Analyzing and Comparing Different Data Driven Methods for Hydrological Forecasting to Understand Dam Construction Requirements

Walter P Moore

Henry Hughes

4A Civil Engineering

May 2023

534 Crimson Court

Oshawa ON L1J 8E2

May 16, 2023

Dr. Rania Al-Hammoud

Associate Chair, Undergraduate Studies

Department of Civil and Environmental Engineering

University of Waterloo

Waterloo, ON, N2L 3G1

Dear Dr. Al-Hammoud:

This report, entitled "Analyzing and Comparing Different Data Driven Methods for Hydrological Forecasting to Understand Dam Construction Requirements" was prepared as my 4A work term report. During my work term, I worked on projects dam reinforcement projects. These dams needed to be reinforced due to higher levels of discharge that initially expected. The purpose of this report is to gain additional understanding into how levels of hydrological discharge are forecasted by using data driven methods.

This report was prepared during my Co-op at Walter P Moore. I would like to thank the team at Walter P Moore, as they allowed me to work on several dam projects which inspired me to develop a report that is focused on how data driven methods for hydrological forecasting can be used to create more resilient infrastructure.

This report focuses on the numerical methods that are Multiple Linear Regression, and Extreme Learning Machines. While not being the most complicated mathematical tools that can be utilized for this type of analysis, these methods are not too challenging to apply, and allow for standard datasets to be used without much difficulty.

This report was written entirely by me and has not received any previous academic credit at this or any other academic institution. I received no outside help with the report.

Sincerely,

Text, letter

Description automatically generated

Henry Hughes  
ID# 20792437

Analyzing and Comparing Different Data Driven Methods for Hydrological Forecasting to Understand Dam Construction Requirements

Walter P Moore

Henry Hughes

4A Civil Engineering

May 2023

# Summary

It is becoming increasingly common that dams around the world are having structural issues due to poorly forecasted demand. This demand exists in the form of discharge that goes through the dams in question. There are several ways that discharge through dams is commonly forecasted. Many of these ways come in the form of empirical equations, and the rest are data driven models. In this report, two data driven models were tested, Multiple Linear Regression, and Extreme Learning Machines. The models were tested on the same dataset and compared to each other. The models performed well, and there was not much that differentiated them when analyzing the forecasts that were generated by the models. This application of statistical techniques shows the importance of

# Acknowledgements

I want to thank Professor John Quilty for allowing me to be a part of his research team, which has inspired me to write about this topic. The work his team does is fascinating, and the reason I am focused on learning more about the topic of applied machine learning and statistical analysis in hydrology.

**Table of Contents**

[Summary v](#_Toc132800602)

[Acknowledgements vi](#_Toc132800603)

[List of Figures viii](#_Toc132800604)

[List of Tables viii](#_Toc132800605)

[List of Appendices viii](#_Toc132800606)

[1.0 Introduction 1](#_Toc132800607)

[1.1. General 1](#_Toc132800608)

[1.2. Background 1](#_Toc132800609)

[1.3. Scope and Objectives 1](#_Toc132800610)

[2.0 Multiple Linear Regression 2](#_Toc132800611)

[2.1. Multiple Linear Regression Formulas 2](#_Toc132800612)

[2.2. Variable Dependency 3](#_Toc132800613)

[2.3. Common MLR Use Cases 3](#_Toc132800614)

[3.0 Extreme Learning Machine 3](#_Toc132800615)

[3.1. ELM Mathematics 4](#_Toc132800616)

[3.2. ELM History and Prior Applications 6](#_Toc132800617)

[4.0 Development of Dataset 7](#_Toc132800618)

[4.1. Dataset of Choice 8](#_Toc132800619)

[4.2. Data Cleaning 10](#_Toc132800620)

[5.0 Application of and Analysis of MLR 11](#_Toc132800621)

[5.1. Data Splitting 12](#_Toc132800622)

[5.2. MLR Results 13](#_Toc132800623)

[6.0 Application and Analysis of ELM 14](#_Toc132800624)

[References 15](#_Toc132800625)

[Appendix A: Original Dataset 16](#_Toc132800626)

[Appendix B: Graphs of Independent Variables 17](#_Toc132800627)

[Appendix C: Graphs of Independent Variables vs Gramenet Discharge 18](#_Toc132800628)

# List of Figures

[Figure 1: ELM Network Structure (AlOmar et al., 2020) 4](#_Toc132800629)

[Figure 2: Barcelona\_Fabra Daily Rainfall 9](#_Toc132800630)

[Figure 3: Heatmap of Missing Data 10](#_Toc132800631)

[Figure 4: Barcelona\_Fabra Daily Rainfall vs Gramenet Discharge 12](#_Toc132800632)

# List of Tables

[Table 1: First Four Explanatory Variables 8](#_Toc132795982)

[Table 2: Final Four Explanatory Variables, and Output Variable 9](#_Toc132795983)

# List of Equations

[Equation 1: MLR General Equation 2](#_Toc132800665)

[Equation 2: Hidden Layer Output Matrix 5](#_Toc132800666)

[Equation 3: Output Weight Equation 5](#_Toc132800667)

[Equation 4: New Forecasts 6](#_Toc132800668)

# Introduction

## General

There are many benefits to understanding how water moves. It is also beneficial to understand how the movement of that water changes over time. It allows cities to be constructed in the right location with the correct water resources assets, ensures proper flood protection measures can be put in place to protect critical infrastructure, and will allow society to function in a way that allows people to utilize the benefits of water, without the drawbacks. This is a challenge that engineers must face, and use technology to better understand, so that moving forward, infrastructure can be created in the most effective way possible to serve individuals around the world.

## Background

Water resources assets are critical for any society to function. This primarily includes the rivers, lakes, and groundwater from which we obtain all our potable water. It also includes physical assets such as dams, water and wastewater treatment plants, desalination plants, and water distribution networks. To ensure these physical assets will be able to last into the future, their owners must be able to forecast their future demand. For dams, the demand is how much upstream discharge will pass through the dam. Accurately forecasting this discharge will allow dams of the correct size to be constructed, increasing their longevity.

## Scope and Objectives

There are several ways that discharge through a dam can be forecasted. Historically, physical methods have been used. These were primarily a function of empirical equations that would be developed from significant amounts of field testing. These equations would allow engineers to forecast discharge through specific locations in rivers so the size of a specific dam could be determined. Recently, data driven methods have been used to forecast hydrological discharge. These methods are more efficient, since they can be done with basic data. Physical methods that require lots of empirical data are difficult to come up with, particularly in a short amount of time. This report will detail and apply two data driven methods; 1. Multiple Linear Regression (MLR), and 2. Extreme Learning Machines (ELM). The results will then be analyzed, and future methods will be explored.

# Multiple Linear Regression

MLR is a simple statistical tool that is used to create forecasts for a specific variable of interest, based on a variable number of explanatory variables. It identifies the linear relationship between a set of input explanatory variables and an output variable. The mathematical model allows the output variable to be forecasted into the future by utilizing a linear equation that projects the dependent variable.

## Multiple Linear Regression Formulas

The general formula for MLR can be shown in Equation 1 below.

Equation : MLR General Equation

In Equation 1, y is the dependent variable, or the output. This variable can be linked to the n independent variables, which are represented by xi. Each of these independent variables are weighted by their coefficients ai. The coefficient ao is a correction term, which allows the curve to be properly fitted. All the other coefficients are what define the relationship between each independent variable xi and the dependent variable y.

To determine whether an independent variable is correlated to the dependent variable in a significant way, a p-value can be used. In linear regression, a p-value of one means that the independent variable in question is perfectly correlated with the dependent variable. A p-value of zero means there is no correlation between the independent and dependent variables. The p-value is calculated based on a t-test of the estimated coefficient, which compares the estimated coefficient to the standard error of the estimate.

The goal of MLR is to determine the values for the coefficients ai such that the function in equation one above best fits the data, which is typically done by minimizing the sum of the squared errors between the actual values of the dependent variable and the predicted values of the dependent variable. This can all be done using basic tools such as Microsoft Excel or Python.

## Variable Dependency

A benefit to MLR is that there is not any connection between the independent variables in the equation. However, it is important to analyze the variables in a way that results in a model that is statistically efficient. This means that multicollinear variables should not be included in the model, and all variables should be statistically significant.

## Common MLR Use Cases

MLR is a common statistical technique that has been used for a wide variety of applications. One reason for this is that MLR is a very accessible tool that anyone can apply if they have a complete dataset. Microsoft Excel has a built-in set of tools for using MLR, meaning that it is often a tool people turn to when they have a dataset and are looking to create a meaningful forecast with it.

# Extreme Learning Machine

An ELM is an algorithm that can be used to develop data forecasts. It is a single layer feedforward neural network. The idea behind an ELM is to randomly select a set of hidden neurons and compute their output weights analytically. ELM algorithms are usually used for regression-based models. While being quite straightforward, ELM has been shown to be an effective technique for forecasting. The main benefit of the ELM method is that it is computationally efficient since it does not require backpropagation to train the neural network (Yaseen et al., 2016).

## ELM Mathematics

An ELM has a few different components, including an input layer, a layer of hidden neurons, and an output layer. The general structure can be seen in Figure 1 below.

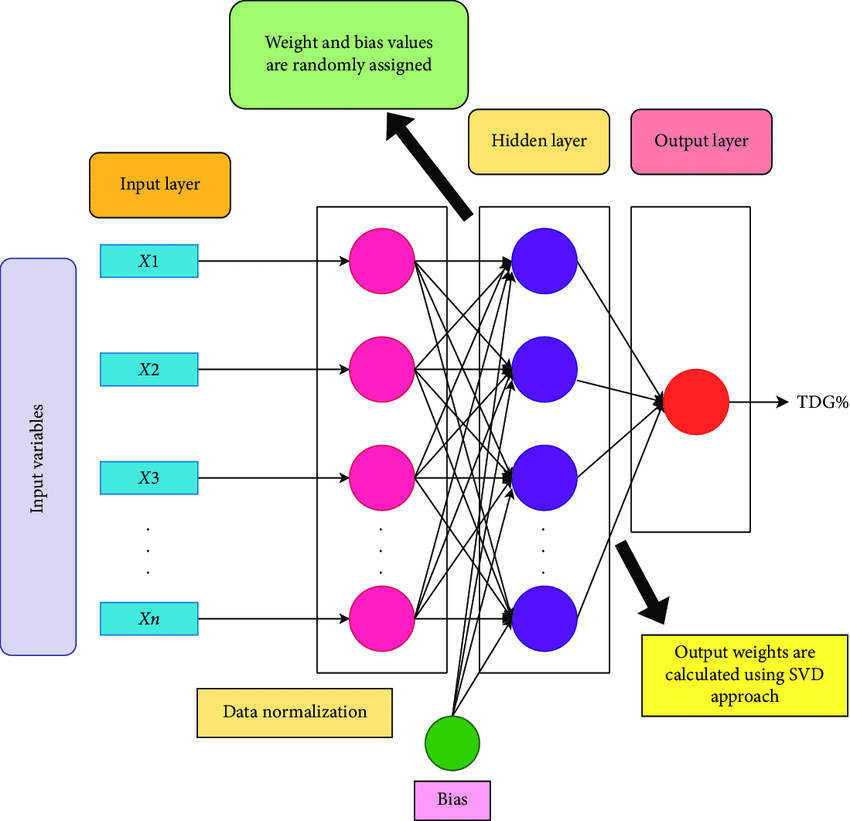


Figure 1: ELM Network Structure (AlOmar et al., 2020)

Looking at the general structure of the ELM begs the question; how are the weights and biases computed? In an ELM, these weights and biases are randomly assigned. After this has been completed, the hidden layer output matrix H can be computed by using an activation function of choice, as shown below in Equation 2. This paper will consider the sigmoid activation function (You, 2023).

Equation : Hidden Layer Output Matrix

The variables and functions in Equation 2 are explained below (You, 2023).

Once the above variables have all been defined, either randomly or through the datasets in question, the H matrix can be computed. Once this has been done, the output weights must be calculated using the Moore Penrose Pseudo Inverse matrix method as shown in Equation 3 below (You, 2023).

Equation : Output Weight Equation

The variables and functions in Equation 3 are explained below.

Once the output weight vector has been computed, predictions can be made using new input data as shown in Equation 4 below (You, 2023).

Equation : New Forecasts

The variables in Equation 4 are explained below.

It should be noted that the same weights and biases used in training will be used during the forecasting process. Once the variables are computed, forecasts can be created for all instances where independent variables exist, but dependent ones do not.

The actual application of the ELM algorithm is straightforward when using Python, as there are functions that exist which will do most of the mathematics once a clear dataset has been obtained.

## ELM History and Prior Applications

ELM was first proposed by Huang et al in 2006. It was proposed since it is a much more computationally efficient method than other neural network algorithms (Yaseen et al., 2016). However, it has been criticized as being less effective than other deep learning techniques. This is mainly because the initial biases and weights are randomly set and are not trained. Additionally, the trainable portion of the neural network follows a similar structure to that of multiple regression. Essentially, the coefficients are determined in a similar way to regression coefficients (Ye, 2020).

ELM has previously been used in hydrology, most notably by Professor John Quilty at the University of Waterloo. Several papers have been published by him that focus on ELM and other machine learning techniques for streamflow forecasting. In a paper that focused on the application of ELM for streamflow forecasting in a semi-arid region of Iraq, it was stated that data driven models are useful in hydrology because they do not require physically based models and parametric assumptions, which can often be hard to develop. This paper tested ELM to combat overfitting and slow training times that were an issue with other types of neural networks (Yaseen et al., 2016).

The paper concluded that ELM did a better job forecasting the data than a Support Vector Regression (SVR) model and a General Regression Neural Network (GRNN) across several metrics, proving its viability in the hydrological space (Yaseen et al., 2016).

# Development of Dataset

To apply the MLR and ELM methods described in the prior sections of this report, one needs data that can be used to create a forecast. I was unable to use data from the project I worked on at Walter P Moore. That specific project focused on data for the Brazos River in Texas and included things such as rainfall and discharge amounts from areas near the dam of interest. Since I cannot use that project data, I have obtained sample data which I outline below in Section 4.1: Dataset of Choice.

## Dataset of Choice

To test the MLR and ELM methods, a dataset was used that was found online. The data was like what was used for the project that Walter P Moore completed. To analyze the data, Python was used, along with the Pandas and Matplotlib packages.

The data is taken from Spain. There are three rainfall measurements, and six discharge measurements. The Gramenet discharge lies below the rest of the measurements of interest in terms of elevation, therefore it is the variable of interest. This is because the water from the other discharges and the precipitation will contribute to the Gramenet discharge. The other explanatory variables will be used to forecast it.

Table 1 shows the data from the first four explanatory variables, and Table 2 shows the data from the final four explanatory variables and the output variable, the Gramenet discharge. They have been split so the tables remain readable. The data continues until December 31, 2010, and it can be seen in Appendix A: Original Dataset.

Table : First Four Explanatory Variables

| **Date** | **Barcelona\_Fabra Daily Rainfall [mm]** | **Barcelona Daily Rainfall [mm]** | **Sabadell\_Aero Daily Rainfall [mm]** | **Garriga Discharge [m^3]** |
| --- | --- | --- | --- | --- |
| 1/1/2003 | 0 | 0 | 0 | 0.254 |
| 1/2/2003 | 0 | 0 |  | 0.254 |
| 1/3/2003 | 0 | 0 |  | 0.246 |
| 1/4/2003 | 0 | 0 | 0 | 0.251 |
| 1/5/2003 | 18.1 | 16.4 | 2.1 | 0.241 |
| 1/6/2003 | 1 | 0.2 | 4.3 | 0.644 |
| 1/7/2003 | 0 | 0.5 |  | 0.317 |
| 1/8/2003 | 2.1 | 1.1 |  | 0.314 |
| 1/9/2003 | 0 | 0 |  | 0.309 |

Table : Final Four Explanatory Variables, and Output Variable

| **Date** | **Castellar Discharge [m^3]** | **Llica Discharge [m^3]** | **Montornes Discharge [m^3]** | **Mogoda Discharge [m^3]** | **Gramenet Discharge [m^3]** |
| --- | --- | --- | --- | --- | --- |
| 1/1/2003 | 0.0327 | 0.155 | 0.3985 | 0.12 | 3.6275 |
| 1/2/2003 | 0.0281 | 0.151 | 0.3745 | 0.106 | 3.9471 |
| 1/3/2003 | 0.0225 | 0.145 | 0.375 | 0.101 | 4.0174 |
| 1/4/2003 | 0.03 | 0.145 | 0.372 | 0.099 | 3.9316 |
| 1/5/2003 | 0.0328 | 0.146 | 0.3686 | 0.097 | 3.7913 |
| 1/6/2003 | 0.0559 | 0.205 | 1.0558 | 0.516 | 7.6332 |
| 1/7/2003 | 0.039 | 0.168 | 0.6 | 0.189 | 4.6152 |
| 1/8/2003 | 0.0386 | 0.163 | 0.519 | 0.159 | 4.7008 |
| 1/9/2003 | 0.0315 | 0.158 | 0.4775 | 0.138 | 4.4481 |

A sample graph of the Barcelona\_Fabra Daily Rainfall [mm] time series can be seen in Figure 2 below. The rest of the graphs for the input variables can be seen in Appendix B: Graphs of Independent Variables.

Chart

Description automatically generated

Figure 2: Barcelona\_Fabra Daily Rainfall

## Data Cleaning

As can be seen in Table 1 and Table 2, the data is not perfect as there are many missing values. It is important to understand what missing values mean, so they can be dealt with in a reasonable manner. These are easy to evaluate in Python. A heatmap can be used to show how many values are missing. This can be seen in Figure 3 below.

A picture containing chart

Description automatically generated

Figure 3: Heatmap of Missing Data

As can be see in Figure 3 above, the Castellar and Montornes measurements are missing a lot of data. As a result of this, they have been entirely removed from the dataset. This leaves the dataset with two less explanatory or input variables. The new data structure can be seen in Table 3 below.

Table : New Data Structure

| **Date** | **Barcelona\_Fabra Daily Rainfall [mm]** | **Barcelona Daily Rainfall [mm]** | **Sabadell\_Aero Daily Rainfall [mm]** | **Garriga Discharge [m^3]** | **Llica Discharge [m^3]** | **Mogoda Discharge [m^3]** | **Gramenet Discharge [m^3]** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1/1/2003 | 0.0 | 0.0 | 0.0 | 0.254 | 0.155 | 0.120 | 3.6275 |
| 1/2/2003 | 0.0 | 0.0 | NaN | 0.254 | 0.151 | 0.106 | 3.9471 |
| 1/3/2003 | 0.0 | 0.0 | NaN | 0.246 | 0.145 | 0.101 | 4.0174 |
| 1/4/2003 | 0.0 | 0.0 | 0.0 | 0.251 | 0.145 | 0.099 | 3.9316 |
| 1/5/2003 | 18.1 | 16.4 | 2.1 | 0.241 | 0.146 | 0.097 | 3.7913 |

Finally, any rows that included null data were removed. This resulted in 2076 data points that can be utilized to run the simulations. To properly run MLR and ELM simulations, complete datasets are key. Even though the dataset used was not complete initially, modifications were done to make the dataset complete and ready for evaluation.

# Application of and Analysis of MLR

MLR is not a technique that can be used on all datasets. For MLR to be effective, there needs to be a linear or consistent relationship between the explanatory or input variables, and the output variables. The reason for this is quite simple and can be demonstrated by analyzing Equation 1: MLR General Equation.

For example, if the output variable y is streamflow discharge, and the input variable x1 is the rainfall in a nearby basin, then as the rainfall increases, the discharge will also increase, assuming the coefficient for the input variable x1 is positive. Therefore, for MLR to be effective, the input variables need to affect the output variable in the same way throughout the dataset. This can be checked by plotting the input variables against the output variable. Figure 4 below shows the relationship between the input variable Barcelona\_Fabra and Gramenet Discharge variables. The Gramenet variable is in orange.

Chart

Description automatically generated

Figure 4: Barcelona\_Fabra Daily Rainfall vs Gramenet Discharge

Generally, as the input variable goes up, so does the output variable. This means that MLR will be applicable to this input variable. Overall, all six of the input variables have similar relationships to the output variable. This is shown in Appendix C: Graphs of Independent Variables vs Gramenet Discharge. Therefore, MLR can be applied to all the variables in question.

## Data Splitting

To apply MLR to a dataset, it is important to split the data. The reasoning for this is quite simple. An equation will be produced that allows the output variable to be forecasted using the input variable. It is important to think about how the equation can be tested. The best way is to have additional data where the input variables can create a forecast, and the forecast can be compared against the actual results. This can be done by splitting the existing data into two groups, which include 1. Training data, and 2. Test data. The training data can be used to generate the coefficients, and the test data can be used to check how effective the MLR equation was.

In Python and Pandas, it is easy to split the dataset. The following code was used, where x and y are the input and output variables that are being split, respectively.

x\_train, x\_test, y\_train, y\_test = train\_test\_split(

    x, y, test\_size=0.2, random\_state=101)

It should be noted that random state means that the results will be reproducible if the code is run again. The test size is the amount of data that is used for testing, so 20% of the data is used for testing and 80% is used for training.

## MLR Results

In Python, there are packages that allow MLR simulations to be easily completed. The following code was used to generate predictions:

# Create a linear regression model and train it using the training data.

model = LinearRegression()

model.fit(x\_train, y\_train)

predictions = model.predict(x\_test)

The model was trained based on the input matrix x\_train which contained independent variables, and a vector y\_train which contained dependent variables. The model was used to forecast the output variables for the set of data x\_test. To see how accurate the model was, it can be plotted as shown in Figure 5 below.

Chart, line chart, histogram

Description automatically generated

Figure 5: MLR Forecast Results

As the figure shows, the forecast is fairly accurate. The specific forecast can be seen in Appendix D.

The results can be further quantified by using two key metrics; 1. Root Mean Squared Error (RMSE), and 2. Mean Absolute Error (MAE). These have been computing using the Python package Sklearn and can be seen in the Table below.

|  |  |
| --- | --- |
| **MLR Metric Results** | |
| **MSE** | **MAE** |
| 2.383 | 0.851 |

# Application and Analysis of ELM

The application of ELM is based entirely off Section 3.1: ELM Mathematics. While performing the required math would be difficult without relevant tools, Python makes the application of ELM on data simple. This report will use a tutorial online that showed how ELM can be applied using common Python packages (You, 2023).

The data that will be used for training is the same as what was used to train the MLR algorithm. This is fully explained in Section 5.1 Data Splitting. Similarly, the same data that was used to test the MLR algorithm will be used to test the ELM algorithm.

It should be noted that the number of hidden nodes is a variable that must be specified. This defined several the mathematical quantities that have been outlined in Section 3.1: ELM Mathematics. To come up with the most effective number of hidden nodes, a loop was created that ran through zero hidden nodes, to 100 hidden nodes. After this was done, it was determined that 62 hidden nodes resulted in the most accurate forecast. The code is included in Appendix D: Python Code.

## ELM Results

The results of the ELM algorithm are visualized below in Figure x. The RMSE and MAE metrics are shown below in Table x.

Chart, histogram

Description automatically generated

|  |  |
| --- | --- |
| **ELM Metric Results** | |
| **MSE** | **MAE** |
| 1.908 | 0.874 |

# Conclusion and Recommendations

From the literature, it is clear data driven forecasting methods have a place in hydrology. This paper outlines and applies two data driven approaches to hydrological forecasting. Both methods were effective in creating a coherent forecast that to the human eye, does a reasonable job when compared to the actual results.

The pros and cons of different machine learning models should be considered by anyone looking to use them for a numerical application. A few things that should be considered include 1. Data complexity, 2. Required speed, 3. Data correlation, and 4. Desired computational efficiency. Once a model is chosen, it can be used to create discharge forecasts that will allow the structural integrity of something like a dam to be quantified.

# References

Include Kaggle data

AlOmar, M. K., Hameed, M. M., Al-Ansari, N., & AlSaadi, M. A. (2020). Data-driven model for the prediction of total dissolved gas: Robust Artificial Intelligence Approach. *Advances in Civil Engineering*, *2020*, 1–20. https://doi.org/10.1155/2020/6618842

Yaseen, Z. M., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J., & El-Shafie, A. (2016). Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. *Journal of Hydrology*, *542*, 603–614. https://doi.org/10.1016/j.jhydrol.2016.09.035

Ye, A. (2020, December 12). *Some call it genius, others call it stupid: The most controversial neural network ever created*. Medium. Retrieved April 19, 2023, from https://towardsdatascience.com/some-call-it-genius-others-call-it-stupid-the-most-controversial-neural-network-ever-created-2224ed22795a

You, J. (2023, February 11). *Extreme learning machine in 3 lines of code*. John You. Retrieved April 19, 2023, from https://johnswyou.github.io/post/elm-in-3-lines/

# Appendix A: Original Dataset

# Appendix B: Graphs of Independent Variables

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# Appendix C: Graphs of Independent Variables vs Gramenet Discharge

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# Appendix D: Python Code