Camera-based Water Stage and Discharge Prediction with Machine Learning.

Andres Eduardo Nowak de Anda A01638430,
Isaac Emanuel García González A01566697,
Samuel Alejandro Diaz del Guante Ochoa A01637592, and
Ernesto Lopez Villarreal A01552124

Instituto Tecnologico de Monterrey, Guadalajara Jalisco, Mexico
Gildardo Sánchez Ante,

Saturday 3rd December, 2022

Abstract. In the Advanced Artificial Intelligence for Data Science II class, a project was carried out in conjunction with the University of Nebraska-Lincoln, which had the objective of obtaining a model that predicts the accumulation and flow of water, this through the use of a database and images of photographs of a weir taken from the same angle over the years (2012-2019). The database was divided into two types of features, the generic features and hand-crafted, the generic features are those obtained through the analysis of the pixels of the images, such as colors, intensity and entropy, obtaining the average and the sum of the values of the above mentioned characteristics, the hand-crafted features, were obtained through the analysis of an expert in the area of hydrology, which identified points of interest in the photos of the weir, such as the region of the weir, the foam generated, texture, etc. There were created two types of artificial intelligence models to analyze the numerical information (MLP, KNN and random forest) and the photographs (CNN, Segmentation-MLP), the results obtained from the MLP model and CNN model were good, but they weren't convincing the predictions look more like noise of a model just trying to get the lowest error instead of understanding the problem, but the Segmentation-MLP model had results that were convincing that this model could work for majority of rivers and could be able to generalize better.

Keywords: Weir · stage · discharge · loss · accuracy · CNN · Segmentation models · MLP · river

1 Introduction

According to Oracle artificial intelligence [1] (AI) refers to those systems or machines that can replicate human intelligence and can improve iteratively from the information they collect. The application of AI is becoming more and more common in our lives, because there are more and more useful applications. According to the Harvard Business Review [2], companies use AI mainly for:

- Detect and deter security intrusions (44%)
- Solve users' technological problems (41%)
- Reduce production management workload (34%)
- Measure internal compliance in the use of approved suppliers (34%)

Within AI there is a discipline known as machine learning, whose algorithms review data and become capable of predicting future behavior. Machine learning uses algorithms to identify patterns in the data, and those patterns are then used to create a data model that can make predictions. With more experience and data, the results of machine learning become more accurate, much in the same way that humans improve with more practice.

This paper describes the development and use of different prediction models with the help of machine learning algorithms in a real problem. The problem is presented by the University of Nebraska-Lincoln where it is requested that, by analyzing images and a database with numerical values, a way to predict flow and water quantity values and find if he variables in the dataset and images given by the University of Nebraska-Lincoln are not going to be sufficient to be able to build a prediction model.

Throughout the paper, different forecasting models will be presented using various sources of information such as numerical data and images.

2 Numerical Models

2.1 Dataset analysis

The database given by the University of Nebraska-Lincoln, contains 42,059 rows and 58 columns of variables, divided into two types of data, these were obtained with collaboration from different academic areas, image engineering and hydrology, dividing the columns into generic and handmade. The generic data are values that were created based on the analysis of the images (*images are in .jpg format, and images are size of wdith: 4288, height: 2848, with color*) and pixels, like colors, intensity and entropy, the handmade features were chosen by the analysis of a hydrology expert, based on what is observed in the photos. This process of selecting characteristics was automated through the use of scripts to obtain the characteristics indicated by the hydrologist. The dependent variables to forecast with the models are stage and discharge Table 1, these variables describe the amount of water and flow that exists in the weir. And the time intervals in which the readings were taken by the sensors vary considerably from a couple of minutes to hours or even days because there wasn't the same amount of photos as sensor readings so only the sensor data that was taken in a range of 5 hours were saved.

The first analysis of the dataset that was done was a correlation graph of all the variables to observe if there was any relationship with the dependent variables and thus have the knowledge of which variables to use or discard for the

creation of the models, eliminating columns with low correlation and lowering the complexity of the final model avoiding unnecessary data. The most popular method for calculating correlation is the Pearson's coefficient, which measures the linear dependence between two variables as long as they are continuous and quantitative. Other heatmaps were also developed using Spearman's Figure 13 and Kendall's Figure 14 coefficients. All three plots show similar results with the Pearson Figure 12 correlation giving the best results. From these correlation graphs two variables were removed 'AreaFeatCount' and 'WwCurveLineMin', because the graphs showed that they didn't have any correlation with the other variables.

Then columns that were not necessary for the prediction of the dependent variables were removed, which in this case were the columns 'Filename', 'Agency', 'SiteNumber', 'TimeZone', 'CalcTimestamp', 'Width' and 'Height' since they do not affect the aforementioned analyses and are data that were obtained at the time of data storage, giving only context to the information and do not give value to the database.

The Standard deviation (SD), variance and coefficient of variation (CV) of all the data was then calculated in order to observe the dispersion of the values in relation to the dependent variables. Values equal to zero were removed to recalculate the values and compare the results of the two calculations Table 2. These values were also calculated by month to see how much they differ by month in all the years Table 3. Comparing the values between the dataset with and without values equal to zero, there is no significant difference while the values between each month is quite different from month to month, but in average the CV was very low for stage and discharge, so the error for our model has to be considered based on the standard deviation of the data.

Range	Stage	Stage without zero	Discharge	Discharge without zero
Min	0	1.37	0	6.73
Max	6.49	6.49	7,920	7,920

Table 1: Min and Max range for Stage and Discharge variables

	Stage	Stage without zero	Discharge	Discharge without zero
SD	0.80	0.80	1,192.27	1,200.88
Variance	0.65	0.65	1,421,513.21	1,442,125.45
CV	0.28	0.28	1.23	1.18

Table 2: Standard deviation, variance and CV for Stage and Discharge variables

	Stage	Discharge
Max SD	1.33	2,197.79
Mean SD	0.39	558.25
Max Variance	1.77	4,830,301.55
Mean Variance	0.29	729,272.88
Max CV	0.41	1.52
Mean CV	0.13	0.63

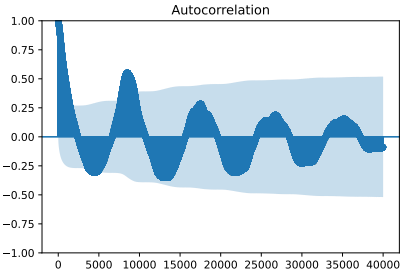
Table 3: Mean and Max values by month of all years for standard deviation, variance and CV for Stage and Discharge variables

For the outliers, an analysis was made of how many values had a large difference in relation to the others in the same column, as they could show significance when making the models and should not be eliminated, in the case of this database, only the values equal to 0 were eliminated because it was observed that in those same observations, the images still show a certain amount of water and in turn flow through the weir, so it was decided to eliminate these values as they showed discrepancy when compared with the photos.

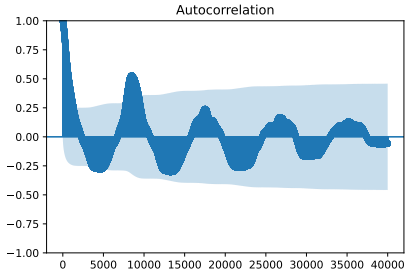
Subsequently, we checked whether the time between each observation was equidistant from each other, in order to verify whether the dataset lends itself to be treated as a time series. If this was not the case, the observations would have to be modified to comply with a time series. Coincidentally, both the observations recorded by the sensors and the camera are the same, but the time between each recording is different, on average from one and a half hours to almost two whole months. To correct this unforeseen issue, filling missing data to make it equidistant by one hour (using forward filling of data and mean for the rest of the values) and removing duplicate time values were done. This also ensures that there is no null data.

Two tests were performed to see if the data were stationary or not, Dickey-fuller and Kwiatkowski-Phillips-Schmidt-Shin (KPSS), the first method showed that the values are stationary and the second method showed that they were not stationary. With each analysis different results were obtained depending on which test will be used, demonstrating the conflict between the values are or are not stationary, but based on the graphs and previous analysis it can be concluded that the KPSS test is probably the correct one, since no stationarity is found in the data.

Autocorrelation Figure 1 and partial autocorrelation plots Figure 2 were created to verify that the data is correlated. The graphs show sinusoidal shapes that serve as evidence that they are not stationary. It was determined that approximately any value of the time series has a lag of approximately 5,000 observations which means it depends on the previous 5,000 data to determine its stage and discharge.

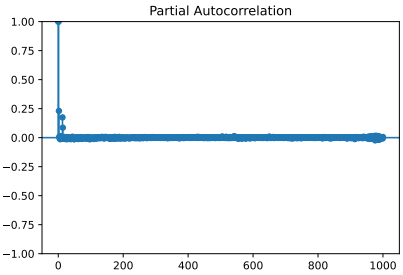


(a) Autocorrelation of stage values with a lag of 40,000.

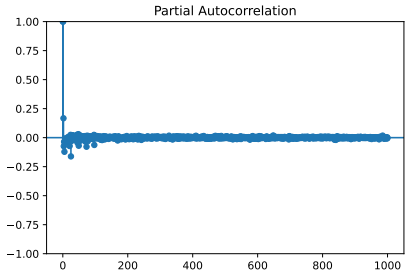


(b) Autocorrelation of discharge values with a lag of 40,000.

Fig. 1: Plots of autocorrelation.



(a) Partial autocorrelation of stage values with a lag of 1,000.



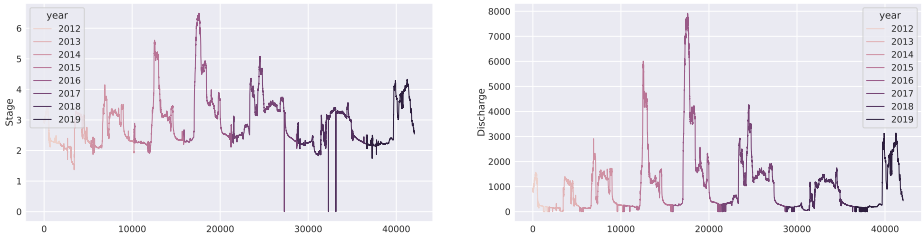
(b) Partial autocorrelation of discharge values with a lag of 1,000.

Fig. 2: Plots of partial autocorrelation.

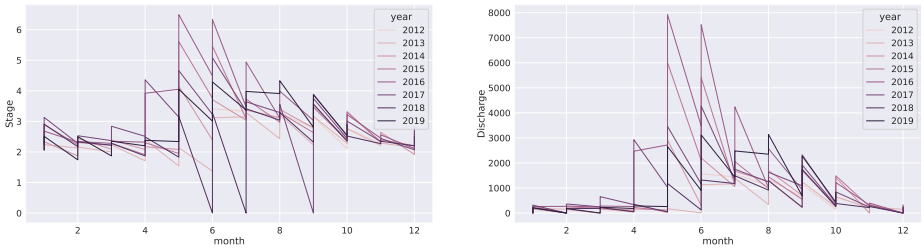
Seasonality Figure 15 Figure 16 plots were made, to verify that the data have a seasonal behavior, the dataset was divided by months, joining in 3 months, by four each year, thus dividing it into the seasons of the year, showing that there is seasonality in the years.

A time series analysis Figure 3 was made in order to obtain through graphs, to observe if there is any relationship with time, these being divided by months of the year, gathering all the data by year and see if there are differences in the months of each year. It was observed that each month there is an increase in the values of the dependent variables, thus demonstrating that they have a similar behavior in all years. Similarly, this same process of time series analysis was performed using only the columns containing generic values. All graphs give similar results for the data set containing both generic and hand-crafted variables. Another interesting thing found in the time series by month is that there is a very big spike at the start of each month, maybe because the weir is

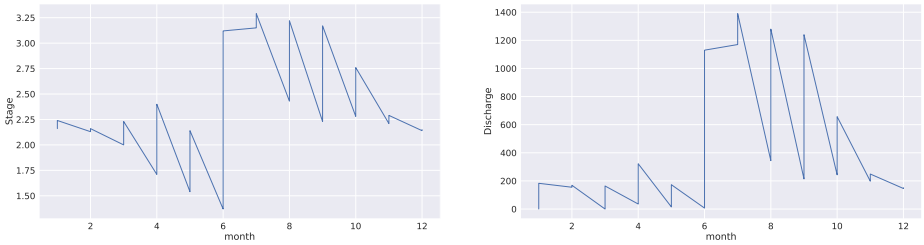
closed at the end of the month and then opened at the begging of the month, this gives us predictable information that can be used to train our models.



(a) Time series through the years of Stage and Discharge. X is all the variables ordered in time, Y axis is the stage and discharge variables.



(b) Time series by month of all the years of the variables Stage and Discharge.



(c) Time series by month in the year 2013 of Stage and Discharge.

Fig. 3: Time series plots.

And lastly because of the standard deviation of the data, the seasonality and pattern of the data, the data was **divided by years for training and testing** (ex. train: 2012 to 2017, test: 2018 to 2019), instead of grabbing data randomly from all the dataset, because grabbing randomly would mean that the training dataset is going to see the values from the testing dataset, they are so close to the average and the pattern is similar for all the data. **The dependent variable to use is only the stage variable**, because stage represents the height of water and discharge the volume in time. Only the stage variable was used, because with a picture it would be difficult to get the information of movement, so only the stage variable is used as output in all the models.

2.2 Models

2.3 General information and preprocessing

General information For these models the data was divided by train (2012 to 2017) and test (2018 to 2019), with (**46 input variables**, combining generic and handcrafted variables and only **1 variable as output**, the stage variable). The training of the model is possible thanks to the cross-validation method, this is used to find the best variables for each model. CV consists of repeating and calculating the arithmetic mean obtained from the evaluation measurements over different partitions.

Preprocessing Standard scaling based on the training data was applied before passing the values to the models

MLP The first model that was used is the multilayer perceptron [9] (MLP) Table 4, this is an artificial neural network formed by multiple layers which serve to solve nonlinear problems, as does a simple perceptron, this model is connected in a unidirectional way, passing the information from left to right with different possible numbers of inputs and outputs, The input layer is where the information is delivered and has no processing, the output layer is the last layer of the artificial neural network, where the results of the processing of the hidden layers will be obtained and the hidden layers are the ones that will process the information, being that unlimited layers can be added for possible better results, these being found between the input and output layer.

Each neuron created in the artificial neural network [10] contains a constant value showing the relative weight which provides the importance of the input within the function of the neuron, in addition, they contain an activation function that defines the output of a node given one or more inputs. The activation function used in this model is the Rectified Linear Unit (ReLU) Equation 1, which transforms the negative input values to a 0 and positive values leaves them as they are. ReLU function

$$ReLU, f(x) = max(0, x) \tag{1}$$

Random Forest for regression The second tested model was the Random Forest Regressor. This type of model is based on a decision tree. The flow of information starts at the root of the tree and follows a path according to the conditions it encounters at each node until it reaches a leaf node (a node that does not point to any other nodes). Another factor that characterizes this type of model is ensemble learning, which consists of the process of using multiple models at the same time, generally trained under the same information, but there exists an alternative that samples the information in each iteration and obtains the average of the results in order to calculate a more accurate forecast.

MLP model	
layer	**Number of neurons**
input	46, ReLU
1st layer	512, ReLU
2nd layer	256, ReLU
3rd layer	128, ReLU
4th layer	128, ReLU
output	1, Linear function

Table 4: MLP model used.

KNN The last implemented model that used the numerical dataset was a variation of the K Nearest Neighbor (KNN) known as KNN regression. This model works on the same principle as the classical version where a number of neighbors are chosen to form nearest neighbor clusters to an observation in the plane. However, instead of generating classification areas, the nearest neighbors will be used to calculate an average that will be the value corresponding to the dependent variable of that observation.

Distance metrics to determine the closest observations include calculations like the Euclidean distance, Manhattan distance or Minkowski distance.

3 Image Models

3.1 General information and preprocessing

General information The dataset was divided in train (2012 to 2016), validation (2017), testing (2018 to 2019), with (**2 input variables** for the two models and, **1 variable as output**, the stage variable).

Preprocessing Neither the input or output variables were scaled for the following models.

3.2 Image dataset analysis

The second part of this study consisted of performing a direct analysis on the images captured at the weir site, making the appropriate modifications and creating predictive models based on the photographs.

Images were originally set at an average resolution of 4288 x 2848 pixels and occupied an storage amount of approximately 240 gigabytes (GB), so the size of the photographs was scaled down to 512 x 512 pixels to decrease the average size of a file. After this reduction, the total dataset was approximately 2 GB.

Just as the first part of the paper, an analysis of the variables along with the images was elaborated to identify any discrepancy between the numerical values obtained by the sensors and the image that belongs to that observation. By examining the rows that have zero water level and comparing it to the corresponding image, consistently some water can be found in each observation without exception. It was decided to remove the observations from the data set that meet this condition.

3.3 Models

CNN model

Preprocessing The content of the photographs was modified so the majority of the image would only contain the river, hiding the top part of the image with a solid black shade. This was done in order to cut out unnecessary information, and hopefully reduce noise that could bias the learning ability of the models.

The first model used for image analysis was a convolutional neural network [3] (CNN), it is a type of artificial neural network and a variation of the multilayer perceptron, based on neurons of the primary visual cortex of a biological brain, using two-dimensional matrices for image classification and/or segmentation. For the extraction of segments and features of images it uses alternating layers of convolutional neurons and downsampling neurons. When processing the data, its dimensionality is reduced and the pixels in each segment are analyzed, thus identifying the color of each pixel of the image on the RGB scale from 0 to 255 and then normalizing them to a scale of 0 to 1.

In the image processing phase, a group of pixels is selected, then it is mathematically analyzed in a scalar product against a smaller matrix called kernel. The process manages to visualize all the input neurons from left to right and from top to bottom, this results in the generation of a new output matrix.

The CNN model uses a Residual Neural Network (ResNet 50 layers Figure 17). This model is a combination of the results of the ResNet model and the variable month when each photo was taken Figure 4 as input to a MLP model that used the ELU Equation 2 function as activation for the layers and a linear function for the output layer.

$$ELU = \begin{cases} x, & \text{for } x >= 0 \\ \alpha(e^x - 1), & \text{for} x < 0 \end{cases} \tag{2}$$

A CNN model was used following the hypothesis statement that maybe the CNN model can find some information in the image that could explain the phenomenon of the stage variable, perhaps the model by itself could find something in the image that could explain the stage of the river. This information would be combined with the month the data was taken, because of the phenomenon

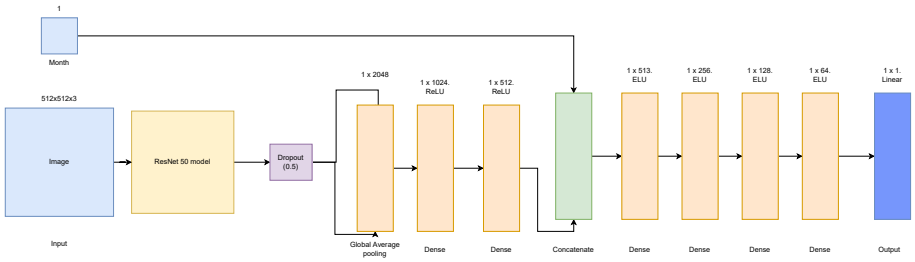


Fig. 4: ResNet-MLP model architecture used.

observed in the time series plots (that the data has a pattern that repeats each month).

Segmentation model and MLP

Important information The river width variable is calculated using Euclidean distance from left shore to right shore, the binary mask and the river area is calculated summing all the 1 ones in the binary mask.

preprocessing For the segmentation model, data augmentation techniques were applied as part of the image processing. Data augmentation consists of methods to artificially increase the number of elements from existing parts by slightly modifying the photograph to achieve a greater diversity of features in order to reduce the impact of obstacles when training machine learning models such as over-fitting. Filter techniques were also implemented in the image to randomly modify the color, saturation, brightness and contrast values, horizontal rotation of some images, cropping and resizing of some sections of the image without changing the image dimension and random rotation. No scaling is done for the input or output variables.

The image segmentation model is a sub-domain of computer vision and digital image processing which aims at grouping similar regions or segments of an image under their respective class labels. Image segmentation is an extension of image classification, where in addition to the classification of objects in images, the model also locates where the object is.

For this model a segmentation model is first trained to be able to do the segmentation of the river images, using a U-Net model with a seresnext101 backbone (the encoder of image for the model), and the output of the segmentation model is a binary mask, where 1 is the river and 0 is everything else. Then from the **binary mask, the river width and river area** is calculated to later be used in an MLP model.

The second model is an MLP model Figure 5 that consists of two inputs, the river width and month and one output, the stage variable. This MLP model uses as activation the ReLU function Equation 1.

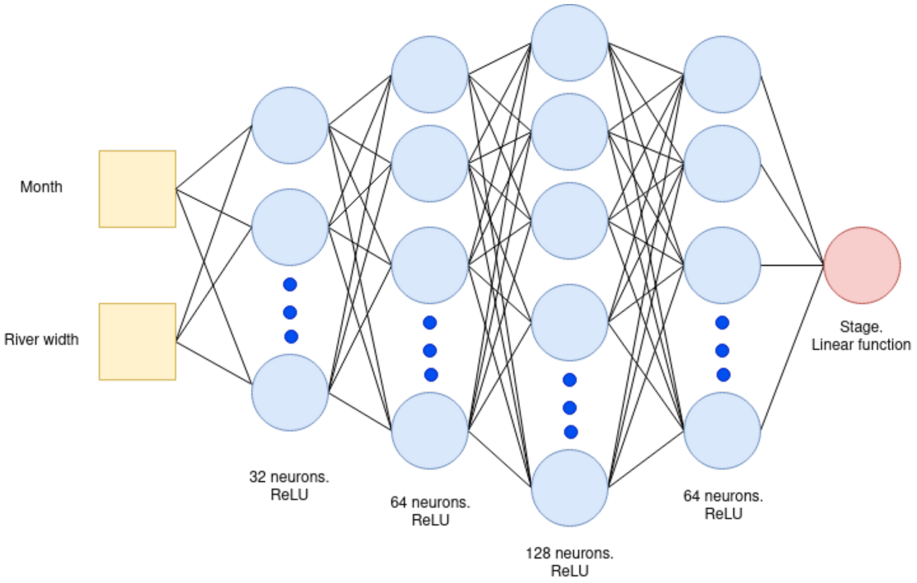


Fig. 5: MLP using data from the segmentation model and month.

The reason this Segmentation-MLP model was used, was because of an hypothesis, a natural river formation [5] is a V-shape figure Figure 18 because of water erosion, so wouldn't that mean that the deeper the river gets, wouldn't the river also get a bigger width and a bigger area thanks to the V-shape of the river. Furthermore, there is a paper [7] that confirms this suspicion. In this paper they calibrate their cameras to get a conversion from pixels in the image to centimeters. Then, they would get the width of the river from the images and with the bathymetry of the river they would calculate the stage of the rivers. In this model instead the stage of the river is calculated based on an MLP model that learns the formula to calculate the stage.

4 Experiments

4.1 Numerical model experiments

General information The inputs were passed with a standard scaler and the models were trained with cross validation using *random search* to find the best combination of parameters for the models. And the output of the models was the stage variable.

1st experiment For the first experiment the (MLP, KNN, random forest) models were trained with **Generic and Hand crafted features**.

2nd experiment For the second experiment the (MLP, KNN, random forest) models were trained with **Generic and Hand crafted features** and making feature selection with the **LASSO** section 9 method.

3rd experiment For the third experiment the (MLP, KNN, random forest) models only **Generic features** were used.

4th experiment For the fourth experiment the (MLP, KNN, random forest) models were trained with only **Generic features** and making feature selection with the **LASSO** section 9 method.

Best model from experiments From the four experiments the best model was the 1st experiment for all the models section 4.1, and this is strange, because majority of the features didn't have that much information, so removing them would make it easier for the model to train and get better results. In the results part this is going to be referenced again to explain the results of the models.

4.2 Image model experiments

General information The output for the CNN and MLP model was the Stage variable

1st experiment with CNN model For this model the size of the images were reduced to a 512x512x3 size, for faster training and because it wasn't necessary the images in the best quality to get information from them. The inputs for the model were the month when the stage value was taken and the image for that stage value. The model was trained using The Adam optimizer section 9 and MSE section 9 as the loss for the model

2nd experiment with Segmentation-MLP model For this model the size of the images were reduced to a 320x320x3 size, for faster training and because the backbone for the segmentation model needed this specific size, the dataset used was the Riwa dataset [4] and 7 masks made by hand of the images of the river at hand. The inputs for the segmentation model was the image with data augmentation and the output was a binary mask, and with this binary mask the river width was calculated using the **euclidean distance** Equation 3 from the left shore to the right shore and the river area was calculated summing up all the ones in the binary mask.

$$\text{Euclidean distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

After training the segmentation model a study was made to see the correlation Figure 6 of the river width and river area with the Stage and Discharge variable. The correlation graph shows that river width and river area has a high correlation with stage and discharge, but the hypothesis is that the correlation could be bigger, but because the model couldn't do a perfect segmentation for images of the river the results weren't perfect

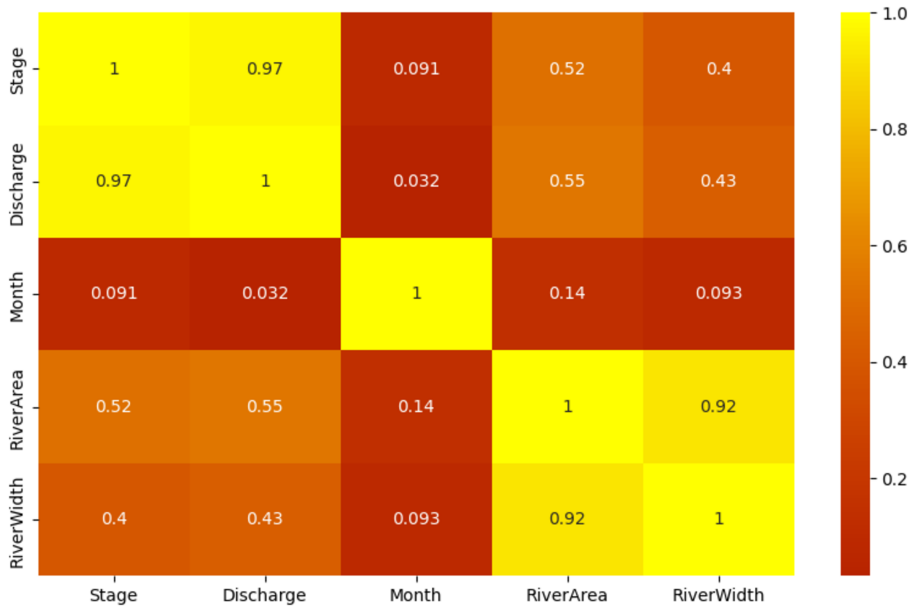


Fig. 6: Pearson correlation of month, river width and area with the stage and discharge.

This images show that the correlations weren't perfect thanks to the bad segmentation and the time series graphs for the river width and stage Figure 9 show that river width Figure 9b has many outliers and that the segmentation weren't perfect because the results vary too much, for the model to work perfectly the segmentation results should be precise in cm, but for this images we didn't find a way to make the segmentations better.

For the second model, the MLP Figure 5, the inputs for the model were the month when the stage value was taken and the river width of the image for that stage value. This model was trained with the Adam optimizer section 9 and MSE section 9 as the loss metric. And the model predicts the stage variable.

(a) River image used in prediction for segmentation model.

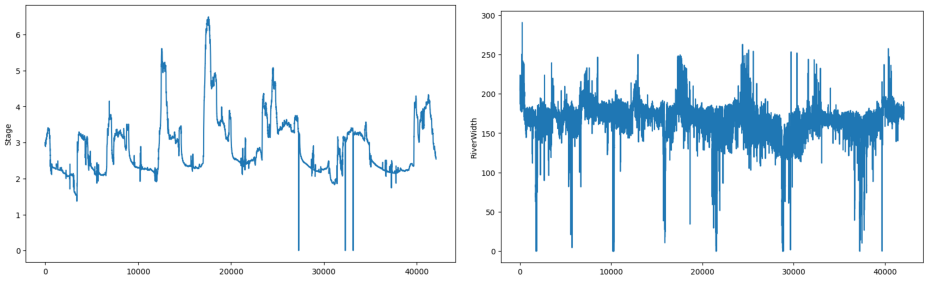
(b) Mask of river for the 1st image.

Fig. 7: Image with a good segmentation.

(a) River image used in prediction for segmentation model.

(b) Mask of river for the 2nd image.

Fig. 8: Image with a bad segmentation.



(a) Time series through the years of Stage. X is all the variables ordered in time, Y axis is the stage variable.

(b) Time series through the years of river width. X is all the variables ordered in time, Y axis is the river width variable.

Fig. 9: Time series plots for stage and river width.

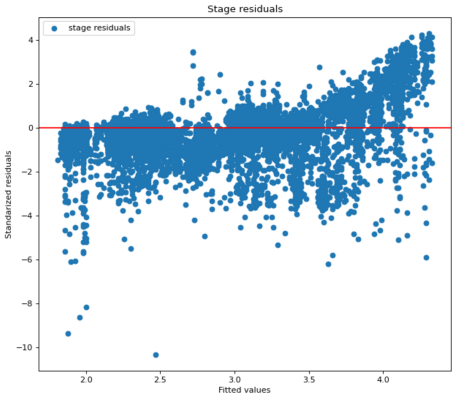
5 Results

When comparing all model results, the Convolutional Neural Network consistently gives the best results. Of these metrics, MAE was chosen to determine the effectiveness of the models Table 5 as it returns the expected forecast error size on average. Alternatively, MAPE was also considered as one of the primary metrics to evaluate models since in some cases the size of the error in MAE can be very large and makes it difficult to interpret the error impact on each model. MAPE scales all the results and converts them into simple to interpret percentages. That said, CNN would be the best model with a MAE of 0.1919 with a MAPE of 0.078. The next closest model to these results was the MLP Regressor with a MAE 0.2223 and a MAPE 0.078. From there the next best option is the MLP Regressor with the Segmentation data (0.2790 MAE and 0.0973 MAPE), followed by the KNN Regressor (0.2821 MAE and 0.1064 MAPE) and with Random Forest Regressor at last with the worst results (0.3127 MAE and 0.1214 MAPE).

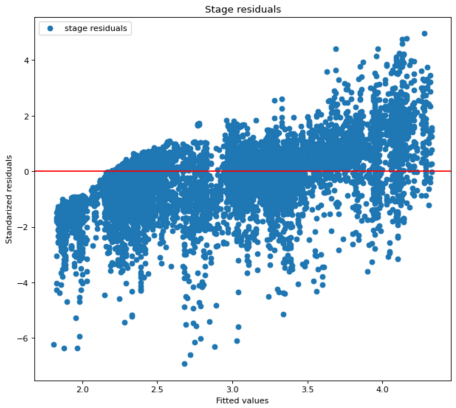
	CNN	MLP Regressor	MLP Regressor with Segmentation	KNN Regressor	Random Forest Regressor
R^2	0.8043055231	0.6911826503	0.4220925667	0.54566232	0.4919470624
MSE	0.07642631698	0.1206052058	0.2258010654	0.1774365638	0.1984144645
MAE	0.1919919813	0.2223266992	0.2790626533	0.2821863754	0.3127794728
MAPE	0.07221717524	0.0780759087	0.09730067528	0.1064906306	0.1214667585

Table 5: Comparison of the models results.

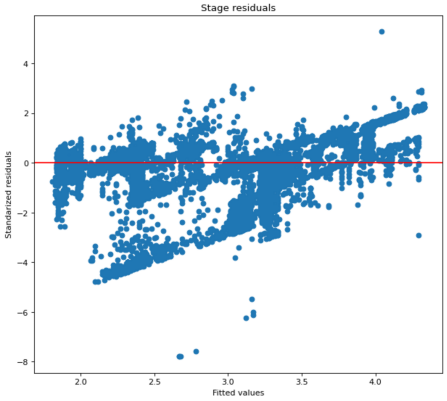
However, after reviewing the results of the model plots Figure 10 Figure 11 the data seems to indicate that there is a better candidate for best model. From the top two models, CNN Figure 11c and MLP Regressor Figure 11b, it can be observed that the predictions generated by this two models have a high level of variability, they have a lot of noise, unlike the MLP Regressor with Segmentation Figure 11d which in general forecasts most of its predictions under a threshold closer to the actual stage and discharge values and it does seem the model is trying to find a formula, instead of just predicting the value that would give the lowest error based on what was learn in the pattern of the training dataset that are very similar to the pattern of the testing dataset.



(a) Stage residuals for MLP regressor model.

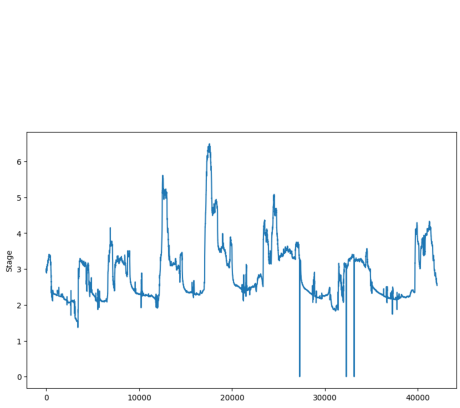


(b) Stage residuals for CNN model.

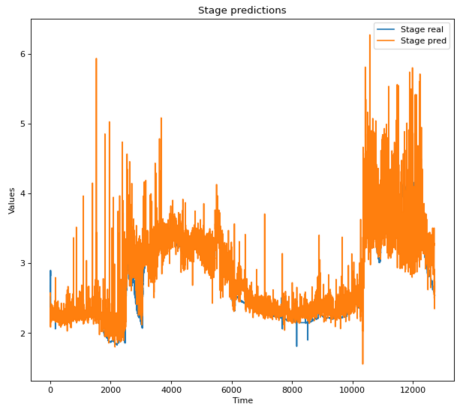


(c) Stage residuals for MLP regressor with segmentation data model.

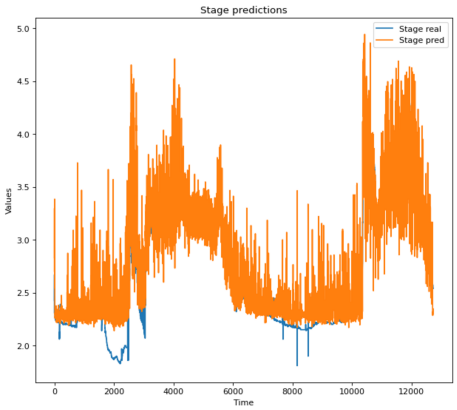
Fig. 10: Prediction residuals from the 3 best models.



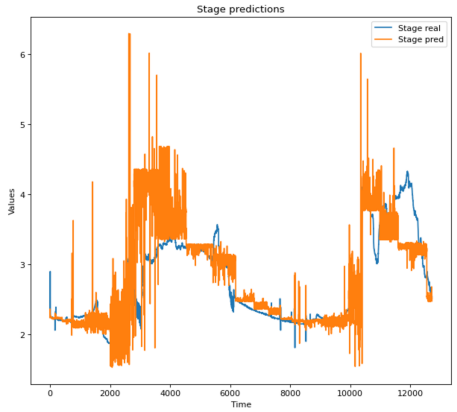
(a) Time series of the ground truth values for stage. X is Time, Y is stage



(b) Time series stage predictions and real values for MLP regressor model.



(c) Time serie stage predictions and real values for CNN model.



(d) Time serie stage predictions and real values for MLP regressor with segmentation data model.

Fig. 11: Predictions from the 3 best models.

As previously said, the data has a relatively low standard deviation, and because the pattern repeats, it appears that the models were able to get good results due to the number of columns in the csv or the degree of diversity in the photos rather than truly comprehending the problem. Theoretically, the CNN model and MLP regressor model will only be capable of producing precise predictions for this particular river as a result, rather than being able to generalize and work for other rivers. The MLP Regressor with Segmentation, on the other hand, is more likely to be able to interpret general image features in a wider range of situations as a result of the manner in which the data is segmented and supplied to the model. As a result, the data suggests that MLP Regressor

with Segmentation is the best model for resolving the current issue, which is predicting the stage of a river.

6 Future work

So it is proved in this research that it is possible to train a model that can learn to predict the stage of a river based on the width of the river and the current month. Nonetheless, in order to prove this hypothesis completely, we need a model that can provide superior segmentation for the photos in the dataset at hand; however, there may be other options beyond simply proving the hypothesis.

- Create 100 masks from the river dataset images to overfit the segmentation model to that river, allowing for improved segmentation precision merely to prove the hypothesis item
- Obtain a separate river dataset that includes photos as well as the stage value; having the images taken from a higher position will make it easier for the model to recognize the edges and segment the river, as well as eliminating reflections from the water of things near the river in favor of reflections from the sky, which will make it easier to segment. Additionally, if the images were taken of a flat section of the river rather than one that displayed the weir, the model would be simpler to divide.

If the hypothesis is confirmed in this way, the next step would be to create a *better segmentation model* that can generalize better, achieve higher precision, and work with images of rivers in the winter because segmenting rivers through ice appears to be more challenging for the model to do. Furthermore, when using this model, the camera should be calibrated to convert pixels to centimeters or inches, which improves precision and simplifies training.

7 Conclusion

At the conclusion of this study, we presented a variety of models for predicting the river stage, and we demonstrate that a segmentation model with an MLP is most likely the most effective model for resolving this issue. By obtaining the width of the river in a picture and the month the image was taken, the Segmentation-MLP model's results demonstrate the potential for the model to be able to predict the stage of a river. But in order to fully support the theory and improve the accuracy of the river's stage, a better segmentation model or an alternative approach to segmenting the river from an image. However, this study demonstrates the prospect of reducing the requirement for sensors to determine a river's stage, and with this research, it is possible to complete any gaps in a river's stage data and receive information from a river in challenging circumstances, such as during a flood.

8 Contributions

- Andres Eduardo Nowak de Anda: Dataset analysis, dataset models (Random Forest Regressor, KNN Regressor), image analysis, image models (CNN, MLP Regressor with Segmentation)
- Isaac Emanuel García González: Dataset analysis, dataset models (MLPRegressor, KNN Regressor), image models (CNN), Report
- Samuel Alejandro Diaz del Guante Ochoa: Dataset analysis, image analysis, image models (CNN, MLP Regressor with Segmentation), Report
- Ernesto Lopez Villareal: Dataset analysis, image analysis, image models (CNN), Report

References

1. ¿qué es la inteligencia artificial (ia)?, https://www.oracle.com/mx/artificial-intelligence/what-is-ai
2. How companies are already using ai (April 2017), https://hbr.org/2017/04/how-companies-are-already-using-ai
3. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network pp. 1–6 (2017). https://doi.org/10.1109/ICEngTechnol.2017.8308186
4. Blanch, X., Wagner, F., Eltner, A.: Riwa dataset (2022). https://doi.org/10.34740/KAGGLE/DSV/4289421, https://www.kaggle.com/dsv/4289421
5. Borneman, E.: How rivers change the landscape (April 2014), https://www.geographyrealm.com/rivers-change-landscape/
6. Cai, S., Wu, Y., Chen, G.: A novel elastomeric unet for medical image segmentation. Frontiers in Aging Neuroscience **14** (2022). https://doi.org/10.3389/fnagi.2022.841297, https://www.frontiersin.org/articles/10.3389/fnagi.2022.841297
7. Eltner, A., Bressan, P.O., Akiyama, T., Gonçalves, W.N., Marcato Junior, J.: Using deep learning for automatic water stage measurements. Water Resources Research **57**(3), e2020WR027608 (2021). https://doi.org/10.1029/2020WR027608, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR027608, e2020WR027608 2020WR027608
8. Mukherjee, S.: The annotated resnet-50 (August 2018), https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758
9. Ramchoun, H., Ghanou, Y., Ettaouil, M., Janati Idrissi, M.A.: Multilayer perceptron: Architecture optimization and training (2016), https://reunir.unir.net/bitstream/handle/123456789/11569/ijimai20164_1_5_pdf_30533.pdf?sequence=1
10. Sharma, S.: Activation functions in neural networks (September 2017), https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6
11. Yan, Q., Iwasaki, T., Stumpf, A., Belmont, P., Parker, G., Kumar, P.: Hydrogeomorphological differentiation between floodplains and terraces. Earth Surface Processes and Landforms **43** (08 2017). https://doi.org/10.1002/esp.4234

9 Glossary

- R^2: Also called coefficient of determination, it is the proportion of total variance of the dependent variable, it reflects the adjustment of a model to the variable to be explained.

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{Y}_t - \overline{Y})^2}{\sum_{t=1}^{T}(Y_t - \overline{Y})^2}$$

- MSE: Mean Squared Error measures the average of the squared errors, the difference between the estimator and what is estimated.

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- MAE: Mean Absolute Error is used to quantify the accuracy of a prediction technique by comparing the predicted values against the observed ones.

$$\frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

- Stage: water level in a river or stream, units: ft.
- Discharge: volumetric flow of water, units: ft^3/s.
- LASSO: Least absolute shrinkage and selection operator is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

$$Loss = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} |w_i|$$

- Adam: is an adaptive learning rate optimization algorithm that's been designed specifically for training deep neural networks.

A Appendix

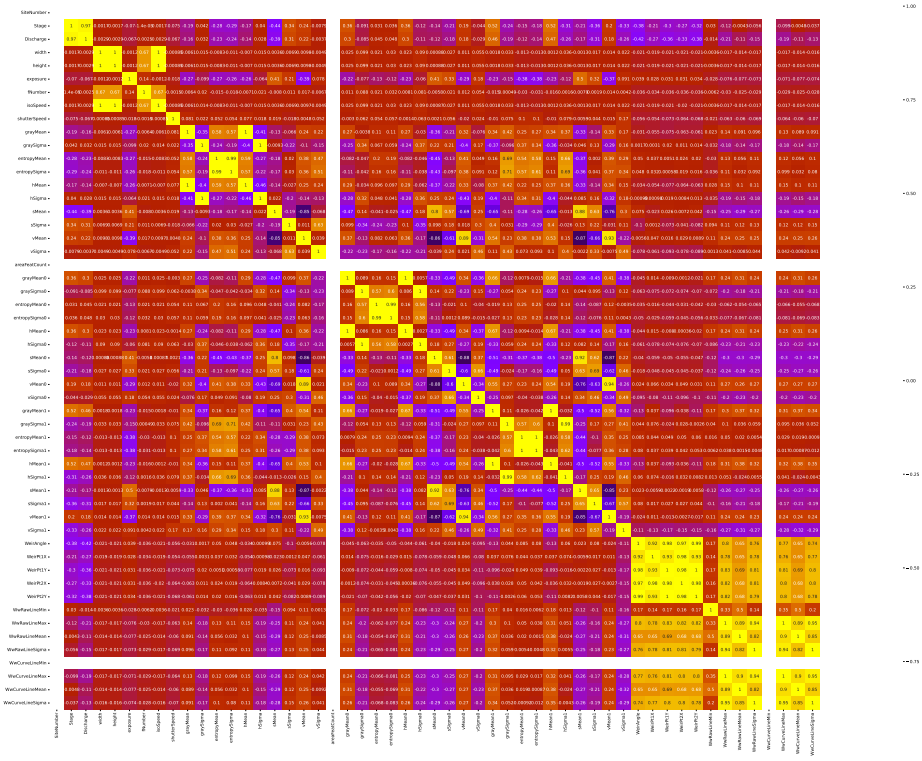


Fig. 12: Pearson correlation of the dataset.

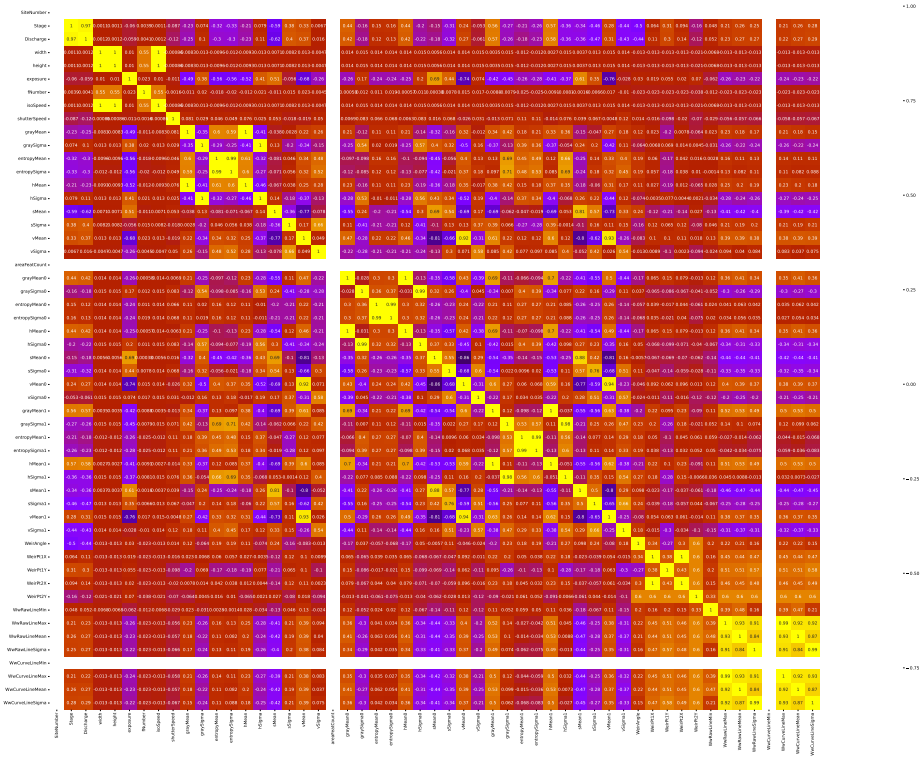


Fig. 13: Spearman correlation of the dataset.

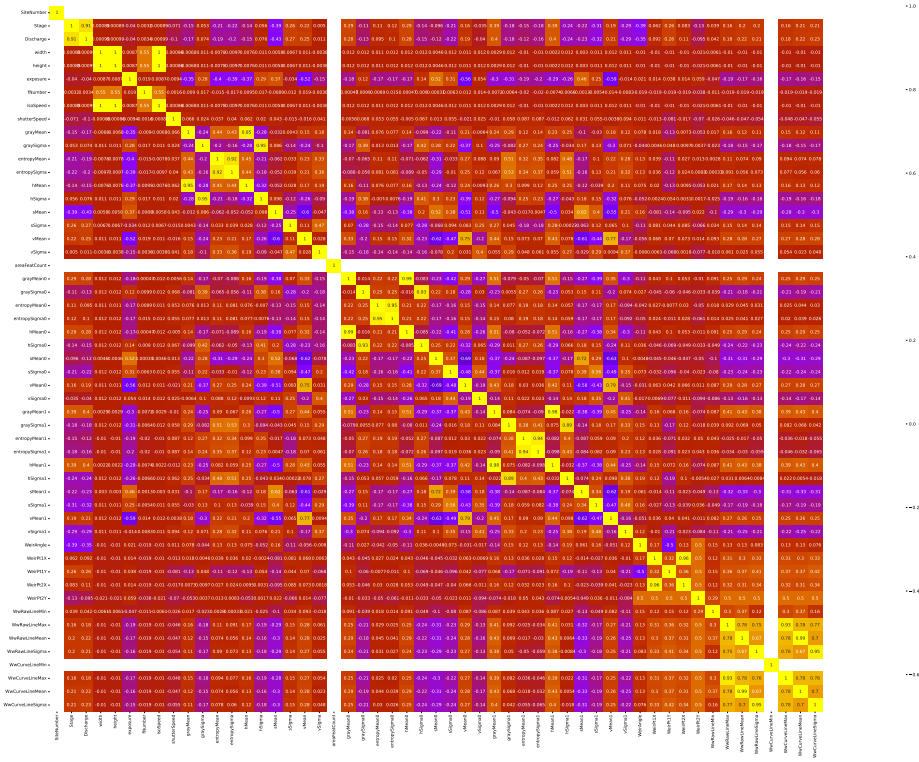


Fig. 14: Kendall correlation of the dataset.

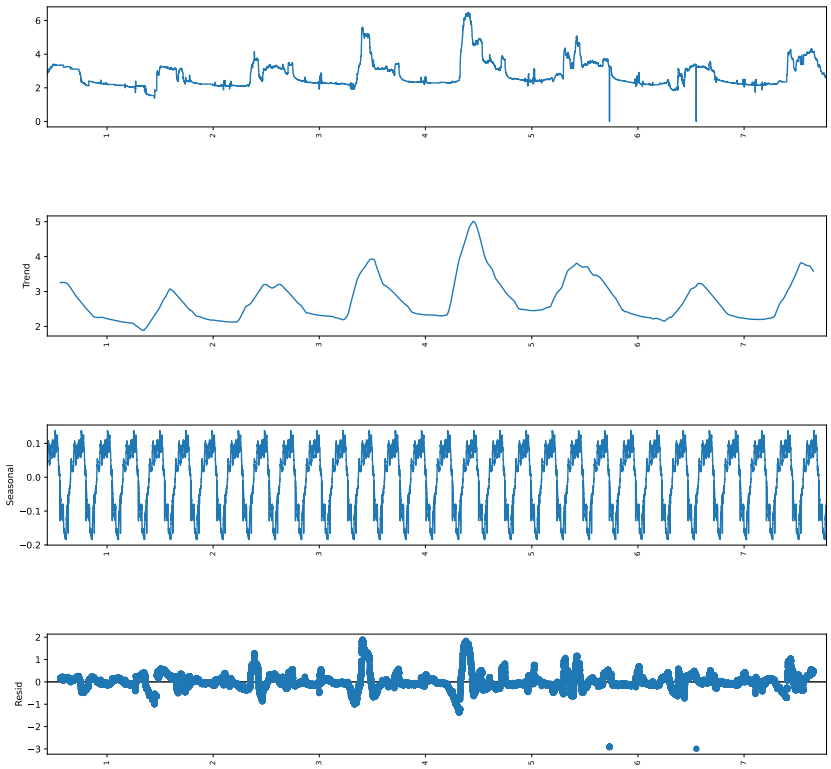


Fig. 15: Seasonality and stationary graph for discharge.

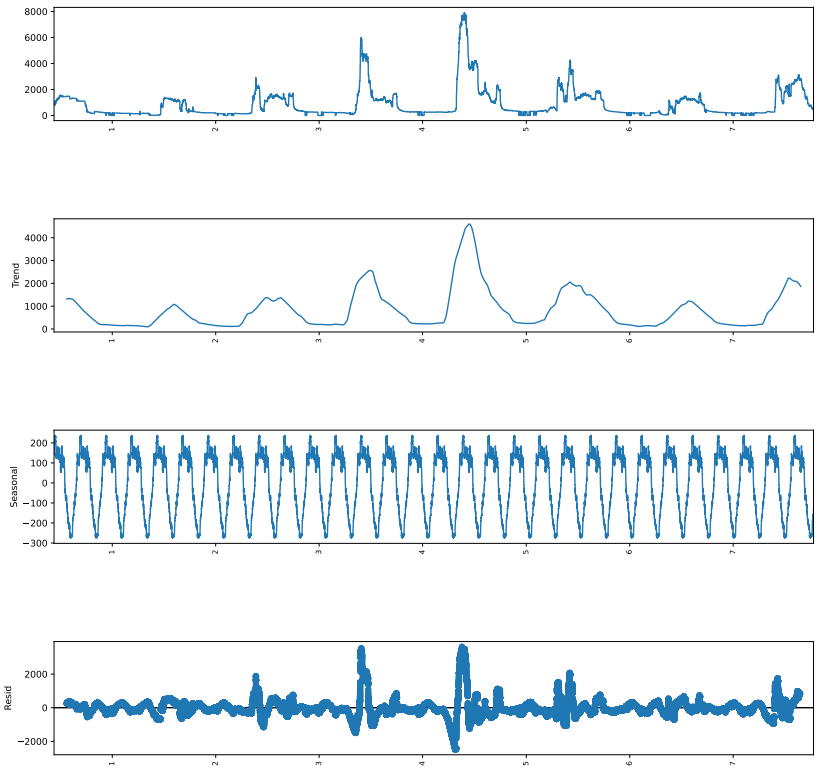


Fig. 16: Seasonality and stationary graph for discharge.

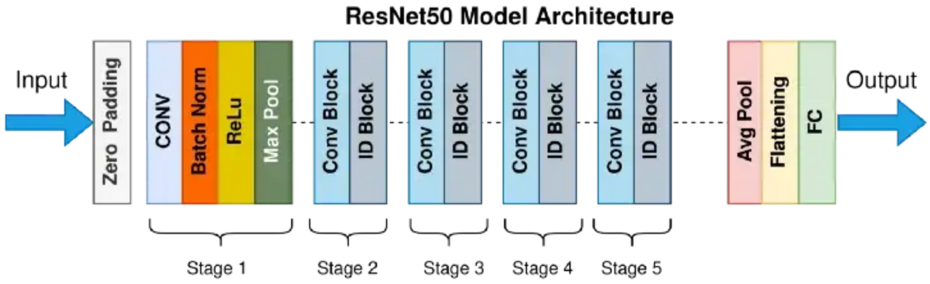


Fig. 17: ResNet50 Architecture [8].

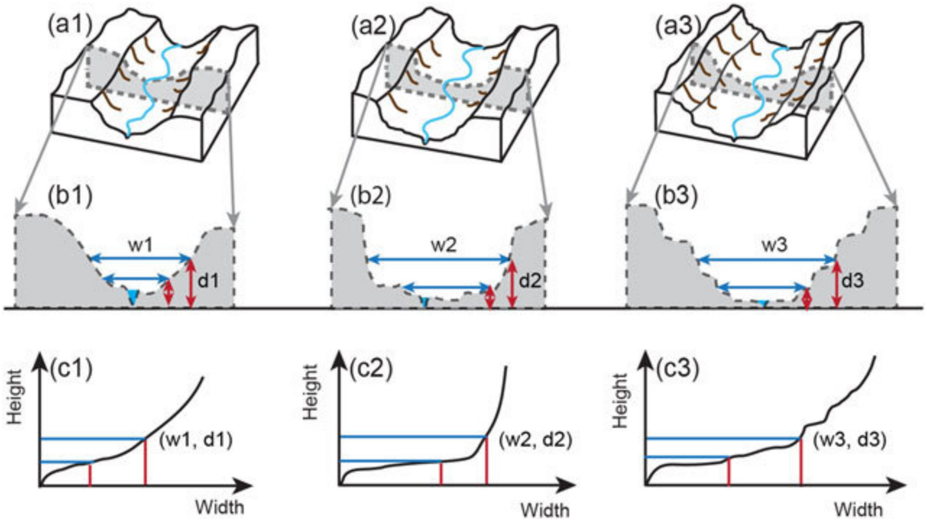


Fig. 18: Illustration of the V-shape in river valleys [11].

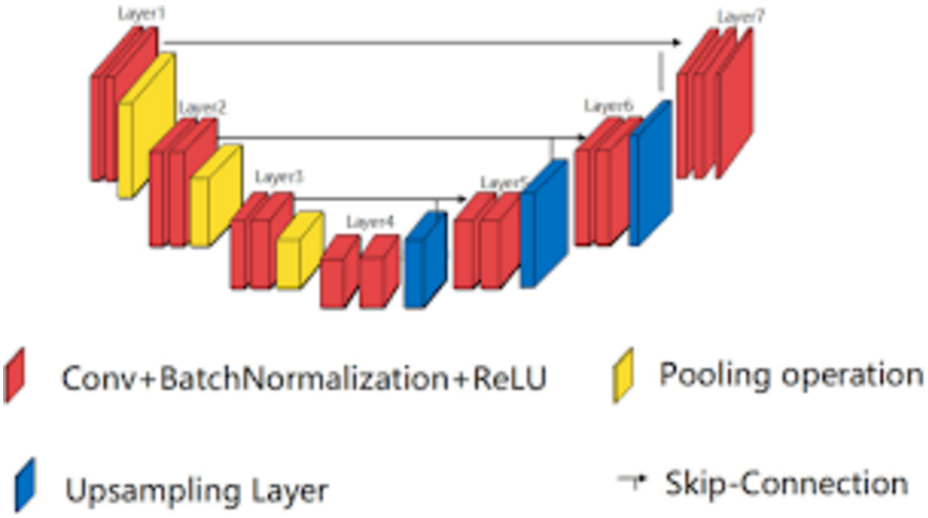


Fig. 19: Ex. illustration of a U-Net model [6].