Master's Thesis of Engineering

Prediction of Metal Removal Efficiency of Passive Treatment Systems for Acid Mine Drainage – Application of Linear Regression, Random Forest, and Artificial Neural Networks

광산배수 자연정화처리시설의 금속 제거효율 예측 연구 – 선형 회귀, 랜덤 포레스트, 인공 신경망 모델의 적용

February 2023

Graduate School of Engineering

Seoul National University

Energy Systems Engineering

Kyeongmin Kim

Prediction of Metal Removal Efficiency of Passive Treatment Systems for Acid Mine Drainage – Application of Linear Regression, Random Forest, and Artificial Neural Networks

광산배수 자연정화처리시설의 금속 제거효율 예측 연구 – 선형 회귀, 랜덤 포레스트, 인공 신경망 모델의 적용

-

지도 교수 정 은 혜

이 논문을 공학석사 학위논문으로 제출함

2023년 2월

서울대학교 대학원

에너지시스템공학부

김 경 민

김경민의 공학석사 학위논문을 인준함

2023년 2월

위 원 장 박 해동

부위원장 정 은 혜

위 원 THOMAS PABST

Abstract

Acid mine drainage (AMD) is a global problem due to the high content of heavy metals and low pH and needs to be monitored and managed by reclamation or treatment systems. The performance of AMD treatment systems is difficult to predict due to the numerous factors associated. Empirical and geochemical models have been developed to predict the AMD treatment. The machine learning-based access can be an alternative when the amount of data and time required to build models are limited. In this study, random forest (RF) and artificial neural network (ANN) model were constructed for predicting the Fe(II) and Mn removal efficiencies of passive systems in 9 abandoned coal mines and compared to the performance of multiple linear regression (MLR) model. Among the three models, the RF model showed the best performance in both predicting the Fe(II) and Mn removal efficiency. According to the sensitivity analysis, the pH of the inflow water, the Fe(II) concentration of the inflow water, and the alkalinity were the most important variables for predicting the Fe(II) removal efficiency. The alkalinity of the inflow water and the pH of the inflow water were important variables to predict Mn removal efficiency.

Table of Contents

Chapter 1. Introduction **1**

1.1. Research Background **1**

1.2. Research Objective **6**

Chapter 2. Background Theory **7**

2.1. Passive Treatment Systems **7**

2.1.1. Oxidation Ponds (OPs) **8**

2.1.2. Aerobic Wetlands (AeWs) **9**

2.1.3. Successive Alkalinity-Producing Systems (SAPS) **10**

2.2. Predictive Models **11**

2.2.1. Multiple Linear Regression (MLR) **11**

2.2.2. Random Forest (RF) **12**

2.2.3. Artificial Neural Networks (ANN) **14**

Chapter 3. Methodology **21**

3.1. Data Description **21**

3.2. Data Sampling **24**

3.3. Setting Variables **25**

3.4. Correlation Analysis **26**

3.5. Data Split **27**

3.6. Model Construction 28

3.6.1. Multiple Linear Regression 28

3.6.2. Random Forest 29

3.6.3. Artificial Neural Networks 31

3.7. Model Evaluation 32

3.8. Variable Importance 33

3.8.1. Random Forest 33

3.8.2. Artificial Neural Network 34

Chapter 4. Result 35

4.1. Data Summary 35

4.2. Correlation Analysis 37

4.2.1. Variables of Fe(II) Dataset 37

4.2.2. Variables of Mn Dataset 39

4.3. Optimization of MLR model 41

4.4. Comparison of Removal Efficiency Prediction 41

4.4.1. Train Dataset Prediction 44

4.4.2. Test Dataset Prediction 47

4.4.2.1. Prediction of Fe(II) Removal Efficiency 47

4.4.2.2. Prediction of Mn Removal Efficiency 51

4.5. Variable Importance 55

4.5.1. Variable Importance in FRE Prediction 55

4.5.2. Variable Importance in MRE Prediction 57

Chapter 5. Discussion **59**

Chapter 6. Conclusion **62**

References **63**

Abstract in Korean **76**

List of Figures

Figure 2.1. The schematic diagram of general passive treatment system in Korea (adapted from Ji et al., 2008).....7

Figure 2.2. The structure of RF model.....13

Figure 2.3. The structure of ANN model with backpropagation.....15

Figure 2.4. Non-linear activation functions.....18

Figure 4.1. Correlation analysis of FRE and explanatory variables.....38

Figure 4.2. Correlation analysis of MRE and explanatory variables.....40

Figure 4.3. Prediction of FRE in train dataset.....45

Figure 4.4. Prediction of MRE in train dataset.....46

Figure 4.5. Prediction of FRE in test dataset.....48

Figure 4.6. Prediction of MRE in test dataset.....52

Figure 4.7. Variable importance in FRE prediction.....56

Figure 4.8. Variable importance in MRE prediction.....58

List of Tables

Table 2.1. Equations of activation functions.....19

Table 3.1. Drinkable water quality guidelines of KME.....23

Table 3.2. Data description of 9 coal mines.....24

Table 3.3. The minimum and maximum values, and intervals of hyperparameters in RF model.....30

Table 3.4. The optimal hyperparameters in RF model.....30

Table 3.5. The minimum and maximum values, and intervals of hyperparameters in ANN model.....31

Table 3.6. The optimal hyperparameters in ANN model.....31

Table 4.1. Summary of 199 datasets for FRE prediction.....36

Table 4.2. Summary of 132 datasets for MRE prediction.....36

Table 4.3. Optimization of MLR model for predicting FRE.....42

Table 4.4. Optimization of MLR model for predicting MRE.....43

Table 4.5. Model performances of FRE prediction.....47

Table 4.6. Observed and predicted FRE in test dataset.....49~50

Table 4.7. Model performances of MRE prediction.....51

Table 4.8. Observed and predicted MRE in test dataset.....53~54

Chapter 1. Introduction

1.1. Research Background

Acid mine drainage (AMD) is a global problem according to the amount of tailings left behind after mining or mineral processing (Matlock et al., 2002; Akcil et al., 2006; McCarthy, 2011; Rezaie and Anderson, 2020). It negatively affects the local ecology when discharged into nearby water systems because of the heavy metals present and the low pH (Lopes et al., 1999; Lei et al., 2010; Equeenuddin et al., 2013). Especially, heavy metals are dangerous due to their persistence, potential toxicity, and ability to accumulate in many body parts (Gray, 1997; Duruibe et al., 2007; Briffa et al., 2020).

AMD is mainly generated by the oxidation of pyrite (FeS_2), which is one of the main minerals in coal and metal ore deposits (Qureshi et al., 2016). Pyrite is oxidized into metallic hydroxides, sulfates, and acid (Eq. (1.1)). It can also be oxidized by ferric ion (Fe3+) as an oxidant when pH is less than 3.5 (Eq (1.2)). AMD may also be produced by other sulfide minerals, such as pyrrhotite (FeS) and chalcopyrite (Cu_2S) (Akcil and Koldas, 2006, Simate and S.Ndlouv., 2014).

$$FeS_2\ (s)+15/4O_2+7/2H_2O \rightarrow Fe(OH)_3\ (s)+2SO_4^{2-}+4H^+ \quad (1.1)$$

$$FeS_2\ (s)+14Fe^{3+}+8H_2O \rightarrow 15Fe^{2+}+2SO_4^{2-}+16H^+ \quad (1.2)$$

The main objective of AMD treatment is to achieve water quality regulation by raising pH and removing heavy metal compounds (Seervi et al., 2017). AMD treatment approaches are classified as active and passive treatment. Active treatments are based on chemical and physical processes such as pH control, absorption or adsorption, and electrochemical concentration. They can be done with portable equipment or in-situ facilities such as high-density sludge, and biological reactor systems (Taylor et al., 2005; Johnson and Hallberg, 2005; Saha and Sinha, 2018).

Passive treatment has been adopted due to low-cost operation, little maintenance, and energy consumption for the rehabilitation of abandoned mine sites (Gazea et al., 1996; Zipper and Skousen, 2014; Clyde et al., 2016). Passive treatment systems are based on chemical and biological techniques to establish reducing conditions which facilitate the precipitation of metal sulfides. Chemical systems involve neutralizing acidic water with alkaline materials such as limestone and steel slag (Skousen et al.,2019). Biological systems utilize several mechanisms including bio-catalyzed oxidation of Fe and Mn, alkali generation by microbiological reduction, and eliminating metals through adsorption and exchange processes interacted with organic substances (Skousen et al., 2017).

Passive treatment systems include aerobic wetlands (AeW), anaerobic wetlands (AnW), vertical flow wetlands (VFW), successive alkalinity producing systems (SAPS), oxic limestone drains (OLD), limestone diversion wells (LDW), permeable reactive barriers (PRB), limestone leach beds (LLB), steal leach beds (SLB), electrochemical covers, and Gas Redox and Displacement Systems (GaRDS) (Taylor et al., 2005).

Selecting suitable AMD treatment systems is affected by numerous important factors such as site characteristics, components, and volumes, and flow rate of discharged water (Hyman and Watzlaf, 1995). These factors influence the standard, effectiveness, and life expectancy of treatment systems and the site feature and removal of heavy metals (Zipper and Skousen, 2014; Seervi et al., 2017). Lab and field tests are frequently used to predict AMD treatments, but the limitations of small-scale and short duration, as well as uncertainties when applying the extrapolated results into real mine sites, must be overcome (Gibert et al., 2002; Heviánková et al.,2014; Igarashi et al.,2020). Building predictive models can be an alternative to addressing the aforementioned issues (Betrie et al., 2013; Cravotta, 2021).

Researchers have tried to find suitable models for predicting the treatment of acid mine drainage over 20 years (Foos, 1997; Amos et al., 2004; Andalaft et al., 2018). First, empirical modeling based on experimental and statistical data was developed. Zipper and Skousen (2010) predicted the alkalinity generation of 5 passive treatment systems AWs, ALDs, VFs, OLCs, and LLBs as functions of water loading and influent acidity. There was a limitation of empirical models that they displayed substantial deviation from actual performance in the case of higher

influent acidity loadings.

There have been several attempts to assess treatment using geochemical modeling. For example, PHREEQC and Minteq.v4 were used for simulating chemical components and neutralizing AMD (Koide et al., 2012). "AMDTreat" was used to calculate the volume of sludge after treatment, the amounts of dissolved metals, and the amount of chemicals needed to reach a target pH (Cravotta et al., 2010). Similar studies were progressed using other tools: AMDTreat 5.0 and PHREEQC (Cravotta et al., 2015) and PHREEQ-N-AMDTreat (Cravotta et al., 2021). PHREEQC was also used for simulating the precipitation of Fe and Al by the remediation process of mixing and neutralization (Nordstrom, 2020). However, the application of geochemical modeling is limited when complex mechanisms that control water composition are unknown and a large amount of precise data, such as mineralogical and hydrological data, is lacking (Nordstrom, 2020; Chen et al., 2020).

Machine learning could be another promising approach for making predictive models because of the good performance with insufficient data, noise insensitivity, and accurate error measurement by avoiding overfitting (Auria and Moro, 2008; Betrie et al., 2013; Yaseen, 2021). There were previous studies applying machine learning-based models for heavy metal removal prediction. For example, an artificial neural network (ANN) model based on the sulfate, COD, alkalinity, and sulfide was employed for predicting the performance of fludized bed reactor (FBR) (Atasoy et al., 2013). Lu et al. (2022) constructed a random forest

(RF) model for predicting the removal of heavy metals using chitosan-based flocculants (CBFs) based on flocculant properties, flocculation conditions, and heavy metal properties. However, the majority of the existing research has focused on analyzing and predicting the performance of a single treatment facility. It is challenging to directly apply the previous prediction models to the cases in Korea because most of AMD treatment facilities consist of 2 or 3 different passive treatment systems (Ji et al., 2008). Therefore, this study attempted to develop machine learning-based predictive models that can predict the metal removal efficiency of AMD treatment facilities in Korea.

1.2. Research Objective

The objective of this study is to compare three predictive models Multiple Linear Regression (MLR), Random Forest (RF), Artificial Neural Network (ANN) and finding the optimal model for predicting the metal removal efficiency of passive systems for acid mine drainage. The mine drainage data were provided from KOMIR (Korea Mine Rehabilitation and Mineral Resources Corporation), and drainage data of 9 mines recorded from 2010 to 2019 were used in this study. The research procedure is as follows. To begin, eight explanatory factors relating to inflow drainage were chosen to explain removal efficiency. Second, three models were constructed, and to avoid overfitting, the optimal hyperparameters of each model were obtained through the tuning process using GridsearchCV of scikit-learn package. Third, the performance of each model was assessed by RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MSE (Mean Squared Error), and determination coefficient R^2. Lastly, the importance of variables in RF and ANN model was calculated and compared. The prediction models and key variables in the study could be used as indicators for the construction and performance evaluation of treatment systems.

Chapter 2. Background Theory

2.1. Passive Treatment systems

Most of AMD treatment facilities in Korea usually consisted of Successive alkalinity-producing systems (SAPS), aerobic wetlands, and oxidation ponds (Fig 2.1).

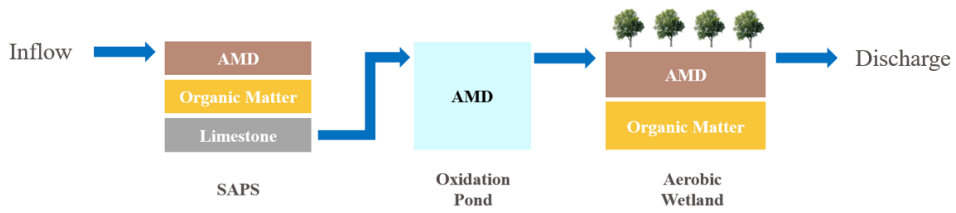


Figure 2.1. The schematic diagram of general passive treatment system in Korea (adapted from Ji et al., 2008)

2.1.1. Oxidation Ponds (OPs)

The main purpose of OPs is to reduce Fe^{2+} with dissolved oxygen to be precipitated as ferric hydroxides by holding acid mine drainage (Lee et al., 2013). Since the rate of Fe^{2+} oxidation in OPs is affected by many factors such as the pH of inflow drainage, the flow patterns, and retention times, the design of OPs should be determined considering the influence of aforementioned variable through the experiments (Ji et al., 2008; Lee and Cheong, 2016)

According to the design guidelines for OPs from the National Coal Board, it was recommended that an oxidation pond have a capacity of 1 L/sec per $100m^2$ and the theoretical holding time was defined as 48 h (Laine and Jarvis, 2003; Lee et al., 2013)

2.1.2. Aerobic Wetlands (AeWs)

AeWs are utilized to hold AMD for the purpose of $\mathbf{Fe^{2+}}$ oxidation and the precipitation of metal hydroxides (Skousen et al., 2017). They are usually shallow basins with wetland plants such as *Typha latifolia* to make wildlife habitats and aesthetics, regulate uniform flow, stabilize the metal precipitation, and maintain the microbial population (Abhishek et al., 2015).

AeWs remove metals by making the AMD flow slow and allowing for $\mathbf{Fe^{2+}}$ oxidation. Fe3+ resulting from the $\mathbf{Fe^{2+}}$ oxidation precipitates as ferric hydroxide and makes other metals co-precipitated, but the metal removal is effective when the pH of inflow drainage is over 6 (Skousen and Ziemkiewicz, 2005; Abhishek et al., 2015). Therefore, alkali materials have to be added if the water is not net-alkaline. Additionally, AeWs should always be connected with other passive treatment systems such as anoxic limestone drains or SAPS in order to get alkaline drainage from them (Skousen et al., 2017).

Mn oxidation, which is slower than the $\mathbf{Fe^{2+}}$ oxidation, is associated with the presence of $\mathbf{Fe^{2+}}$. Since $\mathbf{Fe^{2+}}$ inhibits and counteracts the Mn oxidation, Mn precipitation generally occurs after all of the Fe has been removed (Wildeman et al. 1993; Skousen et al., 2017). Hedin et al., (1994) calculated the removal rates of AeWs as 10–20 g $\mathbf{m^{-2}day^{-1}}$ for Fe and 0.5–1.0 g $\mathbf{m^{-2}day^{-1}}$ for Mn.

2.1.3. Successive Alkalinity-Producing Systems (SAPS)

Kepler and McCleary (1994) developed SAPS to make up for the limitations of using wetlands or limestone drains (ALDs) exclusively. For instance, the ALDs' capacity to produce alkalinity is limited when AMD includes dissolved oxygen. Wetlands have limitations of slow treatment and need for large treatment areas, and the effectiveness is directly influenced by environmental conditions such as rainfall (Kepler and McCleary, 1994; Ordonez et al., 2012). SAPS overcame the limitations of wetlands and limestone drains by adopting only the advantages of both systems (Kepler and McCleary, 1994).

The process of SAPS treatment is as follows; First, ponded AMD flows downstream through the about 0.5 - 1 m organic layer, where dissolved oxygen is eliminated and alkalinity is generated by the bacterial reaction. Second, drainage flows downstream through the about 0.5 - 1m limestone bed. Since dissolved oxygen has been removed in the previous organic layer, alkalinity can be produced successively. Finally, the effluent goes out of the system at the bottom of the limestone bed (Ji et al., 2008; Ordonez et al.,2012).

2.2. Predictive Models

In this study, models for predicting removal efficiency were built using three different approaches: multiple linear regression, random forest, and artificial neural networks.

2.2.1. Multiple Linear Regression (MLR)

The MLR is generally used to determine the relations between the dependent variable and the independent variables, which is a straightforward regression model expressed using linear combinations between dependent and independent variables. The general form of a multiple linear regression model is as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where Y represents the dependent variable, β_0 the intercept $x_1 \sim x_n$ are the independent variables, $\beta_1 \sim \beta_n$ are the coefficients for the independent variables, and ϵ is the estimation error. The MLR model is optimized by minimizing the sum of errors between observed and predicted values (Seber and Lee, 2012)

2.2.2. Random Forest (RF)

Random Forest (RF) is an ensemble machine learning model that has been utilized for creating predictive models in numerous studies since it was first proposed by Breiman (2001). It is suitable for modeling the nonlinear combination of variables because of the ability to manage complicated interactions and resistance to multicollinearity (Breiman, 2001, Biau and Scornet, 2016).

RF is an ensemble model of K decision trees $\{T_1(X),\ T_2(X), \cdots,\ T_K(X)\}$, where X is the set of independent variables $X = \{x_1,\ x_2,\ \cdots,\ x_p\}$ (Fig 2.1). RF model is constructed by the bootstrap aggregating (bagging), which is a technique that prevents the correlation of distinct trees by generating K tree samples from randomized subset of the train dataset $\mathcal{L}$. A total of K predictions are made by tree samples, and they are averaged to determine the output $\hat{y}$ of the RF model (Svetnik et al., 2003; Biau and Scornet, 2016).

In RF model, about one-third of the data in the samples are not used to create each decision tree because of the randomized bootstrap sampling from the original train dataset. The whole set of excluded data is called as out-of-bag dataset (OOB dataset). In contrast, data used for growing trees are referred to as In-Bag dataset. Generated OOBs can be used to estimate the error of RF model instead of applying the time-consuming cross-validation approach (Breiman, 2002; Archer and Kimes, 2008)

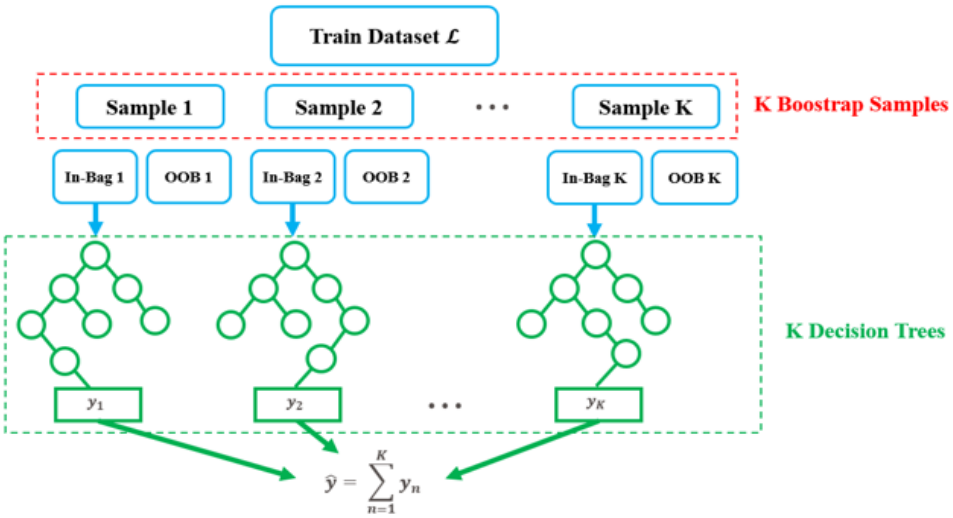


Figure 2.2. The structure of RF model.

2.2.3. Artificial Neural Networks (ANN)

An artificial neural network (ANN) is a model for data processing that draws inspiration from the operations and analyses of the human brain (Hopfield, 1988). ANN has been used for modeling and analysis in many fields, several researchers also have used ANN to predict the chemistry of acid mine drainage; i.e. the concentrations of heavy metals. (Rooki et al., 2011, Ajayi et al., 2021, Mariem et al., 2022).

In particular, a deep neural network (DNN) is an artificial neural network with two or more hidden layers (Fig 2.3). Deep learning is the process of learning this deep neural network (Hinton et al., 2012, Lecun et al., 2015). DNNs can express more sophisticated models and have a greater computing capacity than earlier networks because of the complex linking of layers (Abiodun et al., 2018).

ANN is comprised of one input, one or more hidden, and one output layer in general. The input layer has nodes containing the values of independent variables, and it transmits the input data to the following hidden layer. In one or more hidden layers, the input data is transferred to 2~3 times as many nodes as those in the input layer. The data is then abstracted and transmitted to the output layer. The output layer is where the predicted dependent variables are produced from the abstracted data (Hopfield, 1988).

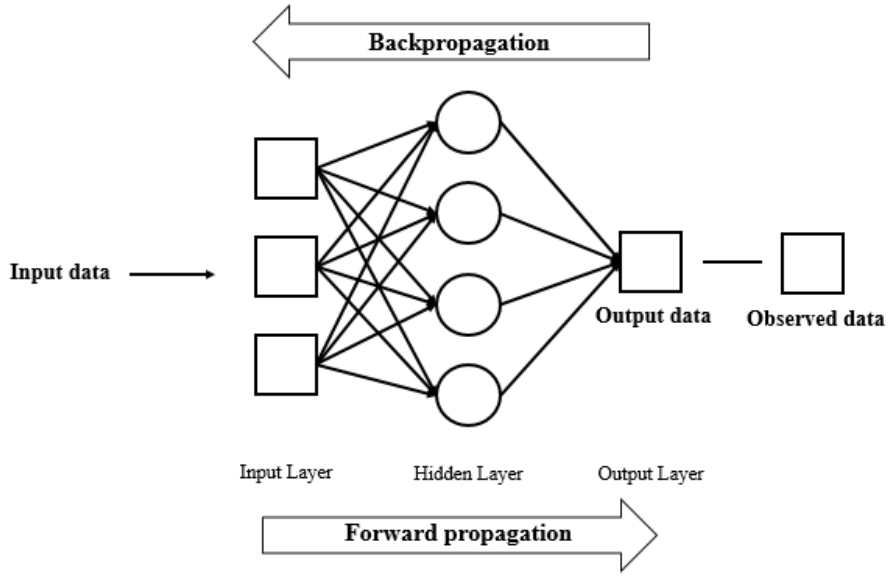


Figure 2.3. The structure of ANN model with backpropagation.

In this study, ANN with forward propagation of data and back propagation of error, also called back-propagation neural network (BPNN), was trained (Fig 2.3). BPNN is a commonly applied neural network since it can be autonomously optimized by minimizing errors using the gradient descent technique (Ruder, 2016).

1) Forward propagation of data

When the number of independent variables in input layer is i, the data in the input layer $X_i = [x_1, x_2, \cdots, x_i]$ will be forwardly propagated as net_1, adding the first bias vector $b^1 = [b_1^1, b_2^1, \cdots, b_i^1]$ to the dot product of the input and the first weight vector $w_{i,h_1} = [w_{i,1}^{h_1}, w_{i,2}^{h_1}, \cdots, w_{i,i}^{h_1}]^T$ (Eqn (2.1)).

Then, an activation function f is applied to make the neural network cope with the complex data and to learn the non-linear mappings between independent and dependent variables (Eqn (2.2)) (Sharma et al., 2017).

$$net_1 = w_{i,h_1} \cdot X_i + b_1 = \sum_{n=1}^{i} (x_n w_{i,n}^{h_1} + b_n^1) \quad (2.1)$$

$$h_1 = f(net_1) \quad (2.2)$$

The processes of data propagation between the hidden layers and between the hidden layer and the output layer are in the same way. The only difference is that activation functions such as Sigmoid, Tanh, and ReLU function (Table 2.1, Fig 2.4) are applied to pass the data between hidden layers, while a linear function is applied between the last hidden layer and the output layer. ReLU function was selected as an activation function in this work since it can deal with the gradient vanishing problem of the Sigmoid and Tanh function (Nair and Hinton, 2010; Ide and Kurita, 2017)

When the number of the hidden layers is l, the vector of the first hidden layer h_1 and the vector of k-th hidden layer h_k, and the vector of predicted output layer Y_o are described as Eqn (2.3) and Eqn (2.4), respectively.

$$h_k = f\left(w_{h_{k-1},\, h_k} \cdot h_{k-1} + b_k\right) \tag{2.3}$$

$$Y_o = w_{h_k,\, o} \cdot h_l + b_{l+1} \tag{2.4}$$

where $w_{h_{k-1},\, h_k}$ is the weight vector connecting (k-1)th and k-th hidden layer and $w_{h_k,\, o}$ is the weight connecting the last hidden layer and the output layer.

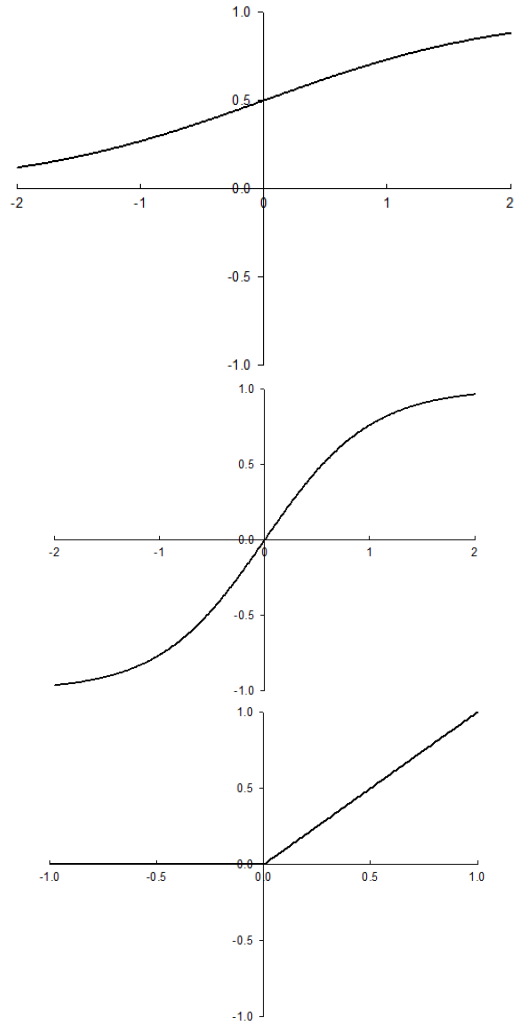


Figure 2.4. Non-linear activation functions (a) Sigmoid. (b) Tanh. (c) ReLU.

Table 2.1. Equations of activation functions

Activation Function	Equation
Sigmoid	$a = \frac{1}{1 + \exp(-z)}$
Tanh	$a = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)}$
ReLU	$a = \begin{cases} 0, & \&z < 0 \\ z, & \&z \geq 0 \end{cases}$

2) Backpropagation of error

Following the generation of the output layer Y_o by feed-forward propagation, the neural network is trained via backpropagation. Backpropagation is the process of using the gradient descent method to update and optimize the weights and biases (Mitchell, 1997). The weights and biases of each connection are modified until the loss function based on the mean squared error (MSE) has been minimized (Eqn (2.5)).

$$C = \sum_{k=1} \frac{1}{2}(y_k - y_{o,k})^2 \qquad (2.5)$$

where y_k and $y_{o,k}$ indicate the *i*-th label of observed output vector Y and the predicted output vector Y_o, respectively.

The loss function and the gradient of it can be expressed as the combination of using all possible weight vector and bias values (Eqn (2.6)). Starting with the initial random vectors, the weight vector and bias with m_k nodes connecting the (*k-1*)-th and *k*-th hidden layer are updated by subtracting a component of the gradient (Eqn (2.7), (2.8)).

$$\nabla C\left(w_{h_{k-1},\ h_k}\right) = \left[\frac{\partial C}{\partial w^{1}_{h_{k-1},\ h_k}}, \frac{\partial C}{\partial w^{2}_{h_{k-1},\ h_k}}, \cdots, \frac{\partial C}{\partial w^{m_k}_{h_{k-1},\ h_k}}\right] \tag{2.6}$$

$$w_{h_{k-1},\ h_k} \leftarrow w_{h_{k-1},\ h_k} - \alpha \nabla C\left(w_{h_{k-1},\ h_k}\right) \tag{2.7}$$

$$b_k \leftarrow b_k - \alpha \frac{\partial C}{\partial b_k} \tag{2.8}$$

where α stands for learning rate which adjusts the optimization rate for avoiding significant deviation of the weight vector and bias due to the large gradient. This procedure is repeated over a significant number of training cycles, or "epochs," until the model converges and the MSE falls below a preset threshold (Mitchell, 1997).

Chap 3. Methodology

3.1. Data description

The mine drainage data of abandoned 123 coal and 128 metal mines in Korea used in this study were obtained from Korea Mine Rehabilitation and Mineral Resources Corporation(KOMIR), Wonju, Korea. They consist of fundamental information about each mine, survey details, treatment systems, number of elements exceeding the water quality standard, and the classification based on wastewater ordinance.

1) Fundamental information about each mine

Fundamental information includes the name, the type of mine, regional information and water system information, and the management number. The regional and water systematic information includes cities, provinces, counties, water system areas, water system main streams, and water system tributaries.

2) Survey details

Survey details include the survey period, the survey year, the survey date, and the survey point. The survey was conducted twice or four times per year. Surveys were performed once in the spring and once in the fall for mines that

were surveyed twice a year. Mines that were surveyed four times a year were investigated on a quarterly basis. For example, the Poongwon mine, which had the most data, recorded a total of 40 data from 2010 to 2019. The subjects of the survey included samples from several locations, including mine drainage, leachate, inflow drainage, and discharge drainage. Especially, water quality analyses of mines with equipped treatment systems were performed on inflow and discharge drainage, which meant drainage before and after passing through the treatment systems, respectively.

3) Treatment systems

Treatment systems were available in 36 coal mines and 8 metal mines out of the total data. Depending on the level of contamination in each mine, the treatment systems were classified as active, semi-active, or natural. In summary, in coal mines, 5 active, 4 semi-active, and 27 natural treatment systems were applied, while in metal mines, 1 active, 4 semi-active, and 3 natural treatment systems were applied.

4) Measured variables

Measured variables consist of flow rate (m^3/d) , temperature (℃), dissolved oxygen (mg/L), electro-conductivity (mS/cm), turbidity (NTU), alkalinity (mg/L), concentration of metals and non-metals (mg/L).

The concentrations of metals and non-metals were based on the result of ICP analysis and element types were as follows: Metals consist of Fe(II), Cr(VI), CN, Fe, Mn, Al, As, Cd, Cr, Cu, Pb, Zn, Hg, Ni, Na, Ca, K, Mg, and non-metals consist of F, Cl, SO_4. The number of polluted elements was recorded by counting the elements that exceeded the Korean Ministry of Environment (KME) established drinkable water quality guidelines (Table 3.2).

Table 3.1. Drinkable water quality guidelines of KME.

Type	Concentration(mg/L)	Type	Concentration(mg/L)
Cr^{6+}	< 0.10	Cr	< 0.50
CN	< 0.20	Cu	< 1.00
Fe(II)	< 2.00	Pb	< 0.10
Fe	< 2.00	Zn	< 1.00
Mn	< 2.00	Hg	< 0.10
Al	< 2.00	Ni	< 0.10
As	< 0.05	F	< 3.00
Cd	< 0.02		

3.2. Data Sampling

The specific data required for analysis were sampled by two criteria. First, mines that had been analyzed four times a year from 2010 to 2019 were re-extracted. Second, mines using treatment systems consisting of a combination of aerobic wetland (AeW), oxidation pond (OP), or SAPS were extracted to predict the performance of passive treatment systems. As a result, 9 coal mines (ST, DS, SS1, SS2, PW, HY, HU, HN, and GJ) were selected for analysis.

Table 3.2. Data description of 9 coal mines.

Name	pH range	Treatment systems	Main Pollutants
ST	2.59 ~ 4.16	SAPS – OP– AeW	**Fe, Mn**, Al, Zn, Al, Ni
DS	2.42 ~ 4.62	SAPS – OP	**Fe, Mn**, Al, Ni
SS1	2.58 ~ 4.47	SAPS – OP – AeW	**Fe, Mn**, Al, Ni
SS2	6.17 ~ 9.05	OP – SAPS – AeW	**Fe**
PW	6.24 ~ 7.57	SAPS – AeW	**Fe**
HY	5.91 ~ 7.52	SAPS – AeW	**Fe**
HU	6.10 ~ 8.08	OP – SAPS – AeW	**Fe, Mn**
HN	5.51 ~ 8.02	OP – SAPS – AeW	**Fe, Mn**, Ni
GJ	5.19 ~ 8.22	OP – SAPS – AeW	**Fe, Mn**

3.3. Setting Variables

All measured variables were used as explanatory variables to predict the removal efficiency (RE). The variables and their abridged terms are as follows:

- IM (Metal concentration in the inflow drainage)
  - IF (Fe(II) concentration in the inflow drainage)
  - IM (Mn concentration in the inflow drainage)
- IpH (pH of inflow drainage)
- FR (Flow rate of the inflow drainge)
- EC (Electro-conductivity of inflow drainage)
- Tur (Turbidity of inflow drainage)
- Alkal (Alkalinity of inflow drainage)
- DO (Dissolved Oxygen of the inflow drainage)
- Temp (Temperature of the inflow drainage)

The removal efficiency of each metal by the treatment systems was calculated by $(C_i - C_o)/C_i$, where C_i, C_o indicate the metal concentration of inflow water and discharge water, respectively. The Fe(II) and Mn removal efficiency (FRE and MRE) were predicted because the main pollutants in the mine used in the analysis were Fe(II) and Mn (Table 3.3).

3.4. Correlation Analysis

Calculating the Pearson correlation coefficient, which shows the linear dependence between two sets of data, is one of the most used techniques for correlation analysis (Dutilleul et al., 2000; Mansson et al., 2004). The Pearson correlation coefficient (r_{xy}) of two variables $X = \{x_1, x_2, \cdots, x_i\}$ and $Y = \{y_1, y_2, \cdots, y_i\}$ means the normalized covariance, which is calculated using covariance and two standard deviations (Eqn (3.1)).

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (3.1)$$

$\bar{x}$, $\bar{y}$: the mean value of elements in each set X and Y

The Pearson correlation coefficient measures the strength of the linear relationship between two variables. The sign of the correlation coefficient can be either positive or negative, denoting that the two variables are directly or inversely related, respectively. The linear relationship between two variables becomes stronger as the absolute value of r_{xy} approaches 1.

3.5. Data Split

The total datasets were 360 for the selected 9 mines. Then, only datasets in which the values of all explanatory variables were completely recorded and whose metal concentration exceeded the drinkable water quality standard were used for analysis. As a result, 199 datasets were utilized for predicting Fe(II) removal efficiency and 132 datasets were utilized for predicting Mn removal efficiency. Finally, each dataset was randomly split into a train dataset (70%) and a test dataset (30%).

3.6. Model Construction

3.6.1. Multiple Linear Regression

The multiple linear regression model was constructed in the forward stepwise method to find the optimal combination of variables among the eight variables IM, IpH, FR, EC, Tur, Alkal, DO, and Temp using the LinearRegression in linear_model package of Scikit-learn. The objective of the forward stepwise method is to minimize the number of independent variables in the regression model while maximizing the model performance (Alicja, 2015). The forward stepwise method was carried out based on the determination coefficient R^2 between the observed and predicted values of datasets. The optimization process is as follows:

1) The determination coefficient R^2 of the model was evaluated by successively adding the variables that had the highest absolute value of Pearson correlation coefficient with the metal removal efficiencies.

2) If R^2 decreased for the test dataset, the addition of variables was stopped and the optimal linear model was selected.

3.6.2. Random Forest

Machine learning (ML) based models such as random forest and neural networks require several hyperparameters that must be set before learning. The model performance is significantly influenced by the hyperparameters (Probst et al., 2019).

In RF model, the number of bootstrapped samples (n_estimators) and the maximum depth of each decision tree (max_depth) were set as the hyperparameters to find the optimal condition (Table 3.4).

The optimal hyperparameters were selected using the GridSearchCV in the model_selection package of Scikit-learn (Fabian et al., 2011). Grid search is a process of selecting the models and hyperparameters on a specified parameter grid by validating all combinations. Grid search is generally performed in conjunction with k-fold cross validation. In this study, the 5-fold cross validation method was used and the R^2 values from 5 iterations were averaged to estimate the performance (Fig 3.1). The hyperparameters recorded in Table 3.5 were the optimal conditions.

Table 3.3. The minimum and maximum values, and intervals of hyperparameters in RF model.

Hyperparameter	Min	Max	Interval
n_estimators	100	1000	100
max_depth	10	50	10

Table 3.4. The optimal hyperparameters in RF model

Model	**n_estimators**	**max_depth**	**Score (mean R^2)**
FRE	500	20	0.8112
MRE	300	20	0.8442

3.6.3. Artificial Neural Networks

For building an ANN model, the number of nodes in each hidden layer and epoch was set as hyperparameters to determine optimal values through the tuning process. The optimal hyperparameters of the ANN model were also selected using the GridSearchCV in the model_selection package of Scikit-learn. The hyperparameters recorded in Table 3.7 were the optimal conditions.

Table 3.5. The minimum and maximum values, and intervals of hyperparameters in ANN model.

Hyperparameter	Min	Max	Interval
Node of 1st Hidden Layer	15	50	5
Node of 2nd Hidden Layer	15	50	5
Epochs	100	500	100

Table 3.6. The optimal hyperparameters in ANN model

Model	**Node of 1st Hidden Layer**	**Node of 2nd Hidden Layer**	**Epochs**	**Score (mean R^2)**
FRE	30	30	400	0.6281
MRE	20	20	300	0.7458

3.7. Model Evaluation

The MAE, MSE, RMSE, and R^2 were used to evaluate and compare the performance of the prediction models. MAE, MSE, and RMSE are parameters based on the difference between the observed and predicted values, and low values mean the accurate prediction of models. The determination coefficient R^2 is a parameter for assessing the similarity between the observed and predicted values. The MAE, MSE, RMSE and R^2 values are defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_p\right| \tag{3.2}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_p\right)^2 \tag{3.3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_p\right)^2} \tag{3.4}$$

$$R^2 = 1 - \frac{\sum\left(y_i - y_p\right)^2}{\sum\left(y_i - \overline{y_i}\right)^2} \tag{3.5}$$

where y_i are the observed values, y_p are the predicted values, $\overline{y_i}$ indicates the mean of observed values, and n is the number of test datasets.

3.8. Variable Importance

3.8.1. Random Forest

About one-third of the data in the samples are not used to build each decision tree due to the randomized bootstrap sampling from the initial train dataset utilized in the RF model. The whole set of excluded data is referred to as out-of-bag dataset (OOB dataset). The OOB (out-of-bag dataset) can be applied to rank the variable importance by calculating the MSE reduction according to the permutation of each variable (Breiman, 2002; Grömping, 2009). For example, the OOBMSE (OOB mean squared error) in tree k is calculated as the average of the squared deviations between the observed values of OOB and the corresponding predictions (Eqn (3.6)). Then the MSE reduction, $\mathrm{OOBMSE_k}\ (\mathrm{X_j}\ \mathrm{permuted}) - \mathrm{OOBMSE_k}$ in OOB dataset is calculated based on the permutation of all variables $\mathrm{X_j}$ in the tree (Eqn (3.7)).

$$\mathrm{OOBMSE_k} = \frac{1}{n_{OOB,k}} \sum_{i=1}^{k} (y_i - \hat{y}_{i,k})^2 \tag{3.6}$$

$$\mathrm{OOBMSE_k}\ (\mathrm{X_j}\ \mathrm{permuted}) = \frac{1}{n_{OOB,k}} \sum_{i=1}^{k} (y_i - \hat{y}_{i,k}(\mathrm{X_j}\ \mathrm{permuted}))^2 \tag{3.7}$$

3.8.2. Artificial Neural Network

SHAP (Shapely Additive explanations) suggested by Lundberg and Lee (2017) was used to evaluate variable importance in the ANN model. ANN inherently involves complexity and interpreting challenges due to the numerous hidden layers and weights connecting them (Olden et al., 2004). SHAP, which evaluates the variable importance based on the order of SHAP value, is one of the promising approaches which can address these problems (Lundberg and Lee, 2017).

SHAP value is referred to as the unique Shapley value of a feature which is calculated using the conditional expectation function $f_x(S) = E[f(x)|x_s]$. When a specific subset of independent variables is S and the total number of independent variables in the model is M, the SHAP value can be expressed as:

$$SHAP_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(M - |S|! - 1)!}{M!} (f_x(S \cup \{j\}) - f_x(S)) \quad (3.8)$$

where N is the full set of independent variables set, and $f_x(S \cup \{j\})$ and $f_x(S)$ are the model output with and without the i-th variable, respectively (Lundberg et al., 2018). About all varieties of neural network models, SHAP values may be computed using Kernel SHAP (Lundberg and Lee, 2017). The calculation was carried out using the SHAP 0.41.0 library.

Chapter 4. Result

4.1. Data Summary

Data about the dependent and explanatory variables used for predicting FRE and MRE was summarized in Table 4.1 and Table 4.2. The content of Fe(II) and Mn in the inflow drainage differed substantially amongst the 9 coal mines. The range of IF (2~128 mg/L) was approximately 5 times larger than the range of IM (2~28mg/L) and the average IF was approximately 3.36 times more than the average IM. The maximum and minimum values of FRE and MRE were the same, but the average FRE was 0.82 and higher than the average MRE, which was 0.56. The FR showed the greatest range and the highest standard deviation, whereas the IpH showed the narrowest range and the lowest standard deviation. Since the difference in the range of each variable was significant, normalized variables were used in the analysis. Using the MinMaxScaler of sklearn.preprocessing package, each variable was normalized so that the lowest value of each variable was 0 and the maximum value of each variable was 1, resulting in all values ranging from 0 to 1.

Table 4.1. Summary of 199 datasets for FRE prediction.

Variable	Unit	Min	Max	Average	Std.Ev
IF	mg/L	2.00	128.00	20.25	27.54
FR	m^3/day	8.00	4434.00	300.96	529.14
Temp	℃	3.40	27.40	14.64	4.61
IpH	-	2.56	9.05	5.31	1.79
DO	mg/L	0.41	11.00	5.67	1.91
EC	mS/cm	0.00	72.10	2.92	10.92
Tur	NTU	0.00	291.80	21.38	45.23
Alkal	mg/L	0.00	313.00	73.74	76.10
FRE	-	0.00	1.00	0.82	0.24

Table 4.2. Summary of 132 datasets for MRE prediction.

Variable	Unit	Min	Max	Average	Std.Ev
IM	mg/L	2.04	28.11	6.05	5.13
FR	m^3/day	0.00	4434.00	433.96	589.55
Temp	℃	3.40	25.80	14.88	3.87
IpH	-	2.70	8.02	5.52	1.46
DO	mg/L	0.41	11.00	5.79	2.02
EC	mS/cm	0.00	1020.00	15.78	100.85
Tur	NTU	0.00	390.50	27.86	53.25
Alkal	mg/L	0.00	674.00	74.17	91.30
MRE	-	0.00	1.00	0.56	0.39

4.2. Correlation Analysis

4.2.1. Variables of Fe(II) Dataset

The heatmap of the Pearson correlation coefficient matrix shows the correlation among input and output variables (Fig 4.1, 4.2). When the color is dark and the absolute value of the Pearson correlation coefficient is near 1, variables have a strong relationship. According to the heatmap of Fig 4.1, IpH was the most closely associated with FRE, whose Pearson correlation coefficient with FRE was 0.67. This indicated a strong positive correlation between the pH of the inflow water and the Fe(II) removal efficiency.

The Alkal was the variable that had the second highest association to FRE, with a Pearson correlation coefficient of 0.52. Same with the correlation between IpH and FRE, the higher alkalinity led to the higher Fe(II) removal efficiency. For the remaining variables, the Pearson correlation coefficients were less than 0.3, making it difficult to explain the linear relationship with FRE.

Some explanatory variables were also correlated. For instance, IpH and IF had a negative correlation with a Pearson correlation coefficient of -0.52, indicating that the concentration of Fe(II) in mine drainage had a negative correlation with the pH of mine drainage. A correlation with the Pearson correlation coefficient of 0.52 was also found between pH and alkalinity.

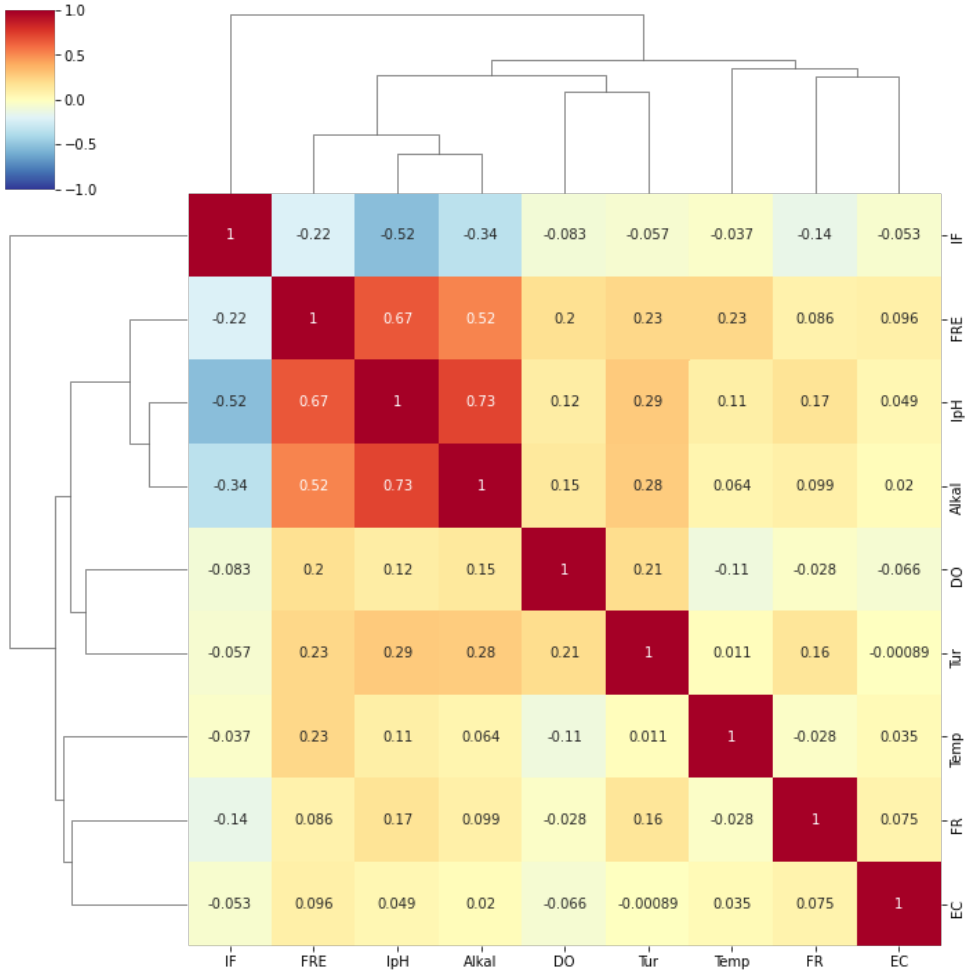


Figure 4.1. Correlation analysis of FRE and explanatory variables.

4.2.2. Variables of Mn Dataset

The three explanatory variables with the highest correlations to MRE were Alkal, IpH, and IM with Pearson correlation coefficients of 0.81, 0.77, and -0.46, respectively. (Fig 4.2). Similar to the result of 4.2.1, IpH and Alkal showed a positive correlation with MRE. Additionally, MRE showed a negative Pearson correlation coefficient to IM, indicating the negative correlation between the Mn removal efficiency and the Mn concentration of the inflow mine drainage.

Explanatory variables, such as IM and IpH, and IM and Alkal showed a high negative correlation with Pearson correlation coefficients of -0.69 and -0.47, respectively. IM was also strongly correlated with EC with a Pearson correlation coefficient of 0.68. Alkal and IpH also showed a weak positive correlation with Tur with the Pearson correlation coefficient of 0.37 and 0.39, respectively.

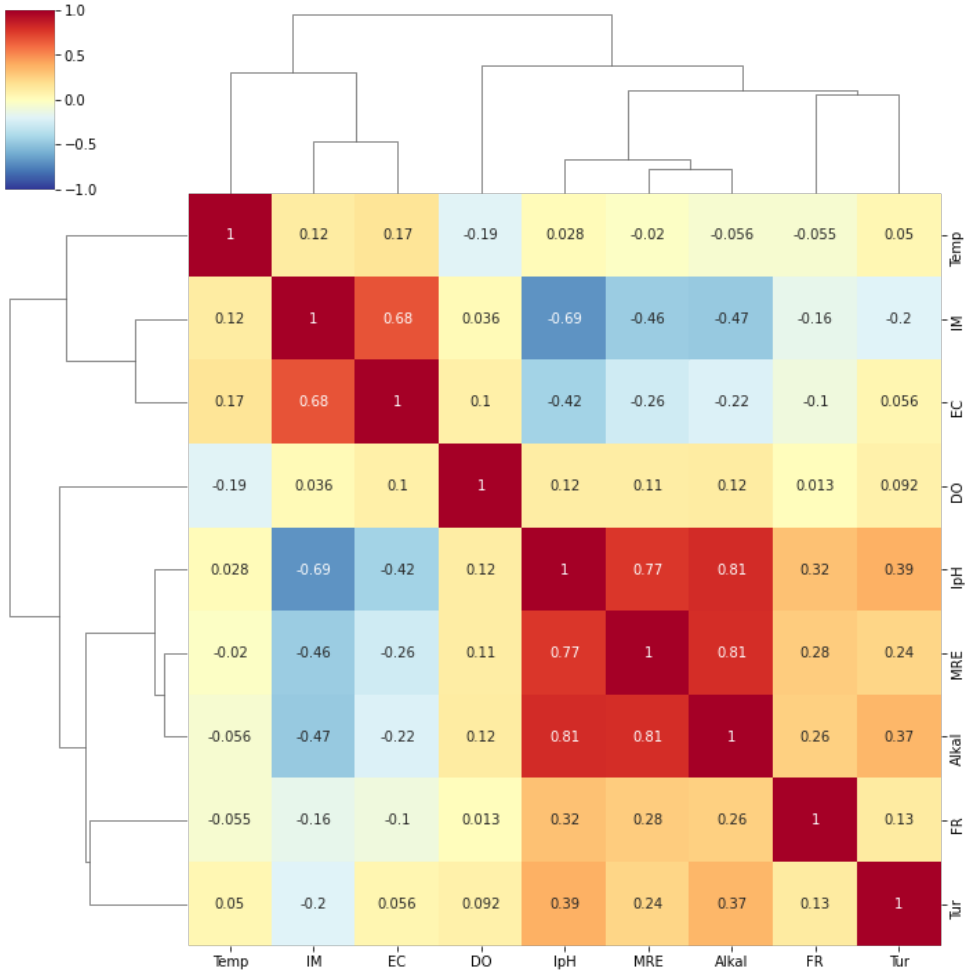


Figure 4.2. Correlation analysis of MRE and explanatory variables.

4.3. Optimization of MLR model

MLR models using Fe(II) and Mn datasets were built by the forward stepwise method, adding the explanatory variables with the order of strong correlation to removal efficiency by referring to the results of correlation analysis. Then, R^2 values between observed and predicted values in the train and test datasets were calculated (Table 4.2, 4.3). From the result, R^2 in the train dataset usually increased as the number of explanatory variables increased, whereas R^2 in the test dataset showed no tendency with the number of explanatory variables.

Using the Fe(II) dataset, the linear model with 6 explanatory variables IpH, Alkal, Tur, Temp, IF, DO, and EC was the optimal condition; R^2 in the test dataset was 0.604. Using the Mn dataset, the linear model with 6 explanatory variables Alkal, IpH, IM, FR, EC, and Tur was the optimal condition; R^2 in the test dataset was 0.733. The optimal MLR models for predicting Fe(II) and Mn removal efficiencies are as follows:

Table. 4.3. Optimization of MLR model for predicting FRE

Explanatory variables	Test R^2
IpH	0.479
IpH, Alkal	0.4945
IpH, Alkal,Tur	0.4948
IpH, Alkal,Tur, Temp	0.538
IpH, Alkal,Tur, Temp, IF	0.565
IpH, Alkal,Tur, Temp, IF, DO	0.600
IpH, Alkal,Tur, Temp, IF,DO,EC	0.604
IpH, Alkal,Tur, Temp, IF,DO, EC, FR	0.586

Table. 4.4. Optimization of MLR model for predicting MRE

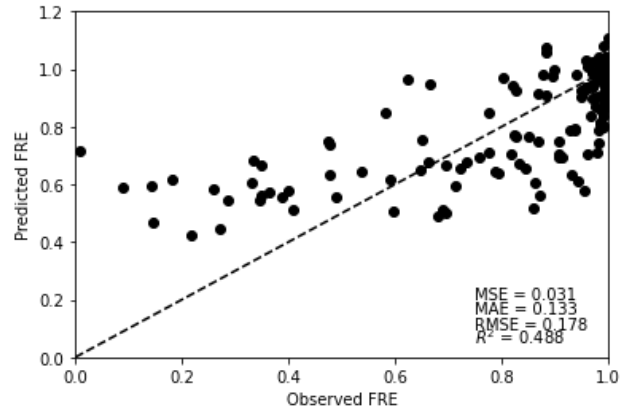
Explanatory variables	Test R^2
Alkal	0.694
Alkal, IpH	0.707
Alkal, IpH, IM	0.714
Alkal, IpH, IM, FR	0.720
Alkal, IpH, IM,FR, EC	0.732
Alkal, IpH, IM,FR, EC, Tur	0.733
Alkal, IpH, IM,FR, EC, Tur, DO	0.727

4.4. Comparison of Removal Efficiency Prediction

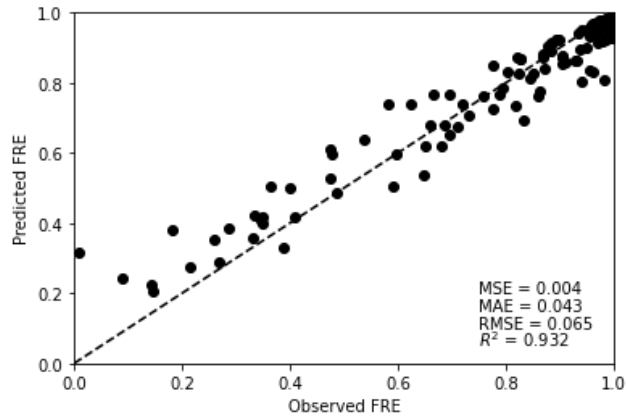
4.4.1. Train Dataset

The FRE and MRE of train datasets were re-predicted using the three trained models and the results were presented in Figure 4.3 and 4.4. The 1:1 line on the graph, which shows where the observed and predicted values are identical, was drawn to show how similar the two values are.

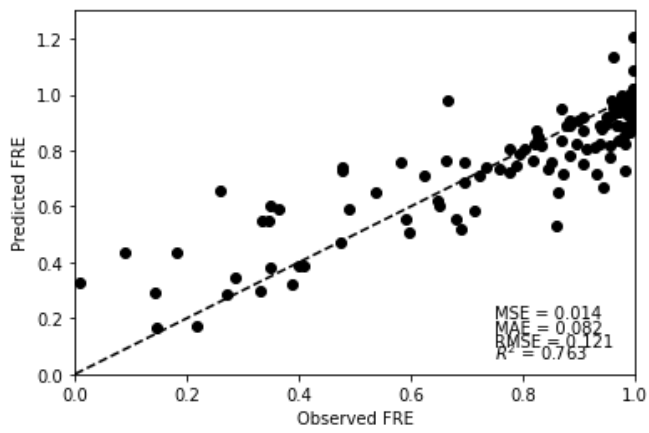
The MLR model showed the lowest performance for predicting the FRE and MRE of train datasets, indicating that the non-linear correlation between variables in the train datasets. R^2 values were 0.488 and 0.690 for predicting the FRE and MRE, respectively (Fig. 4.3a, 4.4a). The model with the best performance was RF, which predicted the FRE and MRE of train datasets with high R^2 values of 0.932 and 0.950, respectively. The RF model outperformed the other models in predicting FRE and MRE across the entire range of removal efficiency, with most of the predicted values that were close to the 1:1 line (Fig 4.3b, 4.4b).



(a)

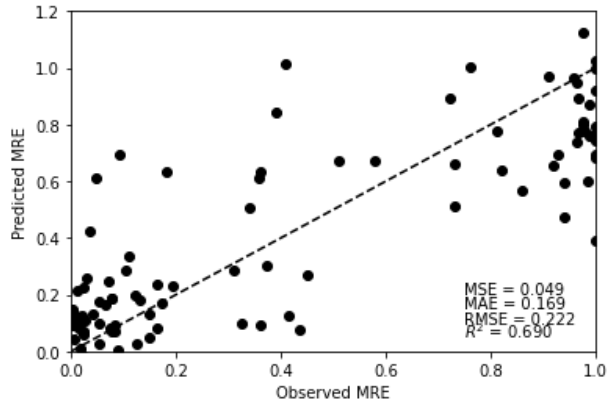


(b)

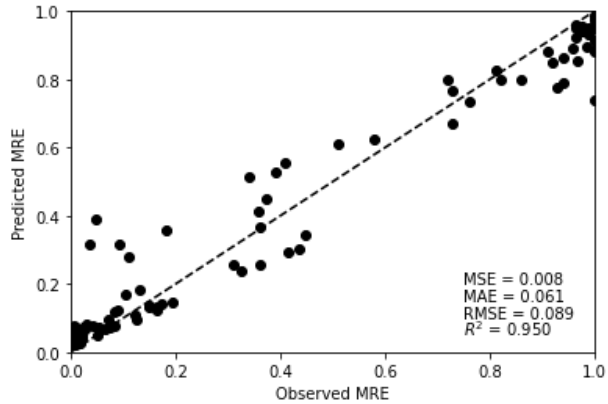


(c)

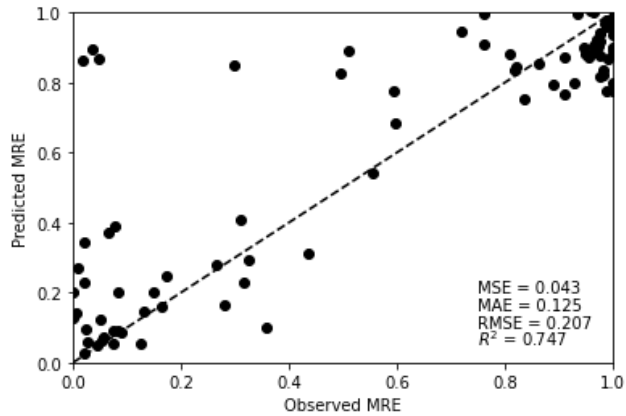
Figure 4.3. Prediction of FRE in train dataset (a) MLR (b) RF (c) ANN



(a)



(b)



(c)

Figure 4.4. Prediction of MRE in train dataset (a) MLR (b) RF (c) ANN

4.4.2. Test Dataset

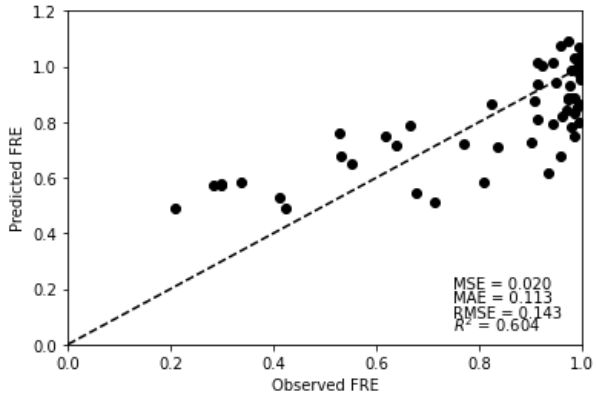
4.4.2.1. Predicting Fe(II) Removal Efficiency

The performances of three models for predicting FRE in the test dataset were evaluated by RMSE, MAE, MSE, and R^2 (Table 4.3 and Figure 4.5). Also, the observed values and predicted values of all elements in test the dataset were described together in a graph (Fig 4.6). According to the results, the RF model showed the most accurate predictions with the lowest RMSE, MAE, and MSE and the highest R^2 (Table 4.3). The ANN model outperformed the MLR model among the remaining two models.

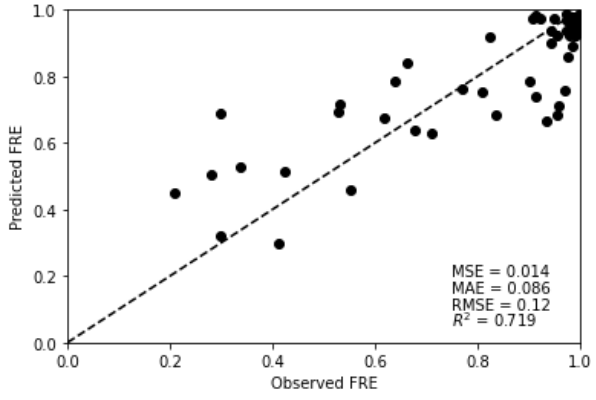
The RF and ANN model performed better than the MLR model at predicting extreme values. Especially, the RF model showed the best prediction of FRE higher than 0.9 since the predicted values did not exceed 1 (Fig. 4.5b, Table 4.5). The ANN model showed a good prediction of FRE lower than 0.4 (Fig 4.6c, Table 4.5).

Table 4.5. Model performances of FRE prediction

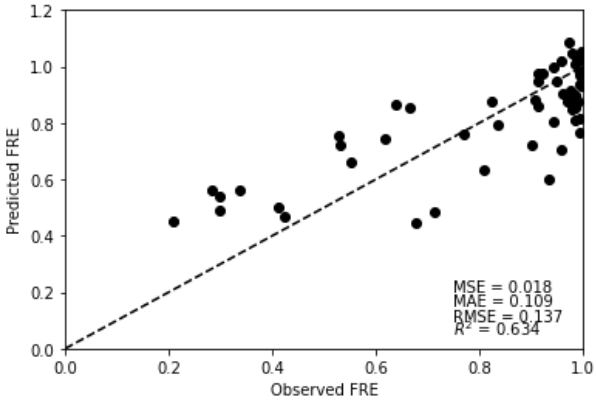
Model	RMSE	MAE	MSE	R^2
MLR	0.143	0.113	0.020	0.604
RF	0.120	0.086	0.014	0.719
ANN	0.137	0.109	0.018	0.634



(a)



(b)



(c)

Figure 4.5. Prediction of FRE in test dataset (a) MLR (b) RF (c) ANN

Table 4.6. Observed and predicted FRE in test dataset

Number	Observed	MLR	RF	ANN
1	0.977	0.93336	0.85953	0.91539
2	0.99	1.024097	0.947522	1.037119
3	0.664	0.787769	0.84141	0.853303
4	0.77	0.720941	0.76453	0.759459
5	0.297	0.57203	0.686638	0.541747
6	0.996	1.028237	0.983192	1.054046
7	0.923	1.003392	0.97193	0.977706
8	0.913	0.93828	0.98363	0.949381
9	0.999	0.982328	0.944472	0.98517
10	0.998	0.992734	0.98329	0.932437
11	0.96	0.81872	0.711272	0.904315
12	0.999	1.005899	0.958602	1.038816
13	0.99	0.855591	0.922548	0.876161
14	0.412	0.531155	0.299898	0.50098
15	0.971	0.845447	0.75743	0.876061
16	0.996	1.000423	0.96815	0.981757
17	0.529	0.761167	0.691912	0.755924
18	0.974	0.888819	0.93624	0.888304
19	0.957	1.074042	0.922826	1.019755
20	0.979	0.888303	0.980888	0.905405
21	0.996	1.036711	0.981722	1.027024
22	0.336	0.585058	0.52577	0.562538
23	0.824	0.866847	0.918176	0.874399
24	0.712	0.511291	0.630852	0.482391
25	0.995	0.987035	0.971804	0.934731
26	0.837	0.708868	0.684216	0.796302
27	0.934	0.618507	0.666146	0.599911
28	0.957	0.676373	0.682382	0.707713
29	0.98	0.984041	0.922128	1.048003
30	0.282	0.572019	0.503892	0.559378
31	0.944	0.794366	0.900426	0.803635

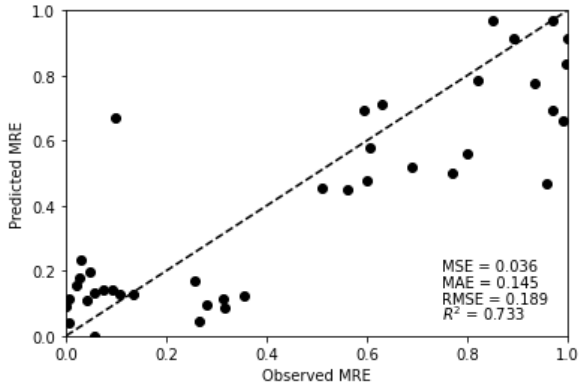
32	0.299	0.579034	0.323278	0.491752
33	0.208	0.490925	0.45198	0.45308
34	0.993	0.863823	0.952184	0.764759
35	0.986	0.752141	0.961854	0.81277
36	0.907	0.875793	0.97621	0.879445
37	0.988	0.991843	0.92281	1.025934
38	0.994	1.068967	0.926452	1.03343
39	0.993	1.001761	0.973322	0.981412
40	0.902	0.727795	0.786798	0.721564
41	0.808	0.582073	0.752002	0.635403
42	0.984	0.888254	0.956598	0.855777
43	0.619	0.747344	0.673518	0.744091
44	0.998	0.953617	0.938048	0.941279
45	0.553	0.650851	0.460548	0.663468
46	0.913	1.014941	0.980944	0.977672
47	0.98	0.784048	0.966068	0.846902
48	0.943	1.015625	0.938122	0.998542
49	0.984	1.030504	0.979768	1.007242
50	0.949	0.943656	0.974616	0.948569
51	0.993	0.796768	0.962268	0.81782
52	0.638	0.717488	0.785338	0.863104
53	0.972	0.883063	0.986526	0.885068
54	0.531	0.679983	0.717268	0.723065
55	0.677	0.547577	0.640264	0.445648
56	0.985	0.834626	0.89336	0.900788
57	0.914	0.809307	0.737958	0.857601
58	0.995	0.974902	0.979856	0.968458
59	0.425	0.491445	0.512038	0.466889
60	0.973	1.092668	0.96843	1.083695

4.4.2.2. Predicting Mn Removal Efficiency

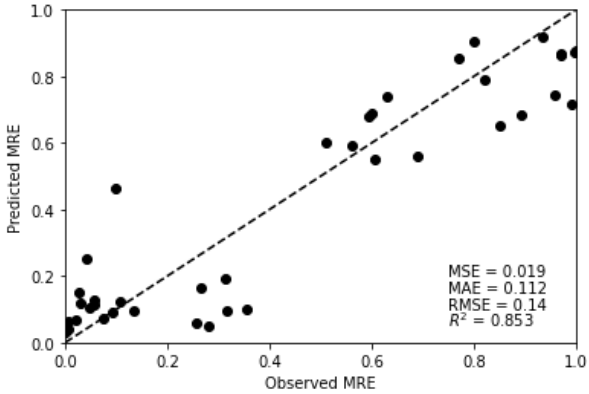
The RF model outperformed the other two models in predicting the MRE in the test dataset with the most accurate predictions, with the lowest RMSE, MAE, and MSE and the highest R^2 (Table 4.5, Figure 4.7, Figure 4.8). The ANN model outperformed the MLR model among the two remaining models. Different from the FRE predictions, all models showed great performances with R2 values exceeding 0.7, and the predictions among models showed no significant difference.

Table 4.7. Model performances of MRE prediction

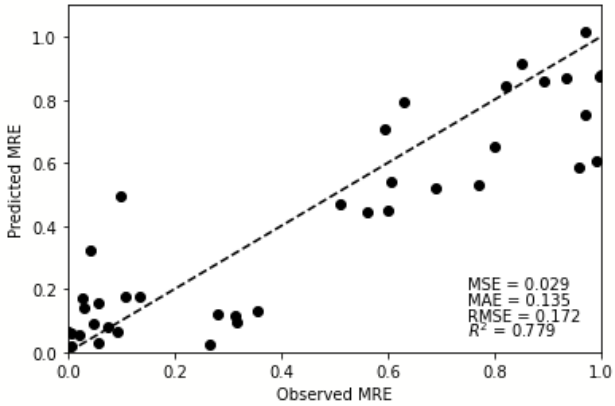
Variable	RMSE	MAE	MSE	R^2
MLR	0.189	0.145	0.036	0.733
RF	0.140	0.112	0.019	0.853
ANN	0.172	0.135	0.029	0.779



(a)



(b)



(c)

Figure 4.6. Prediction of MRE in test dataset (a) MLR (b) RF (c) ANN

Table 4.8. Observed and predicted MRE in test dataset

Number	Observed	MLR	RF	ANN
1	0.56	0.45045	0.59054	0.44284
2	0.8	0.558744	0.902988	0.65128
3	0.031	0.234598	0.119898	0.14168
4	0.607	0.580823	0.550886	0.54265
5	0.317	0.086196	0.095628	0.09288
6	0.971	0.971188	0.86695	1.01785
7	0.098	0.670719	0.461898	0.49332
8	0.595	0.694585	0.681244	0.70485
9	0.107	0.125556	0.122162	0.17489
10	0.256	0.167084	0.058252	0.05568
11	0.056	0.133372	0.129206	0.15665
12	0.96	0.467679	0.746118	0.58572
13	0.85	0.969874	0.650326	0.91529
14	0.6	0.47578	0.690496	0.45119
15	0.267	0.042789	0.166022	0.02484
16	0.69	0.519772	0.561184	0.52182
17	0.99	0.660219	0.714904	0.60551
18	0.049	0.195326	0.105652	0.08827
19	0.77	0.500237	0.854892	0.5324
20	0.135	0.128428	0.095964	0.17441
21	0.97	0.691319	0.865374	0.75273
22	0.892	0.912927	0.683924	0.85718
23	0.043	0.111044	0.252546	0.3233
24	0.936	0.775594	0.91682	0.87026
25	0.005	0.040114	0.041648	0.05817
26	0.001	0.090154	0.032344	0.06439
27	0.027	0.179347	0.150874	0.17077
28	0.091	0.140687	0.090148	0.06725
29	0.005	0.112873	0.063116	0.0175
30	0.282	0.094937	0.050114	0.12115
31	1	0.91483	0.87968	0.88145
32	0.356	0.12165	0.100326	0.13172

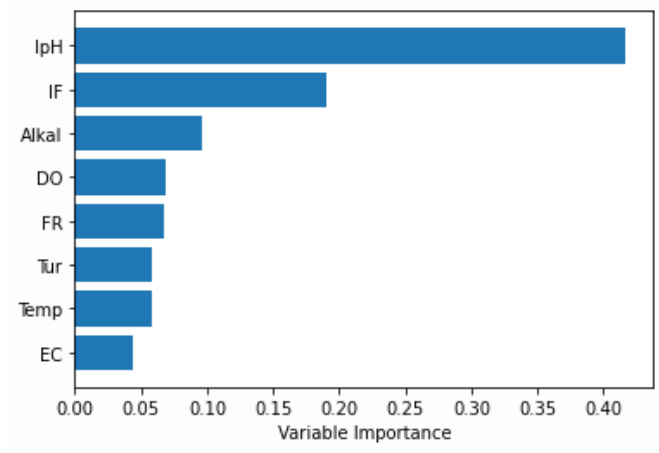
33	0.82	0.783486	0.792188	0.84297
34	0.51	0.454452	0.603518	0.46705
35	0.63	0.713866	0.740254	0.79419
36	0.021	0.156511	0.069068	0.05431
37	0.073	0.143223	0.073524	0.07964
38	0.056	0.003123	0.11398	0.02722
39	0.997	0.837851	0.87247	0.87388
40	0.313	0.116303	0.194074	0.11531

4.5. Variable Importance

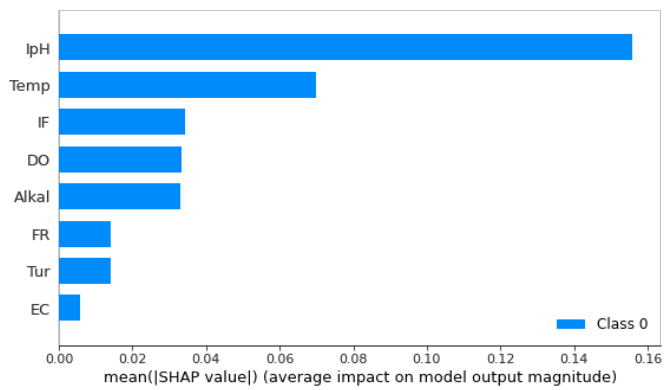
4.5.1. Variable Importance in FRE Prediction

The most important variables defining the two machine learning models RF, ANN for predicting FRE were extracted using the variable importance analysis (Fig 4.9). From this analysis, the relative importance of each variable was compared using the calculated values.

From the result of the analysis, the variable importance in two machine leaning models was different. First, the most important variable in the RF model was IpH, accounting for more than 40% of the total importance. IF and Alkal were followed with the importance of 0.18 and 0.1, respectively. The importance of the remaining variables was less than 0.07. In ANN model, IpH showed the highest SHAP value (0.158). It was followed by other variables Temp, IF, DO, and Alkal with the SHAP value of 0.07, 0.035, 0.034, and 0.033, respectively. The other variables FR, Tur, and EC showed the low importance with the SHAP values lower than 0.02.



(a)



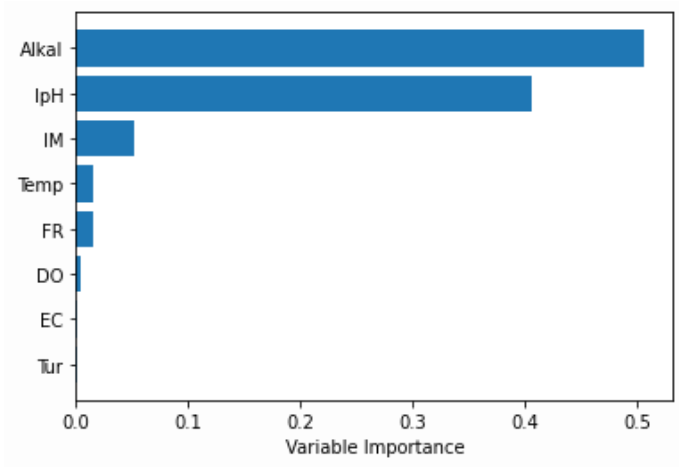
(b)

Figure 4.7. Variable importance in FRE prediction (a) RF (b) ANN

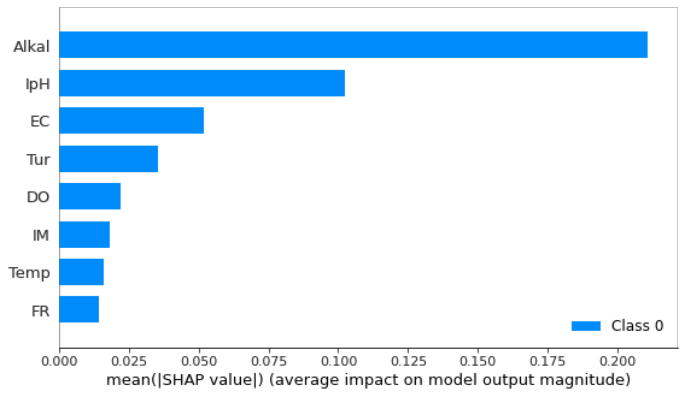
4.5.2. Variable Importance in MRE Prediction

The two most important variables in the RF model was Alkal and IpH. They accounted for more than 90% of the total importance with the variable importance of 0.5 and 0.4, respectively. Whereas, the significance of the remaining variables was less than 0.05 and only contributed approximately 10%.

Alkal and IpH were the two most important variables in the ANN model, with a SHAP value of about 0.22, 0.10, respectively. The two most important variables in RF and ANN model were the same. However, other factors like EC and Tur were also important in defining ANN model.



(a)



(b)

Figure 4.8. Variable importance in MRE prediction (a) RF (b) ANN

Chapter 5. Discussion

In this study, the order of optimal models for predicting the removal efficiency of Fe(II) and Mn was the RF, ANN, and MLR model. This meant that when it comes to predicting the metal removal efficiency, machine learning-based algorithms outperformed the linear regression approach. According to the results from the correlation analysis, only the pH of the inflow drainage was a meaningful explanatory variable showing a clear linear relationship with the Fe and Mn removal efficiency (Pearson correlation coefficient > 0.7). In other words, most explanatory variables had nonlinear correlations with the metal removal efficiencies in this study.

These nonlinear correlations between variables can be more accurately simulated by RF and ANN models than by the MLR model. In RF model, the nonlinearity can be explained by adjusting the number of trees, the node and depth of the single trees (Breiman, 2001). In Neural Networks, the number of hidden neurons and the node of each layer, and the optimization algorithms for optimizing weights are used to learn the model and explain the complexity between variables (Jain et al., 1996).

From the results of variable importance analysis, the pH of the inflow drainage was the most and secondly most important variable in predicting the Fe(II) and Mn removal efficiency, respectively. All 9 coal mines analyzed in this study adopted the series of passive systems for treating the mine drainage. According to the previous studies, two or three passive systems, including successive alkalinity producing systems (SAPS), anoxic limestone drains (ALD), anaerobic wetland, aerobic wetland, and oxidation ponds, have been combined and utilized for treating the abandoned mines in Korea (Sung et al. 1997; Bae et al. 2001; Ji et al., 2008). The performances of aforementioned passive systems were highly dependent on the pH of the inflow drainage since most systems remove metals via pH-controlled precipitation (Ji et al., 2008).

According to previous studies, effective metal removal in aerobic wetlands was possible when the inflow drainage pH was higher than 6 (Skousen and Ziemkiewicz, 2005). Fe and Mn could be precipitated as insoluble compounds (hydroxides) in wetlands by precipitation and reduction, which were sensitive to the pH of the drainage (A.S. Sheoran and V. Sheoran, 2006). In other studies, Fe^{2+} must first be oxidized to Fe^{3+} with the bacterial catalyzing to be precipitated in drainage, since it was too soluble in drainage with low dissolved O_2 up to pH 8 (Robbins and Norden, 1994; Evangelou and Zhang, 1995). And Mn removal was challenging due to the sensitivity to pH, and it was accomplished with the oxidation at pH 8 (Stumm and Morgan, 1981; A.S. Sheoran and V. Sheoran, 2006). In SAPS, the rate of forming iron complexes depended significantly on the pH, ranging from minutes to hours at neutral pH and from months to years at below pH 4 (Kepler ad McCleary, 1994)

Another important variables affecting the performance of RF model was the alkalinity of inflow drainage, which was the second and first important variable in predicting the Fe(II) and Mn removal efficiency, respectively. The relationship between alkalinity and Fe removal efficiency was linked to the properties of passive systems. In SAPS, alkalinity was created by raising the calcium content in order to effectively negate the ability of the Fe^{2+} to form acid. (Kepler ad McCleary, 1994, Ordonez et al., 2012). Alkali materials were also supplied from external sources in the case of other passive systems, such as certain aerobic wetlands or anoxic limestone drains (ALD), when there was not enough alkalinity (Skousen and Larew, 1994). Alkalinity is also associated with Mn removal efficiency, because the most widely used technique for removing Mn from mine drainage was the use of alkaline materials for acidity-neutralization and precipitation (Ayora et al.,2013; Luan and Burgos, 2019; Bryce et al., 2020).

Since the most important variables extracted from the learned model were also major variables in the passive treatment systems, the Fe(II) and Mn removal efficiency of passive systems may be predicted using the machine learning-based RF model. However, performances of RF models are affected by several factors, such as the number of dataset, model optimizing, and explanatory variables (Guyon and Elisseeff, 2003). Especially, the treatment system's design factors such as the size and the hydraulic retention time that might influence the metal removal could not be considered in this study due to the lack of data. If these limitations are further considered, better performance of prediction models will be expected.

Chapter 6. Conclusion

Acid mine drainage (AMD) has to be monitored and managed by reclamation or treatment systems because of the high concentration of heavy metals and low pH. Predicting the performance of AMD treatment systems is challenging due to the many variables involved. In this study, MLR, RF and ANN model were constructed for predicting the Fe(II) and Mn removal efficiencies and the performances were compared by the RMSE, MAE, MSE, R^2. The results are as follows:

The RF model was the optimal model to predict the Fe(II) removal efficiency. Especially, RF performed well to predict high FRE >0.9 since the predicted values did not exceed 1. It was also the optimal model to predict the Mn removal efficiency. But, all models showed great performances with R^2 values exceeding 0.7, and the predictions among models showed no significant differences.

From the result of sensitivity analysis, the pH, the concentration of Fe(II), and the alkalinity of the inflow water were determined to be the most important variables to predict the effectiveness of Fe(II) removal. And the pH, the alkalinity of the inflow water were the most important variables to predict the Mn removal. These variables have been discussed as important factors related to the metal removal in previous studies. However, limited number of factors were importantly taken into account constructing the RF models. The prediction performance of the model is expected to be improved if more important independent variables are

considered.

References

1. Matlock, Matthew M., Brock S. Howerton, and David A. Atwood. "Chemical precipitation of heavy metals from acid mine drainage." *Water research* 36.19 (2002): 4757-4764.

2. Akcil, Ata, and Soner Koldas. "Acid Mine Drainage (AMD): causes, treatment and case studies." *Journal of cleaner production* 14.12-13 (2006): 1139-1145.

3. McCarthy, Terence S. "The impact of acid mine drainage in South Africa." *South African Journal of Science* 107.5 (2011): 1-7.

4. Rezaie, Behnaz, and Austin Anderson. "Sustainable resolutions for environmental threat of the acid mine drainage." *Science of the Total Environment* 717 (2020): 137211.

5. Lopes, I., et al. "Discriminating the ecotoxicity due to metals and to low pH in acid mine drainage." *Ecotoxicology and Environmental Safety* 44.2 (1999): 207-214.

6. Lei, Liang-Qi, et al. "Acid mine drainage and heavy metal contamination in groundwater of metal sulfide mine at arid territory (BS mine, Western Australia)." *Transactions of Nonferrous Metals Society of China* 20.8 (2010): 1488-1493.

7. Equeenuddin, Sk Md, et al. "Metal behavior in sediment associated with acid mine drainage stream: role of pH." *Journal of Geochemical Exploration* 124 (2013): 230-237.

8. Gray, N. F. "Environmental impact and remediation of acid mine drainage: a management problem." *Environmental geology* 30.1 (1997): 62-71.

9. Duruibe, J. Ogwuegbu, M. O. C. Ogwuegbu, and J. N. Egwurugwu. "Heavy metal pollution and human biotoxic effects." *International Journal of physical sciences* 2.5 (2007): 112-118.

10. Briffa, Jessica, Emmanuel Sinagra, and Renald Blundell. "Heavy metal pollution in the environment and their toxicological effects on humans." *Heliyon* 6.9 (2020): e04691.

11. Qureshi, Asif, Christian Maurice, and Björn Öhlander. "Potential of coal mine waste rock for generating acid mine drainage." *Journal of Geochemical Exploration* 160 (2016): 44-54.

12. Akcil, Ata, and Soner Koldas. "Acid Mine Drainage (AMD): causes, treatment and case studies." *Journal of cleaner production* 14.12-13 (2006): 1139-1145.

13. Simate, Geoffrey S., and Sehliselo Ndlovu. "Acid mine drainage: Challenges and opportunities." *Journal of Environmental Chemical Engineering* 2.3 (2014): 1785-1803.

14. Seervi, Vikram, et al. "Overview of active and passive systems for treating acid mine drainage." *Iarjset* 4.5 (2017): 131-137.

15. Taylor, Jeff, Sophie Pape, and Nigel Murphy. "A summary of passive and active treatment technologies for acid and metalliferous drainage (AMD)." *Proceedings of the in fifth Australian workshop on acid mine drainage*. 2005.

16. Johnson DB, Hallberg KB. Acid mine drainage remediation options: a review. Sci Total Environ. 2005;338(2005):3–14.

17. Saha, Sukla, and Alok Sinha. "A review on treatment of acid mine drainage with waste materials: a novel approach." *Global NEST Journal* 20.3 (2018): 512-528.

18. Gazea, B., K. Adam, and A. Kontopoulos. "A review of passive systems for the treatment of acid mine drainage." *Minerals engineering* 9.1 (1996): 23-42.

19. Zipper, Carl, and Jeff Skousen. "Passive treatment of acid mine drainage." *Acid mine drainage, rock drainage, and acid sulfate soils: causes, assessment, prediction, prevention, and remediation* (2014): 339-353.

20. Clyde, Erin J., et al. "The use of a passive treatment system for the mitigation of acid mine drainage at the Williams Brothers Mine (California): pilot-scale study." *Journal of cleaner production* 130 (2016): 116-125.

21. Skousen, Jeffrey G., Paul F. Ziemkiewicz, and Louis M. McDonald. "Acid mine drainage formation, control and treatment: Approaches and strategies." *The Extractive Industries and Society* 6.1 (2019): 241-249.

22. Skousen, Jeff, et al. "Review of passive systems for acid mine drainage

treatment." *Mine Water and the Environment* 36.1 (2017): 133-153.

23. Hyman, D. M., and G. R. Watzlaf. "Mine drainage characterization for the successful design and evaluation of passive treatment systems." *17th Annual National Association of Abandoned Mine Lands Conference, Indiana, October*. 1995.

24. Gibert, O., et al. "Treatment of acid mine drainage by sulphate-reducing bacteria using permeable reactive barriers: a review from laboratory to full-scale experiments." *Reviews in Environmental Science and Biotechnology* 1.4 (2002): 327-333.

25. Heviánková, Silvie, Iva Bestová, and Miroslav Kyncl. "The application of wood ash as a reagent in acid mine drainage treatment." *Minerals Engineering* 56 (2014): 109-111.

26. Igarashi, Toshifumi, et al. "The two-step neutralization ferrite-formation process for sustainable acid mine drainage treatment: Removal of copper, zinc and arsenic, and the influence of coexisting ions on ferritization." *Science of the Total Environment* 715 (2020): 136877.

27. Betrie, Getnet D., et al. "Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques." *Environmental monitoring and assessment* 185.5 (2013): 4171-4182.

28. Cravotta III, Charles A. "Interactive PHREEQ-N-AMDTreat water-quality modeling tools to evaluate performance and design of treatment systems for

acid mine drainage." *Applied Geochemistry* 126 (2021): 104845.

29. Foos, A. "Geochemical modeling of coal mine drainage, Summit County, Ohio." *Environmental Geology* 31.3 (1997): 205-210.

30. Amos, Richard T., et al. "Reactive transport modeling of column experiments for the remediation of acid mine drainage." *Environmental science & technology* 38.11 (2004): 3131-3138.

31. Andalaft, Javier, et al. "Assessment and modeling of nanofiltration of acid mine drainage." *Industrial & Engineering Chemistry Research* 57.43 (2018): 14727-14739.

32. Zipper, Carl E., and Jeffrey G. Skousen. "Influent water quality affects performance of passive treatment systems for acid mine drainage." *Mine Water and the Environment* 29.2 (2010): 135-143.

33. Koide, Ryu, et al. "A model for prediction of neutralizer usage and sludge generation in the treatment of acid mine drainage from abandoned mines: case studies in Japan." *Mine Water and the Environment* 31.4 (2012): 287-296.

34. III, CA Cravotta, et al. "A geochemical module for “AMDTreat” to compute caustic quantity, effluent quality, and sludge volume." *Proceedings America Society of Mining and Reclamation* (2010): 1413-1436.

35. Cravotta, C. A., et al. "AMDTreat 5.0+ with PHREEQC titration module to compute caustic chemical quantity, effluent quality, and sludge volume." *Mine Water and the Environment* 34.2 (2015): 136-152.

36. Kirk Nordstrom, Darrell. "Geochemical modeling of iron and aluminum precipitation during mixing and neutralization of acid mine drainage." *Minerals* 10.6 (2020): 547.

37. Chen, Chong, et al. "A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China." *Scientific reports* 10.1 (2020): 1-13.

38. Auria, Laura, and Rouslan A. Moro. "Support vector machines (SVM) as a technique for solvency analysis." (2008).

39. Yaseen, Zaher Mundher. "An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions." *Chemosphere* 277 (2021): 130126.

40. Gholami, R., et al. "Prediction of toxic metals concentration using artificial intelligence techniques." *Applied Water Science* 1.3 (2011): 125-134.

41. Atasoy, Ayşe Dilek, B. Babar, and E. Sahinkaya. "Artificial neural network prediction of the performance of upflow and downflow fluidized bed reactors treating acidic mine drainage water." *Mine Water and the Environment* 32.3 (2013): 222-228.

42. Lu, Wei, et al. "Detection of heavy metals in vegetable soil based on THz spectroscopy." *Computers and Electronics in Agriculture* 197 (2022): 106923.

43. Ji, Sangwoo, Sunjoon Kim, and Juin Ko. "The status of the passive treatment systems for acid mine drainage in South Korea." *Environmental Geology* 55.6

(2008): 1181-1194.

44. Lee, Dong-Kil, and Young-Wook Cheong. "The relationship between flow paths and water quality in mine water oxidation ponds in South Korea." *Mine Water and the Environment* 35.4 (2016): 469-479.

45. Laine, David M., and Adam P. Jarvis. "Engineering design aspects of passive in situ remediation of mining effluents." *Land Contamination and Reclamation* 11.2 (2003): 113-125.

46. Lee, Dong-kil, et al. "Study on distribution characteristics of some water parameters properties of mine drainage in an oxidation pond, Hwangji-Yuchang coal mine, South Korea." *Environmental earth sciences* 68.1 (2013): 241-249.

47. Roy Chowdhury, Abhishek, Dibyendu Sarkar, and Rupali Datta. "Remediation of acid mine drainage-impacted water." *Current Pollution Reports* 1.3 (2015): 131-141.

48. Wildeman, Thomas R., et al. "Passive treatment methods for manganese: preliminary results from two pilot sites." *Proc, National Meeting of the American Soc of Surface Mining and Reclamation*. 1993.

49. Hedin, Robert S., George R. Watzlaf, and Robert W. Nairn. *Passive treatment of acid mine drainage with limestone*. Vol. 23. No. 6. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, 1994.

50. Kepler, Douglas A., and Eric D. McCleary. *Successive alkalinity-producing systems (SAPS) for the treatment of acidic mine drainage*. Bureau of Mines, 1994.

51. Ordonez, Almudena, Jorge Loredo, and Fernando Pendas. "A successive alkalinity producing system (SAPS) as operational unit in a hybrid passive treatment system for Acid Mine Drainage." *Mine Water and Environment* (2012): 575-580.

52. Seber, George AF, and Alan J. Lee. *Linear regression analysis*. John Wiley & Sons, 2012.

53. Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

54. Biau, Gérard, and Erwan Scornet. "A random forest guided tour." *Test* 25.2 (2016): 197-227.

55. Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling." *Journal of chemical information and computer sciences* 43.6 (2003): 1947-1958.

56. Archer, Kellie J., and Ryan V. Kimes. "Empirical characterization of random forest variable importance measures." *Computational statistics & data analysis* 52.4 (2008): 2249-2260

57. Hopfield, John J. "Artificial neural networks." *IEEE Circuits and Devices Magazine* 4.5 (1988): 3-10.

58. Rooki, R., et al. "Prediction of heavy metals in acid mine drainage using artificial neural network from the Shur River of the Sarcheshmeh porphyry copper mine, Southeast Iran." *Environmental earth sciences* 64.5 (2011): 1303-1316.

59. Ajayi, Toluwaleke, Dina L. Lopez, and Abiodun E. Ayo-Bali. "Using Artificial Neural Network to Model Water Discharge and Chemistry in a River Impacted by Acid Mine Drainage." *American Journal of Water Resources* 9.2 (2021): 63-79.

60. Trifi, Mariem, et al. "Machine learning-based prediction of toxic metals concentration in an acid mine drainage environment, northern Tunisia." *Environmental Science and Pollution Research* (2022): 1-19.

61. Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal processing magazine* 29.6 (2012): 82-97.

62. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.

63. Abiodun, Oludare Isaac, et al. "State-of-the-art in artificial neural network applications: A survey." *Heliyon* 4.11 (2018): e00938.

64. Ruder, Sebastian. "An overview of gradient descent optimization algorithms." *arXiv preprint arXiv:1609.04747* (2016).

65. Sharma, Sagar, Simone Sharma, and Anidhya Athaiya. "Activation functions

in neural networks." *towards data science* 6.12 (2017): 310-316.).

66. Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." *Icml*. 2010.

67. Ide, Hidenori, and Takio Kurita. "Improvement of learning for CNN with ReLU activation by sparse regularization." *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

68. Mitchell, Tom M., and Tom M. Mitchell. *Machine learning*. Vol. 1. No. 9. New York: McGraw-hill, 1997.

69. Dutilleul, Pierre, et al. "The Mantel test versus Pearson's correlation analysis: assessment of the differences for biological and environmental studies." *Journal of agricultural, biological, and environmental statistics* (2000): 131-150.

70. Mansson, Robert, et al. "Pearson Correlation Analysis of Microarray Data Allows for the Identification of Genetic Targets for Early B-cell Factor*[boxs]." *Journal of Biological Chemistry* 279.17 (2004): 17905-17913.

71. Kolasa-Wiecek, Alicja. "Stepwise multiple regression method of greenhouse gas emission modeling in the energy sector in Poland." *Journal of Environmental Sciences* 30 (2015): 47-54.

72. Probst, Philipp, Marvin N. Wright, and Anne-Laure Boulesteix. "Hyperparameters and tuning strategies for random forest." *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3 (2019): e1301.

73. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

74. Grömping, Ulrike. "Variable importance assessment in regression: linear regression versus random forest." *The American Statistician* 63.4 (2009): 308-319.

75. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

76. Olden, Julian D., Michael K. Joy, and Russell G. Death. "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data." *Ecological modelling* 178.3-4 (2004): 389-397.

77. Jain, Anil K., Jianchang Mao, and K. Moidin Mohiuddin. "Artificial neural networks: A tutorial." *Computer* 29.3 (1996): 31-44.

78. Skousen, Jeff, and Paul Ziemkiewicz. "Performance of 116 passive treatment systems for acid mine drainage." *National meeting of the American society of mining and reclamation, Breckenridge, CO*. 2005.

79. Sheoran, A. S., and V. Sheoran. "Heavy metal removal mechanism of acid mine drainage in wetlands: a critical review." *Minerals engineering* 19.2 (2006): 105-116.

80. Robbins, E. I., and A. W. Norden. *Microbial oxidation of iron and manganese in wetlands and creeks of Maryland, Virginia, Delaware, and Washington, DC*.

No. CONF-940930-. Univ. of Pittsburgh, Pittsburgh, PA (United States), 1994.

81. Evangelou, V. P卅, and Y. L. Zhang. "A review: pyrite oxidation mechanisms and acid mine drainage prevention." *Critical Reviews in Environmental Science and Technology* 25.2 (1995): 141-199.

82. Stumm, Werner, and James J. Morgan. *Aquatic chemistry: chemical equilibria and rates in natural waters*. John Wiley & Sons, 2012.

83. Kepler, Douglas A., and Eric D. McCleary. *Successive alkalinity-producing systems (SAPS) for the treatment of acidic mine drainage*. Bureau of Mines, 1994.

84. Skousen, Jeff, and Glenn Larew. "Alkaline overburden addition to acid-producing materials to prevent acid mine drainage." *International Land Reclamation and Mine Drainage Conference*. Vol. 1. Pittsburgh, PA.: Bureau of Mines SP 06B-94, 1994.

85. Ayora, Carlos, et al. "Acid mine drainage in the Iberian Pyrite Belt: 2. Lessons learned from recent passive remediation experiences." *Environmental Science and Pollution Research* 20.11 (2013): 7837-7853.

86. Luan, Fubo, and William D. Burgos. "Effects of solid-phase organic carbon and hydraulic residence time on manganese (II) removal in a passive coal mine drainage treatment system." *Mine Water and the Environment* 38.1 (2019): 130-135.

87. Le Bourre, Bryce, et al. "Manganese removal processes and geochemical behavior in residues from passive treatment of mine

rainage." *Chemosphere* 259 (2020): 127424.

88. Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." *Journal of machine learning research* 3.Mar (2003): 1157-1182.

국문요지

산성광산배수(AMD)는 높은 중금속과 낮은 pH로 인해 세계적인 문제로 대두되었으며 지속적인 모니터링과 처리시설을 이용한 관리를 필요로 한다. 산성광산배수의 처리는 pH, 금속의 농도 등 연관된 수많은 요인으로 인해 예측에 어려움이 있다. 광산 배수 내 중금속의 제거를 예측하기 위한 경험적 및 지구화학적 모델이 개발되어왔으나 모델을 구축하는 데 많은 시간이 소요되고 충분한 양의 데이터를 필요로 하기 때문에, 제한된 정보로도 예측이 가능한 머신 러닝 기반 모델을 구축할 필요성이 존재한다.

본 연구에서는 한국의 9개 폐탄광에 대하여 수동적 처리 시스템의 Fe(II) 및 Mn 제거 효율을 예측하기 위해 RF 및 ANN 모델을 구축하였고, 이를 MLR 모델의 성능과 비교하였다. 이중, RF 모델이 Fe(II) 제거 Mn 제거 효율 예측 모두에서 가장 우수한 성능을 보였다. 민감도 분석 결과, Fe(II) 제거 효율을 예측하는 데 가장 중요한 변수 세 가지는 순서대로 유입수의 pH, 유입수의 Fe(II) 농도 및 알칼리도였다. Mn 제거 효율 예측을 위한 세 가지 중요한 변수는 순서대로 유입수의 알칼리도, 유입수의 pH, 유입수의 Mn 농도로 분석되었다.

핵심어 : 광산배수 정화, 머신러닝, 랜덤 포레스트, 인공신경망, 예측, 민감도 분석

학번 : 2021-27958