo bão thời gian thực, tập trung vào một số cơn bão cuối năm 2020 đổ bộ vào miền Trung Việt Nam.

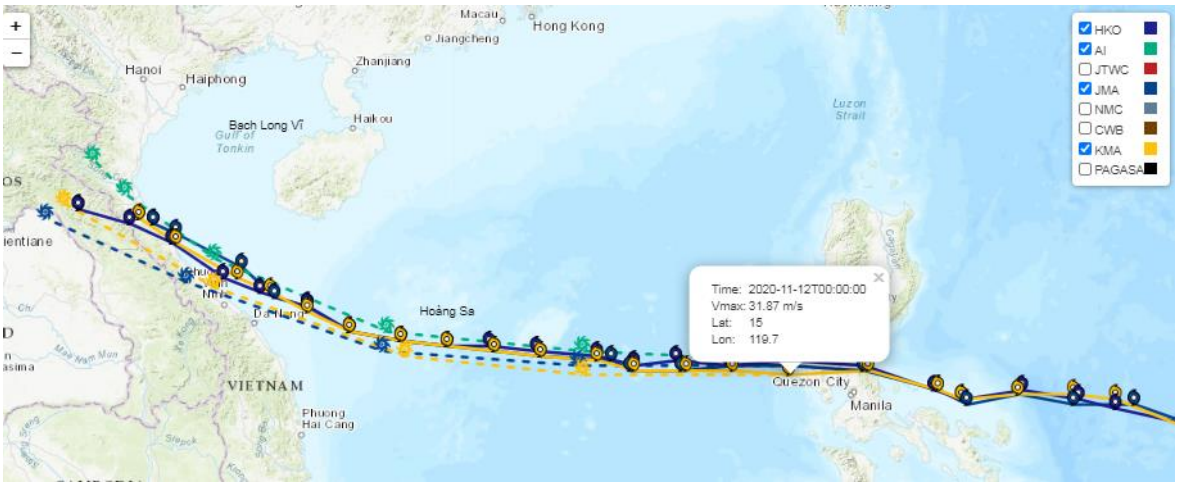
5.5.2.1. Bão VAMCO 2020

Bão VAMCO hoạt động trong thời gian từ 05/11/2020 đến 15/11/2020 với tốc độ gió lớn nhất duy trì trong 10 phút là 155km/h (tốc độ gió lớn nhất duy trì trong 1 phút lên tới 215 km/h), áp suất thấp nhất 950 hPa.



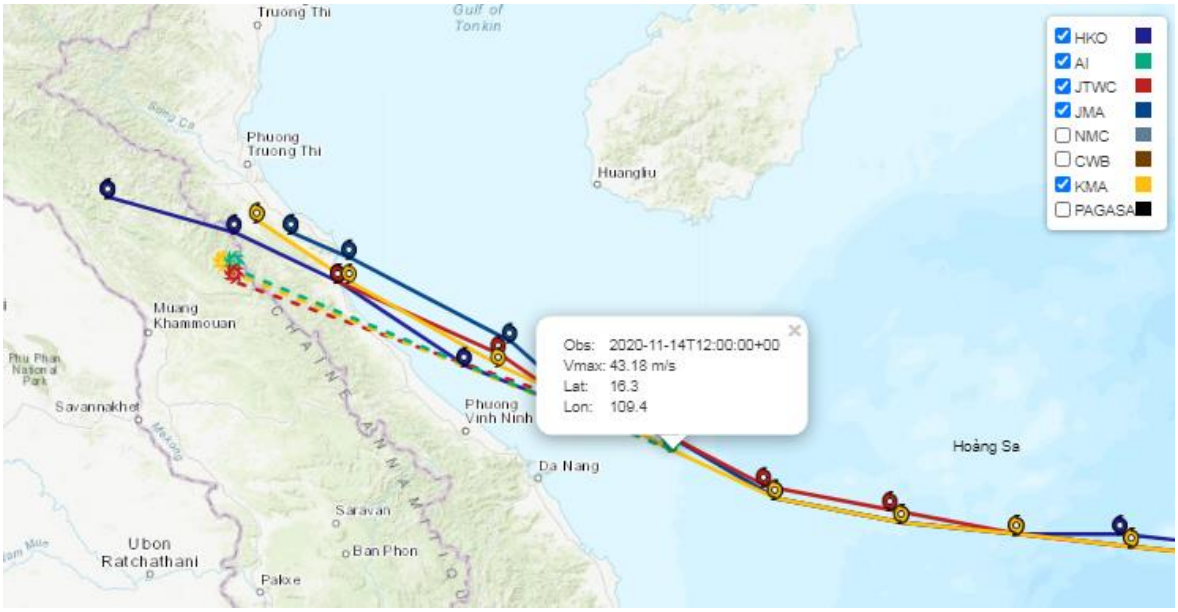
Hình 5.112: Kết quả dự báo bão VAMCO lúc 0h ngày 11/11/20205.113

Bão VAMCO lúc 0h ngày 11/11/2020 ở vị trí 14.7 độ vĩ Bắc 124.8 độ kinh Đông, tốc độ gió lớn nhất 32.9 m/s. Lúc này bão chuẩn bị đổ bộ vào Philippines. Kết quả dự báo của mô hình AI (đường đứt đoạn màu xanh lục) bám sát đường đi thực tế của cơn bão. Điểm lưu ý là trong khi tại thời điểm này, đa số các trung tâm dự báo quốc tế nhận định vị trí đổ bộ của cơn bão (các đường đứt đoạn) hơi thấp hơn so với vị trí đổ bộ thực tế (các đường liền) thì mô hình AI đã chọn được vị trí độ bộ gần sát nhất.



Hình 5.114: Kết quả dự báo bão VAMCO lúc 0h ngày 12/11/20205.115

Bão VAMCO lúc 0h ngày 12/11/2020 ở vị trí 15 độ vĩ Bắc 119.7 độ kinh Đông, tốc độ gió lớn nhất 31.87 m/s. Lúc này bão đã tiến vào biển Đông. Mô hình AI tiếp tục nhận định vị trí đổ bộ của cơn bão hơi cao hơn so với hầu hết các dự báo của các trung tâm quốc tế. Đường đi thực tế của cơn bão (các đường liền) đã cho thấy kết quả dự báo của mô hình AI bám sát với dữ liệu quan trắc và chính xác trong vị trí đổ bộ.

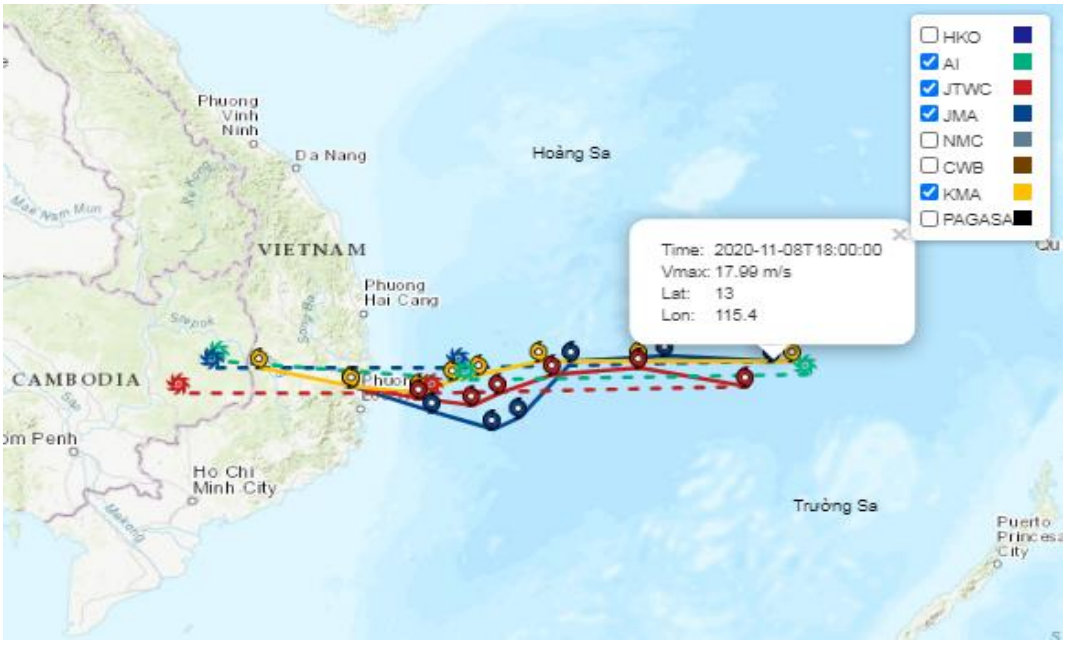


Hình 5.116: Kết quả dự báo bão VAMCO lúc 12h ngày 14/11/20205.117

Bão VAMCO lúc 12h ngày 14/11/2020 ở vị trí 16.3 độ vĩ Bắc 109.4 độ kinh Đông, tốc độ gió lớn nhất 43.18 m/s. Đây là thời điểm trước khi bão đổ bộ khoảng 1 ngày. Kết quả dự báo của mô hình AI (đường đứt đoạn màu xanh) bám sát đường đi thực tế của cơn bão.

5.5.2.2. Bão ETAU 2020

Bão ETAU hoạt động trong thời gian từ 07/11/2020 đến 11/11/2020 với tốc độ gió lớn nhất duy trì trong 10 phút là 85km/h, áp suất thấp nhất 992 hPa. Bão đã gây ra mưa lớn và sạt lở đất tại vùng Quảng Nam đến Phú Yên.

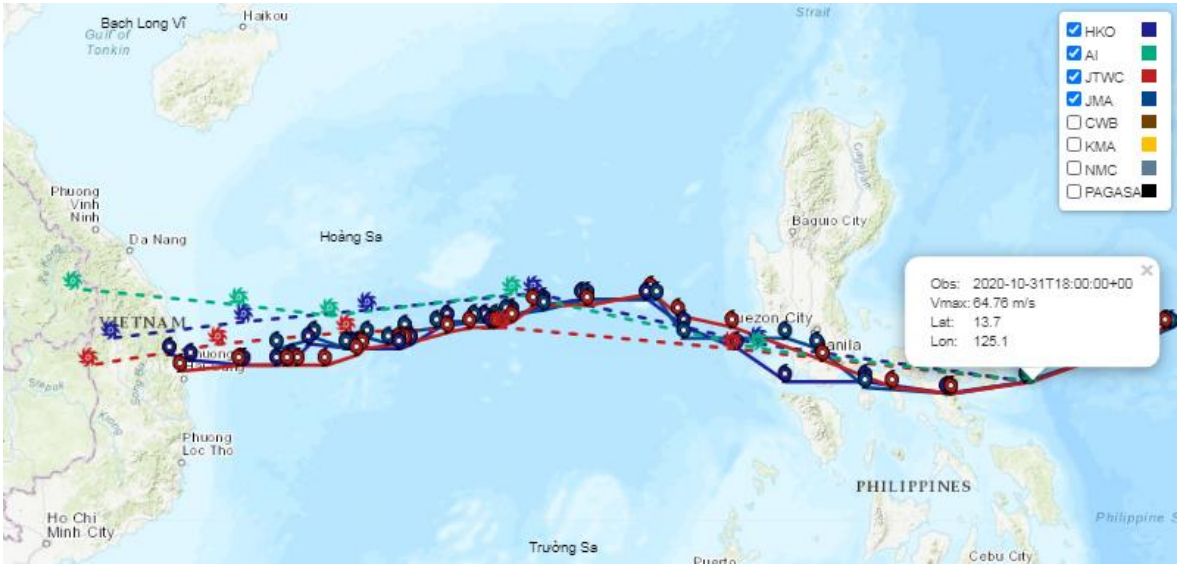


Hình 5.118: Kết quả dự báo bão ETAU lúc 12h ngày 14/11/20205.119

Bão ETAU lúc 18h ngày 08/11/2020 ở vị trí 13 độ vĩ Bắc 115.4 độ kinh Đông, tốc độ gió lớn nhất 17.99 m/s. Có thể thấy cả dữ liệu quan trắc thực tế (các đường liền) và dữ liệu dự báo (các đường đứt đoạn) của các trung tâm quốc tế khá phân tán. Tuy nhiên, kết quả dự báo của mô hình AI (đường màu xanh đứt đoạn) đã bám sát đường đi thực tế của cơn bão và dự báo vị trí đổ bộ phù hợp.

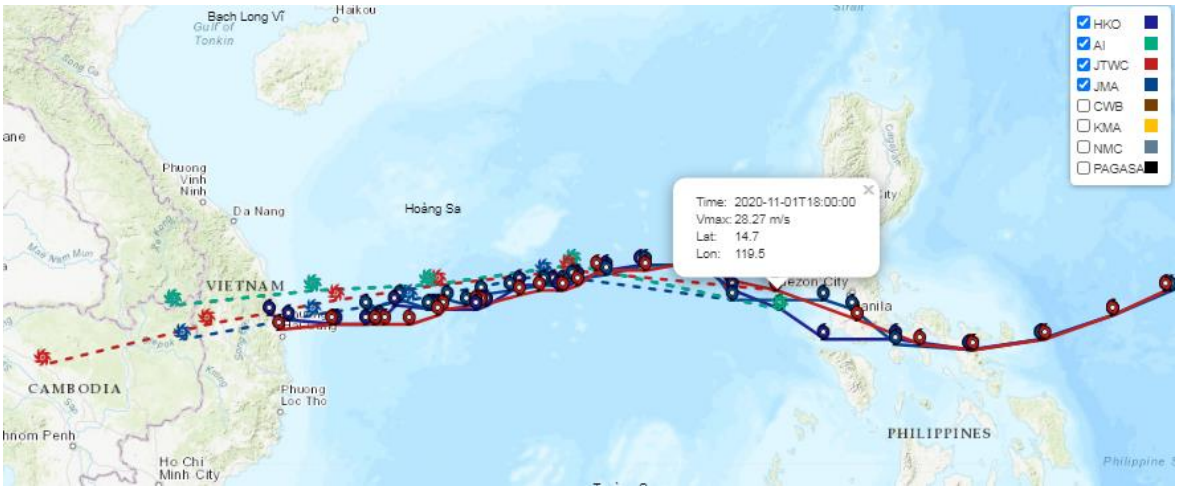
5.5.2.3. Bão GONI 2020

Bão GONI hoạt động trong thời gian từ 26/10/2020 đến 06/11/2020 với tốc độ gió lớn nhất duy trì trong 10 phút là 220 km/h, áp suất thấp nhất 905 hPa.



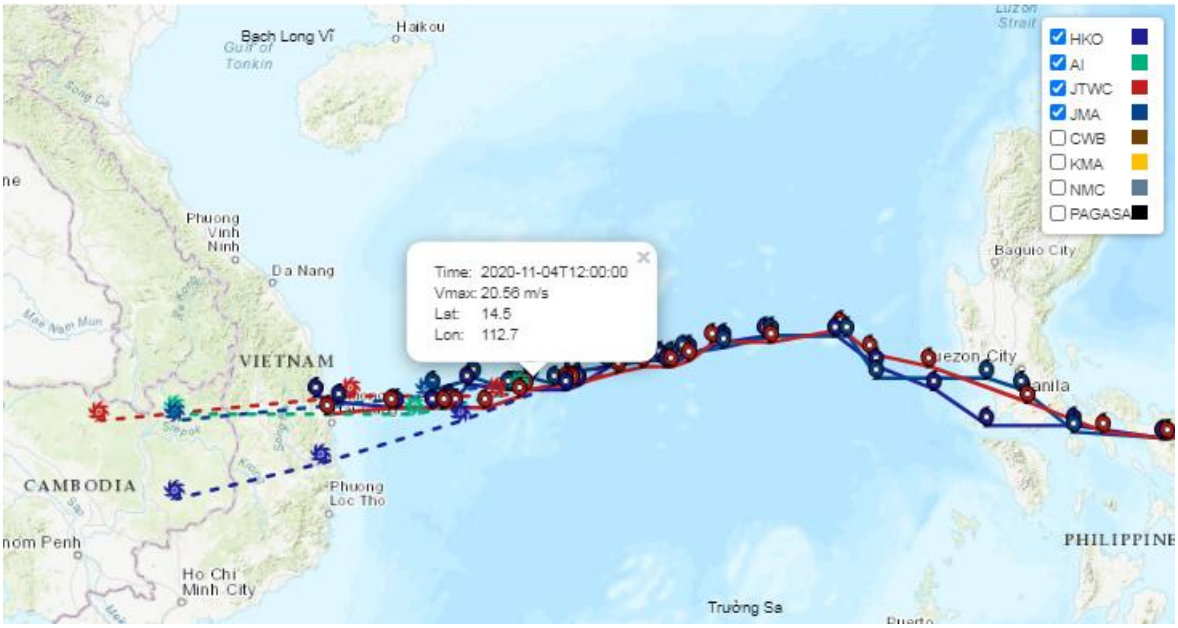
Hình 5.120: Kết quả dự báo bão GONI lúc 18h ngày 31/10/20205.121

Bão GONI lúc 18h ngày 31/10/2020 ở vị trí 13.7 độ vĩ Bắc 125.1 độ kinh Đông, tốc độ gió lớn nhất 64.76 m/s. Lúc này bão chuẩn bị đổ bộ vào Philippines. Kết quả dự báo của mô hình AI (đường màu xanh đứt đoạn) là phù hợp với đường đi thực tế của cơn bão.



Hình 5.122: Kết quả dự báo bão GONI lúc 18h ngày 01/11/20205.123

Bão GONI lúc 18h ngày 01/11/2020 ở vị trí 14.7 độ vĩ Bắc 119.5 độ kinh Đông, tốc độ gió lớn nhất 28.27 m/s. Lúc này bão đã vượt qua Philippines và tiến vào biển Đông, cường độ bão đã giảm. Kết quả dự báo của mô hình AI (đường màu xanh đứt đoạn) bám sát đường đi thực tế của cơn bão.



Hình 5.124: Kết quả dự báo bão GONI lúc 12h ngày 04/11/20205.125

Bão GONI lúc 12h ngày 04/11/2020 ở vị trí 14.5 độ vĩ Bắc 112.7 độ kinh Đông, tốc độ gió lớn nhất 20.56 m/s. Kết quả dự báo của mô hình AI (đường màu xanh đứt) bám sát đường đi thực tế của cơn bão. Mô hình AI đã dự báo vị trí đổ bộ phù hợp.

5.6. Nhận xét, đánh giá các mô hình AI hỗ trợ dự báo KTTV tại các đơn vị tham gia thử nghiệm

Sau thời gian thử nghiệm, ý kiến nhận xét đánh giá kết quả thử nghiệm mô hình AI hỗ trợ dự báo KTTV của các đơn vị tham gia cụ thể như sau:

5.6.1. Về công cụ, mô hình thử nghiệm

Hệ thống Framework tích hợp các module AI hỗ trợ dự báo KTTV nói chung, bão nói riêng được thiết kế khoa học, hợp lý, giao diện phù hợp, thuận tiện cho việc tham khảo dự báo tại các đơn vị dự báo. Cụ thể:

5.6.1.1. Về công cụ quản trị và giám sát hệ thống

Ưu điểm: Các module quản trị và giám sát đã được thiết lập và vận hành đảm bảo an toàn thông tin của hệ thống gồm: (i) Quản trị thông tin nhóm người dùng và người dùng hệ thống; quản trị thông tin quyền thực thi trong hệ thống; quản trị thông tin phân quyền người dùng trong hệ thống; (ii) Các công cụ đã giám sát đầy

đủ quá trình vận hành khai thác mô hình AI của hệ thống Framework của các thành phần (mô hình dự báo, hoạt động máy chủ, hệ thống đường truyền, ..), đối tượng tham gia quá trình vận hành các mô hình dự báo.

Tồn tại và kiến nghị: Đề nghị thiết lập bổ sung và chuẩn hóa mã nhóm người dùng (của các đơn vị thụ hưởng như nhóm Trung tâm Dự báo KTTV quốc gia, nhóm các Đài KTTV khu vực, nhóm dự báo thời tiết, nhóm dự báo thủy văn, hải văn) và mã loại mô hình (mã mô hình dự báo bão, lũ, mưa, …) theo dạng tích/ chọn để thuận tiện cho các đơn vị/ cá nhân tham gia trong quá trình khai thác hệ thống.

5.6.1.2. Về công cụ thiết lập dự báo

Ưu điểm: Đã xây dựng đầy đủ các module công cụ thiết lập dự báo gồm: (i) Module tạo lập, chỉnh sửa các trường dự báo: các trường dự báo bão, trường dự báo nhiệt độ, trường dự báo mưa, trường dự báo lũ, trường dự báo nước biển dâng, …; (ii) Module tạo lập, chỉnh sửa mã và loại mô hình dự báo KTTV (bão, mưa, KKL, lũ, nước biển dâng, …); (iii) Module tạo lập và chỉnh sửa thông tin tham số của mô hình dự báo (Mã tham số của mô hình, mã trạm quan trắc, thời hạn dự báo, …); (iv) Module tạo lập và chỉnh sửa thời gian dự báo (mã thời hạn dự báo, ngày/ giờ dữ liệu bắt đầu - kết thúc, ngày - giờ dự báo, …).

Tồn tại và kiến nghị: Đề nghị chỉnh sửa một số thuật ngữ cho đúng với thuật ngữ quy định về dự báo KTTV hiện hành như: "Khoảng dự báo" là "thời hạn dự báo", …. , và thiết lập bổ sung và chuẩn hóa mã loại và tên loại trường dự báo trong mô hình theo quy định tại Thông tư 06/2016/TT-BTNMT ngày 16/5/2016 quy định về loại bản tin và thời hạn dự báo, cảnh báo KTTV.

5.6.1.3. Về công cụ huấn luyện dự báo và triển khai dự báo

Ưu điểm: Đã xây dựng đầy đủ các module công cụ huấn luyện và các mô hình dự báo cho các lĩnh vực theo đúng thuyết minh đề tài đã được phê duyệt gồm: (i) Tạo lập, chỉnh sửa các module về huấn luyện mô hình (mã huấn luyện mô hình, loại mô hình huấn luyện, trường dự báo, thời gian huấn luyện, …); (ii) Các mô hình dự báo các loại hình thời tiết nguy hiểm như bão, mưa lớn diện rộng, lũ, nước biển dâng do bão, ... theo các lựa chọn (chọn loại mô hình dự báo, thời hạn dự báo, mã trạm quan trắc KTTV, chọn dữ liệu lũ, chọn dữ liệu bão, …).

Tồn tại và kiến nghị: Đề nghị chỉnh sửa một số thuật ngữ cho đúng với thuật ngữ quy định về dự báo KTTV hiện hành, ví dụ "Dự báo khí tượng" là "Dự báo thời tiết". Trên cơ sở kết quả dự báo với các số liệu quá thưa, tiếp tục hiệu chỉnh tối ưu các tham số cấu hình của các mô hình dự báo, đặc biệt là phần dự báo realtime của các mô hình để nâng cao khả năng áp dụng trong nghiệp vụ dự báo thực tế.

5.6.1.4. Nhận xét chung

Các công cụ mô hình AI hỗ trợ dự báo KTTV do đơn vị chủ trì thực hiện chuyển giao về cơ bản đã đáp ứng về yêu cầu dự báo với các yếu tố, thời hạn, nội dung dự báo cụ thể. Đơn vị chủ trì cần tiếp tục hoàn thiện công cụ theo các nội dụng kiến nghị nêu trên.

5.6.2. Về mô hình AI hỗ trợ dự báo bão khu vực Bắc Bộ

Sau thời gian thử nghiệm, ý kiến nhận xét đánh giá kết quả thử nghiệm mô hình AI hỗ trợ dự báo **bão** khu vực Bắc Bộ của các đơn vị cụ thể như sau:

5.6.2.1. Về nội dung, kết quả thử nghiệm

Về khối lượng thử nghiệm: đã hoàn thành thử nghiệm dự báo bão cho 10 cơn bão đổ bộ vào khu vực Bắc Bộ của Việt Nam trong giai đoạn từ 2008 - 2017 (Trung tâm Dự báo KTTV quốc gia: 5 cơn; Đài KTTV khu vực Đồng bằng Bắc Bộ: 5 cơn).

Về CSDL và xử lý dữ liệu bão: Đã tạo lập, lưu trữ đầy đủ các dữ liệu **bão** theo đúng khối lượng (102 cơn bão) và thời gian dữ liệu (10 năm dữ liệu từ 2008-2017) theo yêu cầu; đã có đầy đủ các các công cụ tích hợp truy xuất, phát hiện, xử lý, giảm chiều, chuẩn hóa, phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu bão theo yêu cầu của mô hình huấn luyện dự báo bão.

Về huấn luyện dự báo bão: Trên cơ sở dữ liệu bão được lựa chọn đầu vào, các công cụ của mô hình học máy DL/AI sẽ thực hiện được các bước tự động cấu hình, tham số và triển khai huấn luyện dự báo bão trên dữ liệu đầu vào và tiến hành đánh giá xác định độ tin cậy dự báo của các mô hình. Mô hình dự báo bão đã tham khảo kết quả dự báo bão của các Trung tâm Dự báo trong và ngoài nước (NCHMF, HKO, JTWC, JMA, NMC, CWB, KMA, PAGASA) để hiệu chỉnh nâng cao chất lượng của mô hình và bám sát số liệu quan trắc thực tế.

Về trình diễn kết quả dự báo: Hệ thống đã trình diễn, hiển thị để so sánh, đánh giá kết quả dự đoán bão (theo các dạng bản đồ, biểu đồ đường) của mô hình AI và của các Trung tâm dự báo trong và ngoài nước.

5.6.2.2. Nhận xét chung

Kết quả đạt được:

CSDL của hệ thống đã tích hợp dữ liệu của 103 cơn bão trong giai đoạn 2008-2019 (Lat, Lon, Vmax, Pressure, bán kính gió ảnh hưởng R30KT, R50KT, R64KT, …); dữ liệu dự báo của 8 trung tâm: JMA - Nhật Bản, NMC - Trung Quốc, JTWC - Hải quân Mỹ, KMA - Hàn Quốc, HKO - Hồng Kong, PAGASA-Philippine.

Hệ thống đã phân vùng dữ liệu thành các ô lưới hình vuông cạnh R (km); tự động rò tìm R để đảm bảo đủ dữ liệu của các cơn bão đưa vào tính toán; phân dữ liệu theo mùa (3 tháng một, tương ứng với 4 mùa trong năm).

Đã xác định Trung tâm dự báo trong và ngoài nước có kết quả dự báo tốt nhất cho từng yếu tố vị trí (Lat, Lon), Vmax và xác định Trung tâm dự báo có kết quả dự báo tốt nhất cho cả 2 yếu tố vị trí và Vmax.

Việc đánh giá/ xác định độ tin cậy dự báo của các Trung tâm dự báo có tính đến tham số của từng mô hình AI tự khởi tạo ban đầu cho từng Trung tâm theo ý kiến chuyên gia; đồng thời có module chức năng cho phép chuyên gia điều chỉnh cấu hình và tham số của mô hình.

Kết quả triển khai thử nghiệm mô hình AI hỗ trợ dự báo bão cho thấy kết quả dự báo đường đi của cơn bão và vận tốc gió trong tâm bão cho kết quả khá tốt, chất lượng dự báo bão của mô hình AI ở mức trung bình khá so với kết quả dự báo của các trung tâm dự báo lớn trên thế giới (như JMA - Nhật Bản, NMC - Trung Quốc, JTWC - Hải quân Mỹ, KMA - Hàn Quốc) và bám sát với dữ liệu quan trắc, diễn biến thực tế của các cơn bão.

Tồn tại hạn chế:

Kết quả dự báo bão sử dụng học máy với bài toán xử lý chuỗi thời gian trong một số trường hợp cho kết quả không tốt; nguyên nhân chính là do dữ liệu mỗi cơn bão cụ thể có ít bản ghi (từ 10 tới 15 bản ghi thông tin cơn bão). Kết quả dự báo tại các thời điểm lớn hơn 48 giờ, 72 giờ sai số dự báo lớn (về quỹ đạo bão, giá trị vận tốc gió).

Kiến nghị: Đơn vị chủ trì cần tiếp tục thử nghiệm huấn luyện dự báo với dữ liệu realtime và dữ liệu quá khứ dày hơn (30 đến 50 năm dữ liệu) của nhiều khu vực khác nhau trên phạm vi cả nước (khu vực Trung Trung Bộ, Nam Trung Bộ, Tây Nguyên, Nam Bộ) để hoàn thiện, nâng cao chất lượng của mô hình AI theo các nội dụng kiến nghị nêu trên.

5.6.3. Về mô hình AI hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ

Sau thời gian thử nghiệm, ý kiến nhận xét đánh giá kết quả thử nghiệm mô hình AI hỗ trợ dự báo **nước biển dâng do bão** khu vực ven biển Bắc Bộ và Bắc Trung Bộ của các đơn vị cụ thể như sau:

5.6.3.1. Về nội dung, kết quả thử nghiệm

Về khối lượng thử nghiệm: đã hoàn thành thử nghiệm dự báo cho các đợt nước biển dâng do bão khác nhau tại các trạm khu vực ven biển Bắc Bộ và Bắc Trung Bộ giai đoạn từ 2008 - 2017. Cụ thể, Trung tâm Dự báo KTTV quốc gia và Đài KTTV khu vực Bắc Trung Bộ thử nghiệm 5 đợt nước biển dâng do bão tại trạm Hòn Ngư; Đài KTTV khu vực Đông Bắc và Đồng bằng Bắc Bộ: thử nghiệm 5 đợt nước biển dâng tại trạm Hòn Dấu).

Về CSDL và xử lý dữ liệu nước biển dâng do bão: Hệ thống đã tạo lập, lưu trữ đầy đủ các dữ liệu của các trạm khí tượng hải văn theo yêu cầu; đã có đầy đủ các các công cụ tích hợp truy xuất, phát hiện, xử lý, giảm chiều, chuẩn hóa, phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu nước biển dâng theo yêu cầu của mô hình huấn luyện dự báo nước biển dâng do bão.

Về huấn luyện dự báo nước biển dâng do bão: Trên cơ sở dữ liệu độ cao sóng, chu kỳ sóng, mực nước biển và các dữ liệu khí tượng hải văn được lựa chọn đầu vào, các công cụ của mô hình học máy DL/AI sẽ thực hiện được các bước tự động cấu hình, tham số và triển khai huấn luyện dự báo nước biển dâng do bão trên dữ liệu đầu vào và tiến hành đánh giá xác định độ tin cậy dự báo của các mô hình. Mô hình dự báo dự báo nước biển dâng do bão đã được triển khai trên hệ thống máy chủ cấu hình cao, do đó đã cải thiện được thời gian và chất lượng dự báo. Hệ thống đã thiết lập được các mô hình dự báo lũ theo các phương pháp dự báo nước biển dâng do bão phổ biến là MLP, SVM, kNN, DCT, RF và phương pháp lập trình di truyền GP. Kết quả thử nghiệm đã khẳng định lập trình di truyền GP có kết quả dự báo vượt trội và đáng tin hơn các mô hình sử dụng các phương pháp truyền thống.

5.6.3.2. Nhận xét chung

Kết quả đạt được:

Hệ thống mô hình AI đã thiết lập được các tham số sử dụng trong mô hình dự báo nước biển dâng do bão gồm:

- Tham số khí tượng: tốc độ gió (m/s), hướng gió (độ), khí áp trên mặt biển (hPa) và độ giảm khí áp trên mặt biển (=1013 hPa).
- Tham số hải văn: mực nước bề mặt biển (SS), thủy triều (SSL).
- Tham số cơn bão: kinh độ (LG), vĩ độ (LT), áp suất tâm bão (hPa) và tốc độ gió Max gần tâm bão (m/s).

Hệ thống mô hình AI đã thiết lập được các chỉ số để đánh giá chất lượng mô hình dự báo nước biển dâng do bão gồm:

- NRMSE (normal root mean squared error) là RMSE chuẩn hóa tính theo %,
- CC (correlation coefficien) là hệ số tương quan.

Kết quả thử nghiệm dự báo nước biển dâng cho các cơn bão khẳng định mô hình AI bằng GP bám sát nhất với giá trị quan trắc thực tế. Điều đó cho thấy mô hình GP có khả năng đoán nhận gần đúng nhất dữ liệu quan trắc. NRMSE của 6 phương pháp dự báo trong khoảng từ 6% - 18%, giá trị CC nằm trong khoảng từ 0,75 - 0,97. Kết quả cho thấy phương pháp GP vừa cho kết quả giá trị NRMSE nhỏ (sai số ít nhất) và CC lớn nhất (gần gũi với giá trị thực nhất kể cả các điểm cao) trong số 6 phương pháp.

Tồn tại hạn chế: Hệ thống chưa nghiên cứu thử nghiệm dự báo realtime do thu thập dữ liệu quan trắc khí tượng hải văn theo thời gian thực quá phức tạp.

Kiến nghị: Cần tiếp tục cải tiến mô hình dự báo nước biển dâng do bão theo GP để thu được kết quả dự báo tốt hơn nữa. Ngoài ra, thử nghiệm mô hình dự báo theo GP để áp dụng cho dữ liệu tại các trạm khác, với các cơn bão khác và với thời gian dự báo trước ngắn hơn (12h, 5h) để có được kết quả dự báo phù hợp với yêu cầu thực tế.

5.6.4. Về mô hình AI hỗ trợ dự báo lũ trên hệ thống sông Hồng

Sau thời gian thử nghiệm, ý kiến nhận xét đánh giá kết quả thử nghiệm mô hình AI hỗ trợ dự báo **lũ** trên hệ thống sông Hồng của các đơn vị cụ thể như sau:

5.6.4.1. Về nội dung, kết quả thử nghiệm

Về khối lượng thử nghiệm: đã hoàn thành thử nghiệm dự báo lũ cho 10 đợt lũ khác nhau tại các trạm trên hệ thống sông Hồng (Yên Bái, Vụ Quang) giai đoạn từ 2008 - 2017. Cụ thể, Trung tâm Dự báo KTTV quốc gia: thử nghiệm 5 đợt lũ tại trạm Yên Bái; Đài KTTV khu vực Đồng bằng Bắc Bộ: thử nghiệm 5 đợt lũ tại trạm Vụ Quang).

Về CSDL và xử lý dữ liệu lũ: Hệ thống đã tạo lập, lưu trữ đầy đủ các dữ liệu **lũ** tại 260 trạm thủy văn và thời gian dữ liệu (10 năm dữ liệu từ 2008-2017) theo yêu cầu; đã có đầy đủ các các công cụ tích hợp truy xuất, phát hiện, xử lý, giảm chiều, chuẩn hóa, phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu **lũ** theo yêu cầu của mô hình huấn luyện dự báo **lũ**.

Về huấn luyện dự báo lũ: Trên cơ sở dữ liệu mực nước, lưu lượng được lựa chọn đầu vào, các công cụ của mô hình học máy DL/AI sẽ thực hiện được các bước tự động cấu hình, tham số và triển khai huấn luyện dự báo **lũ** trên dữ liệu đầu vào và tiến hành đánh giá xác định độ tin cậy dự báo của các mô hình. Mô hình dự báo **lũ**

đã được triển khai trên hệ thống máy chủ cấu hình cao, do đó đã cải thiện được thời gian và chất lượng dự báo. Hệ thống đã thiết lập được các mô hình dự báo lũ theo các phương pháp dự báo thủy văn phổ biến là ARIMA, KNN, RF, SVM, LSTM và phương pháp dự báo lai là ARIMA_KNN, ARIMA_RF, ARIMA_SVR và ARIMA_LSTM. Kết quả thử nghiệm đã khẳng định phương pháp kết hợp (lai) ARIMA_RF và ARIMA_KNN có kết quả dự báo vượt trội và đáng tin hơn các mô hình lai khác cũng như các phương pháp truyền thống.

5.6.4.2. Nhận xét chung

Kết quả đạt được:

Hệ thống đã thực hiện chuẩn hóa dữ liệu quan trắc 4 obs/ ngày (vào thời điểm 1h, 7h, 13h, và 19h với bước dữ liệu là 6 giờ/ lần dữ liệu) và 8 obs hoặc 12 obs/ ngày (với các bước dữ liệu là 2-3 giờ/ lần dữ liệu) thành bộ dữ liệu có tần suất là 8 obs/ ngày tại các mốc thời gian là 1h, 4h, 10h, 13h, 16h, 19h, 22h để đưa vào các mô hình huấn luyện dữ liệu dự báo lũ.

Các mô hình dự báo lũ kết hợp là ARIMA_RF và ARIMA_KNN cho các kết quả dự báo có độ tin cậy cao và bám sát dữ liệu thực đo nhất, cụ thể: (i) Thời hạn đến 24 giờ: Kết quả dự báo sát với thực đo; Thời hạn 48 - 72 giờ: kết quả dự báo tốt hơn hẳn các phương pháp dự báo khác; (iii) Thời hạn dự báo đến 5 ngày: kết quả dự báo tốt xu hướng biến đổi của mực nước (các phương pháp khác không dự báo được), nhưng sai số giá trị dự báo các yếu tố lớn. Kết quả thử nghiệm dự báo cụ thể tại trạm Yên Bái và trạm Vụ Quang cho thấy, với các thời hạn dự báo khác nhau (12h, 24h, 48h, 72h) đều có sai số dự báo trong phạm vi sai số dự báo cho phép theo yêu cầu nghiệp vụ hiện hành.

Tồn tại hạn chế: Thời hạn dự báo hạn vừa có sai số dự báo lớn; hệ thống chưa được nghiên cứu và thử nghiệm trên các hệ thống sông ảnh hưởng triều.

Kiến nghị: Cần tiếp tục triển khai thử nghiệm, hiệu chỉnh các mô hình dự báo lũ kết hợp ARIMA_RF và ARIMA_KNN cho các hệ thống sông khác trên phạm vi cả nước, đặc biệt là các trạm quan trắc trên các sông ảnh hưởng triều.

5.6.5. Về mô hình AI hỗ trợ dự báo mưa lớn diện rộng

5.6.5.1. Về nội dung, kết quả thử nghiệm

Về khối lượng thử nghiệm: Đơn vị chủ trì thực hiện đề tài đã thử nghiệm dự báo 10 ngày (từ 06/10 - 16/10/2020) xảy ra mưa lớn diện rộng trên phạm vi cả nước (Bắc Bộ, Trung Bộ và Nam Bộ).

Về CSDL và xử lý dữ liệu mưa: Hệ thống đã tạo lập, lưu trữ đầy đủ các dữ liệu mưa lớn tại 183 trạm KTTV trên phạm vi cả nước và thời gian dữ liệu (10 năm dữ liệu từ 2008-2017) theo yêu cầu; đã có đầy đủ các các công cụ tích hợp truy xuất, phát hiện, xử lý, giảm chiều, chuẩn hóa, phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu mưa theo yêu cầu của mô hình huấn luyện dự báo mưa.

Về huấn luyện dự báo mưa: Trên cơ sở dữ liệu lượng mưa được lựa chọn đầu vào, các công cụ của mô hình học máy DL/AI sẽ thực hiện được các bước tự động cấu hình, tham số và triển khai huấn luyện dự báo mưa lớn trên dữ liệu đầu vào và tiến hành đánh giá xác định độ tin cậy dự báo của các mô hình. Mô hình dự báo mưa lớn đã được triển khai trên hệ thống máy chủ cấu hình cao, do đó đã cải thiện được thời gian và chất lượng dự báo. Hệ thống đã thiết lập được mô hình Gradient Boosting (GB) để huấn luyện dự báo mưa lớn. Kết quả thử nghiệm đã khẳng định phương pháp mô hình Gradient Boosting có kết quả dự báo mưa lớn khả quan và đáng tin cậy.

5.6.5.2. Nhận xét chung

Kết quả đạt được:

Hệ thống đã thực hiện chuẩn hóa dữ liệu quan trắc 4 obs/ ngày (vào thời điểm 1h, 7h, 13h, và 19h với bước dữ liệu là 6 giờ/ lần dữ liệu) để đưa vào các mô hình huấn luyện dữ liệu dự báo mưa lớn diện rộng. Mô hình dự báo **mưa lớn** bằng thuật toán Gradient Boosting (GB) cho các kết quả dự báo có độ tin cậy cao và bám sát dữ liệu thực đo.

Tồn tại hạn chế: Các kết quả triển khai cho thấy mô hình AI để dự báo mưa lớn diện rộng về hiệu năng tương đối tốt. Tuy nhiên, giá trị AUC của mô hình vẫn cần được cải thiện hơn nữa, kết quả dự báo vẫn còn chưa đoán được đúng xu thế của dữ liệu.

Kiến nghị: Nhóm nghiên cứu cần tiếp tục cải tiến phương pháp để thu được kết quả dự báo tốt hơn nữa. Ngoài ra, số tham số phụ thuộc vào số giá trị thời điểm trước cũng cần được điều chỉnh linh hoạt để có được kết quả dự báo phù hợp với giá trị quan trắc thực tế.

5.7. Kết chương 5

Các nội dung Chương 5 trên đã trình bày kết quả chuyển giao, đào tạo sản phẩm, thử nghiệm và đánh giá: (i) hệ thống AI hỗ trợ dự báo bão khu vực Bắc Bộ tại Trung tâm Dự báo KTTV quốc gia, Đài KTTV khu vực Đồng bằng Bắc Bộ; (ii) hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng tại Trung tâm Dự báo

KTTV quốc gia, Đài KTTV khu vực Đồng bằng Bắc Bộ; (iii) Hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ tại Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc Bộ và Bắc Trung Bộ. Kết quả chuyển giao, đào tạo sản phẩm, thử nghiệm và đánh giá đảm bảo khối lượng, nội dung, thành phần theo đúng Thuyết minh đề tài đã được phê duyệt. Ngoài ra, đơn vị chủ trì thực hiện đề tài tự thử nghiệm và đánh giá hệ thống AI hỗ trợ dự báo mưa lớn diện rộng trên phạm vi toàn quốc.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết luận

Sau hơn 2 năm thực hiện, kết quả nghiên cứu của đề tài: Nghiên cứu cơ sở khoa học và giải pháp ứng dụng AI để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng KTTV nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam đã hoàn thành được các mục tiêu đề ra, cụ thể: (i) Đã xác định và đưa ra được cơ sở khoa học và giải pháp ứng dụng của AI để nhận dạng và dự báo một số hiện tượng KTTV nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam; (ii) Đã đề xuất và ứng dụng được một số mô hình AI để nhận dạng và dự báo một số hiện tượng KTTV nguy hiểm (bão, mưa lớn diện rộng, không khí lạnh, lũ, nước biển dâng do bão); (iii) Đã xây dựng được hệ thống nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng KTTV nguy hiểm dựa trên mô hình AI phù hợp và bước đầu triển khai thử nghiệm trong dự báo nghiệp vụ. Kết quả nghiên cứu của đề tài đã cung cấp cơ sở khoa học vững chắc để tiếp tục mở rộng các hướng nghiên cứu mới trong việc ứng dụng AI trong dự báo, cảnh báo KTTV; phát triển công nghệ dự báo KTTV hiện đại qua đó góp phần nâng cao năng lực dự báo, cảnh báo thời tiết nguy hiểm, đáp ứng được yêu cầu của Luật Phòng, Chống thiên tai và Luật KTTV.

Đề tài đã thực hiện đầy đủ các nội dung nghiên cứu và hoàn thành các sản phẩm đề ra trong thuyết minh đăng ký và đạt được một số kết quả khoa học chính như sau:

1) Đã nghiên cứu đánh giá tổng quan về AI và dữ liệu lớn Big data; đã đánh giá hiện trạng, xu thế phát triển và ứng dụng của AI và Big data trong các ngành và lĩnh vực KTTV ở Việt Nam và ở một số nước trên thế giới như ngoài nước Mỹ, Nhật Bản, Hàn Quốc, Trung Quốc, …..

2) Đã xác định được cơ sở khoa học, thực tiễn của các kỹ thuật học máy và AI trong nhận dạng, hỗ trợ dự báo, cảnh báo một số hiện tượng KTTV nguy hiểm gồm: (i) Phương pháp lưu trữ và xử lý dữ liệu; (ii) Phương pháp trích rút đặc trưng về dữ liệu về hiện tượng KTTV; (iii) Các mô hình ML, AI hỗ trợ dự báo KTTV (iv) Phương pháp xác định độ tin cậy của các mô hình ML, AI hỗ trợ dự báo KTTV; Phương pháp ra quyết định dự báo bằng trong công nghệ AI.

3) Đã xây dựng được hệ thống Big data về KTTV phục vụ hệ thống AI dự báo một số các hiện tượng thời tiết nguy hiểm gồm: số liệu khí tượng bề mặt của các trạm trên phạm vi cả nước; số liệu thủy văn của các trạm trên hệ thống sông Hồng; số liệu hải văn của các trạm ven biển Bắc Bộ, Bắc Trung Bộ; số liệu tái phân tích và số liệu viễn thám trong thời gian 10 năm giai đoạn 2008-2017.

4) Đã phát triển được các phương pháp, công cụ cho các mô hình AI dự báo một số hiện tượng KTTV nguy hiểm gồm: (i) các phương pháp,công cụ xử lý dữ liệu KTTV; (ii) Các phương pháp, công cụ trích rút đặc trưng về dữ liệu KTTV; (iii) Các phương pháp, công cụ mô hình ML, AI để huấn luyện dự báo KTTV; iv) Các phương pháp, công cụ để tối ưu hóa cấu hình, tham số và xác định độ tin cậy của các mô hình ML, AI hỗ trợ dự báo KTTV; (v) Các phương pháp, công cụ giải thích và ra quyết định thống kê trong mô phỏng quá trình dự báo KTTV bằng công nghệ AI.

5) Đã nghiên cứu ứng dụng công cụ Cray PE DL Plugin trong bài toán học sâu (DL) để mở rộng quy mô học DL tới một số lượng lớn các node trong hệ thống, qua đó giảm đáng kể thời gian học cho mạng nơ ron và tăng tính hiệu quả của DL khi đưa vào ứng dụng thực tế trong các mô hình AI dự báo KTTV. Nhất là trong giai đoạn hiện nay khi yêu cầu về các thông tin dự báo, cảnh báo KTTV thời gian thực, thời gian cực ngắn trong khoảng 30 phút - 1h.

6) Đã nghiên cứu xây dựng và triển khai các hệ thống AI hỗ trợ dự báo các hiện tượng KTTV nguy hiểm gồm: Hệ thống AI để hỗ trợ dự báo bão, mưa lớn diện rộng và không khí lạnh khu vực Bắc Bộ; Hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng; Hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ; Hệ thống framework tích hợp các module AI dự báo KTTV.

7) Đã tiến hành chuyển giao, đào tạo và thử nghiệm đánh giá khả năng áp dụng của: Hệ thống AI hỗ trợ dự báo bão, mưa lớn diện rộng, không khí lạnh; Hệ thống AI để hỗ trợ dự báo lũ; Hệ thống AI để hỗ trợ dự báo nước biển dâng do bão tại Trung tâm Dự báo KTTV quốc gia, các Đài KTTV khu vực Đông Bắc, Đồng bằng Bắc bộ, Bắc Trung Bộ. Kết quả đánh giá, nhận xét của các đơn vị tham gia thử nghiệm cho thấy:

- *Về hệ thống hệ thống AI hỗ trợ dự báo bão khu vực Bắc Bộ*: Kết quả triển khai thử nghiệm mô hình AI hỗ trợ dự báo bão cho thấy dự báo đường đi của cơn bão và vận tốc gió trong tâm bão cho kết quả tốt, bám sát với thực tế. Kết quả dự báo bão sử dụng với bài toán xử lý chuỗi thời gian một số trường hợp cho kết quả không tốt; nguyên nhân chính do dữ liệu mỗi cơn bão cụ thể có ít bản ghi (từ 10 tới 15 bản ghi thông tin bão); Do đó, để tăng hiệu quả của việc dự đoán, cần bổ sung thêm dữ liệu của từng cơn bão, tăng số lượng bản ghi và kết hợp phương pháp khác.

- *Về hệ thống hệ thống AI hỗ trợ dự báo mưa lớn diện rộng khu vực Bắc Bộ*: Các kết quả triển khai cho thấy mô hình AI để dự báo mưa lớn diện rộng sử dụng phương pháp, công cụ GP vượt trội hơn về hiệu năng so với các phương pháp dự

báo khác (MLP, SVM, kNN, DCT). Tuy nhiên, giá trị CC của mô hình còn tương đối thấp, tức là kết quả dự báo vẫn còn chưa đoán được đúng xu thế của dữ liệu. Vì vậy, trong tương lai, cần tiếp tục cải tiến GP để thu được kết quả dự báo tốt hơn nữa. Ngoài ra, tham số phụ thuộc vào số giá trị thời điểm trước cũng cần được điều chỉnh linh hoạt để có được kết quả dự báo phù hợp với thực tế. Phương pháp GB cho kết quả khá tốt trong dự báo mưa lớn khi hệ thống bắt được nguy cơ có hiện tượng mưa lớn xảy ra, tuy nhiên trong giai đoạn từ đang mưa sang hết mưa thì GB dự báo chưa tốt. Cần bổ sung thêm các dữ liệu dự báo mưa của radar làm đầu vào cho mô hình GB để tăng tính chính xác của mô hình dự báo.

- *Về hệ thống hệ thống AI hỗ trợ dự báo không khí lạnh khu vực Bắc Bộ*: Các kết quả sử dụng KNN - Regression hồi quy tuyến tính cho kết quả dự báo nhiệt độ thời hạn đến 5 ngày khá tốt. Tuy nhiên, dự báo nhiệt độ thời hạn đến 10 ngày lại thiếu chính xác. Vì vậy, cần các nghiên cứu tiếp theo để nâng cao kết quả dự báo. Kết quả chạy thử nghiệm huấn luyện dự báo sử dụng LSTM khá tốt đối với bài toán dự báo không khí lạnh thời hạn đến 5 ngày. Có thể phát triển kỹ thuật này đối với dữ liệu thật. Tuy nhiên, LSTM là mô hình hộp đen, cần xây dựng các phương pháp để giải thích kết quả, mang tính thuyết phục hơn.

- *Về hệ thống AI để hỗ trợ dự báo lũ trên hệ thống sông Hồng*: Kết quả dự báo bằng phương pháp kết hợp ARIMA_RF và ARIMA_KNN cho kết quả vượt trội và đáng tin hơn các mô hình lai khác cũng như các phương pháp truyền thống. Mô hình ARIMA_RF và ARIMA_KNN là các mô hình cần phát triển thử nghiệm để dự báo mực nước trên các trạm thủy văn của các hệ thống sông trên phạm vi cả nước.

- *Về hệ thống AI để hỗ trợ dự báo nước biển dâng do bão khu vực ven biển Bắc Bộ và Bắc Trung Bộ*: Kết quả sử dụng GP để dự báo nước biển dâng do bão cho thấy GP vượt trội hơn về hiệu năng so với các phương pháp dự báo khác (MLP, SVM, kNN, DCT, RF). Vì vậy, nhóm nghiên cứu sẽ tiếp tục cải tiến GP để thu được kết quả dự báo tốt hơn nữa. Ngoài ra, cần tiếp tục sử dụng GP để thử nghiệm áp dụng cho dữ liệu tại các trạm quan trắc khác, với các cơn bão khác và với thời gian dự báo trước ngắn hơn (12h, 5h) để có được kết quả dự báo phù hợp với yêu cầu thực tế.

8) Đề tài đã có 02 bài báo khoa học trên các tạp chí quốc tế Q1, ISI và 06 bài báo trong nước liên quan trực tiếp đến kết quả nghiên cứu của đề tài. Cụ thể:

- 02 bài báo trên Tạp chí chuyên ngành quốc tế Q1, ISI gồm: **Bài 1**: Thi Thu Hong Phan, Xuan Hoai Nguyen. *Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river*. *Advances in Wa-*

*ter Resources, Volume 142,*August 2020, 103656 ,doi.org /10.1016/ j.advwatres. 2020.103656; **Bài 2:** Nguyen Thi Hien, Cao Truong Tran, Xuan Hoai Nguyen, Sooyoul Kim, Nguyen Ba Thuy, Ngo Van Manh and Vu Dinh Phai. *Genetic Programming* for Storm Surge Forecasting. Ocean Engineering, 215, 02-2020,https:// doi.org/10.1016/j.oceaneng.2020.107812.

- 06 bài báo khoa học trên Tạp chí chuyên ngành trong nước: **Bài 1**: Ngô Văn Mạnh, Nguyễn Thị Hiền, Nguyễn Xuân Hoài, Đặng Văn Nam, Nguyễn Việt Huy. *Nâng cao hiệu năng của Deep Learning trong hệ thống tính toán hiệu năng cao Cray-XC40*. Tạp chí Khí tượng Thủy văn, 709, 01-2020, 63-70; **Bài 2**: Đặng Văn Nam, Nông Thị Oanh, Nguyễn Xuân Hoài, Ngô Văn Mạnh, Nguyễn Thị Hiền. *Phát hiện và xử lý ngoại lai cho dữ liệu nhiệt độ tại các trạm quan trắc 3h của Việt Nam*. Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất, 61, 02-2020, 132-146; **Bài 3**: Đặng Văn Nam, Hoàng Quý Nhân, Ngô Văn Mạnh, Nguyễn Thị Hiền. *Các phương pháp chuẩn hóa dữ liệu thủy văn áp dụng cho Trạm 74129 - Yên Bái*. Tạp chí Khí tượng Thủy văn, 714, 06-2020, 18-29; **Bài 4**: Nguyễn Thị Hiền, Trương Tiến Phúc, Ngô Văn Mạnh, Nguyễn Thị Quyên, Hoàng Hải Vân. *Phương pháp dự báo nước biển dâng do bão dựa trên lập trình di truyền*. Tạp chí Khí tượng Thủy văn, 715,07-2020,68-78; **Bài 5**: Nguyễn Thị Hiền, Nguyễn Xuân Hoài, Đặng Văn Nam, Ngô Văn Mạnh. *Dự báo lượng mưa tại một số trạm quan trắc Việt Nam dựa trên lập trình di truyền*. Tạp chí Khí tượng Thủy văn, 707, 11-2019, 1-9; **Bài 6**: Phạm Thị Thương, Nguyễn Xuân Hoài, Nguyễn Thị Hiền, Ngô Văn Mạnh. *Trùng lập cá thể trong lập trình di truyền*. Tạp chí Khoa học và Công nghệ Đại học Tự nhiên, 225,08-2020,61-68.

9) Đề tài đã tạo môi trường học thuật và hỗ trợ đào tiến sĩ, cao học, cụ thể:

- Hỗ trợ đào tạo 01 Tiến sĩ: Nghiên cứu sinh Hoàng Quý Nhân - Đại học Nông lâm- Đại học Thái Nguyên với luận văn "Nghiên cứu ứng dụng Trí tuệ nhân tạo (AI) trong dự báo lũ ngắn hạn trên sông Hồng".

- Hỗ trợ đào tạo 04 Thạc sĩ: (i) Học viên Huỳnh Thị Mỹ Trang - Học viện Kỹ thuật quân sự với luận văn "Phương pháp dự báo hình thái thời tiết bất thường dựa trên phỏng tiến hóa sinh học" (ii) Học viên Trương Tiến Phúc - Học viện Kỹ thuật quân sự với luận văn "Sử dụng lập trình di truyền dự báo nước biển dâng do bão"; (iii) Học viên Nguyễn Đức Thuận - Đại học Tài nguyên và Môi trường Hà Nội với luận văn "Nghiên cứu lập bản đồ phân vùng nguy cơ lũ quét trên địa bàn tính Lai Châu"; (iv) Học viên Tăng Văn An - Đại họcTài nguyên và Môi trường Hà Nội với

luận văn "Nghiên cứu xây dựng phương trình dự báo mưa thời thạn 24 giờ trong mùa mưa cho khu vực tỉnh Nghệ An".

Mặc dù đạt được nhiều kết quả khoa học đáng khích lệ nói trên, tuy nhiên, việc ứng dụng công nghệ AI/ML để hỗ trợ dự báo KTTV là công nghệ phức tạp, lần đầu tiên được triển khai nghiên cứu thực hiện trong ngành KTTV Việt Nam, với nhiều loại hình dự báo và thời hạn dự báo khác nhau, đặc biệt là do thời gian thực hiện ngắn, khối lượng, nội dung nghiên cứu lớn, nhiều loại hình thiên tai, do vậy, kết quả nghiên cứu vẫn còn một số tồn tại, hạn chế sau:

- *Về dữ liệu nghiên cứu*: Thời gian số liệu đầu vào cho các mô hình AI ngắn (10 năm), định dạng fomat, tần suất số liệu quan trắc có sự khác biệt lớn, dẫn đến khó khăn trong quá trình phát hiện, xử lý và chuẩn hóa dữ liệu. Chưa nghiên cứu tích hợp sử dụng các dữ liệu viễn thám (ra đa, vệ tinh) cho các mô hình AI dự báo KTTV.

- *Về các công cụ, phương pháp học máy*: Do đề tài triển khai nghiên cứu rộng (cả 3 lĩnh vực khí tượng, thủy văn, hải văn và với nhiều loại hình thiên tai), vì vậy, việc nghiên cứu phát triển các phương pháp công cụ AI/ML về phân đoạn thời gian theo chu kỳ dữ liệu, công cụ tự động tối ưu cấu hình, tham số của mô hình, công cụ giải thích và ra quyết định dự báo trong các mô hình AI vẫn ở giai đoạn khai phá, học hỏi việc áp dụng từ các lĩnh vực khác (ngân hàng, thống kê,...).

- *Về các mô hình, hệ thống AI hỗ trợ dự báo bão*: Bản ghi dữ liệu mỗi cơn bão cụ thể rất ít (từ 10 tới 15 bản ghi thông tin cơn bão), do đó kết quả dự báo bão sử dụng học máy với bài toán xử lý chuỗi thời gian trong một số trường hợp cho kết quả không tốt.

- *Về các mô hình, hệ thống AI hỗ trợ dự báo mưa lớn diện rộng*: Giá trị hệ số tương quan CC của mô hình còn tương đối thấp, tức là kết quả dự báo vẫn còn chưa đoán được đúng xu thế của dữ liệu quan trắc. Chưa triển khai thử nghiệm các nội dung dự báo khác về mưa lớn diện rộng như: thời gian mưa và cường độ mưa.

- *Về các mô hình, hệ thống AI hỗ trợ dự báo không khí lạnh*: Mới cho kết quả dự báo tương đối tốt đối với dự báo nhiệt độ hạn ngắn (đến 5 ngày). Kết quả dự báo nhiệt độ thời hạn vừa (đến 10 ngày) và hạn dài (tuần, tháng, mùa) thiếu chính xác. Chưa triển khai thử nghiệm dự báo các yếu tố khác trong dự báo không khí lạnh như: gió mạnh, sóng lớn, …

- *Về các mô hình, hệ thống AI hỗ trợ dự báo lũ*: Mới triển khai thử nghiệm để dự báo trị số mực nước tại các trạm thủy văn, chưa thử nghiệm các nội dung dự báo

về trị số lưu lượng, xu thế lũ, biên độ lũ, thời gian xuất hiện đỉnh lũ, diện ngập lụt, độ sâu ngập lụt, thời gian ngập lụt.

- *Về các mô hình, hệ thống AI hỗ trợ dự báo nước biển dâng do bão*: Chưa thực hiện thử nghiệm áp dụng cho dữ liệu tại các trạm quan trắc với các cơn bão khác nhau và với thời gian dự báo trước ngắn hơn (12h, 5h) để có được kết quả dự báo phù hợp với yêu cầu thực tế.

- *Về hệ thống Framework tích các module AI hỗ trợ dự báo KTTV*: Do khối lượng nghiên cứu rộng, nhiều yếu tố và loại hình dự báo và thời gian thực hiện hạn chế, vì vậy, một số công cụ trong các module chức năng của Framework chưa hoàn thành vẫn đang trong thời gian nghiên cứu, hoàn thiện như: tự động thực thi mô hình, thuật toán học máy từ các dữ liệu, tự động hiệu chỉnh tự động các tham số đầu vào, hiệu chỉnh thủ công các tham số đầu vào theo nhận định của dự báo viên, hiển thị các thông tin cảnh báo, gửi tin cảnh báo thông quan email, tin nhắn sms.

- *Về triển khai thử nghiệm và đánh giá hệ thống*: Do thời gian thực hiện đề tài hạn chế so với nội dung, khối lượng và phạm vi nghiên cứu. Các nội dung nghiên cứu phát triển công cụ Big data, công cụ AI/ML và xây dựng, triển khai mô hình hệ thống chiếm tỷ trọng thời gian lớn. Vì vậy, thời gian còn lại để thực hiện thử nghiệm, đánh giá và hiệu chỉnh mô hình, hệ thống AI hỗ trợ dự báo KTTV tương đối ít.

Kiến nghị

Trên cơ sở các kết quả nghiên cứu và các tồn tại nêu trên, nhóm thực hiện đề tài có một số kiến nghị và đề xuất hướng nghiên cứu tiếp theo của đề tài như sau:

- Các đơn vị được chuyển giao, đào tạo sản phẩm tiếp tục phối hợp thử nghiệm, đánh giá khả năng ứng dụng của mô hình hỗ trợ dự báo, cảnh báo bão, mưa lớn diện rộng, không khí lạnh, lũ và nước biển dâng do bão sử dụng AI trong nghiệp vụ thực tế tại đơn vị.

- Nghiên cứu bổ sung công cụ, phương pháp để tích hợp các dữ liệu về yếu tố địa hình và hoàn lưu khí quyển phức tạp của vùng nhiệt đới gió mùa, số liệu viễn thám vào trong các mô hình AI để nâng cao chất lượng dự báo.

- Nghiên cứu, hiệu chỉnh các mô hình AI hỗ trợ dự báo KTTV với các tập số liệu khác nhau để nâng cao độ tin cậy, chất lượng dự báo.

- Nghiên cứu, phát triển mô hình AI dự báo lưu lượng nước đến các hồ thủy điện, đặc biệt là các hồ hồ thủy điện lớn như Lai Châu, Sơn La, Hòa Bình, Tuyên Quang và bị ảnh hưởng của phía thượng nguồn ngoài lãnh thổ Việt Nam.

- Nghiên cứu, phát triển mô hình AI dự báo định lượng mưa, cảnh báo mưa lớn; dự báo số ngày nắng; dự báo lượng bức xạ mặt trời; dự báo thời tiết điểm và dự báo chuyên ngành (du lịch, thể thao, giao thông, …..).

- Ngoài việc phát triển các mô hình AI dự báo KTTV, cần nghiên cứu phát triển các mô hình AI dự báo, cảnh báo và giám sát chất lượng môi trường không khí và nước mặt.

TÀI LIỆU THAM KHẢO

Trong nước:

[1]. https://vi.m.wikipedia.org/wiki/Trituenhantao; Giáo trình Trí tuệ nhân tạo, Trần Ngân Bình. Huynh Tram Vo (2006);

[2]. vi.wikipedia.org › wiki › Dữ_liệu_lớn

[3]. Dương Thái Sơn (2015), Nghiên cứu công nghệ dữ liệu lớn (Big Data) và đề xuất ứng dụng tại Việt Nam; Mã số: 42-15-KHKT-RD

[4]. https://vjst.vn/vn/tin-tuc/2677/phat-trien-tri-tue-nhan-tao-tai-viet-nam--thuc-trang-va-giai-phap.aspx

[5]. https://www.google.com/amp/s/m.thanhnien.vn/cong-nghe/5-ung-dung-noi-bat-nhat-cua-tri-tue-nhan-tao-vao-y-te-987709.amp

[6]. http://m.cafef.vn/co-hoi-ung-dung-tri-tue-nhan-tao-trong-giao-duc-20180710171908344.chn

[7]. https://www.google.com/amp/s/m.thanhnien.vn/cong-nghe/10-ung-dung-thanh-cong-nhat-cua-tri-tue-nhan-tao-trong-kinh-doanh-989117.amp

[8]. http://vneconomy.vn/alibaba-bat-dau-dua-tri-tue-nhan-tao-vao-nong-nghiep-20180611211051921.htm

[9]. https://baomoi.com/trung-quoc-dung-tri-tue-nhan-tao-de-kiem-soat-nguoi-vi-pham-giao-thong/c/25427645.epi

[10]. https://thanhnien.vn/cong-nghe/nong-cuoc-dua-quan-su-hoa-tri-tue-nhan-tao-990707.html

[11]. https://nguoidothi.net.vn/chang-tien-si-dung-ai-bat-mach-nhung-dong-song-de-du-bao-thien-tai-15031.html

[12]. Nghiên cứu ứng dụng mạng no ron thần kinh vào sự báo lũ các sông ở tỉnh Bình Định và Quảng Trị, PGS. TS Lê Văn Nghinh và các cs.

[13]. https://ifactory.com.vn/10-ung-dung-big-data-noi-bat-thay-doi-cach-van-hanh-nen-kinh-te/

[14]. https://www.tractica.com/artificial-intelligence/using-ai-for-more-accurate-weather-forecasting/

[15]. https://global.weathernews.com/news/12254/

[16]. https://www.allerin.com/blog/transforming-weather-forecast-with-ai

[17]. https://nevonexpress.com/mPredicting-Weather-Data-Using-Artificial-Intelligence.php

[18]. https://www.dummies.com/programming/big-data/data-science/big-data-and-weather-forecasting/&prev=search&pto=aue

[19]. https://insights.magestore.com/posts/kien-truc-du-lieu-lon

[20]. Nguyễn Công Hoan, 2015, Tổng quan về dữ liệu lớn (Bigdata)

[21]. https://www.coursera.org/learn/machine-learning

[22]. https://en.wikipedia.org/wiki/Graph_database&prev=search&pto=aue

[23]. https://bigdatauni.com/vi/tin-tuc/phuong-phap-danh-gia-mo-hinh-phan-loai

[24]. (http://hadoop.apache.org)

[25]. https://en.m.wikipedia.org/wiki/Apache_Spark

[26]. https://www.cloudera.com/tutorials/introduction-to-spark-streaming.html

[27]. https://techtalk.vn/he-thong-xu-ly-du-lieu-luong-va-kien-truc.html

[28]. https://www.linkedin.com/pulse/big-data-t%E1%BB%95ng-quan-v%E1%BB%81-elasticsearch-donald-trung-manh-nguyen

[29]. https://techtalk.vn/big-data-tong-quan-ve-elasticsearch.html

[30]. https://tecadmin.net/setup-elasticsearch-on-ubuntu/

[31]. https://bigdatauni.com/vi/tin-tuc/giai-phap-cai-thien-bao-mat-du-lieu-data-security.html

[32]. https://vi.wikipedia.org/wiki/An_to%C3%A0n_th%C3%B4ng_tin

[33]. http://115.78.133.167:81/bitstream/TVHG_07113876976/2021/1/6_phuong_m_nam_6867.pdf: Nguy co mat an ninh an toan thong tin du lieu

[34]. Hoàng Ngọc Thanh và Trần Văn Lăng (2017). Một cách tiếp cận để giảm chiều dữ liệu trong việc xây dựng các hệ thống phát hiện xâm nhập mạng hiệu quả. Hội thảo lần thứ II: Một số vấn đề chọn lọc về an toàn an ninh thông tin.

[35]. Hoàng Thị Thanh Giang, Nguyễn Thị Thuý Hạnh, và Nguyễn Hoàng Huy (2015). So sánh một số thuật toán phân cụm phổ cho dữ liệu biểu diễn Gene. Tạp chí Khoa học và Phát triển, tập 3, số 6, trang 1008-1015.

[36]. Nguyễn Thị Hoàn, 2010, Phương pháp trích chọn đặc trưng ảnh trong thuật toán học máy tìm kiếm ảnh áp dụng vào bài toán tìm kiếm sản phẩm, Khoá luận tốt nghiệp, ĐHCN-ĐHQG Hà Nội.

[37]. Đặng Thị Ánh Tuyết. Tìm hiểu và ứng dụng một số thuật toán khai phá dữ liệu time series áp dụng trong bài toán dự báo tài chính. Khóa luận tốt nghiệp đại học, khoa Công nghệ thông tin - Đại học Công nghệ - Đại học Quốc gia Hà Nội, 2009.

[38]. Nguyễn Thị Hiền Nhã. Sử dụng mô hình ARIMA cho việc giải quyết bài toán dự báo tỷ giả. Luận văn thạc sĩ tin học, Đại học Khoa học Tự nhiên – Đại học Quốc gia TP.HCM, 2002.

[39]. https://cran.r-project.org/ doc/ contrib/ Intro_to_R_Vietnamese.pdf

[40]. https://machinelearningcoban.com/2018/01/14/id3/

[41]. https://techtalk.vn/10-thuat-toan-machine-learning-ma-lap-trinh-vien-can-biet.html

[42]. https://www.stdio.vn/article/gioi-thieu-ve-mo-hinh-svm-D15jcg

[43]. Bùi Minh Tăng và cá ccs, 2009: Nghiên cứu, thử nghiệm dự báo định lượng mưa từ sản phẩm mô hình HRM và GSM. Báo cáo tổng kết đề tài NCKH cấp Bộ

[44]. Áp dụng mô hình AI vào dự báo lưu lượng đến hồ lưu vực sông Ba, Cao Hoàng Hải, Trần Anh Phương, Thái Quỳnh Như, Trần Mạnh Cường, Tạp chí KTTV, 25/9/2019.

Ngoài nước:

[45]. Ian Robinson, Jim Webber, and Emil Eifrem, 2013, “Graph Databases”, O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[46]. Doug Cutting, MapReduce in Nutch, 20 June 2005, Yahoo!, Sunnyvale, CA, USA.

[47]. Doug Cutting, Scalable Computing with MapReduce, 3 August 2005, OSCON, Portland, OR, USA.

[48]. https://bigdatauni.com/vi/tin-tuc/tong-quan-ve-data-visualization-truc-quan-hoa-du-lieu.html

[49]. Tom White, Hadoop Futures, Bristol Hadoop Workshop, August 2009.

[50]. Tom White, Hadoop: The Definitive Guide, First Edition, O'Reilly, United States of America, 2009.

[51]. Atanassov, K., 1986. Intuitionistic fuzzy sets. Fuzzy Sets Systems 20 (1), 87–96.

[52]. Charu C. Aggarwal, 2017, *Outlier Analysis,* Springer International Publishing AG

[53]. https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

[54]. Tamara Munzer, 2014, *Visualization Analysis and Design*, CRC Press

[55]. https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

[56]. https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

[57]. https://www.simplilearn.com/data-preprocessing-tutorial

[58]. http://matplotlib.org/

[59]. G. Jacobson. Space-efficient static trees and graphs. In FOCS, pages 549–554, 1989.

[60]. A. Zheng, Mastering Feature Engineering Principles and Techniques for Data Scientists, O'reilly, 2016.

[61]. Seo, J. H., Lee, Y. H., & Kim, Y. H. (2014). Feature selection for very short-term heavy rainfall prediction using evolutionary computation. Advances in Meteorology, 2014.

[62]. P. Liu and H. Li, Fuzzy Neural Networks - Theory and Applications, World Scientific, 2004.

[63]. Dangeti, P. (2017). Statistics for Machine Learning. Birmingham: Packt Publishing.

[64]. https://www.astera.com/products/centerprise-data/.

[65]. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[66]. https://en.wikipedia.org/wiki/Linear_discriminant_analysis

[67]. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

[68]. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster

[69]. https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy

[70]. Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In Feature extraction (pp. 1-25). Springer, Berlin, Heidelberg.

[71]. Luis Carlos Molina, Luis Belanche, Àngela Nebot: Feature Selection Algorithms, A Survey and Experimental Evaluation, Technical report, Universitat Politècnica de Catalunya Departament de Llenguatges i Sistemes Informátics, France, 2002.

[72]. Huan Liu and Hiroshi Motoda, Computational Methods of Feature Selection, Chapman & Hall/CRC, 2008.

[73]. Ross Ihaka. Time Series Analysis, Lecture Notes for 475.726, Statistics Department, University of Auckland, 2005.

[74]. Ke-lin Du and M.N.S. Swamy, *Neural Network and Statistical Learning*, Springer-Verlag, 2014.

[75]. S. Haykin, *Neural Networks – A Comprehensive Foundation*, 2nd Editions, Prentice-Hall, 1999.

[76]. https://en.wikipedia.org/wiki/Recurrent_neural_network&prev=search&pto=aue

[77]. Davis, D. J. (n.d.). Training Feedforward Neural Networks Using Genetic Algorithms. *Machine Learning*, 762-767.

[78]. Charu C. Aggarwal, Stephen C. Gates , Philip S. Yu, "On the merits of building categorization systems by supervised clustering", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, p.352-356, August 15-18, 1999, San Diego, California, USA

[79]. http://donlnz.github.io/nonconformist/api.html#module-nonconformist.evaluation

[80]. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You?"Explaining the Predictions of Any Classifier"

[81]. http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

[82]. L.A. Zadeh, (1965). "Fuzzy sets." *Information and control*, 8: 338 – 353.

[83]. Enrique Herrera-Viedma, Francisco Herrera, and Francisco Chiclana, "A Consensus Model for Multiperson Decision Making With Different Preference Structures"; IEEE Transactions on Systems, Man and Cybernetics- part a: sysTems and Humans, vol. 32, no. 3, may 2002

[84]. Dipankar Das, Sasikanth Avancha, Dheevatsa Mudigere, Karthikeyan Vaidynathan, Srinivas Sridharan, Dhiraj Kalamkar, Bharat Kaul, Pradeep Dubey, "Distributed Deep Learning Using Synchronous Stochastic Gradient Descent," ArXiv e-prints, Feb.2016.

[85]. Diederick P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," ArXiv e-prints, Dec. 2014..

[86]. Forrest N. Iandola, Khalid Ashraf, Matthew W. Moskewicz, Kurt Keutzer, "FireCaffe: near-linear acceleration of deep neural network training on compute clusters," ArXiv e-prints, Oct. 2015.

[87]. Peter Mendygral, Nick Hill, Krishna Kandalla, Diana Moise, Jacob Balma and Marcel Schongens, "High Performance Scalable Deep Learning with the Cray Programming Environments Deep Learning Plugin," CUG 2018. 2018.

[88]. Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, Tie-Yan Liu, "Asynchronous Stochastic Gradient Descent with Delay Compensation," ArXiv e-prints, Sep. 2016.

[89]. Scpp Hochreiter, Jurgen Schmodhuber, "Long short-term memory", Neural computation 9(8): 1735 – 1780, 1997

[90]. Bhaskaran, B. Sahoo and P. K. Prediction of storm surge and inunda-tion using climatological datasets for the indian coast using soft computingtechniques. Soft Computing, 2019, Vols. 23:12363–12383.

[91]. A. J. Smola and B. Sch. 2004. A tutorial on support vector regression. ¨olkop, Statistics and Computing 14 (3), pp. 199–222.

[92]. Rokach, Lior and Maimon, O. *Data mining with decision trees: theory and applications.* s.l. : World Scientific Pub Co Inc.

[93]. Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York : Springer, 2009.

[94]. Rosenblatt, Frank. x. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Washington DC : Spartan Books, 1961.

[95]. L, Breiman. *Random Forests* . Machine Learning, 2001, Vols. 45 (1): 5–32.

[96]. Hastie, T.T., (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer.

[97]. Khu, S.T., (2004), An evolutionary-based real-time updating technique for an operational rainfall-runoff forecasting model. Proceedings of the 2nd Biennial Meeting of the International Environmental Modelling and Software Society, Manno, Switzerland, 141-146.

[98]. Koza, J.R. (1992), Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA, USA: MIT Press.

[99]. Madsen, H.B. (2000), Data assimilation in rainfall-runoff forecasting. Hydroinformatics 2000, 4th International Conference of Hydroinformatics , (pp. 1-6). Iowa, USA.

[100]. ¨olkop, A.J. (2004), A tutorial on support vector regression. Statistics and Computing, 14(3), 199-222.

[101]. Rokach, L., Maimon, O. (Eds). Data mining with decision trees: theory and applications. World Scientific Publishing Co., Inc. River Edge, NJ, USA, Series in Machine Perception and Artificial Intelligence, 81, pp. 328.

[102]. Rosenblatt, F., (1961), Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Arch Gen Psychiatry. 7(3):218-219.

[103]. Whigham, P.A. (2001), Modelling rainfall-runoff using genetic programming. Mathematical and Computer Modelling, Volume 33, Issues 6–7, March–April 2001, 707-721.

[104]. https://en.m.wikipedia.org/wiki/Missing_data

[105]. https://en.m.wikipedia.org/wiki/Uncertain_data

[106]. https://www.getlore.io/knowledgecenter/data-standardization

[107]. Chapelle, O. Z. (2006). *Semi supervised learning.* MIT Press.

[108]. https://www.tutorialspoint.com/binary-tree-representation-in-data-structures

[109]. S. Kim, Y. Matsumi, S. Pan, and H. Mase A real-time forecast modelusing artificial neural network for after-runner storm surges on the tottoricoast, japan.. Ocean Engineering, 2016, Vols. 122:44–53

[110]. https://www.cbronline.com/internet-of-things/ibm-forecasts-deep-thunder-with-machine-learning-big-data-4924838/

[111]. https://emerj.com/ai-sector-overviews/ai-for-weather-forecasting/

[112]. Elsafi, S.H., (2014), Artificial Neural Networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. Alexandria Eng. J., 53 (3), 655-662.

[113]. https://www.analyticssteps.com/blogs/weather-forecasting-how-do-big-data-analytics-magnify-it

[114]. https://www.mdpi.com/2073-4433/11/7/676/htm (ML dự báo bão)

[115]. https://www.nature.com/articles/s41598-019-49242-6 (ML dự báo mưa)

[116]. https://www.hindawi.com/journals/complexity/2020/8049504/ (ML dự báo chất lượng không khí)

[117]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6891415/ (AI để đo ô nhiễm không khí do giao thông)

[118]. S. Kim, Y. Matsumi, S. Pan, and H. Mase. A real-time forecast modelusing artificial neural network for after-runner storm surges on the tottoricoast, ja-pan.Ocean Engineering, 122:44–53, 2016

[119]. M. M. De Oliveira, N. F. F. Ebecken, J. L. F. De Oliveira, andI. de Azevedo Santos. Neural network model to predict a storm surge.Journal of applied meteorology and climatology, 48(1):143–155, 2009

[120]. Kaboudan, M. A. Genetic programming prediction of stock prices. Com-putational Economics, 2000, Vols. 16(3):207–236.

[121]. Deo, S. Gaur and M. Real-time wave forecasting using genetic program-ming.. Ocean engineering, 2008., Vols. 35(11-12):1166–1172

[122]. Ghani, H. M. Azamathulla and A. A. Genetic programming to predictriver pipeline scour. Journal of Pipeline Systems Engineering and Practice, 2010, Vols. 1(3).

[123]. Koza, John R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA, USA : MIT Press, 1992.

[124]. B. Sahoo and P. K. Bhaskaran. Prediction of storm surge and inunda-tion using climatological datasets for the indian coast using soft computingtech-niques.Soft Computing, 23:12363–12383, 2019

[125]. Breiman, L. (1997). Arcing The Edge.

[126]. Friedman, J. H. (1999). Greedy Function Approximation: A Gradient Boosting Ma-chine.

[127]. Friedman, J. H. (1999). Stochastic Gradient Boosting.

[128]. Hastie, T., Tibshirani, R., Friedman, J. H. (2009). 10. Boosting and Additive Trees. The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 337–384.

[129]. Mason L., Baxter J., Bartlett P. L., Frean M. (1999). Boosting Algorithms as Gradi-ent Descent. In S. A. Solla, T. K. Leen và K. Müller. Advances in Neural Infor-mation Processing Systems 12. MIT Press. pp. 512–518.

[130]. Mason L., Baxter J., Bartlett P. L., Frean M. (1999). Boosting Algorithms as Gradi-ent Descent in Function Space.

[131]. Li C. A Gentle Introduction to Gradient Boosting.

[132]. Ridgeway, G. (2007). Generalized Boosted Models: A guide to the gbm package.

[133]. Tianqi C. Introduction to Boosted Trees.

[134]. Cossock, D. và Zhang, T. (2008). Statistical Analysis of Bayes Optimal Subset Ranking.

PHỤ LỤC

Phụ lục 1: Cấu trúc dữ liệu lưu trữ trong Big data KTTV

Bảng 0.1: Cấu trúc lưu trữ danh mục trạm quan trắc bề mặt

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	StationName	String	x	Tên trạm không dấu
4.	VNStationName	String	x	Tên trạm có dấu
5.	Lat	Double	x	Vĩ độ trạm (°)
6.	Long	Double	x	Kinhy độ trạm (°)
7.	Elevation	Double	x	Độ cao trạm (m)
8.	Country	String	x	Tên nước
9.	Region	String	x	Tên vùng
10.	Province	String	x	Tên tỉnh
11.	District	String	x	Tên quận huyện
12.	Village	String	x	Tên xã phường
13.	IsIsland	Boolean	x	Trạm đảo
14.	NumObsDay	Int	x	Số lần quan trắc trong 1 ngày
15.	Note	String	x	Ghi chú

Bảng 0.2: Cấu trúc lưu trữ số liệu nhiệt độ trung bình ngày

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Ngày số liệu
4.	TT	Double	x	Nhiệt độ trung bình (°C)

Bảng 0.3: Cấu trúc lưu trữ số liệu nhiệt độ tối thấp ngày

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Thời gian số liệu
4.	Tm	Double	x	Nhiệt độ tối thấp ngày (°C)

Bảng 0.4: Cấu trúc lưu trữ số liệu nhiệt độ tối cao ngày

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Thời gian số liệu
4.	Tx	Double	x	Nhiệt độ tối cao ngày (°C)

Bảng 0.5: Cấu trúc lưu trữ số liệu lượng mưa ngày

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Ngày số liệu
4.	RainFallValue	Double	x	Lượng mưa ngày (mm)

Bảng 0.6: Cấu trúc lưu trữ số liệu độ ẩm

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Thời gian số liệu
4.	H	Double	x	Độ ẩm (%)

Bảng 0.7: Cấu trúc lưu trữ số liệu hướng gió và tốc độ gió

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Thời gian số liệu
4.	DD	Double	x	Hướng gió (°)
5.	FF	Double	x	Tốc độ gió (m/s)
6.	WindSpeedMax2s	Double	x	Tốc độ gió lớn nhất 2s (m/s)
7.	WindDirectionMax2s	Double	x	Hướng gió lớn nhất 2s (°)

Bảng 0.8: Cấu trúc lưu trữ số liệu khí áp

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	dtDate	DateTime		Thời gian số liệu
4.	P	Double	x	Khí áp mực trạm (mbar)
5.	P_o	Double	x	Khí áp mực biển (mbar)

Bảng 0.9: Cấu trúc lưu trữ số liệu mưa 1h/ 6h/ 24h của trạm tự động

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	DataDate	DateTime		Ngày số liệu
4.	DataHour	Int		Giờ số liệu
5.	RainFallValue	Double		Lượng mưa 1h/ 6h/ 24h (mm)
6.	flag	Int		Cờ chất lượng

Bảng 0.10: Cấu trúc lưu trữ số liệu bão của Việt Nam

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StormVN Name	String	x	Tên cơn bão
3.	StormVN TLPName	String	x	Tên cảnh báo
4.	StormQT ID	String	x	Mã quốc tế
5.	StormQT Name	String	x	Cờ chất lượng
6.	Track Date	Date	x	Ngày bão
7.	Obs	Int	x	Giờ bão
8.	FCTime Name	String	x	Thời hạn dự báo
9.	Long	Double	x	Kinh độ
10.	Lat	Double	x	Vĩ độ
11.	Pmin	Double	x	Áp suất thấp nhất
12.	Vmax	Double	x	Tốc độ gió lớn nhất
13.	R1	Int	x	Bán kính gió 1
14.	R2	Int	x	Bán kính gió 2
15.	R3	Int	x	Bán kính gió 3

Bảng 0.11: Cấu trúc lưu trữ số liệu bão của các trung tâm quốc tế

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	country	String	x	Nước phát cảnh báo
3.	code	String	x	Mã cơn bão
4.	Name	String	x	Tên cơn bão
5.	Year	String	x	Năm xảy ra
6.	Track Date	Date	x	Ngày bão
7.	Obs	Int	x	Giờ bão

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
8.	FCTime	String	x	Thời hạn dự báo
9.	Long	Double	x	Kinh độ
10.	Lat	Double	x	Vĩ độ
11.	Type	Double	x	Kiểu cảnh báo
12.	30KT	Int	x	Bán kính gió mạnh 30knots
13.	50KT	Int	x	Bán kính gió mạnh 50knots
14.	64KT	Int	x	Bán kính gió mạnh 64knots
15.	Gust	Double	x	Gió giật
16.	Move	Double	x	Tốc độ di chuyển
17.	Press	Double	x	Áp suất khí quyển
18.	V	Double	x	Tốc độ
19.	Vmax	Double	x	Tốc độ tối đa

Bảng 0.12: Cấu trúc lưu trữ lượng mưa ngày

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	StationID	String		Mã trạm
3	dtDate	DateTime		Ngày số liệu
4	RainFallValue	Double	x	Lượng mưa ngày (mm)

Bảng 0.13: Cấu trúc lưu trữ nhiệt độ tối thấp ngày

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	StationID	String		Mã trạm
3	dtDate	DateTime		Thời gian số liệu
4	Tm	Double	x	Nhiệt độ tối thấp ngày (°C)

Bảng 0.14: Cấu trúc lưu trữ danh mục trạm thủy văn

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	StationID	String		Mã trạm
3	StationName	String	x	Tên trạm không dấu
4	VNStationName	String	x	Tên trạm có dấu
5	Lat	Double	x	Vĩ độ trạm (°)
6	Long	Double	x	Kinh độ trạm (°)
7	IsTide	Boolean	x	Có ảnh hưởng triều

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
8	Country	String	x	Tên nước
9	Region	String	x	Tên vùng
10	Province	String	x	Tên tỉnh
11	District	String	x	Tên quận huyện
12	Village	String	x	Tên xã phường
13	Q	Boolean	x	Có số liệu lưu lượng
14	X	Boolean	x	Có số liệu mưa
15	Warming 1	Double	x	Cảnh báo cấp 1
16	Warming 2	Double	x	Cảnh báo cấp 2
17	Warming 3	Double	x	Cảnh báo cấp 3
18	HmaxYear	Array	x	Mảng giá trị theo năm
19	Hmax	Double	x	Mực nước lịch sử lớn nhất

Bảng 0.15: Cấu trúc lưu trữ số liệu mực nước

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	StationID	String		Mã trạm
3	DataDate	DateTime		Ngày số liệu
4	DataHour	Int		Giờ số liệu
5	DataMinute	Int		Phút số liệu
6	WaterLevel	Double		Mực nước (mm)
7	flag	Int		Cờ chất lượng

Bảng 0.16: Cấu trúc lưu trữ số liệu lưu lượng nước

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	StationID	String		Mã trạm
3	DataDate	DateTime		Ngày số liệu
4	DataHour	Int		Giờ số liệu
5	DataMinute	Int		Phút số liệu
6	Discharge	Double		Lưu lượng nước (m^3/s)
7	flag	Int		Cờ chất lượng

Bảng 0.17: Cấu trúc lưu trữ số liệu mực nước trong lũ lụt

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
2	StationID	String		Mã trạm
3	DataDate	DateTime		Ngày số liệu
4	DataHour	Int		Giờ số liệu
5	DataMinute	Int		Phút số liệu
6	WaterLevel	Double		Mực nước (mm)
7	flag	Int		Cờ chất lượng

Bảng 0.18: Cấu trúc lưu trữ số liệu độ cao sóng

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	DataDate	DateTime		Ngày số liệu
4.	WaveHeight1	Double		Độ cao sóng 1(m)
5.	WaveHeight2	Double		Độ cao sóng 2(m)
6.	flag	Int		Cờ chất lượng

Bảng 0.19: Cấu trúc lưu trữ số liệu chu kỳ sóng

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1.	_id	ObjectId		Mã đối tượng
2.	StationID	String		Mã trạm
3.	DataDate	DateTime		Ngày số liệu
4.	WaveFreq	Int		Chu kỳ sóng
5.	flag	Int		Cờ chất lượng

Bảng 0.20: Cấu trúc lưu trữ số liệu mực nước dâng do bão

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	StationID	String		Mã trạm
3	DataDate	DateTime		Ngày số liệu
4	DataHour	Int		Giờ số liệu
5	DataMinute	Int		Phút số liệu
6	WaterLevel	Double		Mực nước (mm)
7	flag	Int		Cờ chất lượng

Bảng 0.21: Cấu trúc lưu trữ đường dẫn file vệ tinh

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
2	filePath	String		Đường dẫn file
3	DataDate	DateTime		Ngày số liệu
4	Channel	String		Tên kênh ảnh

Bảng 0.22: Cấu trúc lưu trữ số liệu tái phân tích của JRA55

STT	Mã trường	Kiểu dữ liệu	Null	Mô tả trường
1	_id	ObjectId		Mã đối tượng
2	Time	String	x	Thời gian số liệu
3	Level	String	x	Mực số liệu
4	gridType	String	x	Kiểu lưới
5	Name	String	x	Tên số liệu
6	Nx	String	x	Số giá trị chiều x
7	Ny	String	x	Số giá trị chiều y
8	dx	String	x	Đơn vị chiều x
9	dy	String	x	Đơn vị chiều y
10	latLonValues	Array	x	Mảng giá trị theo tọa độ

Phụ lục 2: Công văn gửi các đơn vị về phối hợp thực hiện chuyển giao, đào tạo và thử nghiệm, đánh giá hệ thống

TỔNG CỤC KHÍ TƯỢNG THỦY VĂN
TRUNG TÂM THÔNG TIN VÀ DỮ LIỆU KHÍ TƯỢNG THỦY VĂN

Số: 239/CV-TTDL
V/v chuyển giao, đào tạo, vận hành thử nghiệm sản phẩm đề tài mã số BĐKH.34/16-20

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 18 tháng 9 năm 2020

Kính gửi:
- Trung tâm Dự báo khí tượng thủy văn quốc gia;
- Đài khí tượng thủy văn khu vực Đồng Bằng Bắc Bộ;
- Đài khí tượng thủy văn khu vực Bắc Trung Bộ;
- Đài khí tượng thủy văn khu vực Đông Bắc.

Thực hiện đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh bảo một số hiện tượng khí tượng thủy văn (KTTV) nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam" do ThS. Ngô Văn Mạnh làm Chủ nhiệm, trong đó có nội dung "Chuyển giao và đào tạo vận hành hệ thống hỗ trợ dự báo, cảnh báo bão, lũ bằng mô hình sử dụng trí tuệ nhân tạo tại Trung tâm Dự báo KTTV quốc gia và Đài KTTV khu vực Đồng Bằng Bắc Bộ", "Chuyển giao và đào tạo vận hành hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão bằng mô hình sử dụng trí tuệ nhân tạo tại Trung tâm Dự báo KTTV quốc gia, Đài KTTV khu vực Đồng Bằng Bắc Bộ, Đông Bắc và Bắc Trung Bộ". Để thực hiện nội dung nghiên cứu trên, Trung tâm Thông tin và Dữ liệu KTTV kính mong các Quý đơn vị giúp đỡ, phối hợp thực hiện thử nghiệm công nghệ dự báo đã nêu, cụ thể như sau:

1. Về nội dụng thực hiện

- Chuyển giao, đào tạo vận hành hệ thống, bàn giao tài liệu hướng dẫn sử dụng. Tài liệu cụ thể tại Phụ lục kèm theo;

- Thử nghiệm và đánh giá khả năng áp dụng hệ thống hỗ trợ dự báo, cảnh báo bão bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Trung tâm Dự báo KTTV quốc gia ; Đài KTTV khu vực Đồng Bằng Bắc Bộ;

- Thử nghiệm và đánh giá khả năng áp dụng hệ thống hỗ trợ dự báo, cảnh báo lũ bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Trung tâm Dự báo KTTV quốc gia; Đài KTTV khu vực Đồng Bằng Bắc Bộ;

- Thử nghiệm và đánh giá khả năng áp dụng hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Bộ, Bắc Trung Bộ) bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Trung tâm Dự báo KTTV quốc gia; Đài KTTV khu vực Đông Bắc và Bắc Trung Bộ.

2. Công cụ thử nghiệm: Trung tâm Thông tin và Dữ liệu KTTV cung cấp các công cụ phục vụ thử nghiệm bao gồm:

- Website hỗ trợ dữ báo, cảnh báo bão, lũ và nước biển dâng:

o Địa chỉ: http://ai.thoitietnguyhiem.gov.vn/

o Tài khoản (username/pass): dubao/dubao123, dongbac/dongbac123, dongbang/dongbang123, bactrungbo/bactrungbo123

- Tài liệu hướng dẫn sử dụng tại phụ lục kèm theo.

3. Thù lao cán bộ tham gia thử nghiệm: Các cán bộ thuộc các Đài KTTV khu vực tham gia thử nghiệm sẽ được hưởng mức thù lao được phê duyệt của Đề tài.

4. Thời gian thử nghiệm: 01 tháng (dự kiến từ 18/9 - 18/10/2020).

5. Thông tin cán bộ hỗ trợ kỹ thuật thử nghiệm: Ngô Văn Mạnh, điện thoại: 0975594713; Email: manh.ngovan@gmail.com./.

Kính mong được sự quan tâm giúp đỡ của các Quý đơn vị.

Trân trọng cảm ơn./.

Nơi nhận:
- Như trên;
- Vụ KHQT;
- Lưu: VT. KT.6.

GIÁM ĐỐC

TRUNG TÂM THÔNG TIN VÀ DỮ LIỆU KHÍ TƯỢNG THỦY VĂN

Trần Văn Nghĩa

Phụ lục 3: Các Công văn nhận xét đánh giá của các đơn vị tham gia thử nghiệm đánh giá hệ thống

- Công văn của Trung tâm Dự báo KTTV quốc gia nhận xét, đánh giá về kết quả thử nghiệm hệ thống

Ký bởi: Trung tâm Thông tin và Dữ liệu Khí tượng thủy văn
Email: cntt_ttkttvqg@monre.gov.vn
Cơ quan: Tổng cục Khí tượng Thủy văn, Bộ Tài nguyên và Môi trường
Ngày ký: 26-11-2020 09:19:53 +07:00

TỔNG CỤC KHÍ TƯỢNG THỦY VĂN
TRUNG TÂM DỰ BÁO
KHÍ TƯỢNG THỦY VĂN QUỐC GIA

Số: 751 /DBQG

V/v thử nghiệm và đánh giá hệ thống AI hỗ trợ dự báo, cảnh báo KTTV thuộc đề tài mã số BĐKH.34/16-20

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 25 tháng 11 năm 2020

Kính gửi: Trung tâm Thông tin và Dữ liệu khí tượng thủy văn

Phúc đáp Công văn số 239/CV-TTDL ngày 18 tháng 9 năm 2020 của Trung tâm Thông tin và Dữ liệu khí tượng thủy văn (KTTV) về việc chuyển giao, đào tạo và đánh giá thử nghiệm sản phẩm "Hệ thống trí tuệ nhân tạo AI hỗ trợ dự báo, cảnh báo một số các hiện tượng KTTV nguy hiểm" thuộc đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh bảo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam", mã số BĐKH.34/16-20 do ThS. Ngô Văn Mạnh làm Chủ nhiệm, Trung tâm Dự báo KTTV quốc gia có các ý kiến như sau:

1. Trung tâm Dự báo KTTV quốc gia đã được giới thiệu và hướng dẫn khai thác một số kết quả của Đề tài về việc ứng dụng công nghệ AI trong dự báo KTTV trong các buổi hội thảo tại Trung tâm.

2. Đối với hệ thống hỗ trợ dự báo bão: Kết quả dự báo bão sử dụng phương pháp học máy cho bài toán xử lý chuỗi thời gian cho kết quả không tốt do nguyên nhân chính từ việc dữ liệu mỗi cơn bão cụ thể có ít bản ghi (từ 10 tới 15 bản ghi thông tin cơn bão). Bên cạnh đó, các dự báo mới thực hiện thử nghiệm trên các tập dự báo của các Trung tâm dự báo bão của quốc tế, khi đó tính khách quan trong dự báo là hạn chế và nguồn dữ liệu đầu vào là ít (ví dụ khi không có những phát báo quốc tế cho các cơn bão, áp thấp nhiệt đới phát sinh trên Biển Đông mà chỉ Việt Nam phát báo). Do đó, cần xem xét bổ sung đầu vào là các trường khí tượng phân tích và dự báo. Đơn vị chủ trì cần tiếp tục thử nghiệm dự báo với dữ liệu thời gian thực và với tập dữ liệu quá khứ nhiều hơn.

3. Đối với hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Bộ): Dựa trên cơ sở dữ liệu độ cao sóng, mực nước biển và các dữ liệu khí tượng hải văn, hệ thống đã có khả năng đưa ra các kết quả liên quan đến dự báo nước biển dâng do bão. Tuy nhiên, hệ thống chưa thử nghiệm dự báo nước biển dâng do bão trong điều kiện thời gian thực với những thay đổi nhanh của quỹ đạo và cường độ bão.

4. Đối với hệ thống hỗ trợ dự báo, cảnh báo lũ: Phương pháp học máy đã thực hiện được các bước tự động cấu hình, tham số và thử nghiệm dự báo lũ dựa

trên dữ liệu đầu vào cho 05 đợt lũ tại trạm Yên Bái đoạn từ 2008-2017. Bên cạnh các kết quả khả quan đã nhận được, hệ thống cần triển khai thử nghiệm, hiệu chỉnh mô hình dự báo lũ kết hợp (ví dụ: ARIMA_RF) cho các hệ thống sông khác trên phạm vi cả nước và cả việc xem xét đến sự ảnh hưởng của triều đến các hệ thống sông áp dụng

Kết luận và kiến nghị chung:

- Các sản phẩm được giới thiệu trong khuôn khổ Hội thảo cho thấy có thể sử dụng để tham khảo trong nghiệp vụ dự báo KTTV; các mô-đun tính toán AI của đề tài có thể ứng dụng tại Trung tâm Dự báo KTTV quốc gia.

- Đề nghị Trung tâm Thông tin và Dữ liệu KTTV tiến hành tập huấn và chuyển giao đầy đủ các hệ thống hỗ trợ dự báo bằng công nghệ AI là sản phẩm của đề tài nêu trên để việc vận hành, khai thác tại Trung tâm Dự báo KTTV quốc gia được hiệu quả và có những đánh giá khả năng áp dụng hệ thống được đầy đủ hơn.

- Ngoài ra, việc chuyển giao đầy đủ cũng sẽ cho phép các dự báo viên, nghiên cứu viên của Trung tâm Dự báo KTTV quốc gia có thể thực hiện các thử nghiệm phù hợp với bài toán dự báo nghiệp vụ hơn.

Trung tâm Dự báo KTTV quốc gia gửi Quý Trung tâm để tổng hợp./.

Nơi nhận:
- Như trên;
- PTCT Hoàng Đức Cường (để báo cáo);
- Giám đốc Trung tâm (để báo cáo);
- Lưu: VT, STVT. T. 05.

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

TRUNG TÂM DỰ BÁO KHÍ TƯỢNG THỦY VĂN QUỐC GIA
TỔNG CỤC KHÍ TƯỢNG THỦY VĂN

Trung tâm Dự báo Khí tượng thủy văn quốc gia
Email: kttvtu@monre.gov.vn
Cơ quan: Tổng cục Khí tượng Thủy văn, Bộ Tài nguyên và Môi trường
Ngày ký: 25.11.2020 16:47:19 +07:00

Hoàng Phúc Lâm

- Công văn của Đài KTTV Khu vực Đồng bằng Bắc Bộ nhận xét, đánh giá về kết quả thử nghiệm hệ thống

TỔNG CỤC KHÍ TƯỢNG THỦY VĂN
ĐÀI KHÍ TƯỢNG THỦY VĂN
KHU VỰC ĐỒNG BẰNG BẮC BỘ

Số:805 / CV-ĐKVĐBBB
V/v thử nghiệm và đánh giá hệ thống AI hỗ trợ dự báo, cảnh báo KTTV thuộc đề tài mã số BĐKH.34/16-20

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 18 tháng 11 năm 2020

Kính gửi: Trung tâm Thông tin và Dữ liệu khí tượng thủy văn

Phúc đáp Công văn số 239/CV-TTDL ngày 18 tháng 9 năm 2020 của Trung tâm Thông tin và Dữ liệu Khí tượng thủy văn (KTTV) về việc chuyển giao, đào tạo, vận hành thử nghiệm sản phẩm "Hệ thống trí tuệ nhân tạo AI hỗ trợ dự báo, cảnh báo một số các hiện tượng KTTV nguy hiểm" thuộc đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh bảo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam", mã số BĐKH.34/16-20 do ThS. Ngô Văn Mạnh làm Chủ nhiệm. Đài Khí tượng Thủy văn khu vực đồng bằng Bắc Bộ đã tiếp nhận và thử nghiệm các hạng mục như sau:

1. Hệ thống hỗ trợ dự báo, cảnh báo bão bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế.

2. Hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Bộ) bằng mô hình sử dụng trí tuệ nhân tạo.

3. Hệ thống hỗ trợ dự báo, cảnh báo lũ bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại một số lưu vực sông chính thuộc khu vực đồng bằng Bắc Bộ.

Sau thời gian vận hành thử nghiệm các hệ thống nói trên tại Phòng Dự báo KTTV, Đài KTTV khu vực đồng bằng Bắc Bộ có các ý kiến nhận xét đánh giá về sản phẩm của đề tài nêu trên, cụ thể như sau:

1. Về vận hành hệ thống hỗ trợ dự báo, cảnh báo bão bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Đài KTTV khu vực đồng bằng Bắc Bộ:

- *Về khối lượng thử nghiệm*: đã triển khai và hoàn thành thử nghiệm dự báo cho 5 cơn bão trong giai đoạn 2008 - 2017 trong đó hệ thống AI được xây dựng dựa trên 10 năm dữ liệu từ 2008-2017.

- *Về kết quả thử nghiệm:* Hệ thống hiển thị kết quả dự đoán bão của mô hình AI và của các Trung tâm dự báo trong và ngoài nước cho thấy phù hợp với thực tế và có sai số nằm trong phạm vi cho phép

2. Về vận hành hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Bộ) bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Đài KTTV khu vực đồng bằng Bắc Bộ:

- *Về khối lượng thử nghiệm:* đã triển khai và hoàn thành thử nghiệm dự báo cho 05 đợt nước biển dâng do bão khác nhau sử dụng số liệu quan trắc tại các trạm Hòn Dấu giai đoạn từ 2008 - 2017. Hệ thống AI được chuyển giao đã tạo lập, lưu trữ đầy đủ các dữ liệu lũ tại các trạm khí tượng hải văn và thời gian dữ liệu (06 năm dữ liệu từ 2013-2018)

- *Kết quả thử nghiệm:* số liệu dự báo nước biển dâng cho các cơn bão bằng hệ thống AI đạt kết quả tốt với giá trị quan trắc thực tế. Tuy nhiên, Cần tiếp tục cải tiến mô hình dự báo nước biển dâng dô bão với thời gian dự báo trước ngắn hơn (12h,5h) để có được kết quả dự báo phù hợp với thực tế.

3. Về vận hành hệ thống hỗ trợ dự báo, cảnh báo lũ bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Đài KTTV khu vực đồng bằng Bắc Bộ:

- *Về khối lượng thử nghiệm:* đã triển khai và hoàn thành thử nghiệm dự báo cho 05 đợt lũ trong giai đoạn từ 2008 - 2017 trong đó hệ thống thử nghiệm được thiết lập dựa trên các mô hình dự báo lũ theo các phương pháp dự báo thủy văn phổ biến là ARIM, KNN, RF, SVM, LSTM và phương pháp dự báo lai là ARIMA_KNN, ARIMA_RF, ARIMA_SVR và ARIMA_LSTM.

- *Kết quả thử nghiệm*: Hệ thống AI cho các kết quả dự báo có độ tin cậy cao và bám sát dữ liệu thực đo nhất, cụ thể: (i) Thời hạn đến 24 giờ: Kết quả dự báo sát với thực đo; Thời hạn 48 - 72 giờ: kết quả dự báo tốt hơn hẳn các phương pháp dự báo khác; (iii) Thời hạn dự báo đến 5 ngày: kết quả dự báo tốt xu hướng biến đổi của mực nước (các phương pháp khác không dự báo được), nhưng sai số giá trị dự báo các yếu tố lớn. Kết quả thử nghiệm đã khẳng định phương pháp kết hợp (lai) ARIMA_RF và ARIMA_KNN có kết quả dự báo vượt trội và đáng tin hơn các mô hình lai khác cũng như các phương pháp truyền thống. Tuy nhiên, vẫn cần tiếp tục triển khai thử nghiệm, hiệu chỉnh các mô hình dự báo lũ kết hợp ARIMA_RF và ARIMA_KNN cho các hệ thống sông khác trên phạm vi cả nước, đặc biệt là các trạm quan trắc trên các sông ảnh hưởng triều.

Trên đây là ý kiến nhận xét đánh giá của Đài Khí tượng Thủy văn khu vực đồng bằng Bắc Bộ gửi Trung tâm Thông tin và Dữ liệu khí tượng thủy văn để biết và tổng hợp kết quả thử nghiệm thực tế./.

Nơi nhận:
- Như trên;
- Lưu PDB, VP, D05.

GIÁM ĐỐC

Võ Văn Hòa

- Công văn của Đài KTTV Khu vực Đông Bắc nhận xét, đánh giá về kết quả thử nghiệm hệ thống

TỔNG CỤC KHÍ TƯỢNG THỦY VĂN
ĐÀI KHÍ TƯỢNG THỦY VĂN
KHU VỰC ĐÔNG BẮC

Số: 489/CV-ĐKVĐB

V/v thử nghiệm và đánh giá hệ thống AI hỗ trợ dự báo, cảnh báo KTTV thuộc đề tài mã số BĐKH.34/16-20

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hải Phòng, ngày 17 tháng 11 năm 2020

Kính gửi: Trung tâm Thông tin và Dữ liệu Khí tượng Thủy văn

Phúc đáp Công văn số 239/CV-TTDL ngày 18/9/2020 của Trung tâm Thông tin và Dữ liệu Khí tượng Thủy văn (KTTV) về việc chuyển giao, đào tạo, vận hành thử nghiệm sản phẩm "Hệ thống trí tuệ nhân tạo AI hỗ trợ dự báo, cảnh báo một số các hiện tượng KTTV nguy hiểm" thuộc đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam", mã số BĐKH.34/16-20 do ThS. Ngô Văn Mạnh làm Chủ nhiệm; sau thời gian vận hành thử nghiệm hệ thống, Đài KTTV khu vực Đông Bắc có các ý kiến nhận xét đánh giá về sản phẩm của đề tài nêu trên, cụ thể như sau:

1. Về công cụ, mô hình phục vụ thử nghiệm:

- Hệ thống Framework tích hợp các module AI hỗ trợ dự báo **bão, nước biển dâng do bão** được thiết kế khoa học, hợp lý, giao diện phù hợp, thuận tiện cho việc tham khảo dự báo tại các đơn vị dự báo.

- Bộ công cụ được thiết lập và vận hành đảm bảo an toàn thông tin, quản trị hệ thống phân quyền người dùng.

Kiến nghị:

Đề nghị thiết lập bổ sung và chuẩn hóa mã nhóm người dùng (như nhóm Trung tâm Dự báo KTTV quốc gia, nhóm các Đài KTTV khu vực, nhóm dự báo thời tiết, nhóm dự báo thủy văn, hải văn).

2. Về vận hành hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Bộ) bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Đài KTTV khu vực Đông Bắc

- Về khối lượng thử nghiệm: Đài đã sắp xếp bố trí cán bộ dự báo, cán bộ kỹ thuật phối hợp chặt chẽ với đơn vị chủ trì đề tài triển khai và hoàn thành thử nghiệm dự báo cho 05 đợt nước biển dâng do bão khác nhau sử dụng số liệu quan trắc tại các trạm Hòn Dáu giai đoạn từ 2008 - 2017 theo đúng khối lượng, chất lượng và tiến độ theo yêu cầu.

- Về CSDL và xử lý dữ liệu nước biển dâng do bão: Hệ thống AI được chuyển giao đã tạo lập, lưu trữ đầy đủ các dữ liệu lũ tại các trạm khí tượng hải

văn và thời gian dữ liệu (06 năm dữ liệu từ 2013-2018) theo yêu cầu; đã có đầy đủ các các công cụ tích hợp truy xuất, phát hiện, xử lý, giảm chiều, chuẩn hóa, phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu nước biển dâng theo yêu cầu của mô hình huấn luyện dự báo nước biển dâng do bão.

- Về huấn luyện dự báo nước biển dâng do bão: Trên cơ sở dữ liệu độ cao sóng, mực nước biển và các dữ liệu khí tượng hải văn được lựa chọn đầu vào, các công cụ của mô hình AI đã thực hiện được các bước tự động cấu hình, tham số và triển khai huấn luyện dự báo nước biển dâng do bão.

- Hệ thống mô hình AI đã thiết lập được các tham số sử dụng trong mô hình dự báo nước biển dâng do bão gồm: Tham số khí tượng: tốc độ gió (m/s), hướng gió (độ), khí áp trên mặt biển (hPa) và độ giảm khí áp trên mặt biển (=1013 hPa); tham số hải văn: mực nước bề mặt biển (SS), thủy triều (SSL); tham số cơn bão: kinh độ (LG), vĩ độ (LT), áp suất tâm bão (hPa) và tốc độ gió Max gần tâm bão (m/s).

- Hệ thống mô hình AI đã thiết lập được các chỉ số để đánh giá chất lượng mô hình dự báo nước biển dâng do bão gồm: NRMSE (normal root mean squared error) là RMSE chuẩn hóa tính theo %; CC (correlation coefficien) là hệ số tương quan.

- Kết quả thử nghiệm dự báo nước biển dâng cho các cơn bão khẳng định mô hình AI bằng GP bám sát nhất với giá trị quan trắc thực tế.

Tồn tại hạn chế:

Hệ thống chưa nghiên cứu thử nghiệm dự báo nước biển dâng do bão realtime với dữ liệu quan trắc khí tượng hải văn theo thời gian thực.

Kết luận: Đài KTTV khu vực Đông Bắc đã phối hợp chặt chẽ với Trung tâm Thông tin và Dữ liệu KTTV thực hiện đầy đủ các nội dung về tiếp nhận chuyển giao, đào tạo và vận hành sản phẩm theo yêu cầu. Sản phẩm hệ thống AI hỗ trợ dự báo nước biển dâng do bão vùng biển Bắc Bộ và Bắc Trung Bộ do đơn vị chủ trì xây dựng và chuyển giao về cơ bản đã hỗ trợ tương đối tốt cho hoạt động nghiệp vụ dự báo nước biển dâng do bão của đơn vị.

Trên đây là ý kiến nhận xét đánh giá của Đài KTTV khu vực Đông Bắc gửi quý Trung tâm để tổng hợp và nghiên cứu phương án chỉnh sửa hoàn thiện các kiến nghị và tồn tại hạn chế.

Trân trọng phúc đáp./.

Nơi nhận:
- Như trên;
- Lưu: PDB, VP.S3

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

BỘ TÀI NGUYÊN VÀ MÔI TRƯỜNG
ĐÀI KHÍ TƯỢNG THỦY VĂN KHU VỰC ĐÔNG BẮC
TỔNG CỤC KHÍ TƯỢNG THỦY VĂN

Phan Văn Đoàn

- Công văn của Đài KTTV Khu vực Bắc Trung Bộ nhận xét, đánh giá về kết quả thử nghiệm hệ thống

TỔNG CỤC KHÍ TƯỢNG THỦY VĂN
ĐÀI KHÍ TƯỢNG THỦY VĂN
KHU VỰC BẮC TRUNG BỘ

Số: 467 /CV-ĐKVBTB

V/v thử nghiệm và đánh giá hệ thống AI hỗ trợ dự báo, cảnh báo KTTV thuộc đề tài mã số BĐKH.34/16-20

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Tp Vinh, ngày 20 tháng 11 năm 2020

Kính gửi: Trung tâm Thông tin và Dữ liệu khí tượng thủy văn

Phúc đáp Công văn số 239/CV-TTDL ngày 18/9/2020 của Trung tâm Thông tin và Dữ liệu Khí tượng thủy văn (KTTV) về việc chuyển giao, đào tạo, vận hành thử nghiệm sản phẩm "Hệ thống trí tuệ nhân tạo AI hỗ trợ dự báo, cảnh báo một số các hiện tượng KTTV nguy hiểm" thuộc đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh bảo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam", mã số BĐKH.34/16-20 do ThS. Ngô Văn Mạnh làm Chủ nhiệm;

Sau thời gian vận hành thử nghiệm hệ thống, Đài KTTV khu vực Bắc Trung Bộ có các ý kiến nhận xét đánh giá về sản phẩm của đề tài nêu trên, cụ thể như sau:

1. Về công cụ, mô hình phục vụ thử nghiệm

Hệ thống Framework tích hợp các module AI hỗ trợ dự báo **nước biển dâng do bão** do đơn vị chủ trì thực hiện đề tài chuyển giao cho Đài đã xây dựng đầy đủ các module công cụ huấn luyện và các mô hình dự báo cho các lĩnh vực theo đúng thuyết minh đề tài được phê duyệt gồm:

- Chức năng tạo lập, chỉnh sửa các module học máy về huấn luyện mô hình: Mã huấn luyện mô hình, loại mô hình huấn luyện, trường dự báo, thời gian huấn luyện, …;
- Có đầy đủ các mô hình AI nhận dạng và huấn luyện dự báo các loại hình thời tiết nguy hiểm gồm bão, mưa lớn diện rộng, lũ, nước biển dâng do bão, ... theo các tùy chọn của đối tượng khai thác sử dụng như: chọn loại mô hình dự báo, thời hạn dự báo, mã trạm quan trắc KTTV, chọn dữ liệu lũ, chọn dữ liệu bão, ….

Kiến nghị:

Đề nghị chỉnh sửa một số thuật ngữ cho đúng với thuật ngữ quy định về dự báo KTTV hiện hành, ví dụ "Dự báo khí tượng" là "Dự báo thời tiết".

2. Về vận hành hệ thống hỗ trợ dự báo, cảnh báo nước biển dâng do bão (ven biển Bắc Trung Bộ) bằng mô hình sử dụng trí tuệ nhân tạo vào thực tế tại Đài KTTV khu vực Bắc Trung Bộ

- Về khối lượng thử nghiệm: Đài đã sắp xếp bố trí cán bộ dự báo, cán bộ kỹ thuật phối hợp chặt chẽ với đơn vị chủ trì đề tài triển khai và hoàn thành thử nghiệm dự báo cho 05 đợt nước biển dâng do bão khác nhau sử dụng số liệu quan trắc tại các trạm Hòn Ngư giai đoạn từ 2008 - 2017 theo đúng khối lượng, chất lượng và tiến độ theo yêu cầu.

- Về CSDL và xử lý dữ liệu nước biển dâng do bão: Hệ thống AI được chuyển giao đã tạo lập, lưu trữ đầy đủ các dữ liệu lũ tại các trạm khí tượng hải văn và thời gian dữ liệu là từ 2013-2018;

- Các công cụ tích hợp truy xuất, phát hiện, xử lý, giảm chiều, chuẩn hóa, phân cụm, phân hạng và lựa chọn đặc trưng dữ liệu nước biển dâng được thử nghiệm cho kết quả đúng theo yêu cầu của mô hình huấn luyện dự báo nước biển dâng do bão.

- Về huấn luyện dự báo nước biển dâng do bão: Mô hình dự báo dự báo nước biển dâng do bão đã được triển khai trên hệ thống máy chủ cấu hình cao, do đó đã cải thiện được thời gian và chất lượng dự báo.

Tồn tại hạn chế:

Hệ thống chưa nghiên cứu thử nghiệm dự báo nước biển dâng do bão realtime với dữ liệu quan trắc khí tượng hải văn theo thời gian thực.

Kết luận: Đài KTTV khu vực Bắc Trung Bộ đã phối hợp chặt chẽ với Trung tâm Thông tin và Dữ liệu KTTV thực hiện đầy đủ các nội dung về tiếp nhận chuyển giao, đào tạo và vận hành sản phẩm theo yêu cầu. Sản phẩm hệ thống AI hỗ trợ dự báo nước biển dâng do bão vùng biển Bắc Bộ và Bắc Trung Bộ do đơn vị chủ trì xây dựng và chuyển giao về cơ bản đã hỗ trợ tương đối tốt cho hoạt động nghiệp vụ dự báo nước biển dâng do bão của đơn vị.

Đài KTTV khu vực Bắc Trung Bộ trân trọng phúc đáp và đề nghị quý Trung tâm nghiên cứu phương án chỉnh sửa hoàn thiện các kiến nghị và tồn tại hạn chế.

Trân trọng phúc đáp./.

Nơi nhận:
- Như trên;
- Lưu: DB, VP.H(2).

GIÁM ĐỐC

BỘ TÀI NGUYÊN VÀ MÔI TRƯỜNG
ĐÀI KHÍ TƯỢNG THỦY VĂN KHU VỰC BẮC TRUNG BỘ
TỔNG CỤC KHÍ TƯỢNG THỦY VĂN

Nguyễn Văn Lượng