ACS ES&T | Water

pubs.acs.org/estwater

Article

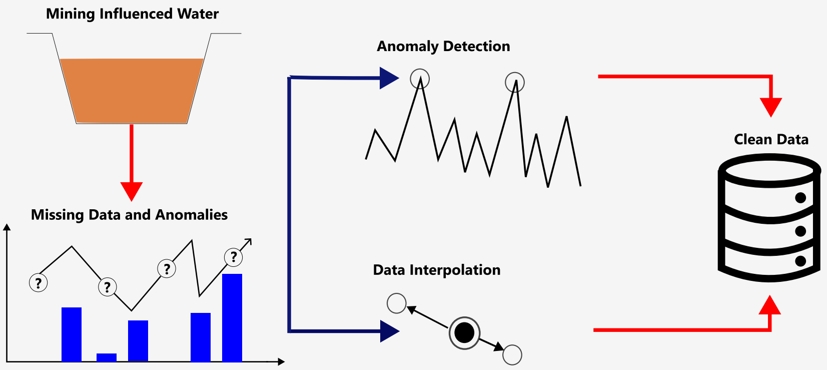
Exploring Advanced Statistical Data Analysis Techniques for Interpolating Missing Observations and Detecting Anomalies in Mining Influenced Water Data

Kagiso S More and Christian Wolkersdorfer*

Cite This: https://doi.org/10.1021/acsestwater.3c00163

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Collecting mining influenced water (MIW) quality data can result in incomplete data sets with missing values and anomalies, making it challenging to use the data for optimizing mine water management. This work explores advanced statistical data analysis approaches for addressing missing data interpolation and anomaly detection in MIW data sets. The study compares the performance of five different interpolation techniques and four different anomaly detection techniques using supervised and unsupervised machine learning algorithms developed using Python 3.8.16. The results of the study demonstrate that the radial basis function, spline, and *k*-nearest-neighbors interpolation techniques, along with the predictive confidence interval level anomaly approach based on gradient boosting regression trees, perform best for missing data interpolation and anomaly detection, respectively. Thorough application of these advanced techniques can improve the accuracy and reliability of mine water quality data, which is crucial for making conclusions on the safety of the environment, public health, and effective MIW management. This paper highlights the importance of developing effective methods for addressing missing data and anomalies in MIW data sets, which can ultimately lead to improved treatment plant optimization.

KEYWORDS: *Supervised and Unsupervised Machine Learning Algorithms, Python Programming Language, Mining Influenced Water Data, Missing Data Interpolation, Anomaly Detection*

1. INTRODUCTION

Statistical data analysis plays a crucial role in the mining industry, especially when it comes to managing mining influenced water (MIW). First introduced by Schmiermund and Drozd,[1] the term MIW was defined by McLemore[2] as any water resulting from mining activities, regardless of its characteristics, including but not limited to acid/rock mine drainage (AMD/ARD). The data set for this study is based on circumneutral mine water from an abandoned gold mine with an average pH of 6.55 (based on proton activity) and an electrical conductivity of 3474 μS/cm, mainly due to a sulfate concentration of 2434 mg/L and an Fe concentration of 175 mg/L (Table 1). Mine water management is a complex and challenging task that requires the collection and analysis of large amounts of data from various sources.[3,4] Therefore, statistical data analysis is used to make sense of these data and make informed decisions about how to manage MIW effectively. In many cases, mine water data contain missing data and anomalies due to several factors such as missed sampling days or equipment malfunctioning. Missing data and anomalies are the main factors contributing negatively in MIW chemistry interpretations; thus, advanced statistical data analysis approaches need to be applied. Data play a critical

Special Issue: Applications of Artificial Intelligence, Machine Learning, and Data Analytics in Water Environments

Received: April 3, 2023
Revised: July 13, 2023
Accepted: July 13, 2023

Table 1. Mine Water Quality Data Set from Shaft No. 9 of the Randfontein Gold Mine from 2016-03-07 to 2021-07-13 (Westrand Mine Water Pool)[a]

parameter	n	$\bar{x}$	σ	min	max
acidity, mg/L $CaCO_3$	1123	406	337	48	1484
alkalinity, mg/L $CaCO_3$	1123	155	55	70	298
EC, mS/m	1123	347	47	187	497
Fe, mg/L	1111	175	158	14	668
Mn, mg/L	1111	28.0	7.0	10.0	45.2
pH	1123	6.5	0.3	5.8	9.4
SO_4, mg/L	989	2436	303	1833	3184
temp, °C	1123	19.7	2.1	9.5	26.2
turbidity, NTU	1116	22	36	0.7	275

[a] Definitions: n, number of measurements; $\bar{x}$, average; σ, standard deviation; min, minimum value; max, maximum value. pH average calculated as $-\log_{10}[\sum C_i/n]$, where C is the proton activity (www.wolkersdorfer.info/pH_en); measured values and units as reported by the plant (from More and Wolkersdorfer[10]).

role in monitoring, control, optimization, maintenance and research activities at mine water treatment plants.[3,4] So, missing data and anomalies can result in biasness and give wrong interpretations which may lead to wrong plant optimization decisions such as developing bias forecasting models. Therefore, addressing missing data and anomalies in mine water data is crucial for maintaining data quality and ensuring that data can be relied on to make further treatment plant optimization decisions.

Mine water management involves the collection, treatment, and discharge of water that is encountered during mining operations. This water can include groundwater, surface water, and water that are used in the mining process.[5−7] Managing this water is essential to ensuring that it does not harm the environment or human health. However, due to the complex nature of mining operations, managing MIW requires advanced statistical data analysis approaches as opposed to the current methods (summary in Text S1 in the Supporting Information). Mine water management involves dealing with a wide range of environmental factors, including water quality, geological context, hydrological characteristics, and operational variables, making this a complex and multidimensional problem. Current statistical methods used in mine water management tend to focus on a single parameter, using linear or quadratic interpolation to fill in missing data and focus on the z-score method or visual detection of outliers or anomalies. Compared to advanced statistical data analysis techniques, these methods are therefore less effective in the treatment and management of MIW due to their limited ability to capture and process complex patterns and relationships in the data.[8,9] This manuscript presents techniques that use all the data in a given data set to uncover insights and patterns in the specific parameter of interest. As a result, these insights can be used to improve mine water management, predict mine water chemistry, and provide effective, sustainable water management.[10] Consequently, this highlights the potential of advanced statistical data analysis methods in addressing the environmental challenges associated with mine water treatment and management.

Collection or measurements of mine water data are conducted at various times under different conditions, and in most cases missing data occur due to challenges that are referred to as the missingness mechanism.[11−13] Mine water data can also contain unexpected observations in its data set, which are referred to as anomalies. When statistical data analysis techniques are applied, missing data interpolation is normally conducted in parallel with anomaly detection. These steps are crucial and must be conducted thoroughly, especially when the data need to be used to build models. Missing data and anomalies can have serious consequences in mine water management, e.g., missing data can lead to inaccurate conclusions about mine water quality, while anomalies can make it difficult to identify and respond to potential environmental risks. Therefore, it is imperative to address missing data and anomalies in MIW data to ensure that they can be used to make informed decisions about mine water management.

Exploratory data analysis is conducted on the data before they can be used on the selected algorithms to detect anomalies and interpolate the missing observations. This method is conducted to understand the behavior of the data and as a guidance to apply appropriate techniques such as the probabilistic methods (e.g., ref 14) or numerical statistical modeling (e.g., ref 10). This paper explores several advanced numerical statistical methods for interpolating missing data, such as k-nearest-neighbors (kNN) interpolation, radial basis function (RBF) interpolation, spline interpolation, multivariate adaptive regression splines (MARS), and artificial neural network (ANN) interpolation. In addition, advanced anomaly detection methods are explored, such as one-class support vector machines (SVMs), isolation forest, long short-term memory (LSTM) autoencoder-based anomaly detection, and predictive confidence interval level (PCIL) approach with regression models.

Although several authors have used the term “numerical statistical models”, it has not yet been defined by any of them.[15−17] Numerical models were introduced in the 1970s with computers or the digital revolution. They solved real-world problems with mathematical equations describing the relevant processes. A common feature of all of these “mathematical” solutions is the use of numerical methods to solve the governing equations in an iterative approach. Similar to numerical models, modern statistical approaches require the use of sophisticated algorithms and computations to deal with large data sets consisting of multiple variables. Because machine learning belongs to this group of leading-edge approaches, these statistical methods can be termed “numerical statistical models”.

Statistical data analysis approaches are used to make sense of the large amounts of data that are collected during mine water management and uncover insights that can be used to improve operations. Additionally, these approaches allow mining houses to make informed decisions about mine water management and reduce the environmental effects of mining operations in the context of integrated water resource management (IWRM). With the increasing amount of data being generated on a daily basis, the importance of these approaches will continue to play a crucial role in mining operations. Therefore, components that form a substantial part of these approaches, such as data interpolation and anomaly detection, need to be taken seriously, and their techniques must be advanced to allow the improvement of data accuracy and reliability. The goal of this paper is to present for the first time comprehensive advanced statistical data analysis methods to interpolate missing data and detect anomalies in the MIW data set.

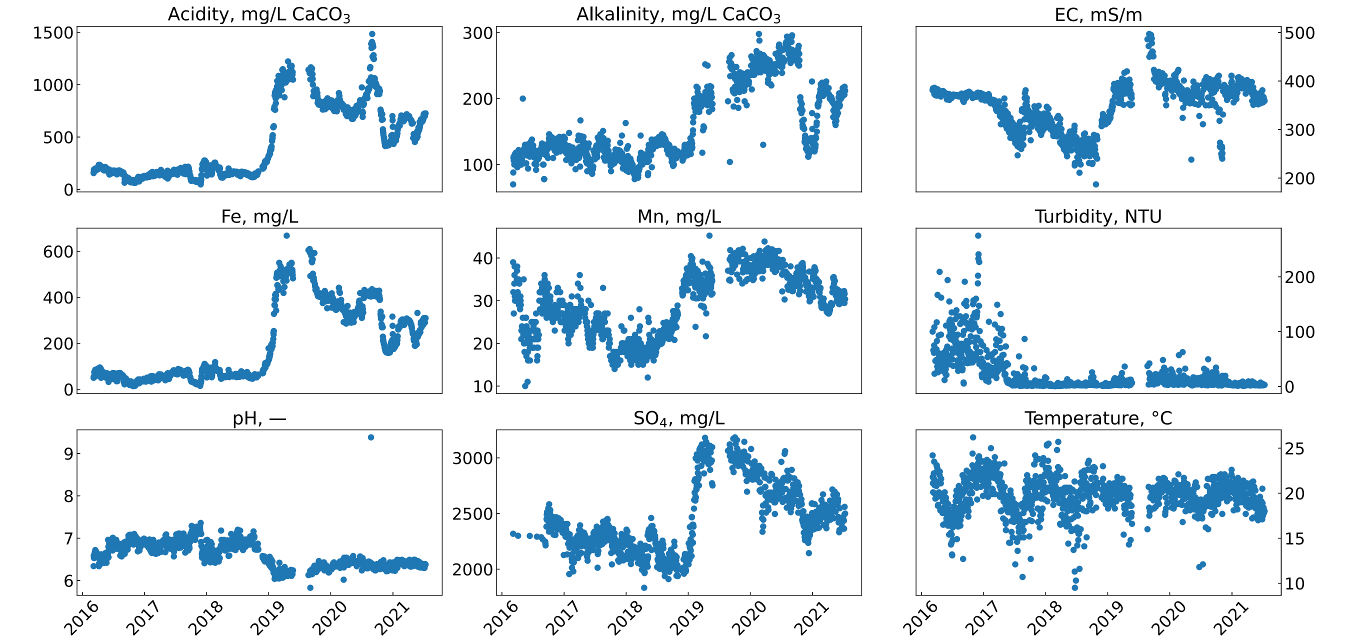


Figure 1. Westrand mine water pool data from 2016-03-07 to 2021-07-13, including possible anomalies and missing data.

2. METHODS

2.1. MIW Data. This study uses the historical data from the Westrand mine water pool gathered between the years 2016 and 2021, i.e., all the explored techniques were tested with the same data set. The full data set consists of nine parameters, i.e. electrical conductivity (EC), alkalinity, acidity, Mn, Fe, SO_4, pH, turbidity, and temperature (Figure 1 and Table 1). It contains the missing observations and was tested for anomalies. Python programming language was used to develop the missing data interpolation and anomaly detection techniques.

2.2. Data Exploration. Statistical modeling depends on the results of the statistical tests applied on the data set before models can be built.[18] Statistical tests applied on the data set include stationary tests, normality tests, and data transformation. For the stationarity of the data set, the augmented Dickey Fuller (ADF) test was used, which resulted in p-values of the parameters being greater than 0.05 (except for temperature), meaning that the data set is nonstationary. Numerical statistical models can be particularly useful in analyzing nonstationary data, unlike the probabilistic models, and thus this paper explored several numerical statistical modeling techniques. Nonstationary data often contain trends that may require modeling capabilities to capture and understand the underlying dynamics. Numerical statistical models can also help simulate and predict the behavior of nonstationary systems by incorporating relevant factors and variables (e.g., refs 19 and 20).

In identifying the data that has been sampled from a normal distribution, Kolmogorov−Smirnov, Shapiro−Wilks, and Anderson−Darling normality tests were conducted.[21−23] The p-values for the parameters were below 5% using the Kolmogorov−Smirnov and Shapiro−Wilks tests, meaning that the data do not follow a normal distribution. For the same reason, data did not follow a normal distribution applied with the Anderson−Darling test when the test statistics were all above the critical values at α = 0.15, 0.10, 0.05, 0.025, and 0.01 (0.574, 0.654, 0.784, 0.915 and 1.088). The data set showed a normal, bimodal, multimodal, and log-normal distribution. The goal was to use normally distributed data to build missing data interpolation and anomaly detection models to avoid the random or unpredictable variations that may be present in the data; therefore, the natural log-normal transformation was used.

The indicator parameters at the Westrand mine water pool are Fe and acidity. However, for this study, the main focus was on the Fe parameter. Therefore, the data were compared using the Pearson correlation coefficient (Figure S1) and data distribution statistical charts (Figure 2). This was conducted to select the parameters that have a good relationship with the Fe parameter so they can be used as inputs for the models. Therefore, from the statistical graphical charts, acidity, alkalinity, pH, SO_4, and Mn displayed a good relationship with Fe and thus were used as inputs. However, some of the algorithms used (e.g., splines and RBF) only require a single parameter to perform computations. Therefore, these algorithms were only modeled with the target output, Fe. This approach is not entirely accurate, as the parameters in mine water influence each other; thus, it is recommended that the algorithms used for statistical data modeling must have both inputs and the target output/s. Nonetheless, these algorithms are included in this study, so they can be compared with others.

3. RESULTS AND DISCUSSION

3.1. Long Short-Term Memory Autoencoder-Based Anomaly Detection. A long short-term memory (LSTM) autoencoder-based anomaly detection approach is a machine learning model that uses a type of artificial neural network (ANN) known as an autoencoder to detect anomalies in complex data sets, i.e. the type of ANN used in this approach consists of an encoder and a decoder submodel.[24] The model is trained on a complex data set, and the LSTM network is used to learn the normal patterns and trends in the data. In the LSTM model, input values are downscaled and reconstructed, captured as a latent vector, and decoded. If the reconstruction

https://doi.org/10.1021/acsestwater.3c00163

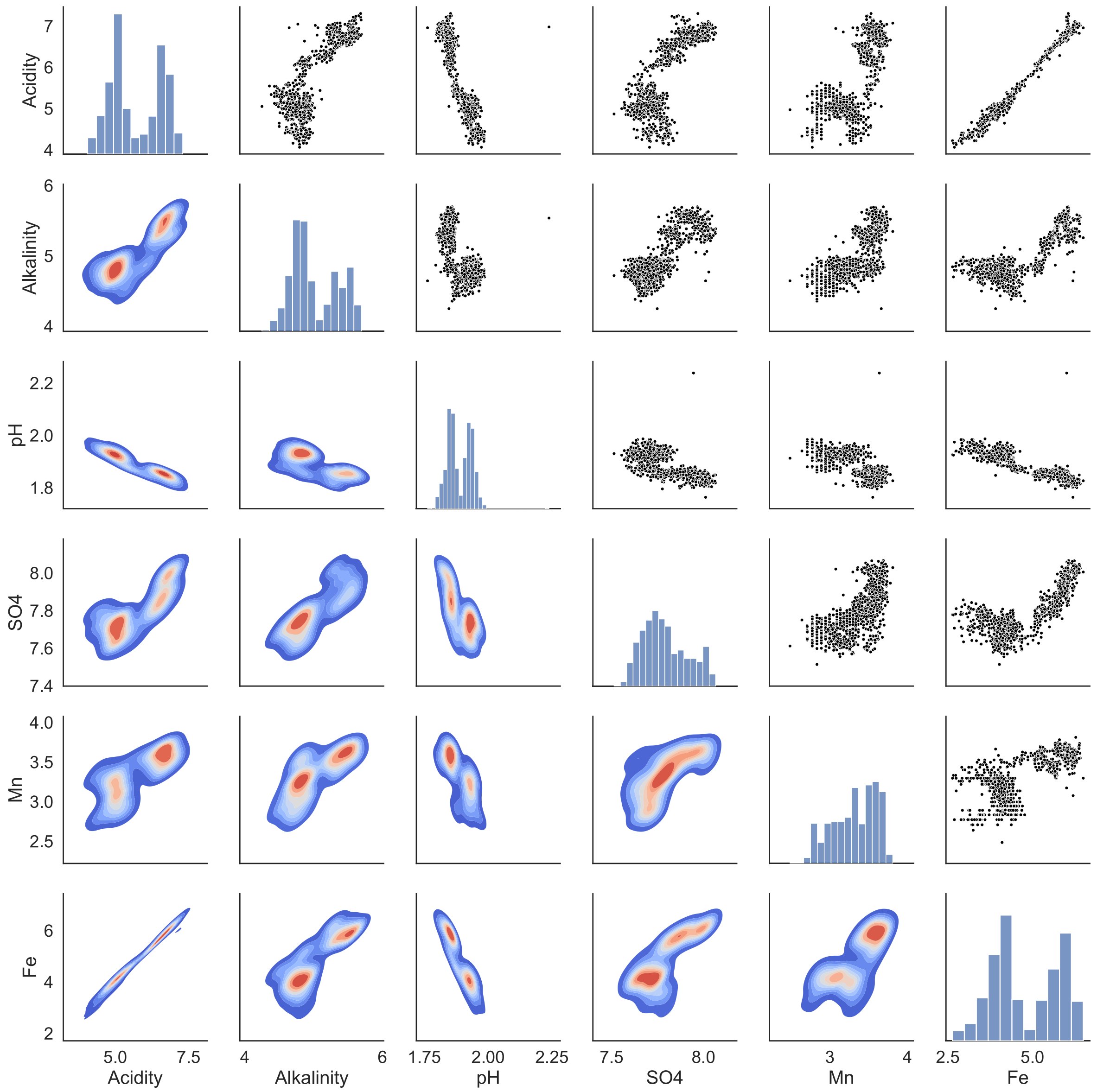


Figure 2. Transformed data distribution of parameters to be used in building data interpolation and anomaly detection models.

error is high, it indicates that the input is an anomaly and deviates from the normal behavior learned by the model (e.g., ref 25). An LSTM autoencoder consists of five layers: input sequence, encoder, latent vector, decoder, and reconstructed sequence.[24,26,27] This type of model can be useful for detecting unusual patterns or events in mine water data and potentially flagging them for further investigation. In this study, the LSTM autoencoder-based anomaly detection model was built using Tensorflow version 2.2.0 as the backend and Keras version 2.4.3 as the core model development module. It should be noted that the latest versions of the aforementioned modules do not work on this technique as it was developed for earlier versions of Python (i.e., Python 3.7 was used for this method instead of version 3.8).

The LSTM autoencoder structure used in this study contained 16 memory units with 80% of the data being used for the training set and 20% for testing and validation purposes. This structure takes in the input data and creates a compressed version of it. Input parameters were compressed into a single feature vector, and for the output to regenerate a dimension similar to the original input, the repeat vector layer was added to convert the feature tensor from one-dimensional to two-dimensional. Therefore, the final decoder output layer provided the reconstructed input data. This LSTM autoencoder model was fitted over 30 epochs with a batch size of 32 and was further compiled using the adaptive moment estimation (Adam) optimizer and mean absolute error (MAE) for calculating the loss function (Figure S2A). After the loss function in the training and testing sets was examined, a suitable threshold value for identifying anomalies was computed. Therefore, the reconstruction loss in training and testing sets (Figure S2B,C) was computed to determine when the observations of the parameters crossed the anomaly threshold (Figure S2D,E). Finally, the model did not detect any anomalies in the target output (Fe concentrations).

3.2. Predictive Confidence Interval Level with Regression Models for Anomaly Detection. The predictive confidence interval level (PCIL) is a process of fitting a nonlinear regression model to the data with different prediction bands to detect anomalies. In this approach, the

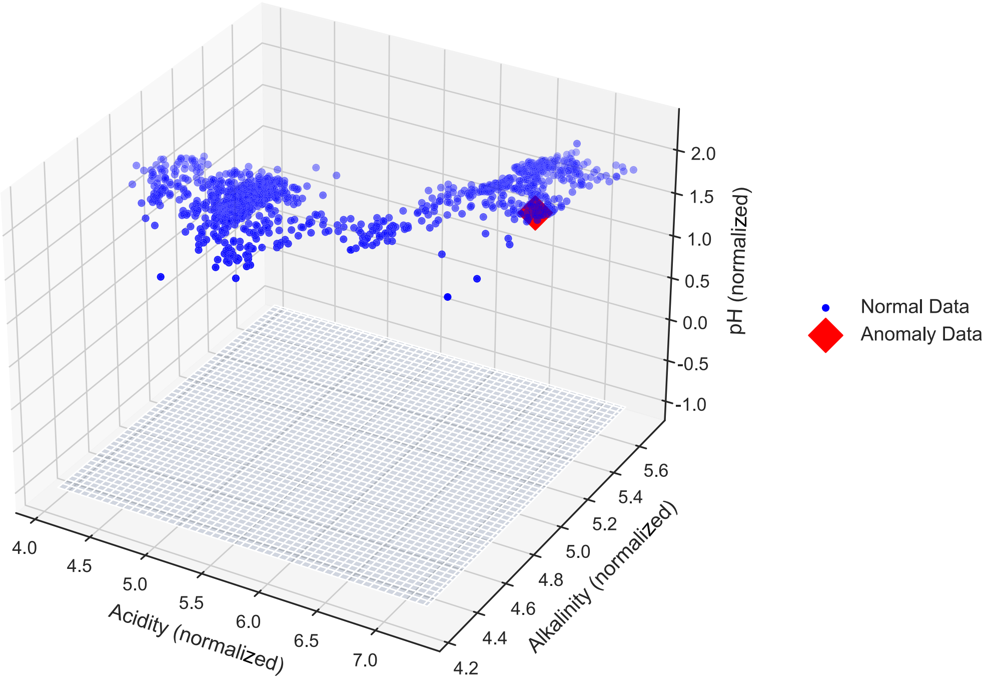


Figure 3. Fe data plotted with its anomalies in relation to acidity, alkalinity, and pH as model inputs using the one-class SVM model. Normalized data (natural logarithmic and standard scaler normalization) were used to develop the model.

Scikit-learn library version 1.2.1 was used to build a gradient boosting regression tree model with three prediction bands: α = 0.10, α = 0.50, and α = 0.90 (Figure S3). α = 0.10 and α = 0.90 represent the lower and upper bounds of the data, and data that plotted below the lower band and above the upper band were considered anomalies. For the midprediction, α was set to 0.50 with the least-squares loss function. Therefore, values plotted on the α = 0.50 prediction band can be used to replace possible anomalies.

Gradient boosting regression tree model structure buildup consisted of 100 trees, a maximum depth of 2, and a learning rate of 0.05, and data were split into 20% of the test size. The loss function of the gradient boosting model was changed to quantiles with three prediction intervals α used. This kind of configuration ensures that the model performs predictions that correspond to percentiles.

3.3. One-Class Support Vector Machines for Anomaly Detection. One-class support vector machine (SVM) is an anomaly detection model that is unsupervised. In contrast to the traditional supervised SVM, this model does not require labeled data during training.[28] Rather, it establishes the boundary for normal data and considers anything outside of this boundary to be an anomaly (e.g., refs 29 and 30). One-class SVMs are useful for problems where only one class of data is available for training and can be applied in a variety of fields such as fraud detection, intrusion detection, water quality analysis, and fault diagnosis.

One-class SVMs can be useful in detecting anomalies in mine water data, and these models can be trained on a data set of normal operating conditions for the mine water system and can be used to identify any anomalous behavior in the system. Anomalies may include variations in mine water quality, mine water flow rates, or pressure that could indicate a potential problem or failure in the system. When these anomalies are identified early, maintenance and repair can be carried out before more serious problems can occur.[31,32] One-class SVM can be a useful tool for mine operators to monitor the conditions of the water systems and prevent costly downtime.

In building the one-class SVM model, Scikit-learn version 1.2.1 was used as the core library. Anomalies were detected in the Fe data, and model inputs used were acidity, alkalinity, and pH. A 3D anomaly visualization was taken into consideration when selecting model inputs; thus, only three parameters showing a strong relationship with Fe were considered. The target column, Fe, has continuous numerical values, and the one-class SVM works with binary data. Therefore, the threshold values were defined to determine what values of the anomaly column should be considered as "normal" and "anomalies". Lower and upper threshold values were determined by using the quartiles, i.e. $Q_1 - 1.5(Q_3 - Q_1)$ for the lower and $Q_3 + 1.5(Q_3 - Q_1)$ for the upper threshold.

For the model buildup, the test size was set to be 20%, and the model was initialized using a ν value (a parameter that controls the number of training points that are considered to be support vectors) of 0.1. Furthermore, a kernel function (which is used to transform the input data to a higher dimensional space) of RBF was applied and a γ value (a parameter used to control the width of the Gaussian normal distribution used to compute the similarity between two points in the transformed space) of 0.1 was used. In the SVM models, ν, kernel and γ are the parameters used for configuration.[33] The parameters can be changed to enhance the performance of the model. Finally, the model only detected one anomaly (668 mg/L) (Figure 3), and this is strongly convincing, as it is the largest value from the Fe data (Figure 1).

3.4. Isolation Forest for Anomaly Detection. Isolation forest is an unsupervised machine learning algorithm that uses a decision tree algorithm[34,35] to identify anomalies by separating them from the rest of the data.[36] This is accomplished by randomly selecting a feature and then selecting a split value between the minimum and maximum values of that feature. Anomalous data points will have shorter paths in the resulting trees due to this random partitioning, making them stand out from the rest of the data (e.g., refs 36−38). Unlike typical anomaly detection methods, which start by defining what is considered normal and then flagging anything that falls outside of that definition, isolation forest does not have a predefined notion of normal behavior. Instead, it explicitly separates anomalous points from the rest of the data set.

Isolation forest can be used for anomaly detection in the mine water data. Since anomalies in mine water can be caused by various factors such as changes in water quality, flow rates, or pressures, isolation forest can be used to identify these anomalous patterns based on their different characteristics compared to normal operating conditions.

In building the isolation forest model, Scikit-learn version 1.2.1 was used as the main library. Anomalies are detected on the Fe concentration, with acidity and pH being used as input parameters. The model was compiled by using 100 decision trees to build the forest, a contamination value of 0.01, and a random state of 42, and the data were split into 80% training size and 20% testing size. The distribution of anomaly scores computed by the isolation forest model for the input data is indicated by the anomaly score histogram (Figure S4A). An anomaly score is a measure of how isolated a data point is, with isolated points having a higher absolute score and nonisolated points having a lower absolute score.[39] Therefore, the histogram shows how many data points have anomaly scores above a threshold. This histogram is shown alongside a scatter plot (Figure S4B) of the input data, with anomalies highlighted in red. Thus, this allows for a visual comparison of the anomaly scores with the actual data points and shows how the algorithm is able to identify anomalies in the data.

Similar to the one-class SVM model, the isolation forest model detected only one observation of Fe concentration as an anomaly. According to the isolation forest model architecture, this conclusion makes statistical sense, because the detected anomalous observation is isolated from the rest of the population (Fe concentration data). However, the visual representation of the scatter plot implies that there are two anomalies. Different model parameters were applied to see if the model can detect other anomalies and supplement what can be seen in the scatter plot. It should be noted that all of the model architectures used still detected only one anomaly.

3.5. *k*-Nearest-Neighbors Interpolation. *k*-nearest-neighbors (*k*NN) interpolation is a method of estimating values for missing data points based on the values of the *k* nearest neighbors. The *k* value in the *k*NN is the number of nearest neighbors to consider for predicting the class label of a sample. The idea is that the majority of the *k* nearest samples to a given test sample belong to the same class, so the class label of the test sample can be assigned accordingly.[40,41] For mine water data, *k*NN interpolation is hardly used due to the complex nature of the data sets. However, this approach can be used to accurately interpolate and predict the missing values for mine water quality parameters such as pH, EC, and TDS.[42] It can be applied by identifying the *k* closest data points to the missing value based on some distance metric, such as Euclidean distance, and then averaging the values of those *k* points to estimate the missing value (e.g., ref 43). *k*NN interpolation is advantageous as it is relatively simple to implement, analyzes data robustly, and can provide good results when the data have a high degree of spatial autocorrelation. However, in certain cases, it can be sensitive to the choice of *k*, and the results may be affected by anomalies in the data; thus, robust ways to detect anomalies are introduced in this study. In addition, it is imperative to note that in mine water data, it may be necessary to consider additional factors such as the type of mining operation and the geological setting of the area to ensure that the interpolated values are accurate.

*k*NN interpolation method was built by using the "KNeiborsRegressor" class from the Scikit-learn library version 1.2.1. The model was built using a test size of 20%, and a random state of 10 and further compiled with 5 nearest neighbors, a "distance" weight parameter which uses the inverse of the distance as the weight of each neighbor, and a "Minkowski" metric, which uses the Minkowski distance as the metric to determine the nearest neighbors. Finally, the model performed relatively well with a mean absolute error (MAE) of 0.06957, a mean squared error (MSE) of 0.01002, and an r^2 value of 0.9968 between the measured and predicted/interpolated concentrations (Figure S5). Therefore, the trained model was used to predict and interpolate the Fe concentration.

3.6. Radial Basis Function Interpolation. Radial basis function (RBF) interpolation is a technique used for approximating a multivariate function from a set of scattered data points. This interpolation approach makes use of a radial basis function such as linear, cubic, multiquadric, or Gaussian as a weighting function to give a value to a point in space based on its distance from known data points.[44,45] The coefficients of RBF are determined by the values of the function at the known points; therefore, this can further be applied to determine the value of the function at any other point. In the RBF method, the interpolating function is defined as a linear combination of RBFs centered at the data points. RBF can be used to estimate unknown parameter values in mine water management based on a limited number of available measurement data, making it a useful interpolation technique. The RBF technique was developed using the SciPy.interpolate class from the SciPy library version 1.9.3. There are several functions that can be used to perform interpolation using the RBF technique, and the suitable one for this study was the linear function (Figure S6).

3.7. Spline Interpolation. Spline interpolation is a technique for estimating the missing observations of a function between known points.[46,47] In MIW data, spline interpolation can be used to estimate the missing concentrations and values of the mine water physicochemical parameters based on a small number of available measurement data. This approach works by constructing a piecewise polynomial function that goes through the data points and is continuous and smooth at the control points. Control points are the points at which the polynomial functions are joined and used to determine the shape of a spline curve.[48,49] A interpolate.CubicSpline class from the SciPy version 1.9.3 library was used to perform spline interpolation, and this resulted in piecewise polynomial functions being formed and interpolation taking place (Figure 4).

3.8. Multivariate Adaptive Regression Spline Interpolation. Multivariate adaptive regression splines (MARS) is a type of nonparametric regression technique that is often used to model complex relationships between multiple input variables and a single target output.[50,51] This approach is useful in interpolating the missing mine water data, as it incorporates the relationships between all of the available

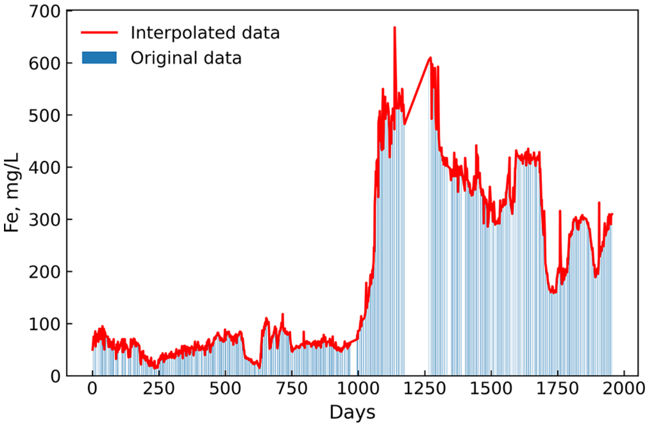


Figure 4. Spline interpolation for Fe concentrations of the Westrand mine water pool data from 2016 to 2021.

parameters to interpolate a single target output. The MARS technique works by constructing a piecewise linear function that is adapted to the data through recursive splitting of the input space into regions, where the linear function within each region is defined by a set of basis functions.[50] Therefore, the resulting model can then be used to predict the concentration or values of the target output based on the values of the measurement variables. MARS can provide a powerful and flexible way to model complex relationships in mine water data.

MARS model was built using the pyGAM library. pyGAM is a flexible and user-friendly library that makes it easy to fit and predict with generalized additive models (GAMs). A test size of 20% and 10 splines were utilized to fit the model, and MAE and MSE values of 0.07067 and 0.00988 were produced, implying that the model performed relatively well on the fitted data. Statistical analysis results were relied on and used Fe as the target output, and the model finally predicted and interpolated Fe concentrations (Figure S7).

3.9. Neural Network Interpolation. Similar to MARS, neural networks are also used to model the relationship between a set of input variables and a single output variable. Neural network interpolation is a supervised regression machine learning technique that works by constructing a multilayer feedforward neural network that maps the input variables to the output variable.[52,53] This network is trained on the available data, and the resulting model can be used to estimate the target output based on the values of the input variables (e.g., refs 54−56). Neural network interpolation often gives good results, especially when the data set is nonlinear such as MIW data. A neural network structure consists of neuron that forms layers, and the layers are connected by the weights that are associated with the bias. The network has an input layer in which raw data are initially introduced into the network, hidden layer(s) to perform the network's computations, and an output layer in which the final prediction for the network takes place.

Three core libraries, Scikit-learn" version 1.2.1, TensorFlow version 2.10.0, and Keras version 2.10.0, were used to build the neural network interpolation model. From the statistical analysis results, Fe was used as the target output, while alkalinity, acidity, SO_4, pH, and Mn were used as model inputs. The neural network structure consisted of two hidden layers with four and eight neurons, and a rectified linear (ReLU) activation function was used for both the hidden and output layers. A test size of 20% was used for the model and was compiled using the mean squared error (MSE) loss function and the adaptive moment estimation (Adam) optimizer. Furthermore, the model was fitted using 50 epochs and a batch size of 32 and yielded very low error values (training MSE, 0.1254; test MSE, 0.1239; validation MSE, 0.1239; Figure S8A), indicating that the model performed well on the supplied data. Therefore, the model was used for prediction/interpolation, and the results were compared with the measured data (Figure S8B).

4. MODEL COMPARISON

4.1. Anomaly Detection Models. The anomaly detection techniques applied in this study identified about 0.5% of the total samples as anomalies (Figure 5). This value is close to that of Wolkersdorfer,[57] who identified that 0.6% of his 8611 individual measurements of various parameters in MIW can be classified as outliers. The ground truth regarding the detected anomalies in the data set is unknown, making it impossible to determine the exact nature of each individually measured data point. Furthermore, the definition of anomalies can vary based on subjective judgments and statistical thresholds (see discussion in ref 57). Therefore, it is important to carefully consider and understand that anomaly detection methods depend on specific assumptions and thresholds defined by the researcher and methods applied.

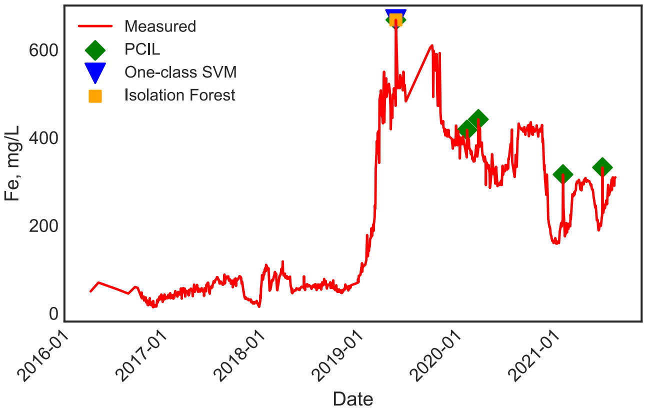


Figure 5. Comparison of the three anomaly detection models used to detect anomalies in Fe concentrations of the Westrand mine water pool from 2016 to 2021. The LSTM autoencoder is not included because it did not detect any anomalies.

From the results obtained, the PCIL approach using a gradient boosting regression tree model appears to be the most effective model for detecting anomalies, as it was able to detect five anomalies from the Fe concentration. On the other hand, the LSTM autoencoder did not detect any anomalies. This may be because LSTM autoencoders are more suitable for detecting anomalies in temporal data with a clear pattern or trend,[25,58] whereas the data used in this study do not have a clear pattern or trend that the LSTM model could detect. The one-class SVM and isolation forest models both detected only one anomaly, which suggests that they may not have been as effective as the PCIL approach in detecting anomalies in the data. However, it is worth noting that the performance of these models can be highly dependent on the anomaly threshold set, specific parameters, and settings used during training and they may perform differently with different data sets.

It is worth noting that selecting anomaly threshold values such as in the autoencoder and one-class SVM models has great potential to influence how the algorithms perform. Choosing an anomaly threshold value involves determining the point at which data points are classified as anomalies or normal. Various techniques can be applied to set the threshold values, including the statistical methodological approaches, receiver operating characteristic curve, and precision-recall curve, and the threshold can also be set by using the knowledge and experience of working with the data set. This study used the subject matter knowledge in the autoencoder model and the statistical methodological approach in the one-class SVM model. The statistical method approach involves using measures such as mean, standard deviation, and percentiles to define a threshold based on deviations from the normal data distribution. Therefore, the results obtained are accurate only for the methods used, and different results can be obtained by using different approaches of selecting the anomaly threshold.

4.2. Missing Data Interpolation Models.

Interpolation is a technique used to estimate data points based on the available data. In the context of mine water management, accurate and reliable estimates of concentrations are essential for the decision-making process. For this study, statistical methods were applied to select the target output to be used in the models. Therefore, the missing concentrations of Fe were then predicted and interpolated by using various techniques. The ANN, *k*NN, MARS, RBF and spline techniques were applied to fill in the gaps in Fe concentrations. In addition, the performance of each method was evaluated based on its ability to accurately predict missing data. Real (measured) data was used to train the algorithms, and the resulting model was used to predict the missing, new data in the gaps. The predicted new data were then compared to the real data to calculate the values of the evaluation metrics. Therefore, the results indicated that spline, RBF, and *k*NN performed relatively well, with MARS performing slightly well, while ANN did not perform well, as its predictions and interpolations are slightly far away from the original data (Figure 6).

5. CONCLUSIONS

This study presented advanced statistical data analysis approaches for missing data interpolation and anomaly detection in an MIW data set. Five interpolation techniques (*k*NN, RBF, spline, MARS, and ANN) and four anomaly detection techniques (one-class SVMs, isolation forest, LSTM autoencoder-based anomaly detection, and PCIL approach using gradient boosting regression tree model) were evaluated and compared for their accuracy and effectiveness. The results of the study show that the RBF, spline, and *k*NN interpolation techniques outperformed other interpolation techniques for missing data estimation, while the PCIL approach technique performed best for anomaly detection. These findings demonstrate the importance of using advanced statistical data analysis approaches for addressing missing data and anomalies in MIW data set.

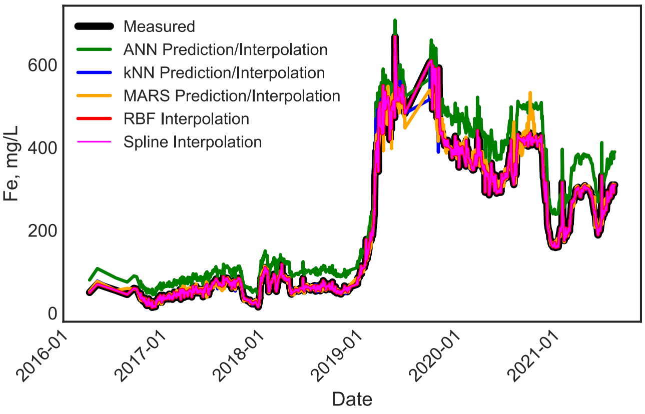


Figure 6. Comparison of the five missing data interpolation models used to interpolate Fe concentrations of the Westrand mine water pool from 2016 to 2021.

The implications of this study are important for ensuring the safety of the environment and public health in mining areas. Accurate and reliable data analysis is essential for identifying potential risks and developing effective mitigation strategies. The advanced statistical data analysis approaches presented in this study provide a powerful tool for improving the accuracy and reliability of water quality data analysis in mine water treatment plants.

Future research can further improve the performance of the interpolation and anomaly detection techniques by exploring alternative algorithms and parameter optimization. Additionally, software embedded with the explored algorithms would be beneficial and advantageous to the industry so that non-programmers can easily implement these techniques. The application of these techniques can be extended to other environmental data sets beyond MIW, such as wastewater data, air quality, soil quality, and weather data. In conclusion, the study contributes to the advancement of statistical data analysis approaches for environmental data analysis, highlighting the potential of these techniques to support decision-making processes and promote sustainable resource management.

■ ASSOCIATED CONTENT

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestwater.3c00163.

Description of existing techniques for missing data interpolation and anomaly detection including additional literature on these techniques and supplementary figures describing Pearson correlation matrix of the data set, LSTM autoencoder model performance, predictive confidence interval level approach for Fe, isolation forest anomaly score, *k*NN model prediction, RBF interpolation, MARS technique, and ANN model performance (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Christian Wolkersdorfer − *SARChI Chair for Mine Water Management, Department of Environmental, Water and Earth Sciences, Tshwane University of Technology, Pretoria 0001, South Africa;* orcid.org/0000-0003-2035-1863; Email: christian@wolkersdorfer.info

Author

Kagiso S More − *SARChI Chair for Mine Water Management, Department of Environmental, Water and*

Earth Sciences, Tshwane University of Technology, Pretoria 0001, South Africa; orcid.org/0000-0003-2803-7983

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsestwater.3c00163

Author Contributions

Conceptualization: K.S.M. Funding acquisition: C.W. Software: K.S.M. Supervision: C.W. Writing—original draft: K.S.M. Writing—review and editing: C.W. CRediT: **Kagiso S More** conceptualization (lead), data curation (lead), formal analysis (lead), investigation (equal), methodology (lead), software (lead), validation (equal), visualization (equal), writing-original draft (lead), writing-review & editing (equal); **Christian Wolkersdorfer** conceptualization (supporting), data curation (supporting), formal analysis (supporting), funding acquisition (lead), methodology (supporting), project administration (lead), resources (lead), supervision (lead), validation (equal), visualization (supporting), writing-review & editing (equal).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the National Research Foundation (NRF Grant UID 86948) South Africa under the SARChI Chair for Mine Water Management and the Tshwane University of Technology (TUT) for funding this project and supporting this research. Additional thanks go to the Council for Geosciences (CGS), South Africa, and the relevant authorities and mine operator Sibanye Gold for providing us with historical mine water data. Parts of this paper are based on the Ph.D. thesis of K.S.M.

■ REFERENCES

(1) Schmiermund, R. L.; Drozd, M. A. Acid mine drainage and other mining-influenced waters (MIW). In *Mining environmental handbook: Effects of Mining on the Environment and American Environmental Controls on Mining*; Marcus, J. J., Ed.; Imperial College Press: 1997; pp 599−617.

(2) McLemore, V.T. *Basics of Metal Mining Influenced Water: Management Technologies for Metal Mining Influenced Water*; SME: 2008.

(3) Qi, C.-c. Big data management in the mining industry. *Int. J. Miner. Metall. Mater.* **2020**, *27*, 131−139.

(4) More, K. S.; Wolkersdorfer, C.; Kang, N.; Elmaghraby, A. E. Automated measurement systems in mine water management and mine workings — A review of potential methods. *Water Resour. Ind.* **2020**, *24*, 100136.

(5) Wolkersdorfer, C. *Water Management at Abandoned Flooded Underground Mines – Fundamentals, Tracer Tests, Modelling, Water Treatment*; Springer: 2008.

(6) Seervi, V.; Yadav, H.; Srivastav, S.; Jamal, A. Overview of active and passive systems for treating acid mine drainage. *Int. Adv. Res. J. Sci. Eng. Technol.* **2017**, *4*, 131−137.

(7) Skousen, J.; Zipper, C. E.; Rose, A.; Ziemkiewicz, P. F.; Nairn, R.; McDonald, L. M.; Kleinmann, R. L. Review of Passive Systems for Acid Mine Drainage Treatment. *Mine Water Environ.* **2017**, *36*, 133−153.

(8) Rong, Y. Statistical Methods and Pitfalls in Environmental Data Analysis. *Environ. Forensics.* **2000**, *1*, 213−220.

(9) Kwak, S. K.; Kim, J. H. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.* **2017**, *70*, 407−411.

(10) More, K. S.; Wolkersdorfer, C. Application of machine learning algorithms for nonlinear system forecasting through analytics — A case study with mining influenced water data. *Water Resour. Ind.* **2023**, *29*, No. 100209.

(11) Fielding, S.; Fayers, P. M.; Ramsay, C. R. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health Qual. Life Outcomes.* **2009**, *7*, 57.

(12) Little, T. D.; Jorgensen, T. D.; Lang, K. M.; Moore, E. W. G. On the joys of missing data. *J. Pediatr. Psychol.* **2014**, *39*, 151−162.

(13) Newman, D. A. Missing data: Five practical guidelines. *Organ. Res. Methods.* **2014**, *17*, 372−411.

(14) Nalecki, P.; Gowan, M. Mine Water Management—Dynamic, Probabilistic Modelling Approach. *10th International Mine Water Association Congress*, Wolkersdorfer, C., Ed.; 2008; pp 533−536.

(15) Guo, W.; Xu, T., Application of a Numerical Statistical Model to Estimate Potential Oil Spill Risk, In *Mathematical Modelling and Numerical Simulation of Oil Pollution Problems*; Ehrhardt, M., Ed.; Springer: 2015; pp 127−140. DOI: 10.1007/978-3-319-16459.

(16) Mouazen, A. M.; Ramon, H. A numerical—statistical hybrid modelling scheme for evaluation of draught requirements of a subsoiler cutting a sandy loam soil, as affected by moisture content, bulk density and depth. *Soil Tillage Res.* **2002**, *63*, 155−165.

(17) Mikhaĭlov, G.; Medvedev, I.y.N. Optimization of Weighted Monte Carlo Methods with Respect to Auxiliary Variables. *Sib. Math. J.* **2004**, *45*, 331.

(18) Wackerly, D.; Mendenhall, W.; Scheaffer, R.L. *Mathematical Statistics with Applications*, 7th ed.; Thomson: 2014.

(19) Bui, X.-N.; Nguyen, H.; Choi, Y.; Nguyen-Thoi, T.; Zhou, J.; Dou, J. Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. *Sci. Rep.* **2020**, *10*, 9939.

(20) Hrnjica, B.; Bonacci, O. Lake Level Prediction using Feed Forward and Recurrent Neural Networks. *Water Resour. Manage.* **2019**, *33*, 2471−2484.

(21) Pestman, W.R. *Mathematical Statistics*; De Gruyter: 2009. DOI: 10.1515/978311020

(22) González-Estrada, E.; Cosmes, W. Shapiro−Wilk test for skew normal distributions based on data transformations. *J. Stat. Comput. Simul.* **2019**, *89*, 3258−3272.

(23) Davis, J.C. *Statistics and Data Analysis in Geology*, 3rd ed.; Wiley: New York, 2002.

(24) Skansi, S. *Introduction to Deep Learning: from logical calculus to artificial intelligence*; Springer: 2018. DOI: 10.1007/978-3-319-73004-2.

(25) Nguyen, H. D.; Tran, K. P.; Thomassey, S.; Hamad, M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *Int. J. Inf. Manage.* **2021**, *57*, No. 102282.

(26) Manaswi, N.K. *Deep learning with applications using python: Chatbots and Face, Object, and Speech Recognition with TensorFlow and Keras*; Apress: 2018. DOI: 10.1007/978-1-4842-3516.

(27) Michelucci, U. *Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks*; Springer: 2018. DOI: 10.1007/978-1-4842-3790-8.

(28) Hejazi, M.; Singh, Y. P. One-Class Support Vector Machines Approach to Anomaly Detection. *Appl. Artif. Intell.* **2013**, *27*, 351−366.

(29) Harrou, F.; Dairi, A.; Taghezouit, B.; Sun, Y. An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine. *Solar Energy.* **2019**, *179*, 48−58.

(30) Saari, J.; Strömbergsson, D.; Lundberg, J.; Thomson, A. Detection and identification of windmill bearing faults using a one-class support vector machine (SVM). *Measurement.* **2019**, *137*, 287−301.

(31) Stefaniak, P.; Wodecki, J.; Zimroz, R. Maintenance management of mining belt conveyor system based on data fusion and advanced analytics. Paper presented at the 6th International Congress on Technical Diagnostics;2018; pp 465−476. DOI: 10.1007/978-3-319-62042-8_42.

(32) Grzesiek, A.; Zimroz, R.; Śliwiński, P.; Gomolla, N.; Wyłomańska, A. Long term belt conveyor gearbox temperature data analysis – Statistical tests for anomaly detection. *Measurement.* **2020**, *165*, No. 108124.

(33) Steinwart, I.; Christmann, A. *Support vector machines*; Springer: 2008. DOI: 10.1007/978-0-387-77242-4.

(34) Wagner, H. N. R.; Köke, H.; Dähne, S.; Niemann, S.; Hühne, C.; Khakimova, R. Decision tree-based machine learning to optimize the laminate stacking of composite cylinders for maximum buckling load and minimum imperfection sensitivity. *Compos. Struct.* **2019**, *220*, 45–63.

(35) Romero, M. P.; Chang, Y.-M.; Brunton, L. A.; Parry, J.; Prosser, A.; Upton, P.; Rees, E.; Tearne, O.; Arnold, M.; Stevens, K.; et al. Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. *Prev. Vet. Med.* **2020**, *175*, No. 104860.

(36) Lesouple, J.; Baudoin, C.; Spigai, M.; Tourneret, J.-Y. Generalized isolation forest for anomaly detection. *Pattern Recognit. Lett.* **2021**, *149*, 109–119.

(37) Chen, H.; Ma, H.; Chu, X.; Xue, D. Anomaly detection and critical attributes identification for products with multiple operating conditions based on isolation forest. *Adv. Eng. Inform.* **2020**, *46*, No. 101139.

(38) Togbe, M. U.; Chabchoub, Y.; Boly, A.; Barry, M.; Chiky, R.; Bahri, M. Anomalies detection using isolation in concept-drifting data streams. *Computers.* **2021**, *10*, 13.

(39) Mensi, A.; Bicego, M. Enhanced anomaly scores for isolation forests. *Pattern Recognit.* **2021**, *120*, No. 108115.

(40) Sahoo, A.; Ghose, D. K. Imputation of missing precipitation data using KNN, SOM, RF, and FNN. *Soft Comput.* **2022**, *26*, 5919–5936.

(41) Liu, G.; Zhao, H.; Fan, F.; Liu, G.; Xu, Q.; Nazir, S. An enhanced intrusion detection model based on improved kNN in WSNs. *Sensors.* **2022**, *22*, 1407.

(42) Betrie, G. D.; Tesfamariam, S.; Morin, K. A.; Sadiq, R. Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques. *Environ. Monit. Assess.* **2013**, *185*, 4171–4182.

(43) Khampuengson, T.; Wang, W. Novel Methods for Imputing Missing Values in Water Level Monitoring Data. *Water Resour. Manage.* **2023**, *37*, 851–878.

(44) De Marchi, S.; Santin, G. A new stable basis for radial basis function interpolation. *J. Comput. Appl. Math.* **2013**, *253*, 1–13.

(45) Karimi, N.; Kazem, S.; Ahmadian, D.; Adibi, H.; Ballestra, L. V. On a generalized Gaussian radial basis function: Analysis and applications. *Eng. Anal. Bound. Elem.* **2020**, *112*, 46–57.

(46) Averbuch, A. Z.; Neittaanmäki, P.; Zheludev, V. A. *Spline and spline wavelet methods with applications to signal and image processing*; Springer: 2014. DOI: 10.1007/978-3-319-22303-2.

(47) Kineri, Y.; Wang, M.; Lin, H.; Maekawa, T. B-spline surface fitting by iterative geometric interpolation/approximation algorithms. *Comput. Aided Des.* **2012**, *44*, 697–708.

(48) Keller, W.; Borkowski, A. Thin plate spline interpolation. *J. Geod.* **2019**, *93*, 1251–1269.

(49) Sharifi, E.; Saghafian, B.; Steinacker, R. Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *J. Geophys. Res. Atmos.* **2019**, *124*, 789–805.

(50) Friedman, J. H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–67.

(51) Saadi, A. M.; Msilini, A.; Charron, C.; St-Hilaire, A.; Ouarda, T. B. Estimation of the area of potential thermal refuges using generalized additive models and multivariate adaptive regression splines: A case study from the Ste-Marguerite River. *River Res. Appl.* **2022**, *38*, 23–35.

(52) Krenker, A.; Bester, J.; Kos, A., Introduction to the artificial neural networks. In *Artificial neural networks — methodological advances and biomedical applications*; Suzuki, K., Ed.; InTech Open: 2011. DOI: 10.5772/15751.

(53) Osisanwo, F.; Akinsola, J.; Awodele, O.; Hinmikaiye, J.; Olakanmi, O.; Akinjobi, J. Supervised machine learning algorithms: classification and comparison. *Int. J. Comput. Trends Technol.* **2017**, *48*, 128–138.

(54) Ma, L.; Huang, C.; Liu, Z.-S.; Morin, K.A.; Aziz, M.; Meints, C. Artificial neural network for prediction of full-scale seepage flow rate at the equity silver mine. *Water Air Soil Pollut.* **2020**, *231*, 179.

(55) Rooki, R.; Doulati Ardejani, F.; Aryafar, A.; Bani Asadi, A. Prediction of heavy metals in acid mine drainage using artificial neural network from the Shur River of the Sarcheshmeh porphyry copper mine, Southeast Iran. *Environ. Earth Sci.* **2011**, *64*, 1303–1316.

(56) Singh, K. P.; Basant, A.; Malik, A.; Jain, G. Artificial neural network modeling of the river water quality — a case study. *Ecol. Modell.* **2009**, *220*, 888–895.

(57) Wolkersdorfer, C. *Hydrogeochemische Verhältnisse im Flutungswasser eines Uranbergwerks: die Lagerstätte Niederschlema/Alberoda [Hydrogeochemical conditions in the mine water of a flooding uranium mine – The Niederschlema/Alberoda deposit] [Doctor of Natural Sciences: Dissertation]*, Clausthal University of Technology: 1996.

(58) Lindemann, B.; Maschler, B.; Sahlab, N.; Weyrich, M. A survey on anomaly detection for technical systems using LSTM networks. *Comput. Ind.* **2021**, *131*, No. 103498.