TỔNG QUAN ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY TRONG DỰ BÁO LŨ

Đinh Nhật Quang, Tạ Quang Chiểu
Trường Đại học Thuỷ lợi
Trịnh Trần Tiểu Long
Trường Đại học Cantabria

Tóm tắt: *Với tầm quan trọng của việc dự báo và cảnh báo lũ, các nghiên cứu tập trung vào ứng dụng các mô hình học máy vào các bài toán dự báo lũ đang ngày càng được quan tâm. Trong bài báo này, chúng tôi tổng quan lại các bài nghiên cứu gần đây về ứng dụng của học máy trong lĩnh vực về dự báo lũ, dự đoán mực nước, lưu lượng, độ sâu ngập,... cùng với đó là các chỉ số thường dùng để đánh giá độ tin cậy của các mô hình học máy. Các nghiên cứu đã cho thấy, khác với các mô hình toán, mô hình học máy cần ít thông số đầu vào, tốn ít thời gian để mô phỏng và không đòi hỏi nhiều kiến thức về mô phỏng ngập lụt mà vẫn đưa ra kết quả dự đoán có độ chính xác tốt. Bên cạnh đó, nhóm nghiên cứu cũng đã chỉ ra một số hạn chế của việc ứng dụng các mô hình học máy, từ đó đề ra những gợi ý về hướng nghiên cứu cần thực hiện để tối ưu hoá các mô hình học máy trong dự báo lũ.*
Từ khóa: *Học máy, Học sâu, Dự báo lũ, Dự đoán Mực nước, Dự đoán Lưu lượng.*

Summary: *In accordance with the great significance of flood prediction and warning, there has been much research focusing on machine learning models applications (data-driven models) in flood prediction problems. In this paper, we reviewed recent research on the application of machine learning in flood prediction, water-level prediction, discharge prediction, flood depth prediction, etc along with adopted popular indicators for the evaluation of the reliability of machine learning models' performance. These studies have shown that, unlike traditional numerical models, machine learning models require fewer input parameters, and less simulation time, and do not require extensive knowledge of flood modeling, while still providing good precision prediction results. Besides, the research group has also identified and highlighted some limitations and challenges in the application of machine learning models, along with suggestions for future research orientations to optimize machine learning models in flood prediction.*
Keywords: *Machine Learning, Deep Learning, Flood Prediction, Water Level Prediction, Discharge Prediction.*

1. ĐẶT VẤN ĐỀ

Lũ lụt là một hiện tượng thiên nhiên hằng năm gây ra nhiều thiệt hại nặng nề đối với cơ sở hạ tầng, hoa màu cũng như nền kinh tế của các nước trên toàn thế giới. Nguyên nhân chủ yếu gây ra lũ lụt là do các tác động của biến đổi khí hậu, do tác động của con người, do mưa lớn làm mực nước trên sông tăng nhanh, khiến nước không kịp thoát [1]. Trong hơn 27 năm qua, lũ lụt là nguyên nhân gây ra cái chết cho hơn 175.000 người, và gây ảnh hưởng nặng nề về kinh tế ước tính lên đến 2,2 tỉ đô trên toàn

Ngày nhận bài: 10/7/2023
Ngày thông qua phản biện: 25/7/2023
Ngày duyệt đăng: 02/8/2023

cầu [2]. Việc ứng phó với lũ lụt rất quan trọng, đặc biệt tại các nước đang phát triển, khi các biện pháp phòng ngừa và giảm thiểu thiên tai còn hạn chế và các đồng bằng nơi thường phải hứng chịu lũ lụt thường tập trung đông dân cư [3]. Do đó việc dự báo lũ, mực nước và lưu lượng trên sông đặc biệt trên các sông chưa có hoặc có ít trạm quan trắc thuỷ văn, rất quan trọng trong việc cảnh báo lũ cho người dân và chính quyền địa phương.

Đến nay đã có nhiều dự án cũng như nghiên cứu của các nhà khoa học trong và ngoài nước ứng dụng các mô hình dựa trên tính chất vật lý (*physically-based model*) có độ chính xác cao vào việc dự đoán mực nước, lưu lượng hay dòng chảy đến trên các con sông, hồ chứa, … như mô hình MIKE, HYDRO River, HEC-HMS, SOBEK, EFDC, … Tuy nhiên các mô hình loại này còn tồn tại nhiều hạn chế bởi vì chúng cần rất nhiều các thông số đầu vào trong mô hình như mực nước, lưu lượng, bốc hơi, tốc độ thấm, độ ẩm của đất, … và đặc biệt cần rất nhiều thời gian để mô phỏng. Hơn thế, để thiết lập, mô phỏng và phân tích các kết quả đầu ra của các mô hình dựa trên tính chất vật lý đòi hỏi sự tham gia của các chuyên gia trong lĩnh vực thuỷ văn, thuỷ lực,… Do đó tính ứng dụng thực tế của các mô hình này vào việc cảnh báo lũ theo thời gian chưa cao. Hướng phát triển trong tương lai của ngành thuỷ văn và quản lý tài nguyên nước đó là tìm ra phương pháp để tích hợp quản lý tài nguyên nước dựa trên các mô hình toán truyền thống vào các mô hình học máy để trực tiếp xử lý, phân tích và lấy thông tin từ các nguồn dữ liệu lớn [4]. Do đó trong những năm gần đây, học máy (*machine learning*) đã thu hút nhiều sự quan tâm, chú ý của các nhà thuỷ văn học và được ứng dụng rộng rãi trong nhiều lĩnh vực nhờ khả năng quản lý dữ liệu lớn.

Ngày nay, với sự phát triển của công nghệ thông tin thì các thuật ngữ học máy hay học sâu (*deep learning*) không còn quá xa lạ với chúng ta. Học máy được ứng dụng vào nhiều ngành nghề, lĩnh vực khác nhau của xã hội trong đó có lĩnh vực quản lý tài nguyên nước. Trong Bảng 2 liệt kê một số bài báo tổng quan về ứng dụng phương pháp học máy trong lĩnh vực quản lý tài nguyên nước nói chung và dự báo lũ nói riêng. Các bài tổng quan đã cho thấy sự phát triển nghiên cứu và ứng dụng học máy vào các bài toán trong lĩnh vực quản lý tài nguyên nước và quản lý rủi ro thiên tai. Tuy nhiên các bài bài báo này chủ yếu được viết bằng tiếng Anh khiến cho nhiều độc giả, nhà khoa học trong lĩnh vực thuỷ văn, tài nguyên nước ở Việt Nam có thể chưa tiếp cận được hay chưa có cái nhìn sâu đến kỹ thuật tiên tiến và hiện đại này. Hơn nữa, các tác giả hiện nay phần đa tập trung nhiều vào việc làm sáng tỏ cấu trúc thuật toán trong học máy hơn là tập trung vào các ứng dụng của chúng, khiến người đọc với nền tảng về lĩnh vực thủy văn, tài nguyên nước khó tiếp cận.

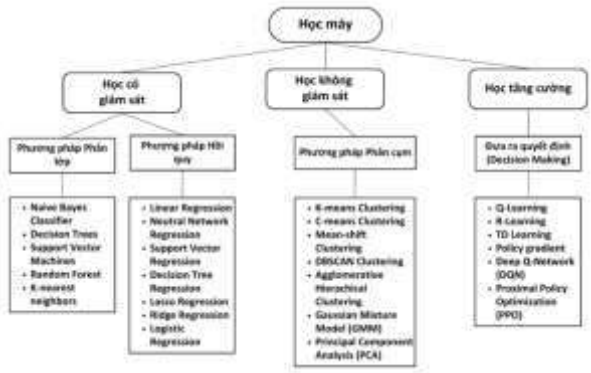
Bảng 2: Các bài báo tổng quan gần đây về ứng dụng học máy trong bài toán dự báo lũ

Lĩnh vực nghiên cứu	Tóm tắt	Trích dẫn
Mô phỏng ngập lụt	Đánh giá tính chính xác của mô hình định hướng dữ liệu (*data-driven model*) sử dụng học máy trong việc mô phỏng ngập lụt thông qua việc so sánh với các mô hình dựa trên tính chất vật lý truyền thống	[5]
Xây dựng bản đồ ngập lụt	Đánh giá tổng quan về những ứng dụng của học sâu trong việc xây dựng bản đồ ngập lụt và hướng nghiên cứu trong tương	[6]

Lĩnh vực nghiên cứu	Tóm tắt	Trích dẫn
	lai.	
Dự đoán mực nước	Đánh giá tổng quan về các điểm mạnh của một số thuật toán trong học máy trong bài toán dự đoán mực nước	[7]
Dự đoán lưu lượng	Đánh giá tổng quan về các mô hình trí tuệ nhân tạo (*Artificial Intelligence – AI*) được sử dụng trong lĩnh vực dự đoán lưu lượng nhằm góp phần cải thiện và tối ưu hoá trong việc quản lý và vận hành hồ chứa	[8]
Dự đoán lũ, lượng mưa, chất lượng nước và mực nước ngầm	Đánh giá tổng quan cho thấy các mô hình định hướng dữ liệu thể hiện tốt và chính xác hơn các mô hình toán trong các bài toán về dự báo lũ, lượng mưa, chất lượng nước và mực nước ngầm, đặc biệt trong bài toán dự báo ngắn hạn	[9]
Dự báo dòng chảy	Đánh giá tổng quan về những phát triển về ứng dụng học máy trong gần 2 thập kỷ trong việc mô phỏng thuỷ văn và dự báo dòng chảy tại các lưu vực không có trạm quan trắc	[10]
Mô phỏng dòng chảy và dự báo lũ	Đánh giá tổng quan các mô hình học máy được ứng dụng chính trong lĩnh vực thuỷ văn học như dòng chảy mặt, chất lượng nước, quá trình thoát nước, vận chuyển bùn cát, mực nước ngầm, ngập lụt,...	[11]
Dự đoán sự thay đổi của mực nước trong hồ	Đánh giá một cách có hệ thống về các thuật toán trong học máy sử dụng cho việc dự đoán dao động và sự thay đổi mực nước trong hồ	[12]

Việc ứng dụng học máy vào các bài toán dự báo lũ là một phương pháp khá mới và hiện đại ở Việt Nam cũng như trên thế giới. Trong bài báo này, nhóm nghiên cứu tổng quan các bài báo khoa học được công bố từ năm 2020 đến nay trên các cơ sở dữ liệu khoa học uy tín (bao gồm Google Scholar, Science Direct, Web of Science, Scopus, IEEE Xplore, Springer Link,…) nhằm cập nhật và đem đến cho người đọc những kiến thức và nghiên cứu mới nhất về chủ đề này.

2. CÁC THUẬT TOÁN HỌC MÁY SỬ DỤNG TRONG DỰ BÁO LŨ



Hình 1: Các nhóm thuật toán chính trong học máy

Phương pháp học máy (*Machine Learning*) là một lĩnh vực của trí tuệ nhân tạo (*Artificial Intelligence*) và tập trung vào việc xây dựng các mô hình và thuật toán có khả năng học hỏi và tự điều chỉnh dữ liệu. Mục tiêu chính của học máy là cho máy tính tự động "học hỏi" từ những dữ liệu mà không cần phải được lập trình một cách cụ thể. Các thuật toán học máy có thể được chia thành 3 nhóm chính gồm học có giám sát (*Supervised Learning*), học không giám sát (*Unsupervised Learning*) và học tăng cường (*Reinforcement Learning*) [9] (Hình 1).

Học có giám sát là nhóm phổ biến nhất trong các thuật toán học máy. Chúng huấn luyện mô hình bằng cách sử dụng một tập dữ liệu gồm các tham số đầu vào đã được gán nhãn, có cặp dữ liệu đầu vào và đầu ra tương ứng với mục đích nhằm giúp mô hình có thể "học hỏi" để đưa ra dự đoán đầu ra cho các tập dữ liệu mới. Các thuật toán học có giám sát được phân ra thành hai loại chính: i) phân lớp (*classification*) – khi các nhãn của dữ liệu đầu vào được chia thành một số hữu hạn lớp (miền là giá trị rời rạc); và ii) hồi quy (*regression*) – khi nhãn không được chia thành các nhóm mà là một giá trị thực cụ thể (miền là giá trị liên tục) [13]. Trái với học có giám sát, học không giám sát không yêu cầu dữ liệu đã được gán nhãn. Mô hình được huấn luyện để tìm hiểu cấu trúc, mẫu hoặc nhóm trong dữ liệu mà không có đầu ra tương ứng. Các thuật toán học không có giám sát thường được sử dụng để phân cụm (*clustering*) dữ liệu, phân tích thành phần chính (*principal component analysis*), hoặc khám phá cấu trúc dữ liệu [14]. Học tăng cường là một hình thức học máy nơi mô hình học thông qua tương tác liên tục với một môi trường và nhận điểm thưởng (thu được trong quá trình tương tác của mô hình với môi trường để khích lệ mô hình thực hiện các hành động có lợi để đạt được mục tiêu – *reward*) dựa trên hành động của nó. Mục tiêu của phương pháp này là tìm ra cách tối đa hóa điểm thưởng tích lũy trong thời gian dài thông qua việc tìm hiểu và tối ưu hóa chính sách hành động (quy tắc chiến lược xác định cách mà mô hình chọn hành động trong một tình huống cụ thể - *action policy*) [15]. Bảng 3 trình bày một cách tổng quan về sự khác nhau giữa ba phương pháp chính của học máy.

Với những đặc điểm và tính ứng dụng của các phương pháp học máy kể trên, bài báo sẽ tập trung tổng quan về ứng dụng của các thuật toán hồi quy, phân lớp cũng như học sâu thường được sử dụng trong các bài toán dự báo dữ liệu theo chuỗi thời gian (*time series prediction*) như dự đoán mực nước, lưu lượng, ….

Bảng 3: So sánh ba phương thức học chính trong học máy

Phương thức học	Loại dữ liệu	Huấn luyện mô hình	Lĩnh vực áp dụng		Các thuật toán tiêu biểu
Học có giám sát	Dữ liệu có gán nhãn	Huấn luyện mô hình từ dữ liệu đã được gán nhãn	Bài toán hồi quy (*Regression*): thường sử	Bài toán phân loại (*Classification*): thường sử	Linear Regression, Neutral Network

			dụng trong các bài toán dự đoán	dụng trong phân loại nhị phân và các lớp khác nhau	Regression, Random Forest Regression, ...
Học không có giám sát	Dữ liệu không gán nhãn	Huấn luyện mô hình với dữ liệu không gán nhãn, để mô hình tự tìm ra cấu trúc dữ liệu	Phân cụm (*clustering*) và giảm số chiều của dữ liệu (*dimension reduction*)		GMM, PCA, t-SNE, K-means, C-means, Phân cụm phân cấp (*Hierarchical clustering*),...
Học tăng cường	Dữ liệu không xác định trước	Mô hình học hỏi qua tương tác liên tục với môi trường xung quanh	Lý thuyết trò chơi (*Game theory*)		Q-Learning, DQN, Policy gradient, PPO,...

2.1. Các thuật toán học máy dùng cho bài toán dự báo lũ

Với sự phát triển của các thuật toán trong học máy, ngày nay nhiều nhà khoa học, chuyên gia trong lĩnh vực thuỷ văn và quản lý tài nguyên nước đã nghiên cứu và áp dụng các thuật toán kể trên vào các bài toán thực tế trong dự báo và cảnh báo lũ (Bảng 4).

Bảng 4: Các nghiên cứu gần đây về việc ứng dụng phương pháp học máy trong dự báo lũ

Lĩnh vực và quy mô nghiên cứu	Thuật toán	Kết quả	Trích dẫn
Dự đoán mực nước (theo ngày)	Hồi quy tuyến tính (LR), Support Vector Machine (SVM), Ensemble Regression (ER), Xgboost, Tree Regression (TR), Gaussian Process Regression (GPR)	GPR đã đem lại kết quả tốt nhất so với các mô hình sử dụng các thuật toán còn lại	[16]
Dự đoán mực nước (theo ngày)	LR, Random Forest Regression (RFR) và Light Gradient Boosting Machine Regression (LGBMR)	LR đã thể hiện độ chính xác cao hơn so với 2 mô hình còn lại với chỉ số R^2, NSE, MAE và RMSE lần lượt là 0,959; 0,958; 6,67 cm và 12,2 cm	[17]
Dự đoán mực nước	Single-output Long-Short Term Memory (LSTM SO) và	LSTM ED cho thấy sự chính xác hơn so với LSTM SO và kết quả của nó cũng	[18]

Lĩnh vực và quy mô nghiên cứu	Thuật toán	Kết quả	Trích dẫn
(theo giờ)	Encoder-Decoder Long-Short Term Memory (LSTM ED)	cho thấy xu thế tương đồng so với các nghiên cứu cũ sử dụng các mô hình Encoder-Decoder	
Dự đoán mực nước (theo giờ)	Support Vector Regression (SVR), RFR, Multilayer Perceptron Regression (MLPR) và LGBMR	LGBMR thể hiện tốt nhất với sai số mực nước khi đạt đỉnh lũ và thời gian mực nước đạt đỉnh lũ so với số liệu quan trắc lần lượt là 0,22 m và 2h	[19]
Dự đoán lưu lượng (theo tuần)	Artificial Neural Network (ANN), Fuzzy Logic và Adaptive Neuro Fuzzy Inference System (ANFIS)	Mô hình ANFIS kết hợp với thuật toán "*hybrid training*" cho ra kết quả tốt nhất với các chỉ số NSE, R^2, MSE và RMSE lần lượt là 0,968; 97,066%, 0,00034 m^3/s và 0,018 m^3/s	[20]
Dự đoán lưu lượng (theo ngày)	LSTM, Bayesian Neural Network (BNN), LSTM with Monte Carlo Dropout (LSTM-MC) và Bayesian LSTM (BLSTM)	BLSTM thể hiện kết quả vượt trội hơn so với các mô hình khác về độ tin cậy, sắc nét và về hiệu suất tổng thể của dự báo	[21]
Dự đoán lưu lượng (theo tháng)	ANN, Convolutional Neural Network (CNN), LSTM	Dùng phương pháp phân tích phổ đơn lẻ (SSA) và phân rã Mùa vụ bằng phương pháp Loess (STL). Kết quả cho thấy các mô hình dựa trên SSA (SSA-ANN, SSA-CNN và SSA-LSTM) đều thể hiện vượt trội hơn với mô hình dựa trên STL. Và đặc biệt mô hình SSA-ANN thể hiện tốt nhất so với các mô hình còn lại với chỉ số NSE = 0,9045 và chỉ số Willmott WI = 0,9764	[22]
Dự đoán đường quan hệ giữa mực nước và lưu lượng (theo ngày)	SVM, Extreme learning machine (ELM) và ANN	Cả 3 mô hình đều thể hiện tốt, tuy nhiên SVM vẫn nhỉnh hơn trong dự đoán hệ số lưu lượng (*Discharge Coefficient* - C_d) với R^2 và RMSE lần lượt là 0,95 và 0,01	[23]
Dự đoán lưu lượng (theo tuần)	CNN-LSTM, CNN, LSTM và Deep Neural Network (DNN)	Mô hình CNN-LSTM vượt trội hơn các mô hình AI khác, với 84% lỗi giá trị dự đoán lưu lượng Q dưới 0,05 m^3/s, trong khi ở mô hình LSTM và DNN lần lượt là 80% và 66% lỗi giá trị dự đoán Q dưới 0,05 m^3/s	[24]
Bản đồ ngập lụt (sau mỗi 30s)	CNN	Mô hình thể hiện tốt trong việc dự đoán vùng ngập lụt dựa vào thuật toán CNN trong học sâu, để xử lý các ảnh thu được từ camera giám sát với tỷ lệ chính xác lên tới 92,7%	[25]
Bản đồ	Logistic Regression, SVM, K-	Kết quả cho thấy XGBoost đã thể hiện	[26]

Lĩnh vực và quy mô nghiên cứu	Thuật toán	Kết quả	Trích dẫn
ngập lụt (theo các trận lũ lịch sử)	nearest neighbor (KNN), Adaptive Boosting (AdaBoost), XGBoost.	tốt nhất so với các mô hình còn lại trong việc dự đoán vùng bị ngập và vùng không bị ngập tại Ấn Độ với giá trị trung bình cao nhất của diện tích dưới đường cong ROC là 0,83	
Bản đồ ngập lụt (theo các trận lũ lịch sử)	CNN	Kết quả cho thấy việc tích hợp CNN và phương pháp xử lý ảnh *Regional Growing* (RG) cho kết quả về vùng ngập chính xác hơn so với phương pháp RG đơn thuần (92,4% so với 91,8%)	[27]

2.2. Một số phương pháp học có giám sát thường sử dụng trong dự báo lũ

Học có giám sát là quá trình xây dựng một mô hình dự đoán hoặc phân loại dựa trên tập dữ liệu huấn luyện đã được gán nhãn. Trong học có giám sát, chúng ta có một tập dữ liệu huấn luyện bao gồm các mẫu dữ liệu đã được gán nhãn với kết quả mong muốn. Mục tiêu là tìm mối quan hệ giữa biến độc lập (các đặc trưng) và biến phụ thuộc (kết quả dự đoán hoặc nhãn), để sau đó có thể ước lượng hoặc dự đoán giá trị kỳ vọng của biến phụ thuộc cho các dữ liệu chưa được gán nhãn. Một số phương pháp học có giám sát thường được sử dụng trong bài toán dự báo lũ được miêu tả dưới đây.

2.2.1. Hồi quy tuyến tính

Hồi quy tuyến tính (*Linear Regression* - LR) là một trong những phương pháp quan trọng và phổ biến nhất trong học máy. Mục tiêu của hồi quy tuyến tính là xây dựng một mô hình dự đoán có thể tìm ra các hệ số tối ưu để dự đoán giá trị biến mục tiêu dựa trên giá trị biến đầu vào [19]. Trong bài toán dự báo chuỗi thời gian, hồi quy tuyến tính có thể được áp dụng bằng cách sử dụng các đặc trưng (*features*) của chuỗi thời gian để dự đoán giá trị tương lai của chuỗi thời gian. Ví dụ như trong bài toán về dự báo mực nước, mô hình sẽ học từ các đặc trưng của chuỗi số liệu mực nước, lượng mưa, bốc hơi hay lưu lượng đến, … để đưa ra dự đoán về mực nước tại vị trí và thời điểm muốn dự đoán. Mô hình hồi quy tuyến tính cố gắng tìm một đường thẳng hoặc siêu mặt phẳng tốt nhất để phù hợp với dữ liệu chuỗi thời gian và dự đoán giá trị tiếp theo dựa trên mối quan hệ tuyến tính của các đặc trưng.

2.2.2. Random Forest Regression

RFR là một phương pháp học máy có giám sát được sử dụng trong bài toán hồi quy. Nó là một thuật toán tổng hợp dựa trên việc kết hợp nhiều cây quyết định để tạo ra một mô hình dự báo chính xác và ổn định. RFR hoạt động bằng cách xây dựng một tập hợp các cây quyết định đơn lẻ. Mỗi cây trong RFR được xây dựng bằng cách lấy một mẫu ngẫu nhiên từ tập dữ liệu huấn luyện và sử dụng một phương pháp gọi là "*bagging*" để lựa chọn một tập con của các thuộc tính tại mỗi nút của cây. Quá trình này giúp tạo ra sự đa dạng và khả năng tổng hợp thông tin từ các cây con. Khi có một dữ liệu mới cần dự báo, RFR sử dụng các cây quyết định để đưa ra dự đoán riêng lẻ. Kết quả cuối cùng của RFR được tính bằng cách lấy trung bình (hoặc trung vị) của các dự đoán từ các cây [17]. Điều này giúp cân nhắc các dự đoán từ các cây khác nhau và tạo ra một dự báo tổng thể chính xác hơn.

2.2.3. Light Gradient Boosting Machine Regression

LGBMR là phương pháp được phát triển dựa

trên kỹ thuật "*boosting*" và sử dụng thuật toán *gradient boosting*. Mô hình được cho là có hiệu quả tốt hơn, thời gian xử lý chạy mô hình nhanh hơn và tốn ít bộ nhớ hơn do nó sử dụng phương pháp phát triển theo lá cây (*leaf-wise tree growth*) [28] (*Hình 2*).

2.2.4. Support Vector Machine

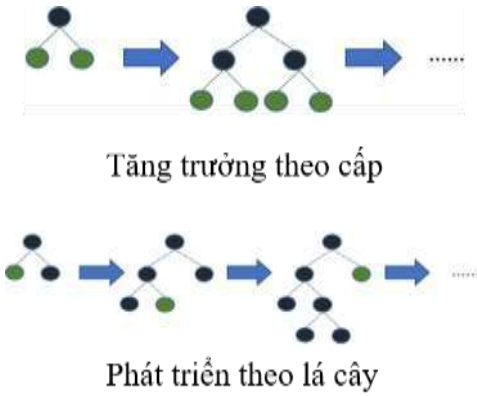
SVM là một kỹ thuật phi tuyến tính trong lĩnh vực học máy và có thể hoạt động tốt trong các bài toán về phân loại, hồi quy và dự báo chuỗi thời gian [29]. Phương pháp SVM được sử dụng rộng rãi trong lĩnh vực về dự báo lũ và thuỷ văn vì chúng phù hợp cho cả các vấn đề về tuyến tính và phi tuyến tính, và cũng nổi tiếng với khả năng tổng quát hoá mạnh mẽ. Tuy nhiên, nhược điểm chính của thuật toán này là thời gian huấn luyện có thể lớn đối với tập dữ liệu chứa nhiều đối tượng (nếu không sử dụng thuật toán cụ thể để thực hiện tối ưu hiệu quả) [30].

2.3. Một số phương pháp học học sâu sử dụng trong dự báo lũ

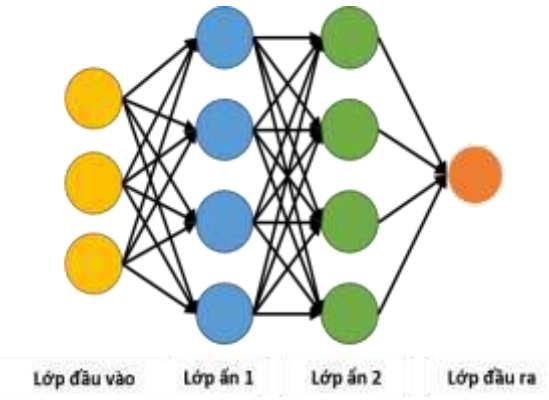
Mô hình học sâu (*Deep Learning*) hay mô hình các mạng nơron, là một nhánh của học máy được phát triển dựa trên hệ thống nơron trong não bộ của con người. Do đó các mô hình học sâu có thể xử lý và hiểu sự tương tác phức tạp hơn của các dữ liệu đầu vào. Về cấu trúc, các mô hình học sâu sẽ gồm ba lớp chính, cụ thể là lớp đầu vào (*input layer*) để mô hình nhận dữ liệu, các lớp ẩn (*hidden layers*) để thực hiện các phép tính và biến đổi dữ liệu thông qua các nơron nhân tạo, và lớp đầu ra (*output layer*) để sản xuất và đưa ra kết quả dự đoán của mô hình (*Hình 3*).

2.3.1. Recurrent Neural Network

RNN là một loại mạng thần kinh nhân tạo được sử dụng trong học sâu để xử lý dữ liệu chuỗi hoặc dữ liệu có tính tuần tự [31]. RNN có khả năng mô hình hóa và hiểu các mối quan hệ phụ thuộc thời gian trong dữ liệu. Một đặc điểm quan trọng của RNN là khả năng lưu trữ thông tin từ các bước trước đó và sử dụng nó để ảnh hưởng đến các bước kế tiếp. Điều này giúp RNN xử lý các chuỗi dữ liệu có độ dài thay đổi và phụ thuộc vào ngữ cảnh. Cấu trúc cơ bản của một RNN bao gồm một chuỗi các "đơn vị" (*units*) nằm trong các lớp. Mỗi đơn vị RNN nhận đầu vào từ bước thời gian hiện tại và trạng thái ẩn từ bước thời gian trước đó, sau đó tính toán trạng thái ẩn mới. Quá trình này được lặp lại qua từng bước thời gian trong chuỗi dữ liệu.



Hình 2: Cấu trúc của phương pháp LGBMR



Hình 3: Cấu trúc của mô hình học sâu

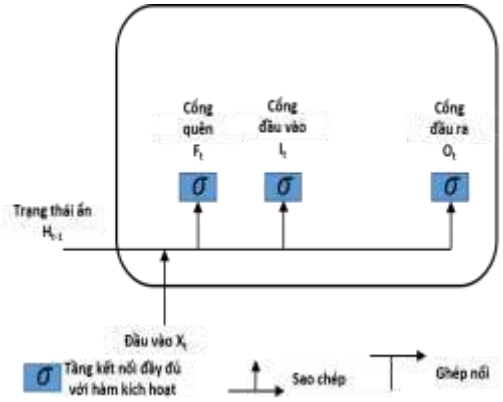
2.3.2. Long Short Term Memory

LSTM là một biến thể của RNN đang được sử dụng rộng rãi trong các bài toán về dự báo chuỗi thời gian [32]. Bằng cách sử dụng các cơ chế cổng (*gate mechanism*) để kiểm soát luồng thông tin trong quá trình tính toán trạng thái ẩn

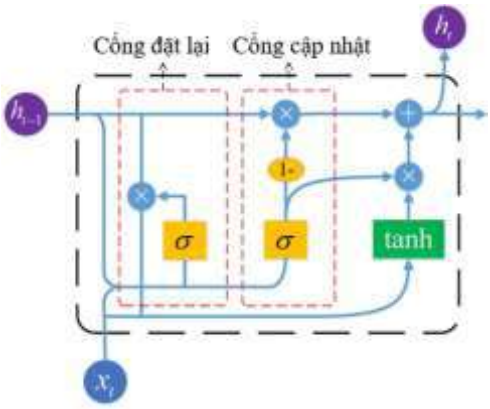
đã giúp LSTM có thể tăng khả năng mô hình học và hiểu các phụ thuộc dài hạn trong chuỗi dữ liệu hơn nhiều so với phương pháp RNN. Các cổng này gồm: Cổng quên (*Forget gate*) để quên thông tin không cần thiết từ trạng thái ẩn trước đó, cổng đầu vào (*Input gate*) để quyết định thông tin mới nào sẽ được lưu trữ vào trạng thái ẩn và cổng đầu ra (*Output gate*) để quyết định phần nào của trạng thái ẩn sẽ được chọn làm đầu ra (Hình 4). Các cổng này cho phép LSTM lưu trữ thông tin quan trọng từ các bước thời gian trước đó và điều chỉnh lưu trữ và truyền thông tin tùy theo ngữ cảnh. Điều này giúp LSTM xử lý dữ liệu có tính tuần tự và mô hình hóa các phụ thuộc dài hạn trong dữ liệu chuỗi.

2.3.3. Gated Recurrent Unit

GRU là một loại kiến trúc RNN được đề xuất nhằm giải quyết vấn đề mất mát thông tin xa (*long-term information loss*) trong các mạng RNN truyền thống. GRU có khả năng học các phụ thuộc dài hạn trong dữ liệu chuỗi thời gian mà không bị ảnh hưởng nhiều bởi vấn đề "*gradient vụn*". Cấu trúc của GRU bao gồm các cổng để kiểm soát thông lượng thông tin trong mạng (Hình 5). Các cổng này bao gồm: Cổng cập nhật (*Update gate*) để xác định mức độ cập nhật thông tin mới vào trạng thái ẩn của GRU. Cổng đặt lại (*Reset gate*) để quyết định xem thông tin trong quá khứ có cần được đặt lại hay không. GRU có thể lưu giữ thông tin quan trọng từ quá khứ và sử dụng nó để ảnh hưởng đến trạng thái hiện tại. Điều này giúp GRU vượt qua vấn đề mất mát thông tin xa và giữ được khả năng dự báo cho các chuỗi thời gian dài [33].



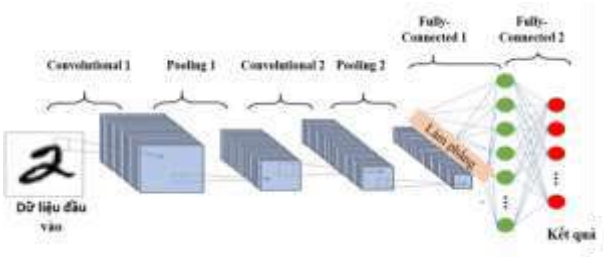
Hình 4: Cấu trúc của thuật toán LSTM



Hình 5: Cấu trúc của thuật toán GRU

2.3.4. Convolutional Neural Network

CNN là một kiến trúc mạng nơron nhân tạo được sử dụng chủ yếu trong xử lý dữ liệu không gian như hình ảnh và video. Tuy nhiên, mô hình CNN cho bài toán xây dựng bản đồ ngập lụt, dữ liệu được chuyển thành một ma trận đầu vào 2 chiều, trong đó trục thứ nhất đại diện cho thời gian và trục thứ hai đại diện cho các biến đặc trưng. Các lớp "*Convolutional*" trong CNN được sử dụng để trích xuất các đặc trưng cục bộ từ dữ liệu chuỗi thời gian. Các lớp "*Pooling*" giúp giảm kích thước không gian của đặc trưng trích xuất, trong khi lớp "*Activation*" tạo tính phi tuyến cho mô hình. Cuối cùng, các lớp "*Fully Connected*" được sử dụng để dự đoán giá trị tiếp theo dựa trên các đặc trưng đã được trích xuất (Hình 6).



Hình 6: Cấu trúc của thuật toán CNN

3. CÁC BƯỚC XÂY DỰNG MÔ HÌNH HỌC MÁY

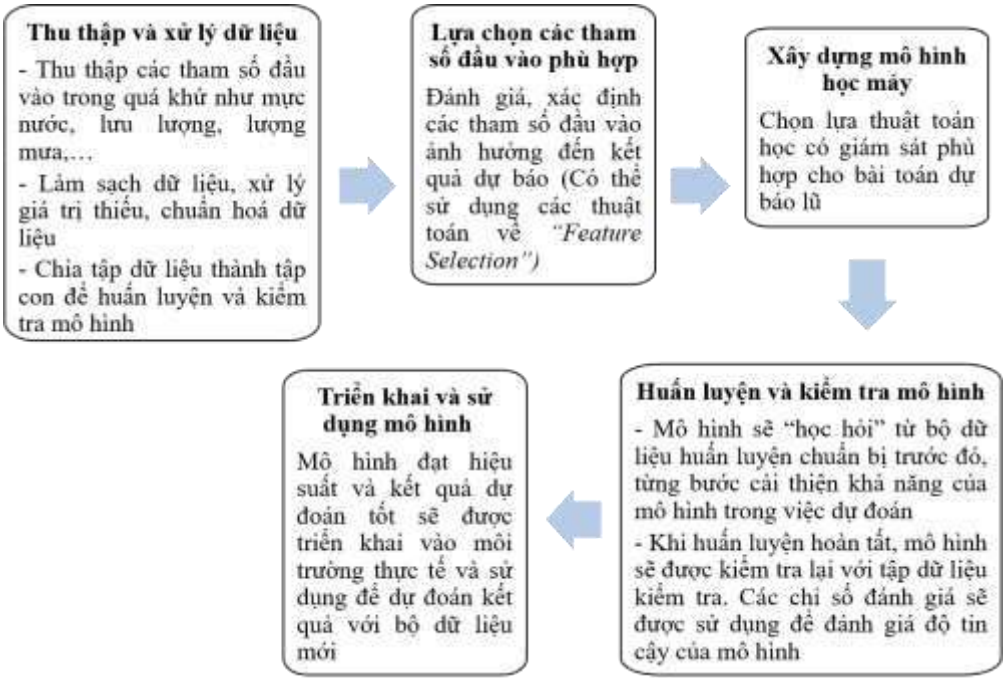
Trong các bài toán của học máy nói chung và trong các bài toán về dự báo lũ nói riêng, việc mô phỏng mô hình học máy để dự đoán ra kết quả mong muốn sẽ gồm 5 bước chính, đó là i) Thu thập và xử lý dữ liệu – các tham số đầu vào vẽ được thu thập và tổng hợp lại, cùng với đó các giá trị ngoại lai (*outliers*) sẽ được loại bỏ, xử lý; ii) Lựa chọn các tham số đầu vào phù hợp – trong bước này thực hiện chọn các tham số (biến) đầu vào có ảnh hưởng đến kết quả dự báo (*outcome*); iii) Lựa chọn thuật toán học máy – lựa chọn thuật toán phù hợp cho bài toán; iv) Huấn luyện mô hình – tập dữ liệu sẽ được chia ra 80% tập dữ liệu cho quá trình huấn luyện mô hình (hoặc 70%), tại đây mô hình sẽ tự học và tìm ra mối liên hệ giữa các tham số từ đó đưa ra giá trị dự đoán tương ứng. Sau đó tập dữ liệu còn lại sẽ được đưa vào mô hình để tiến hành kiểm tra mô hình dựa vào các chỉ số nhằm đánh giá độ chính xác của giá trị dự đoán với giá trị thực đo; và v) Triển khai và sử dụng mô hình –mô hình đạt được hiệu suất tốt sẽ được đưa vào môi trường thực tế và sử dụng để dự đoán kết quả trên tập dữ liệu mới (Hình 7).

Tính chính xác và độ tin cậy của các mô hình dự báo thường được đánh giá dựa trên các chỉ số đánh giá như hệ số xác định (R^2), sai số trung bình tuyệt đối (MAE), sai số trung bình phương (MSE), sai số trung bình phương căn bậc hai (RMSE), độ lệch chuẩn trung bình hoá (MBE). Ngoài ra, hai chỉ số mới là sai số đỉnh mực nước (PWE) và sai số thời gian mực nước đạt đỉnh cũng rất phù hợp trong việc đánh giá hiệu quả của mô hình dự đoán (Bảng 5).

Bảng 5: Các chỉ số thường được sử dụng để đánh giá độ chính xác của mô hình

Chỉ số đánh giá	Phương trình	Tóm tắt
Sai số đỉnh mực nước	$PWE = \hat{H}^p - H^p$	Đo sai số giữa mực nước đạt đỉnh giữa giá trị dự đoán và thực đo (nếu PWE càng gần 0 thì mô hình càng chính xác)
Sai số thời gian mực nước đạt đỉnh	$ETP = \hat{T}^p - T^p$	Đo sai số giữa thời gian mực nước dự đoán đạt đỉnh với thời gian mực nước thực đo đạt đỉnh (nếu ETP càng gần 0 thì mô hình càng chính xác)

Trong đó, H^p, $\hat{H}^p$, T và $\hat{T}^p$ lần lượt giá trị mực nước thực đo và dự đoán đạt đỉnh, thời gian mực nước thực đo và dự đoán đạt đỉnh.



Hình 7: Các bước chạy mô hình học máy trong bài toán dự báo lũ

4. KẾT LUẬN

Việc dự báo và cảnh báo lũ luôn là nhiệm vụ cấp bách cần phải thực hiện để hỗ trợ cho việc đưa ra quyết định của chính quyền địa phương. Tuy nhiên, những hạn chế của các mô hình dựa trên tính chất vật lý truyền thống (như cần nhiều dữ liệu, tham số đầu vào để mô phỏng chi tiết và sát nhất với thực tế, đòi hỏi nhiều thời gian mô phỏng và cũng như các chuyên gia trong lĩnh vực để có thể xây dựng và vận hành mô hình) khiến cho việc dự báo lũ theo thời gian thực chưa khả thi. Do đó, bài báo này đã cung cấp cho người đọc cái nhìn tổng quan về hướng tiếp cận mới và hiệu quả, đó là sử dụng các mô hình định hướng dữ liệu, cụ thể là các mô hình ứng dụng học máy, học sâu vào trong dự báo lũ. Để đánh giá hiệu quả của mô hình, bài báo cũng liệt kê các chỉ số thường được dùng để đánh giá hiệu quả và độ tin cậy của các mô hình. Nhìn chung, các mô hình sử dụng phương pháp hồi quy và học sâu đều đưa ra các dự đoán có độ chính xác cao và hiệu quả. Từ các kết quả của các nghiên cứu thì đây sẽ là một cách tiếp cận mới, hiệu quả trong dự đoán và giảm nhẹ thiệt hại gây ra do lũ lụt.

Tuy việc áp dụng các mô hình học máy, học sâu vào bài toán dự báo lũ mang nhiều tiềm năng và cơ hội, nhưng cũng đặt ra một số thách thức như đòi hỏi lượng lớn dữ liệu, chất lượng của dữ liệu, hiệu quả của mô hình, độ chính xác của mô hình, cụ thể như dưới đây:

- Dữ liệu đầu vào: Một trong những thách thức lớn đó là sự khan hiếm dữ liệu lũ chính xác và đầy đủ. Bên cạnh đó cũng cần phải phát triển thêm các phương pháp xử lý, nội suy hay sinh thêm dữ liệu bị thiếu để có thể đảm bảo tính chính xác và độ tin cậy của mô hình dự báo. Đặc biệt điểm hạn chế ở các mô hình học sâu là phải cần một tập dữ liệu lớn hơn rất nhiều so với hồi quy hay phân lớp để có thể đưa ra những dự báo chính xác;

- Độ phức tạp của mô hình: Do mối quan hệ giữa các tham số đầu vào là phi tuyến tính và rất phức tạp, do đó cần thêm nhiều nghiên cứu để phát triển các kiến trúc mạng nơron và các thuật toán tối ưu hoá phù hợp với bài toán dự báo lũ. Gần đây một số nghiên cứu đã kết hợp các thuật toán học máy lại đã đem lại một số

bước tiến mới trong việc dự báo lũ;

- Kỹ thuật phản hồi: Trong bài toán dự báo lũ, thông tin thời gian thực và sự phản hồi nhanh chóng là rất quan trọng. Do đó cần phát triển các hệ thống tự động và theo thời gian thực để có thể đưa ra các cảnh báo cho chính quyền và người dân một cách kịp thời nhất;

- Kết hợp dữ liệu đa nguồn: Kết hợp dữ liệu từ nhiều nguồn khác nhau như mô hình thủy văn, thuỷ lực, radar, và các trạm quan trắc mưa, mực nước tự động có thể cải thiện khả năng dự báo lũ. Nghiên cứu về kỹ thuật tích hợp dữ liệu đa nguồn và kết hợp các phương pháp học máy và học sâu sẽ tạo ra kết quả tốt hơn.

TÀI LIỆU THAM KHẢO

[1] G. Blöschl *et al.*, "Changing climate both increases and decreases European river floods," *Nature*, vol. 573, no. 7772, Art. no. 7772, Sep. 2019, doi: 10.1038/s41586-019-1495-6.

[2] S. N. Jonkman, "Global Perspectives on Loss of Human Life Caused by Floods," *Nat Hazards*, vol. 34, no. 2, pp. 151–175, Feb. 2005, doi: 10.1007/s11069-004-8891-3.

[3] L. Alfieri *et al.*, "A global network for operational flood risk reduction," *Environmental Science & Policy*, vol. 84, pp. 149–158, Jun. 2018, doi: 10.1016/j.envsci.2018.03.014.

[4] F. Ghobadi and D. Kang, "Application of Machine Learning in Water Resources Management: A Systematic Literature Review," *Water*, vol. 15, no. 4, Art. no. 4, Jan. 2023, doi: 10.3390/w15040620.

[5] F. Karim, M. A. Armin, D. Ahmedt-Aristizabal, L. Tychsen-Smith, and L. Petersson, "A Review of Hydrodynamic and Machine Learning Approaches for Flood Inundation Modeling," *Water*, vol. 15, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/w15030566.

[6] R. Bentivoglio, E. Isufi, S. N. Jonkman, and R. Taormina, "Deep learning methods for flood mapping: a review of existing applications and future research directions," *Hydrology and Earth System Sciences*, vol. 26, no. 16, pp. 4345–4378, Aug. 2022, doi: 10.5194/hess-26-4345-2022.

[7] W. J. Wee, N. B. Zaini, A. N. Ahmed, and A. El-Shafie, "A review of models for water level forecasting based on machine learning," *Earth Sci Inform*, vol. 14, no. 4, pp. 1707–1728, Dec. 2021, doi: 10.1007/s12145-021-00664-9.

[8] K. S. M. H. Ibrahim, Y. F. Huang, A. N. Ahmed, C. H. Koo, and A. El-Shafie, "A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting," *Alexandria Engineering Journal*, vol. 61, no. 1, pp. 279–303, Jan. 2022, doi: 10.1016/j.aej.2021.04.100.

[9] H. Mosaffa, M. Sadeghi, I. Mallakpour, M. Naghdyzadegan Jahromi, and H. R. Pourghasemi, "Chapter 43 - Application of machine learning algorithms in hydrology," in *Computers in Earth and Environmental Sciences*, H. R. Pourghasemi, Ed., Elsevier, 2022, pp. 585–591. doi: 10.1016/B978-0-323-89861-4.00027-0.

[10] Y. Guo, Y. Zhang, L. Zhang, and Z. Wang, "Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review," *WIREs Water*, vol. 8, no. 1, p. e1487, 2021, doi: 10.1002/wat2.1487.

[11] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," *Journal of Hydrology*, vol. 598, p. 126266, Jul. 2021, doi: 10.1016/j.jhydrol.2021.126266.

[12] S. R. Sannasi Chakravarthy, N. Bharanidharan, and H. Rajaguru, "A systematic review on machine learning algorithms used for forecasting lake-water level fluctuations," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 24, p. e7231, 2022, doi: 10.1002/cpe.7231.

[13] C. Crisci, B. Ghattas, and G. Perera, "A review of supervised machine learning algorithms and their applications to ecological data," *Ecological Modelling*, vol. 240, pp. 113–122, Aug. 2012, doi: 10.1016/j.ecolmodel.2012.03.001.

[14] X. Wu, X. Liu, and Y. Zhou, "Review of Unsupervised Learning Techniques," in *Proceedings of 2021 Chinese Intelligent Systems Conference*, Y. Jia, W. Zhang, Y. Fu, Z. Yu, and S. Zheng, Eds., in Lecture Notes in Electrical Engineering. Singapore: Springer, 2022, pp. 576–590. doi: 10.1007/978-981-16-6324-6_59.

[15] R. Nian, J. Liu, and B. Huang, "A review On reinforcement learning: Introduction and applications in industrial process control," *Computers & Chemical Engineering*, vol. 139, p. 106886, Aug. 2020, doi: 10.1016/j.compchemeng.2020.106886.

[16] A. N. Ahmed *et al.*, "Water level prediction using various machine learning algorithms: a case study of Durian Tunggal river, Malaysia," *Engineering Applications of Computational Fluid Mechanics*, vol. 16, no. 1, pp. 422–440, Dec. 2022, doi: 10.1080/19942060.2021.2019128.

[17] Quang Đ. N., Chiểu T. Q., Huệ Đ. T., and Ngân N. T. K., "Prediction of Water Level in Kien Giang river using Regression-Based Models," *1*, no. 80, Art. no. 80, Nov. 2022.

[18] T. Kusudo, A. Yamamoto, M. Kimura, and Y. Matsuno, "Development and Assessment of Water-Level Prediction Models for Small Reservoirs Using a Deep Learning Algorithm," *Water*, vol. 14, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/w14010055.

[19] W.-D. Guo, W.-B. Chen, S.-H. Yeh, C.-H. Chang, and H. Chen, "Prediction of River Stage Using Multistep-Ahead Machine Learning Techniques for a Tidal River of Taiwan," *Water*, vol. 13, no. 7, Art. no. 7, Jan. 2021, doi: 10.3390/w13070920.

[20] R. Tabbussum and A. Q. Dar, "Performance evaluation of artificial intelligence paradigms—artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction," *Environ Sci Pollut Res*, vol. 28, no. 20, pp. 25265–25282, May 2021, doi: 10.1007/s11356-021-12410-1.

[21] F. Ghobadi and D. Kang, "Multi-Step Ahead Probabilistic Forecasting of Daily Streamflow Using Bayesian Deep Learning: A Multiple Case Study," *Water*, vol. 14, no. 22, Art. no. 22, Jan. 2022, doi: 10.3390/w14223672.

[22] H. Apaydin, M. Taghi Sattari, K. Falsafian, and R. Prasad, "Artificial intelligence modelling integrated with Singular Spectral analysis and Seasonal-Trend decomposition

using Loess approaches for streamflow predictions," *Journal of Hydrology*, vol. 600, p. 126506, Sep. 2021, doi: 10.1016/j.jhydrol.2021.126506.

[23] S. Li, J. Yang, and A. Ansell, "Discharge prediction for rectangular sharp-crested weirs by machine learning techniques," *Flow Measurement and Instrumentation*, vol. 79, p. 101931, Jun. 2021, doi: 10.1016/j.flowmeasinst.2021.101931.

[24] S. Ghimire, Z. M. Yaseen, A. A. Farooque, R. C. Deo, J. Zhang, and X. Tao, "Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks," *Sci Rep*, vol. 11, no. 1, Art. no. 1, Sep. 2021, doi: 10.1038/s41598-021-96751-4.

[25] J. Hou *et al.*, "A deep learning technique based flood propagation experiment," *Journal of Flood Risk Management*, vol. 14, May 2021, doi: 10.1111/jfr3.12718.

[26] R. Madhuri, S. Sistla, and K. Srinivasa Raju, "Application of machine learning algorithms for flood susceptibility assessment and risk management," *Journal of Water and Climate Change*, vol. 12, no. 6, pp. 2608–2623, Apr. 2021, doi: 10.2166/wcc.2021.051.

[27] L. Hashemi-Beni and A. A. Gebrehiwot, "Flood Extent Mapping: An Integrated Method Using Deep Learning and Region Growing Using UAV Optical Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2127–2135, 2021, doi: 10.1109/JSTARS.2021.3051873.

[28] G. Ke *et al.*, "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 3149–3157.

[29] K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Using support vector machines for time series prediction," 1999, pp. 243–253. doi: 10.1515/9783110915990.1.

[30] U. Thissen, R. van Brakel, A. P. de Weijer, W. J. Melssen, and L. M. C. Buydens, "Using support vector machines for time series prediction," *Chemometrics and Intelligent Laboratory Systems*, vol. 69, no. 1, pp. 35–49, Nov. 2003, doi: 10.1016/S0169-7439(03)00111-4.

[31] T. Robinson and F. Fallside, *The utility driven dynamic error propagation network*. 1987.

[32] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, Jan. 2021, doi: 10.1016/j.procir.2021.03.088.

[33] W. Zheng and G. Chen, "An Accurate GRU-Based Power Time-Series Prediction Approach With Selective State Updating and Stochastic Optimization," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13902–13914, Dec. 2022, doi: 10.1109/TCYB.2021.3121312.