Application of Artificial Intelligence (AI) to predict mine water quality, a case study in South Africa

Emmanuel Sakala,[1] Obed Novhe †,[1] Viswanath Ravi Kumar Vadapalli[1]

[1]*Council for Geoscience, PVT Bag 112, Pretoria 0001, South Africa, esakala@geoscience.org.za, vvadapalli@geoscience.org.za*

Abstract

This study investigates the use of artificial neural networks (ANNs) for mine water quality prediction. The water chemistry of the Witkranz discharge site was used to develop the prediction system. Parameters such as rainfall, air temperature, depth to water table and discharge pH were used as training inputs, while sulfate was used as the training output. A graphical user interface (GUI) was developed by combining long-short-term memory nets (LSTM) for each of the four input parameters and an ANN combining the LSTM outputs to predict future sulfate values. The system was tested using historical data with over 99% training accuracy.

Keywords: artificial neural network, long-short-term memory, prediction system, training accuracy

Introduction

Mining water-related pollution is one of the major environmental challenges in South Africa. Historical coal mining has substantially changed the landscape in a number of provinces at the same time altering the hydrology and hydrogeology. The pollution of the groundwater is associated with acid mine drainage (AMD). Many interventions have been proposed and implemented, adapted to the volumes and quality of the AMD. However, in order to design sustainable future solutions to AMD, knowledge about future mine drainage conditions is indispensable.

The application of artificial neural networks (ANNs) to the fields of water engineering, ecological and environmental sciences have gained momentum in the last two decades (Najah et al. 2011). Chen and Mynett (2003) and Lee et al. (2003) have successfully used data-driven modelling techniques for the prediction of freshwater and seawater quality alike. Najah et al. (2011) have demonstrated that ANNs can be used to predict river water quality, based on historical data. Therefore, ANNs may be an alternative to current methods of mine-water quality prediction. ANN captures the embedded spatial and unsteady behaviour inherent in the AMD-affected area. The architecture and non-linear nature of ANNs make them more suitable than other modelling techniques. Owing to the correlations and interactions between water quality parameters, it is interesting to investigate whether a domain-specific mechanism governing observed patterns exists to prove the predictability of these variables (Najah et al. 2011).

The current study intends investigating the application of ANNs to predict the water quality with a focus on sulfate concentrations from one of the abandoned underground coal mines in Mpumalanga Province, South Africa, based on readily available monitoring and historical water quality data. The study area is the Witkranz discharge site, situated in the town of Carolina. Geologically, the area forms part of the Ermelo coalfield (south of Carolina). There is very little information about the mining history of the area. However, an old mine plan obtained from a local mining company shows mining activities in a portion of land to the east of the main discharge point, where both underground and open-cast methods have been used. Researchers from the Council for Geoscience (CGS) have been working to implement a passive treatment plant at this discharge point and have collected geochemical data over four years. The current project, promoting the application of historical data to predict future mine water quality, may serve as a prototype for future projects.

Methods

The methodology used in this study involves a cascading approach where one-step feeds into the next sequentially, until the last step has been reached. The sequence of the methodology is as follows: Hydrogeochemistry data > Identifying controlling factors > Gather all controlling factors > Pre-process all datasets > Design LSTM model > Carry out prediction tests using model > ANN model evaluation.

Hydrogeochemistry data

Hydrogeochemistry data used in this research were collected from the Witkranz discharge point (fig. 1) from 1 November 2014 to 13 June 2018 using conventional best-practice groundwater sampling guidelines, as described by Weaver et al. (2007). A total of forty-six (46) mine water samples were collected and analysed at the CGS laboratory according to the required parameters.

Input-control factors for sulfate concentrations in mine water

The first step in the development of a predictive mine water quality system involves understanding the possible factors controlling water quality, in this case, sulfate concentrations of the mine water. Based on the literature survey and our understanding of AMD generation, the following important controlling factors were considered as inputs to the system: precipitation, soil temperature, water table levels and water pH. These datasets for the site were either obtained from various governmental organisations or derived from variables, or measured on site.

Output — sulfate

Sulfate concentrations will be used to determine output labels (values) for the given input features as these are a good indicator of AMD pollution (Sakala et al. 2018).

System development

The GUI and other system developments were carried out using the Python open-source programming language. The programs are built using several Python built-in libraries: Tkinter, Numpy, Matplotlib, OSGeo, Tensorflow, Sklearn and Pandas.

Fig. 1 illustrates the development of the ANN system for the prediction of mine water quality. The system development may be broken down into data pre-processing and the prediction system architecture (input prediction — LSTM and ANN systems).

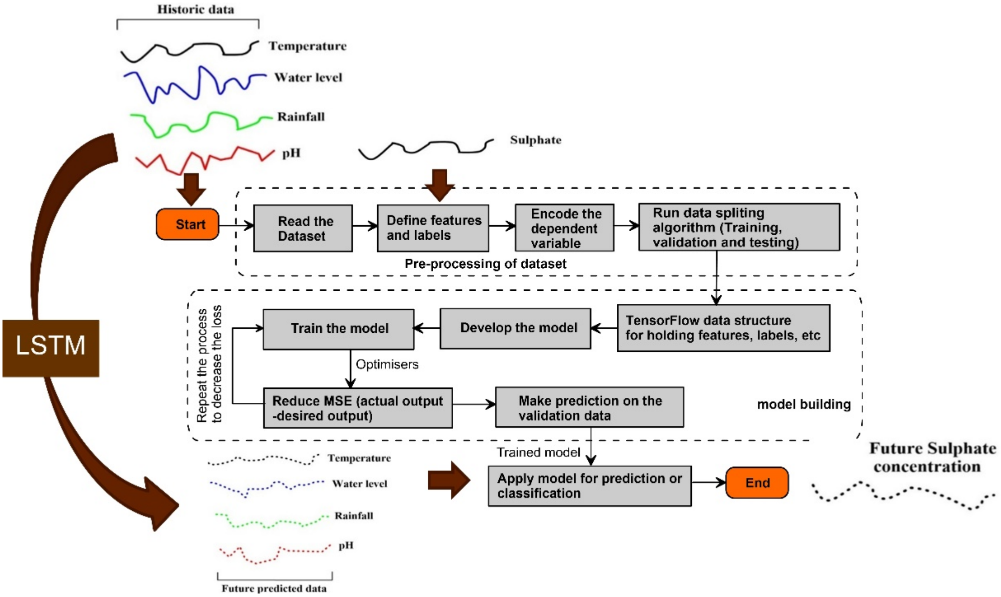

Historic data
Temperature
Water level
Rainfall
pH
Sulphate
Start
Read the Dataset
Define features and labels
Encode the dependent variable
Run data spliting algorithm (Training, validation and testing)
Pre-processing of dataset
LSTM
Repeat the process to decrease the loss
Train the model
Develop the model
TensorFlow data structure for holding features, labels, etc
Optimisers
Reduce MSE (actual output -desired output)
Make prediction on the validation data
model building
Trained model
Apply model for prediction or classification
End
Future Sulphate concentration
Temperature
Water level
Rainfall
pH
Future predicted data

Figure 1 System overview.

Pre-processing

Pre-processing data involves reconfiguring the data to a format which can be used by the artificial intelligence (AI) system. The format in which data are gathered requires “*cleaning*” and formatting before the informaton can be used. The inputs (rainfall, temperature, pH, depth to water table) and output (sulfate) datasets are pre-processed in Microsoft Excel to attain a format which can be used by the prediction program.

Prediction system architecture

The design of the system architecture involves the generation of a system that uses historical data (features and output labels) to train/optimise its parameters and can be used as predicted labels for unseen/future data. This process involves the following steps:

- Designing the prediction system architecture using a sequential machine learning algorithm (long short-term memory — LSTM),
- Training and validation of the developed system.

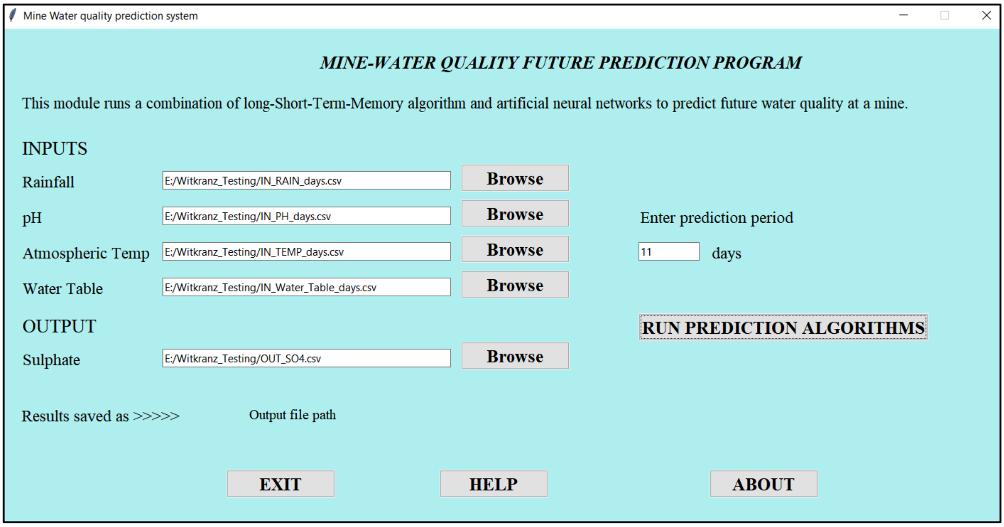
Results and discussion

The GUI comprises a launch window and a processing window. The launch window is the main window from where the program is launched and from where the help files and information about the program can be accessed. The processing window prompts the user to load the four input parameters (rainfall, atmospheric temperature, pH values and depth to water table) and the output sulfate values using the “Browse” button. Fig. 2 shows the GUI for the processing module of the prediction program.

LSTM training

One of the most difficult problems associated with programming LSTM systems entails determining whether the LSTM model is performing well on the sequence prediction problem. That is, the model may be obtaining a good prediction score and / or a good model fit but in fact the model underfitted (where the model fits the data too well causing low bias but high variance) or overfitted (where the model does not fit the data well enough causing high bias but low variance) the training data. This can be overcome by using diagnosis plots to monitor the training process both for the training and validation datasets and stopping the training process. Early stopping involves stopping the training process once the performance of the validation dataset stops increasing (i.e. the cost begins to increase steadily instead of decreasing).

After the early stopping points for each of the four parameters have been determined, the LSTMs for each input is run to predict future values for each dataset. Fig. 3 shows a graphic representation of parameters


Mine Water quality prediction system
INPUTS
OUTPUT
Browse
Browse
Browse
Browse
Browse
RUN PREDICTION ALGORITHMS
EXIT
HELP
ABOUT

MINE-WATER QUALITY FUTURE PREDICTION PROGRAM

This module runs a combination of long-Short-Term-Memory algorithm and artificial neural networks to predict future water quality at a mine.

Rainfall E:/Witkranz_Testing/IN_RAIN_days.csv
pH E:/Witkranz_Testing/IN_PH_days.csv
Atmospheric Temp E:/Witkranz_Testing/IN_TEMP_days.csv
Water Table E:/Witkranz_Testing/IN_Water_Table_days.csv
Sulphate E:/Witkranz_Testing/OUT_SO4.csv

Enter prediction period 11 days

Results saved as >>>>> Output file path

Figure 2 *Prediction processing window of the GUI.*

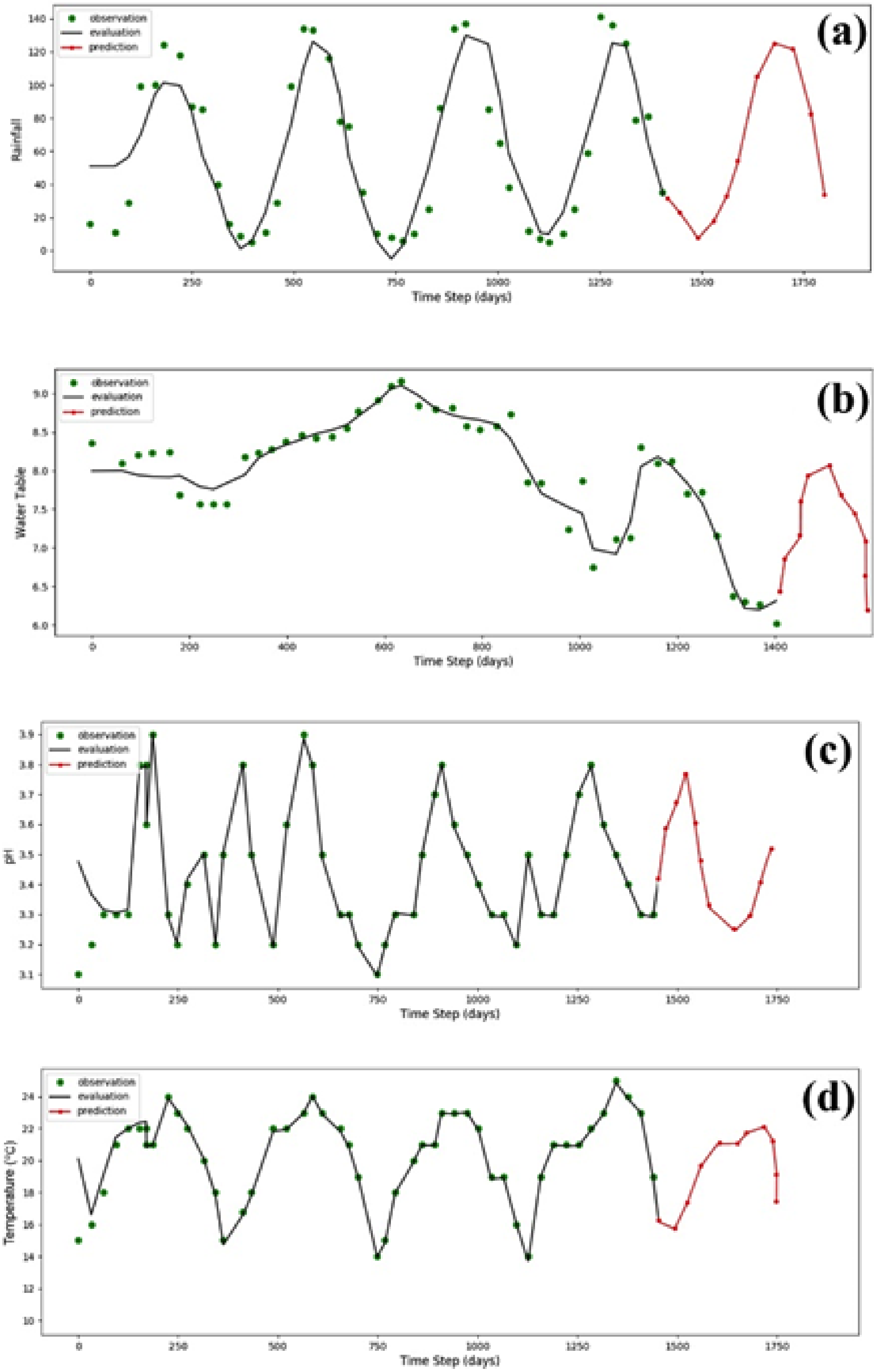


***Figure 3** Testing the input predictions for a) rainfall, b) water table, c) water pH, d) air temperature and input parameters for the Witkranz discharge site.*

against days with results for the training and prediction. The observation values are shown as green dots, the LSTM fitting model as a black line and the output prediction data as a red line for the four input parameters. Training was done using 1400 days of historical information and 350 days of future prediction.

ANN network system

Using the historical input parameters (rainfall, water table, temperature and pH), the ANN is trained, and the trained ANN model is used to predict the future sulfate concentrations by feeding the LSTM predicted future values of each input into the model.

Fig. 4 shows the schematic layout of the ANN system used in the development of the prediction system. The four inputs (which are outputs of the LSTMs for each input) are fed into the ANN input layer which, in turn, is linked to the output layer via the hidden layer. The hidden layer is used to adjust the node connections to establish the relationship between the inputs and the output.

The parameters used for the training are: Tensorflow (Gradient Descent Optimiser) with a value of 0.01 and a batch size value of 100. When the ANN system code is executed, the loss and accuracy graphs are plotted, as shown in Fig. 5. The graph shows the decaying mean square errors (MSE) as the number of training epochs increases which flattens as the number of training epochs increases. This corresponds to the time when the ANN is termed "trained network" with an MSE value of 0.054. The training accuracy increases sharply in the first 200 epochs and flattens around 99% accuracy, showing that the ANN system has been well trained.

The trained ANN model was used to predict sulfate values into the future from the last date of the historical data. Predictions were done from June 2018 to May 2019. The orange line in fig.6 shows the predicted sulfate values from the model. It is encouraging to note that the sulfate values are progressively decreasing over time. One possible reason is that the sulfide-bearing minerals may be depleted as the acid-producing reactions have been continuing for several years. As a part of future work, the system will be validated by collecting samples from the discharge site and will be refined by training the network with more data. Furthermore, the control processes affecting the sulfate concentrations in mine water will also be revisted and the system will be updated accordingly. The idea is to develop a preliminary AI prototype prediction system that can be reviewed and refined as when new data and information become available, a goal which was indeed achieved, as presented in this article.

Conclusions

Development of a future mine water quality prediction system should take the factors affecting the generation of AMD into consideration. The factors considered for the current study are rainfall, air temperature, water discharge pH and depth to water table. Sulfate concentrations were used as a proxy for the water quality index. The ANN system was designed and tested using data from

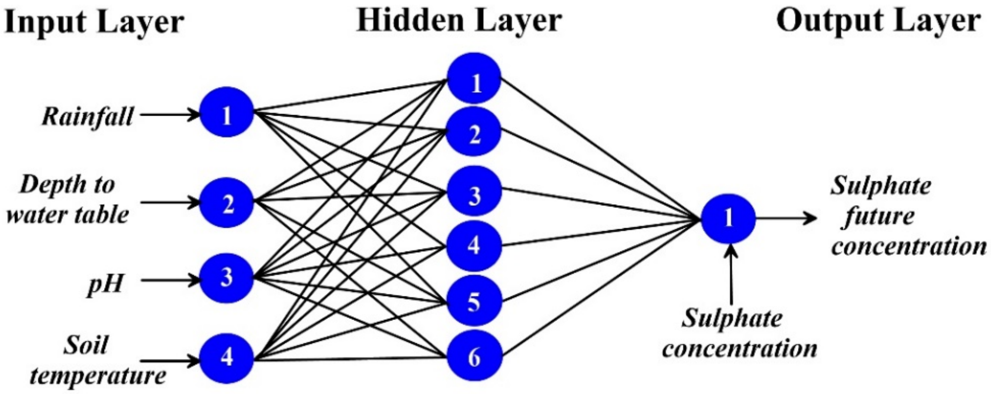


Figure 4 *Machine learning ANN system for the prediction of mine water quality.*

the Witkranz discharge site and over 99% training accuracy was obtained for a sample size of 46.

The effect of the water quality prediction system development are as follows:

- The system can be scaled up and used for the prediction of the future water quality of mine water, groundwater and surface water at regional and national scales. This would help in the management of these natural resources and raising awareness among the public and industry.
- The developing such a system will create more innovative possibilities and support the development of predictive mindsets with many benefits for the scientists involved and humanity at large.

Acknowledgements

The authors would like to thank the Council for Geoscience for providing financial assistance to carry out this research as a part of its Annual Technical Programme.

References

Chen QW, Mynett AE (2003) Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. Ecol Modell, 1–2: 55–67

Lee JHW, Huang Y, Dickmen M and Jayawardena AWN (2003) Neural network modelling of coastal algal blooms. Ecol Modell, 159:179–201

Najah A, El-Shafie A, Karim OA, Jaafar O, El-shafie AH (2011) An application of different artificial intelligence techniques for water quality prediction. Int. J. Phys. Sci. 6(22):5298–5308, doi: 10.5897/IJPS11.1180

Sakala E, Fourie F, Gomo M, Coetzee H (2018) GIS-based groundwater vulnerability modelling: A case study of the Witbank, Ermelo and Highveld Coalfields in South Africa. J. Afr. Earth Sci. 137:46–60, doi: 10.1016/j.jafrearsci.2017.09.012.

Weaver JMC, Cave L, Talma AS (2007) Groundwater sampling: A comprehensive guide for sampling methods. WRC Report No: TT 303/07, Water Research Commission, Pretoria.

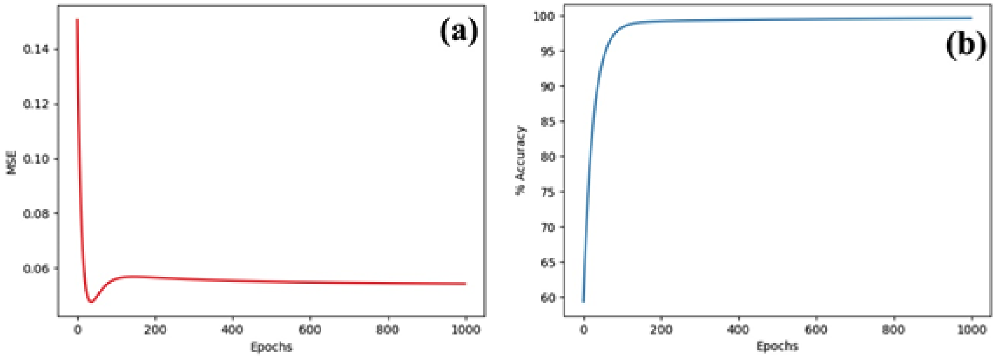


Figure 5 *Designed ANN system performance checks during training a) MSE b) training accuracy.*

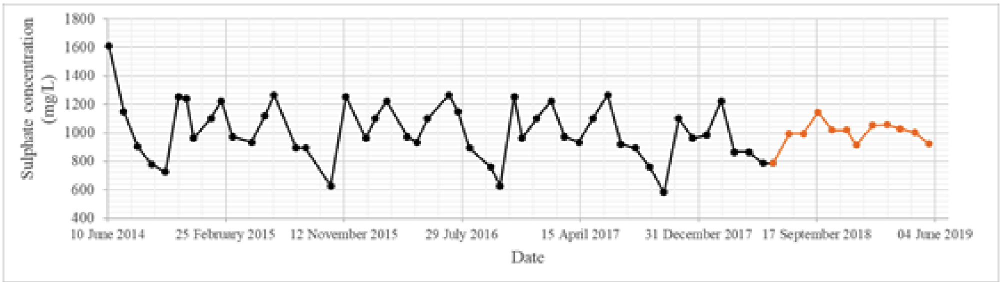


Figure 6 *Prediction values for sulfate using the developed prototype.*