

Course Name: Watershed Hydrology

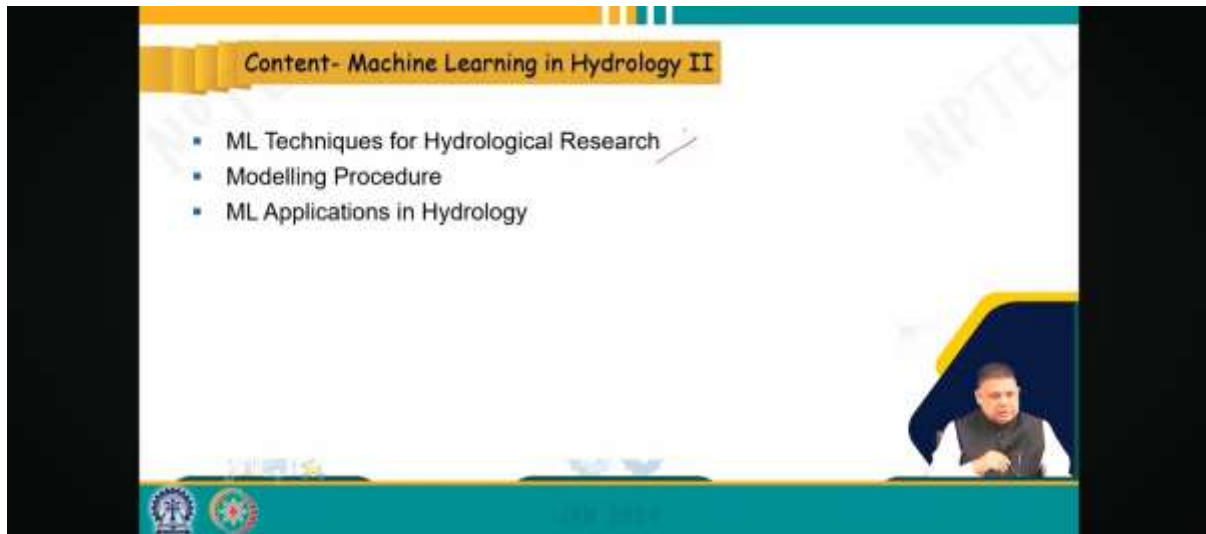
Professor Name: Prof. Rajendra Singh

Department Name: Agricultural and Food Engineering

Institute Name: Indian Institute of Technology Kharagpur

Week: 08

Lecture 40: Machine Learning in Hydrology II



Hello friends. Welcome back to this online certification course on Watershed Hydrology. I am Rajendra Singh, a professor in the Department of Agriculture and Food Engineering at the Indian Institute of Technology Kharagpur. We are in module 8 and this is lecture 5 which is the last lecture of this module. We will be continuing with the discussion on machine learning hydrology. This is part 2, as we started this in the previous lecture. In this lecture, we will talk about machine learning techniques used in hydrological research.

We will discuss the modelling procedure and examine some machine learning applications done in hydrology through research papers. Now, regarding machine learning algorithms, as we discussed in the previous lecture, we talked about artificial intelligence, its role in our daily life, as well as machine learning and deep learning. So, machine learning techniques or algorithms are increasingly being applied in hydrological research to model complex relationships, improve predictions, and enhance understanding of hydrological processes. We have talked about hydrological models; we saw that we have conceptual models, physically based models, but we also saw that we have empirical models. Machine learning is a type of black box or empirical model where physical processes are not considered, but they have certain advantages, such as easily modelling complex relationships compared to other kinds of models, which is why they are quite popular.

ML Techniques for Hydrological Research

ML Algorithms

- Machine learning (ML) techniques are increasingly being applied in hydrological research to model complex relationships, improve predictions, and enhance understanding of hydrological processes.
- Here are some ML techniques commonly used in hydrological research:
 - ✓ Regression methods
 - ✓ Neural networks
 - ✓ Decision trees
 - ✓ Support vector machines
 - ✓ Nearest neighbours
- All these are Supervised Learning Methods

Some machine learning techniques commonly used in hydrological research are listed here: regression methods, neural networks, decision trees, support vector machines and nearest neighbours. Incidentally, all these are supervised learning methods. As we discussed previously when we classified machine learning algorithms, there are supervised learning techniques, unsupervised learning, and reinforcement learning. So, there are three different types, but we will focus more on supervised learning and discuss the methods that fall under supervised learning methods.

ML Techniques for Hydrological Research

Regression Methods

Linear Regression

Logistic Regression

- Establish a relationship between dependent and independent variables based on the concept of Maximum Likelihood Estimation
- Predict the dependent variable

Handwritten notes: Rainfall - Runoff (P) vs R =

Starting with the regression method which is basically establishing a relationship between dependent and independent variables based on the concept of maximum likelihood estimation. So, obviously, it is very common in what we do best-fit relationships, and I think we have been doing this right from class 6 onwards, maybe, and obviously, we always have a dependent and independent variable, and then try to find out their relationship by plotting their data and then getting the best possible relationship using the maximum likelihood or least square method we use. And then the relationship could be linear or logistic or power or exponential or whatever, and once we have fitted the relationship, then obviously, for any independent variable, we can get the dependent variable. For example, we also discuss the example of rainfall versus runoff, which is the principal goal of hydrological science anyway, and then we say if we call this H_p if we call this R , then obviously, we can fit a relationship between R and P , which we can say

$$A_p + C \text{ or } A_p R = AP^n$$

So, the relationship could be of any type.

So, once we have a long length of data, then we can fit this kind of relationship, and then of course, R is the dependent and P is the independent variable here. So, next time when we have a value of P, then we can predict the value of R using the relationship that we have fitted. So, that is a regression method, the simplest of the machine learning techniques. Then of course, we have the most popular one under neural networks and out of neural networks also the most popular one is artificial neural network ANN, which consists of artificial neurons that are a copy of human brain neurons. We discussed that we have deep learning, and I also mentioned the term ANN was being used right from the mid-90s or so, but in the same times, we have been knowing it more popularly under deep learning algorithms.

ML Techniques for Hydrological Research

Neural Networks

Artificial Neural Network (ANN)

- Consists of artificial neurons, which are a copy of human brain neurons.
- Artificial neurons connect in a neural network by weights to perform the tasks.
- ✓ A Multilayer Perceptron (MLP) is a class of feed-forward ANN in the form of deep learning.
- ✓ The term MLP refers to networks composed of multiple layers of perceptrons (neurons) with back-propagation for training.

Input data is fed in a forward direction (input to output layer) – no feedback

Feed Forward

Input layer Hidden layer Output layer

Weights

Back-Propagation

Used during training to minimise the error between the predicted output and the actual output

So, where are the neurons that mimic the human neurons? That means, their cognitive skills and artificial neurons connect in a neural network by weights to perform the task, and typically, a multilayer perceptron (MLP) is a class of feedforward ANN in the form of deep learning. So, multilayer perceptron—multi-layer because we use different layers. So, we have an input layer which has all the inputs provided into the system we are modelling, then we have an output layer. Now the output could be a single variable or multiple variables, similarly, the input could be a single variable or multiple variables. So, these are the layers where the number of neurons is put and these neurons basically represent the output or input numbers.

If you have 5 inputs, then you will have 5 neurons in the input layer; if you have one output, then you will only have one output neuron in the output layer, but in between, there are a number of hidden layers, and each of these hidden layers has a number of neurons in them, and that is basically where we fit the relationship, and that is basically where we decide the weights, the relationship mathematically function where the weights are assigned to transfer the information from one layer to the next layer, that is from input to the output layer. So, multilayer perceptron has a large number of layers, and the term MLP refers to networks composed of multiple layers of perceptions. Just now we discussed neurons with BAP propagation for learning. So, there are two important terms being used here: that is a multilayer perceptron is a feedforward ANN. So, which is we know that is feedforward and that simply means that input data is fed in a forward direction, that is input to output layer, and there is no feedback. That

means, information flows only in one direction, that is input you provide and then obviously, you adjust the weights and finally, you reach the output layer where we basically match the observed and simulated output, maybe runoff, maybe stream flow, whatever and that depends on the system you are modelling.

The other term which comes here is BAP, backpropagation, which is used in training. As mentioned, backpropagation is used in training to minimize the error between the predicted output and actual output. So, you remember, say, supervised learning. So, there we provide the inputs as well as the measured output, that is, which we wish to match basically through simulation. So, that is basically in order for this error to be minimized, that is, predicted output you have and simulated output you have, or the actual output or the observed output you have. So, basically, you match them and you try to minimize the error, and that way in backpropagation, you adjust the weights. So, adjusting the weight while adjusting the weights, you can go into backward direction, but as for information flow is concerned, that is only in the forward direction. That is why it is a feedforward ANN. That is what we mentioned: feedforward MLP is a feedforward ANN which uses backpropagation for training. So, that is artificial neural network, and an important thing is there are no feedbacks in the system.

The image shows a presentation slide with the following content:

- ML Techniques for Hydrological Research**
- Neural Networks**
- Recurrent Neural Network (RNN)**
- In ANNs, inputs and outputs are typically independent, but for tasks like predicting the next word in a sentence, retaining information about previous words becomes essential.
- RNN is a Neural Network where the output from the previous step is fed as input to the current step through recurrent connections on hidden states.

The diagram illustrates an RNN architecture with three layers: an input layer (4 nodes), a hidden layer (3 nodes), and an output layer (3 nodes). A 'Recurrent connection' box highlights a feedback loop from the hidden layer back to itself. The presenter is visible in the bottom right corner of the slide.

The next type of neural network is RNN or recurrent neural network. In ANNs, input and outputs are typically independent, but for tasks like predicting the next word in a sentence, retaining information about previous words becomes essential. So, sometimes, it may be as if you are modelling a type of system, it may be necessary to know what happened in the previous step.

In giving an example of a sentence, in reservoir simulation, quite often we use 'that' because in reservoir operation, you supply certain water, and then you also would like to know how much water is available still. So, that information goes into carrying out the water balance. That is the recurrence neural network application. It's a neural network where the output from the previous system is fed as input to the current step through recurrent connections on the hidden state. Obviously, you have the input layer, you have the output layer, and you have the hidden layer, but at the hidden layer, you have recurrent connection where there is a possibility of giving feedback; feedback information is provided. You remember in ANN, we said that it is a feed forward, but no feedback, but here there is a possibility of providing the feedback into the

system. So, it learns at a faster rate. That is another way or another type of neural network, that is recurrent neural network.

ML Techniques for Hydrological Research

Neural Networks

Long Short Term Memory (LSTM)

- In RNN, weight updates are based on the error derivative, but sometimes gradients become too small, hindering weight changes.
- LSTM, a class of RNN, was developed by Hochreiter and Schmidhuber (1997) to deal with the above vanishing gradient problem.
- LSTM has self-connected units that preserve and retrieve values (forward pass) or gradients (backward pass) at the required time step.

Connection to control cell states (change with each time step, i.e., over short term), and maintain information (say, learned weights) in memory for a long term

Then we have another type of neural network which is referred to as long short-term memory (LSTM), which is getting a lot of popularity nowadays. So, what happens is that in RNN, weight updates are based on the error derivative; that is you want to minimize error.

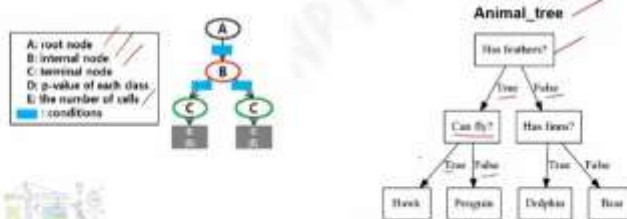
So, error derivatives are used, but sometimes the gradient becomes too small, hindering weight change. Because if you want to minimize the error, as the error reduces, the gradient value becomes too small and then it might be very difficult to change the weight basically. So, in that case, LSTM has been developed, which is a class of RNN, which is basically developed to take care of the vanishing gradient problem. It was developed by Hirsch Reiter and Smith-Zebo in 1997 to deal with the vanishing gradient problem which we might face in RNN.

So, it has self-connected units that preserve and retrieve values in the forward pass or gradients in the backward pass as required at each time step. It has the input layer, the output layer, and the feedback mechanism. Within the hidden layer, it also has a provision of self-connected units which preserve and retrieve values and gradients. Basically, this hidden layer with self-connected unit's controls cell states, which change with each time step. These changes are quick but are kept in memory for a long time, hence the name long short-term memory (LSTM). As mentioned, LSTM is gaining quite a popularity in neural network or machine learning applications.

ML Techniques for Hydrological Research

Decision tree

- > Decision tree is the **most powerful and popular** tool for classification and regression.
- > A Decision tree has a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents **an outcome of the test**, and each leaf node (terminal node) holds a class label.



Then, the next class of machine learning technique is decision trees which is probably the most popular tool for classification and regression problems. It has a flow chart-like tree structure, where each internal node denotes a test or an attribute, each branch represents an outcome of the test, and each leaf node or terminal node holds a class level. It has a root node, internal nodes, and terminal nodes, along with certain conditions and a number of cells. Let's take an example, the simplest example being an animal tree.

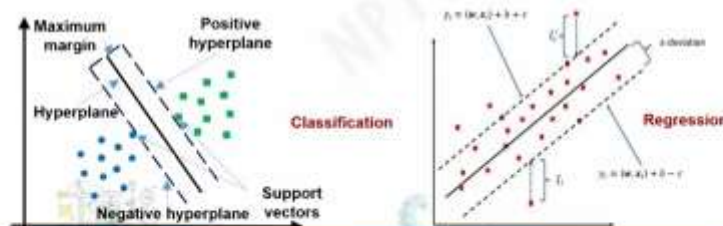
Obviously, you may raise the question: does it have feathers? There are two possibilities: true or false. If it is true, the next question may be: can it fly? Again, two answers could be there: true or false. If it is true, you label it as a hawk; if it cannot fly, then you may say it is a penguin.

So, and on the other hand, the same question, when has it got feathers? False. Then the next question could be, has it got fins? If it is true, then it is a dolphin. If it is false, then it could be a bear. So, this is how that different nodes in a tree are related, and they keep the information flowing. And as mentioned, a test on any attribute is provided at each internal node, and then the branch node represents the outcome, like here true or false. Then each leaf node, then again branch and finally, you have the terminal node where you will provide the label. So, this is a decision tree. Example techniques under this category are random forest, fuzzy decision tree and gradient boosting. Random forest is of course, popularly used nowadays.

ML Techniques for Hydrological Research

Support Vector Machine (SVM)

- SVM is a novel type of learning machine.
- Goal of the SVM algorithm is to create the **best line or decision boundary** that can segregate **n-dimensional space into classes**. This best decision boundary is called a **hyperplane**.
- SVM chooses the **extreme points/vectors (support vectors)** that help in creating the hyperplane, which can be used for classification or regression.



Then in the next category, we have support vector machine (SVM), which is a novel type of learning machine. The goal of SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes. This best decision boundary is called a hyperplane. SVM chooses the extreme points or vectors, which are referred to as support vectors, that help in creating the hyperplane, which can be used for classification or regression. So, it can be used both for classification as well as regression analysis. So, obviously as you can see, you have two types of articles or variables, and of course, based on the distance, we provide the support vector, that is, the positive hyperplane or the negative hyperplane you can see here. And obviously, in between is the hyperplane which represents the center of these two positive or negative hyperplanes.

Similarly, on the regression side, we only know that in the least squares, we also use error margins. So, basically, here these hyperplanes act as 5 percent or 10 percent error lines, and then you have the best-fit line. So, it is the best-fit line, and then you have the hyperplane or support vectors basically. So, obviously, if it is a straight line, all three could have a straight-line relationship. So, this SVM is used for classification and regression by using the positive and negative hyperplanes, and obviously, the hyperplane to distinguish between the type of articles, if it is classification or to have a relationship if it is regression.

ML Techniques for Hydrological Research

Nearest Neighbours

k-Nearest Neighbours (KNN):

- Simplest Machine Learning algorithms based on Supervised Learning technique, mostly used for the Classification problems.
- KNN algorithm at the training phase just stores the dataset, and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The diagram illustrates the KNN process. On the left, 'Before KNN', a 2D plot shows two clusters: 'Category A' (green dots) and 'Category B' (orange dots). A 'New data point' (blue dot) is located between them. An arrow labeled 'KNN' points to the right, 'After KNN', where the 'New data point' is now assigned to 'Category B' because it is closer to the 'Category B' cluster.

The next class is nearest neighbours, and the k nearest neighbour KNN is one example. It is the simplest machine learning based on supervised learning technique mostly used for classification problems, and the KNN algorithm at the training phase just stores the data. When it gets new data, then it classifies that data into a category that is much similar to the new data. So, obviously, before KNN, you have data class grouped into two different categories, category A and category B, and suppose you provide new data, then it tries to match the characteristics and then decides that okay, this data belongs to either category A or category B through the nearest neighbourhood technique. So, if you compare various kinds of techniques here, then you already talked about different algorithms like regression method, nearest neighbour, neural networks, support vector machine, decision tree, and of course, there are ensembles where one or more could be combined together. So, obviously, their prediction speed, training speed, memory usage, and running required tuning will differ and of course, that is why they are used for a particular purpose.

ML Techniques for Hydrological Research

| Algorithm | Prediction speed | Training speed | Memory usage | Required tuning | Assessment |
|-------------------------|------------------|----------------|-----------------|-----------------|---|
| Regression method | Fast | Fast | Small | Minimum | Small problems with linear decision boundaries |
| Nearest neighbor | Moderate | Minimal | Medium | Minimum | Lower accuracy and easy to use |
| Neural networks | Moderate | Slow | Medium to large | Maximum | Popular for classification and forecasting |
| Support vector machines | Slow | Slow | Medium | Some | Good for binary problems, high dimensional data |
| Decision trees | Fast | Fast | Small | Some | Generalist, but overfitting |
| Ensembles | Moderate | Slow | Varies | Some | Higher accuracy, small to medium sized datasets |

For example, if we talk about the regression method, the prediction speed is fast, training speed is fast, memory usage is small, and tuning required is minimum. That is why it could be used for small problems with linear decision boundaries. So, fitting lines, as we already discussed. On the other hand, if we talk about neural networks, then the prediction speed is moderate, training speed is slow, memory usage is medium to large, and required tuning is maximum because we have the input layer, the output layer, they can be more than one hidden layer and the number of neurons in the hidden layers. That is why the tuning, the weight assignment, will take a lot of time, and also the memory will, of course, depend on how many neurons are there, how many layers are being used, how many neurons are there in each, and that is why the training speed is also slow, but then it is popular for classification and forecasting. So, similarly, each of these algorithms has its own applications and, of course, their advantage in this one is compared if you compare on the scales of prediction speed, training speed, memory usage or required training. So, this is, in a nutshell, about various algorithms which are used in machine learning.

ML Techniques for Hydrological Research

Explainable Machine Learning

- To explain or interpret the black-box model operation, a subdomain of ML, **Explainable Machine Learning (eXML)**, has obtained significant attention recently. The eXML can be broadly classified as **local** and **global** explanations.
 - Global explanations** describe the whole model behaviour by explaining unrealistic relations between the prediction and input.
 - Local explanations** describe the reasons for a particular prediction and rank the variables by their impact on model outputs. **Shapley additive explanations (SHAP)** and **Local Interpretable model-agnostic explanations (LIME)** are the popular local explanation methods used in ML applied studies.

Today

```

graph LR
    A[Training Data] --> B[Machine Learning Phases]
    B --> C[Learned Function]
    C --> D[Decision]
    D --> E[Test]
    E --> F[Model Performance]
          
```

eXML

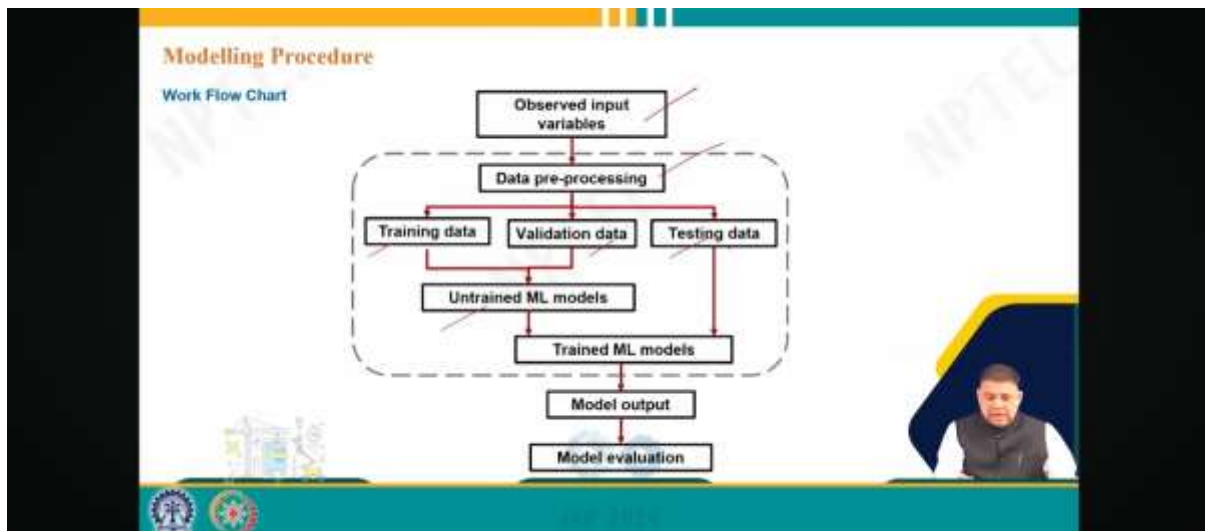
```

graph LR
    A[Training Data] --> B[Machine Learning Phases]
    B --> C[Explainable Model]
    C --> D[Explainable Output]
    D --> E[Test]
    E --> F[Model Performance]
          
```

Now, another type of machine learning is coming into a big way, that is explainable machine learning. Basically, it is used to explain or interpret the black box model operation, that is a subdomain of ML. Explainable machine learning, XML, which is in short form, has obtained significant attention recently, and these can be broadly classified as local and global

explanations. Now, we already discussed in hydrological modelling also that we have black box models where the physical structure is not taken into account, and that is why there is a lot of uncertainty. You do not know what is happening, we cannot explain anything, and that is the same is true for machine learning also. So, in order to take care of this drawback, this explainable machine learning is coming, is being developed in a big way, and as it can be brought classified as local and global.

Global explanations describe the whole model behaviour by explaining unrealistic relationship between prediction and output where input where is local explanations describe the regions for a particular prediction and rank the variables by their impact on model outputs and the in the local explanations SHAPLAE that is SHAPLAE additive explanation which is briefly called SHAP and local interpretable model agnostic explanation which is short form is LIME are the popular local explanation methods used in ML ah applied studies especially SHAP is being used in a big way nowadays. So, the advantage is that when nowadays when we are using, we are giving training data we have machine learning process we have learned function and then we have come up with some recommendation. So, basically certain questions are remained unanswered that why did you do that why not something else why do you succeed ah when do you fail when can I trust you how do I correct an error. So, there are so many questions ah related to explainability of the results that remain unanswered wherein if we provide or use EXML along with your machine learning technique then we can answer many of these question that I understand why because ah this SHAP or LIME, they provide you explanation then I understand why not I know when you succeed when you fail when to trust and when you add. So, all those questions can be answered in a better way if you go for explainable machine learning and that is why it is getting popular.



Now, we go to modelling procedure and ah modelling procedure basically we have observed input variables we then we have data pre-processing then we divide data into training data validation data and testing data. The training data and validation data is fed into untrained machine learning model ah where you really tune the model parameters. So, that you get the trained machine learning model and then you test them trained machine learning model using the testing data and then finally, you will get the model output and then the results will be evaluated. So, this is the nutshell the modelling procedure ah very similar to calibration validation we do only thing is that here training and validation is taken together and testing is

similar to validation which we discussed earlier. So, otherwise most of the process remain the same.

Now, coming to the modelling procedure, variables to be considered for input can be either univariate or multivariate approach. For example, when we say "unit variable," that is using a single time-independent variable for prediction. So, using only stream flow for stream flow prediction. Suppose we are predicting stream flow, then we use only the previous stream flow values to predict the stream flow. Whereas, in a multivariate approach, we will be using multiple time-independent variables for prediction.

Modelling Procedure

Variables to be considered for input

- **Uni-variate approach** "Using a single time-dependent variable for prediction"
Ex: Using only streamflow for streamflow prediction
- **Multi-variate approach** "Using a multiple time-dependent variables for prediction"
Ex: Using precipitation, temperature, Humidity & streamflow for streamflow prediction

✓ It usually depends on available data and the degree of complexity.

Data Pre-processing

- **Handling missing data:** Missing data can be nullified by deleting the particular row or placing the mean value in place of missing value.
- **Encoding Categorical data:** As machine learning model completely works on mathematics and numbers, it is necessary to encode the categorical variables into numbers.
- **Feature scaling:** It is a technique to standardise or normalise the independent variables of the dataset in a specific range, so that no variable dominates the other variable.

So, obviously, for the same stream flow prediction, we may use precipitation, temperature, humidity, and of course, stream flow. Multiple inputs could be used for making the prediction, and that is why I said that the number of layers in the input layer as well as the output layer could differ depending upon what kind of approach you are adopting. And of course, it depends on the data availability and the degree of complexity you want in yours because more input layer nodes mean, your neurons, that means, more weight assignment, more time to tune and train the model. Then of course, we have to do some kind of data processing as we said and many times we may have to handle missing data; we already discussed missing data earlier, the same is true here also. So, missing data can be nullified by deleting the particular row or placing the mean value in the place of the missing value.

So, either you can completely ignore the data or you can use the mean value in place of the missing value. Then obviously, we need to encode categorical data. So, a machine learning model completely works on mathematics, that is numbers. So, it is necessary to encode the categorical variables into numbers. So, many times if you are trying to model, say, where the subjects are there, women, men and women both are there.

Modelling Procedure

Splitting the Dataset into the Training, Validation and Test sets

- Data splitting is crucial for enabling accurate performance evaluation and preventing overfitting to ensure robust generalisation to new data.

It is essential for an unbiased evaluation of prediction performance.

"We always train the model with the training set, fine-tune parameters of model using the validation set, and evaluate performance on unseen data with the test set"

The diagram illustrates the process of splitting a dataset. A horizontal bar labeled 'Dataset' is divided into three segments: 'Training set' (shaded grey), 'Validation set' (shaded blue), and 'Test set' (shaded light blue). Each segment has a diagonal line across it, suggesting a cut or division. Above the bar, a double-headed arrow spans the entire length, labeled 'Dataset'. Below the bar, there are some decorative icons and a small inset image of a person in the bottom right corner.

So, their sex or gender could be one definition. So, obviously, that is not in number. So, that has to be encoded into numbers basically. Then you have feature scaling. So, standardize or normalize the independent variables of the dataset in a specific range.

So, that no variable dominates the other variable. So, basically, we normalize the data. So, that data becomes independent of the units or the large variations. Then of course, we have to split the dataset into training, validation, and test sets, and data splitting is crucial for enabling accurate performance evaluation and preventing overfitting to ensure robust generalization of new data and of course, it is essential for unbiased evaluation. So, basically we convert we break the data into 3 groups: training set, validation set, and test set.

So, we always train the model with the training set, fine-tune parameters of the model using the validation set, and evaluate performance on unseen data with the test set. So, all 3 have different purposes and of course, for training we require a longer length of record and validation and test set a little smaller. So, maybe if you have data length, 70 percent data could be used for training, 15 percent for validation, and the rest 15 percent for testing the model. Then coming to training and validation of the model, so obviously, during the training ML algorithm launches the training dataset which helps it acquire necessary knowledge and patterns.

In validation, we adjust the weights of the model parameters until optimal performance is achieved, and this helps optimize the model's performance and selecting the current parameter that will be modified to inflate ML models key to attaining accurate correlation. So, basic, very similar to the calibration process which we do in say conceptual or physically based models, and a set of parameters that are selected based on their influence on the model architecture are called hyperparameters, and the process of identifying the hyperparameters by tuning the model is called parameter tuning. And typical hyperparameters, let us say if we talk about LSTM model, then there is a learning rate. In fact, any of most of the artificial neural networks will have similar hyperparameters, maybe the algorithm could differ, then there will be batch size the data will be broken into different pages we epoch dropout rate that how much variation you want from one run to another, and there are optimization algorithms like SGD, RMS, PROP, ADAM which we use in the case of LSTM. So, SGD stands for Stochastic Gradient Descent, Random Mean Squared Propagation for RMS, PROP, and ADAM is for Adaptive Learning Rate algorithm.

Modelling Procedure


Training and Validation of model

- During the training, the ML algorithm learns from the training dataset, which helps it acquire necessary knowledge and patterns.
- In validation, we will adjust weights of model's parameters until optimal performance is achieved during training. This helps optimise the model's performance and ensures it generalises well to new data.
- ✓ Selecting the correct parameter that will be modified to influence the ML model is key to attaining accurate correlation.
- ✓ The set of parameters that are selected based on their influence on the model architecture are called hyperparameters.
- ✓ The process of identifying the hyperparameters by tuning the model is called parameter tuning.

Ex- Hyperparameter of LSTM Model

| Hyperparameter | Range | Optimum Value |
|------------------------|-------------------------|---------------|
| Learning rate | [0.001, 0.01, 0.1, 0.2] | 0.001 |
| Batch size | [8, 12, 32] | 32 |
| Epochs | [10, 20, 30] | 30 |
| Dropout rate | [0.2, 0.3, 0.4] | 0.3 |
| Optimization algorithm | [SGD, RMSprop, Adam] | Adam |

(Stochastic Gradient Descent; Random mean squared propagation; Adaptive learning rate algorithm)




So, these parameters we tune during the validation run and then finally, we are testing the model with unseen data like we did in the validation case and of course, then we finally use some evaluation technique already we discussed a lot of evaluation techniques and methods that in using the statistical coefficients or visualization that means, we here also we can use R square in CRM, SEP bias and of course, the visualization like earlier we terms time series plot or scatter plot we can use. And then of course, forecasting or predicting the variable comes once we have the model then can forecaster predict the variable.

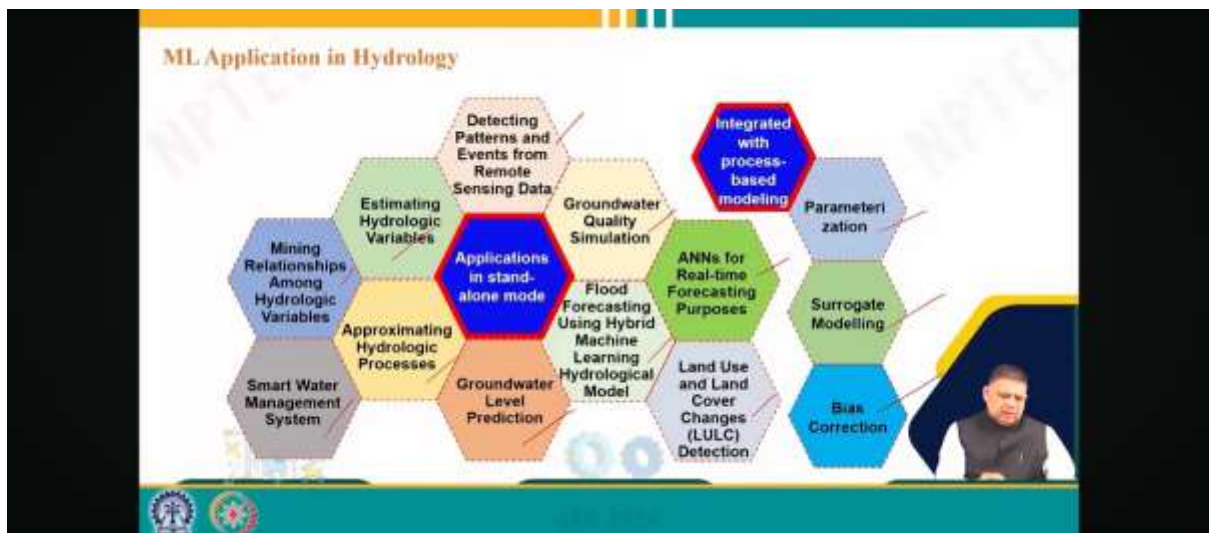
Modelling Procedure

Testing the Model

- Testing the ML model after validation helps ensure that the model's performance is assessed on completely unseen data, providing a more robust evaluation of its generalisation capabilities..
- Identifying the measures of success based on what the model is intended to achieve is critical for justifying correlation.
- The model performance can be evaluated by using statistical coefficients (Ex: R^2 , NSE, RMSE, PBIAS, etc.) and visualisation.
- ✓ Forecasting or predicting the variables for a given time can be done using this trained and tested model.



Now coming to ML applications in hydrology, we find nowadays application of ML in almost all domains like detecting patterns and events from remote sensing data, groundwater quality simulation, flood forecasting using hybrid machine learning, hydrological model, groundwater level prediction, approximately hydrological processes, estimating hydrological variables. Then mining relationships among hydrological variables or smart water management systems have also been developed using machine learning. Similarly, we have NNs for real-time forecasting purposes, land use, land cover change analysis all these and of course, ML models are being used integrated with process-based modelling for parameterization for surrogate modelling or for bias correction.



So, I mean almost all studies in hydrology and machine learning techniques are being used nowadays. Now, coming to some specific applications in hydrology, Nearing and others in 2020 provided a detailed commentary on the role of application of ML in hydrological science. And the study offers some potential perspectives and preliminary examples that show the role of hydrological sciences in the age of machine learning. For example, in prediction in ungauged basins, hypothesis testing, uncertainty analysis, forecasting and prediction. This is the paper we are talking about.

This paper is published in Water Resources Research. The title is "What Role Does Hydrological Science Play in the Age of Machine Learning?" So, anybody who is trying to initiate into this area should definitely read this paper. Similarly, ensemble technique is being used in a big way. So, using multimodal ensembles for various fields. This one talks about the climate models to reproduce observed monthly rainfall and temperature using machine learning methods.

ML Application in Hydrology

Generation of Ensembles

- ML techniques are also getting popularity in Ensemble Modelling, i.e., in the generation of multi-model ensembles.
- Wang et al. (2018) used four techniques, i.e., Random Forest (RF), Support Vector Machine (SVM), Bayesian Model Averaging (BMA), and the Arithmetic Ensemble mean (AE) in reproducing observed monthly rainfall and temperature.

RESEARCH ARTICLE

Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia

Bin Wang¹, Libing Zheng¹, Di Li Liu^{1,2}, Fan Li¹, Anthony Clark³, Qing Yu^{1,2*}

So, a lot of techniques are being used like here in this paper Wang and others have used four techniques: random forest, SVM, Bayesian model averaging, and arithmetic ensemble mean for reducing the observed monthly rainfall and temperature. Then ML has also been applied for the prediction of stream flow and here the paper is on near real-time forecasting of reservoir inflow using explainable machine learning and short-term weather forecast. And that is for stream flow and then for the prediction of floods, also, a lot of researchers have used it, and this is a review paper on flood prediction using machine learning models which will give you a detailed analysis of all kinds of results that have been done. Similarly, machine learning has also been applied for the prediction of droughts. This is an example paper and it has also been used for land use and land cover prediction.

ML Application in Hydrology

Prediction of Streamflow

- ML methods require minimum data and can quickly model the linear and nonlinear relationships between a streamflow and its parameters and yield good performance.

ORIGINAL RESEARCH | Published: 21 June 2022

Near-real-time forecasting of reservoir inflows using explainable machine learning and short-term weather forecasts

Ashim Sabu^{1*}, Ashish Mishra¹, Subhasanku Mishra¹, Maheshwar K. Sanku¹, Dinesh Kumar¹

Frontiers in Environmental Science and Earth Science, 11:794549 | DOI: 10.3389/fenv.2022.794549

Prediction of Floods

- ML models are used for prediction with greater accuracy in flood prediction.

Flood Prediction Using Machine Learning Models: Literature Review

By: Ashim Sabu^{1*}, Ashish Mishra¹, Subhasanku Mishra¹, Maheshwar K. Sanku¹, Dinesh Kumar¹

¹ Department of Computer Science (CS), Netaji Subhas University of Science and Technology (NSUST), Kolkata, West Bengal, India

² Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong, China

* Author to whom correspondence should be addressed

So, this is a paper in the international realm of remote sensing. So, obviously, this is just to show that machine learning is being used in a big way in hydrology. So, all those who are looking for research in the future, I mean probably this should be a learning platform one should use. So, with this, we come to the closer of this lecture where we discuss the various machine learning algorithms that are used in hydrology. We also talked about the modelling process which is used in machine learning and we also try to show some of the applications which people are doing in machine learning in hydrology. So, please give your feedback and raise your questions or doubts, we shall be able to answer, happy to answer them.

ML Application in Hydrology

Prediction of Droughts

- ML methods are gradually replacing conventional approaches in drought prediction, revolutionising the accuracy and reliability of forecasts.

Applying machine learning for drought prediction using data from a large ensemble of climate simulations

Elizaveta Feltschik^{1,2,3} and Ralf Ludwig²

¹Center for Digital Technology and Management, Munich, Germany

²Department of Geography, Ludwig-Maximilians-University of Munich, Munich, Germany

³Technical University of Munich, Munich, Germany

Prediction of LULC

- ML algorithms are also getting popularity in Land use and Land cover (LULC) predictions.



International Journal of Remote Sensing

<http://www.tandfonline.com/doi/full/10.1080/01448795.2018.1528888>

Spatio-temporal analysis of land use and land cover change: a systematic model inter-comparison driven by integrated modeling techniques

Enrah-Gour, Anshika Mittal, Anshu Bhatnagar, et al. *Hydrology & Earth System Sciences*

