Developing the Internet of Mine Water

by

Kagiso Samuel More

Submitted in fulfilment of the requirements for the degree

DOCTOR OF PHILOSOPHY

in the

Department of Environmental, Water and Earth Sciences

FACULTY OF SCIENCE

TSHWANE UNIVERSITY OF TECHNOLOGY

Supervisor: Prof Dr habil Ch Wolkersdorfer

Co-Supervisor: Dr M Lupankwa

July 2022

Declaration

I hereby declare that the thesis submitted for the degree Ph.D. Science (Water Science and Technology), at Tshwane University of Technology, is my own original work and has not previously been submitted to any other institution of higher education. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references.

Kagiso Samuel More

Plagiarism Declaration

1. I hereby declare that this thesis is submitted in fulfilment of the requirements for the degree Ph.D. Science (Water Science and Technology) at Tshwane University of Technology.
2. This thesis is my own original work and has not been partially or wholly copied/translated from the academic work of another person.
3. This thesis has not previously been submitted to any other institution of higher education in South Africa or any other country.
4. I declare that this thesis fully complies with the stipulations of the TUT Policy on Plagiarism. Each scholarly contribution and all the sources cited or quoted have been properly attributed and are indicated and acknowledged by means of a comprehensive list of references.
5. I understand that the University may impose disciplinary actions against me should a reasonable suspicion exist that this thesis is not my own unaided work or that I have failed to properly acknowledge the source of the information in my thesis work.

I confirm that I have been granted permission by the academic publishing companies and the Tshwane University of Technology to include presented conference papers and journal articles (published, in press, under review or under supervisory review) based on this thesis work in my thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publications.

Kagiso Samuel More

Full Names & Surname

210042091

Student/Staff Number

July 2022

Date

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof Ch Wolkersdorfer and co-supervisor, Dr M Lupankwa, for their continuous support of my doctoral study, for their patience, motivation and immense knowledge. Most importantly, I would like to thank Prof Ch Wolkersdorfer for the opportunity afforded me to be part of such a remarkable mine water research team.

Besides my supervisors, I would like to thank my friends and family for their insightful comments and encouragement, the institution, Tshwane University of Technology, for allowing me to use their laboratory facilities to conduct my research. I would also like to thank Sibanye Stillwater Gold Mine, the Council for Geosciences and the Department of Water and Sanitation, South Africa for providing us with historical mine water data for the Eastrand and Westrand mine water treatment plants. Technical expertise from Prof Adewale Owolawi and Mr Chester Muyeza in the development of *e*Metsi application is hereby acknowledged.

Financial assistance provided by the National Research Foundation (NRF Grant UID 121723, 138770 and 86948) South Africa under the SARChI Chair for Mine Water Management at the Tshwane University of Technology in respect of the costs of this study and supporting this research is hereby acknowledged.

Acknowledgement of Software Sources

By far the majority of the analyses in this thesis have been conducted in Python 3.7.1 and R 4.1.1. Python was used to develop and run machine learning models, and was also used to perform statistical analyses. Further statistical analyses were computed using R. Both the software were used in an Anaconda 4.11.0 platform. Spyder 5.1.5 and RStudio 1.1.456 IDEs were used to write the programming codes for Python and R, respectively. For the graphical user interface (GUI), Qt designer was used together with PyQt5 to develop the application. InstallForge 1.4.2 was used to create an installer for the GUI application.

Presentations and Publications

The work contained in this thesis has previously been presented at conferences and published in journal articles as indicated below.

Journal outputs

- ***More, K. S.** & Wolkersdorfer, Ch. (2022): Intelligent Mine Water Management Tools — *e*Metsi and Machine Learning GUI. *Mine Water Environ*. Under Review.
- ***More, K. S.** & Wolkersdorfer, Ch. (2022): Data Analytics behind Forecasting of Nonlinear Systems Using Machine Learning Models — an Example with Mining Influenced Water Data. *Water Resour. Ind.* Under Review
- ***More, K. S.** & Wolkersdorfer, Ch. (2022): Predicting and Forecasting Mine Water Parameters Using a Hybrid Intelligent System. *Water Resour. Manage*. doi:10.1007/s11269-022-03177-2
- ***More, K. S.**, Wolkersdorfer, Ch., Kang, N. & Elmaghraby, A. S. (2020): Automated Measurement Systems in Mine Water Management and Mine Workings – A Review of Potential Methods. *Water Resour. Ind.* Article 100136, doi:10.1016/j.wri.2020.100136

*****: peer reviewed

Conference outputs

- **More, K. S.** & Wolkersdorfer, Ch. (2022): Application of a Hybrid Neural Network Model to Forecast Mining Influenced Water Parameters. Oral Presentation and Extended Abstract Publication. In: Bezuidenhout, J.: Navigating the Course. Water Institute of Southern Africa Conference Sandton, South Africa (Sandton Convention Centre, 28–30 September 2022).

- **More, K. S.** & Wolkersdorfer, Ch. (2021): Application of Artificial Intelligence Systems in Mine Water Management – An Introduction to two Effective Predictive Models. Oral Presentation and Extended Abstract Publication. In: Stanley, P.; Wolkersdorfer, Ch.; Wolkersdorfer, K.: Mine Water Management for Future Generations. International Mine Water Association Congress. p. 365–367; Cardiff, Wales, United Kingdom (Natural Resources Wales, The Coal Authority, Welsh Government, Cardiff University, 12–16 July 2021)
- **More, K. S.** & Wolkersdorfer, Ch. (2019): Disruptive Technologies in Mine Water Management – The Future. Oral Presentation and Full Paper Publication. In: Khayrulina, E.; Wolkersdorfer, Ch.; Polyakova, S.; Bogush, A.: Mine Water – Technological and Ecological Challenges. International Mine Water Association Congress. p. 597–602; Perm, Russia (Perm State University, 15–19 July 2019)

Other outputs

- Wolkersdorfer, Ch. & **More, K. S.** (2021): Digital Technologies for Mine Water Management. Water & Sanitation Africa, 16(4): 34–35
- **More, K. S.** & Wolkersdorfer, Ch. (2021): Application of Machine Learning Models and Time Series Forecasting to Predict Electrical Conductivity and pH of Mine Water. Oral Presentation. National Science Week, 6 August 2021, Tshwane University of Technology, Pretoria

Extended Abstract

Mining influenced water (MIW), especially acid mine drainage (AMD), but also circum-neutral and alkaline drainage, is a threat to mining communities in South Africa and beyond. It is therefore imperative that MIW is properly treated before it can be discharged into receiving water courses for potential domestic and industrial use. Mine water quality changes over time and is unpredictable unless proper techniques are applied. Technology relating to mine water management in the mining industry is often outdated and therefore needs to be improved. This thesis introduces the use of artificial intelligence (AI) and the Internet of Things (IoT) techniques to optimise mine water management.

The study is divided into four sections: in the first section, the IoT techniques were explored by using radio-frequency identification (RFID) technology in conjunction with mobile and website applications in mine water sampling (*e*Metsi — *e*-Tag based Mine Water Evaluation, Testing, Sampling and Identification. *e*Metsi means electronic water in Setswana language). The mining industry and many other industries are still widely using hand labelled or barcoded sample containers for identification and tracking purposes, and this has proven to be a time-consuming practice being prone to errors. In addition, when samples are taken, there is currently no direct communication between the sampling site, laboratory information management system and the sampling institution. This further slows down exchange of data and can result in high response times to regulating the plant's parameters. Fast communication between the samplers, the laboratory and the clients is therefore important for keeping to environmental targets. This study explored the practice of incorporating Near-Field Communication (NFC) tags to sampling bottles,

usage of a NFC mobile application for recording on-site parameters during sampling and display of this data on a website application.

The second and third sections focused on AI supported prediction and forecasting of mine water parameters using hybrid intelligent models. In the second section, the hybrid model consisted of the long short-term memory (LSTM), gradient boosting regression tree and random forest regression tree models. Using Python programming language within an Anaconda platform, machine learning (ML) models were built using historical data from 2016 to 2021 to predict electrical conductivity (EC) and pH of MIW from the “Acid Mine Drainage” treatment plant in Springs, South Africa. The random forest and gradient boosting models were deployed as a web application using Flask application programming interface (API) and Heroku cloud platforms for prediction of EC and pH. The LSTM model was used to forecast data that was therefore supplied to the web application to predict the values of EC and pH for 60 days.

In the third section, the task was to compare the neural networks with regression trees and select the best performing model to forecast mine water parameters. This approach explored the multivariate LSTM, artificial neural network (ANN), deep neural network (DNN), random forest and gradient boosting models. Westrand gold mine’s treatment plant in Randfontein, South Africa was used as an example in this study, in which historical mine water data (2016–2021) from shaft № 9 is used to build the ML models. The multivariate LSTM model was used to forecast data that were supplied to the best performing model to forecast Fe and acidity for 60 days. Additionally, this section explored robust statistical analysis for time series data, where various ways to interpolate the missing data and detect anomalies were tested.

The last section focused on the machine learning graphical user interface (GUI). The developed GUI can load CSV data, perform data pre-processing and visualisation on the loaded data. Several ML algorithms (gradient boosting regression tree, random forest regression tree and linear regression) are embedded in the GUI. Instead of writing long Python code lines, the ML models can be built in just a few clicks using the GUI.

This thesis with the tested AI and IoT techniques will ensure that the mine water treatment plants find efficient ways to treat and manage MIW. In addition, the presented techniques can be applied to optimise wastewater treatment plants. When put all together, the main purpose of the explored techniques was to predict and forecast the chemistry of MIW so the treatment plant operators can better plan the chemical, electricity and working power that they need to treat and manage MIW efficiently.

Contents

List of Figures

List of Tables

Glossary of Terms, Abbreviations and Acronyms

2D QR Codes	Two Dimensional Barcodes
Adam	Adaptive Moment Estimation is an algorithm for optimisation technique for gradient descent
ADF Test	Augmented Dickey-Fuller test is a statistical test used to test whether a given time series is stationary or not
AI	Artificial Intelligence
AIC	The Akaike Information Criterion is a mathematical method for evaluating how well a model fits the data it was generated from

AMD	Acid Mine Drainage
Anaconda	A distribution of Python and R programming languages for scientific computing, that aims to simplify package management and deployment
ANFIS	Adaptive Neural-Fuzzy Inference System
ANN	Artificial Neural Networks
API	Application Programming Interface, which is a software intermediary that allows two applications to talk to each other
AR	Autoregressive is a statistical analysis model that predicts future behaviour based on past values of the series
ARIMA	Autoregressive Integrated Moving Average is a statistical analysis model that uses time series data to understand it and predict future trends
ARMA	Autoregressive Moving Average is a statistical analysis technique that combines AR and MA techniques
CSS3	Cascading Style Sheets 3
DNN	Deep Neural Networks
DOI	Digital Object Identifier
EC	Electrical Conductivity
*e*Metsi	e-Tag based Mine Water Evaluation, Testing, Sampling and

	Identification
Executable	An executable file (exe file) is a computer program that can run in Microsoft Windows
Flask	A web framework for Python
FMF	Fuzzy Membership Function
FMV	Fuzzy Membership Value
Git	A version control system that enables the management and tracking of the source code history
GitHub	A provider for internet hosting of software development and version control using Git
GUI	Graphical User Interface
Heroku	A cloud platform on which applications can be deployed, managed and scaled
HF	High Frequency spectrum
HTML5	Hyper Text Mark-up Language 5
IDE	Integrated Development Environment software
IIoT	Industrial Internet of Things
iMineWa	Intelligent Mine Water Management
IMWMP	Integrated Mine Water Management Plan

Industry 3.0	The third industrial revolution. The automation of processes using logic processors and information technology
Industry 4.0	The fourth industrial revolution. Transformation of industry through the intelligent networking of machines and processes with the help of information and communication technology
IoMW	Internet of Mine Water
IoT	Internet of Things
Java	A general-purpose computer programming language designed to produce programs that will run on any computer system
JavaScript	An object-oriented computer programming language designed to produce programs that will run on any computer system
LF	Low Frequency spectrum
LIMS	Laboratory Information Management System
LSTM	Long Short-Term Memory networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems
M2M	Machine to Machine
MA	Moving Average is a statistical analysis technique that forecast a series based solely on the past errors in the series
MAE	Mean Absolute Error

MIW	Mining Influenced Water
ML	Machine Learning
MLP	Multilayer Perceptron is a type of feedforward neural network
MSE	Mean Square Error
NFC	Near-Field Communication
NURBS	Non-Uniform Rational Basis Spline are mathematical representations of three-dimensional geometry used to describe curves and surfaces
Python	A high-level general-purpose programming language
ReLU	Rectified Linear Unit is an activation function in deep learning models. The function output the input directly if it is positive, otherwise, it will output zero
RFID	Radio-Frequency Identification
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network is a type of artificial neural network which recognises data's sequential characteristics and use patterns and trends to predict future events
SI	Swarm Intelligence
Spyder	An open-source cross-platform integrated development environment for scientific programming in the Python language

SQL	Structured Query Language
TDS	Total Dissolved Solids
TIR	Thermal Infra-Red
UAVs	Unoccupied Aerial Vehicles (“drones”)
UHF	Ultra-High Frequency spectrum
URN	Uniform Resource names
WSN	Wireless Sensor Network
XML	Extensible Mark-up Language

CHAPTER 1

1 Introduction

1.1 Brief summation of the thesis

The thesis has been organised into nine chapters, each with detailed information and thorough data analysis. A summation of each section is presented below:

Introduction

The introduction displays a general background on the use of digital technologies in the mining industry. A brief problem statement, hypothesis outline and proposed solutions are also given.

Literature Review

In the literature review, the use of digital technologies in the mining industry is evaluated in detail explaining them with clear and current examples. This chapter touches on the persisting mining influenced water (MIW) problem in the mines due to the aging technology. It also gives an overview on the use of machine learning (ML) methodologies, i.e. random forest, artificial neural networks (ANN) and gradient boosting regression tree. A thorough overview of other predictive ML models is presented in this section. This literature review chapter is based on the following published research outputs:

- More, K. S., Wolkersdorfer, Ch., Kang, N. & Elmaghraby, A. S. (2020): Automated Measurement Systems in Mine Water Management and Mine Workings – A Review of Potential Methods. *Water Resour. Ind.* Article 100136, doi: 10.1016/j.wri.2020.100136

- More, K. S. & Wolkersdorfer, Ch. (2019): Disruptive Technologies in Mine Water Management – The Future. Oral Presentation and Full Paper Publication. In: Khayrulina, E.; Wolkersdorfer, Ch.; Polyakova, S.; Bogush, A.: Mine Water – Technological and Ecological Challenges. International Mine Water Association Congress. p. 597–602; Perm, Russia (Perm State University, 15–19 July 2019)

Methodology

The methodology chapter outlines the procedures for the radio frequency identification (RFID)–based mine water sampling process and protocol, and the application of ML methodologies. It gives a detailed structure of the random forest algorithm, ANN and gradient boosting technique for mine water management. Usage of Python programming language and its libraries is also given in full detail under this chapter.

***eMetsi* – RFID Controlled Mine Water Sample Management**

*e*Metsi is an application developed for electronic tracking and identification of mine water samples and sampling results. The eMetsi application chapter explain the development and use of near-field communication (NFC) mobile application and supporting website application. These applications are used for mine water sample management. The main purpose of this chapter was to utilise the internet of things (IoT) techniques to share sampling data in real-time. *e*Metsi chapter forms part of the following research output:

- More, K. S. & Wolkersdorfer, Ch. (2022): Intelligent Mine Water Management Tools — *e*Metsi and Machine Learning GUI. *Mine Water Environ*. Under Review.

Review of Some of the Techniques for Missing Data Interpolation and Anomaly Detection in Time Series

This chapter about missing data and anomalies gives a brief review of different methods that can be applied in time series for handling missing data and properly detecting anomalies. Old and new methods have been thoroughly discussed, and some of the methods are applied in this thesis. This thesis is largely based on data analytics techniques, and this chapter introduces the beginning of the aforementioned methodology. It should be noted that data was not used in this chapter to explain the discussed techniques. However, the majority of the discussed techniques were applied in the following chapters of the thesis.

Developing Artificial Intelligence Systems – Web Application for Predictive Analysis

The AI web application chapter explains the usage of ML regressors for predictive analysis. ML regression model evaluation is thoroughly explained with different evaluation metrics. This chapter also goes into details of deploying trained and tested ML models using Flask application programming interface (API) and Heroku cloud platform. Data supporting this study are available as an online resource (www.wolkersdorfer.info/disskgmore). Most parts of this AI web application chapter are based on the following published research output:

- More, K. S. & Wolkersdorfer, Ch. (2022): Predicting and Forecasting Mine Water Parameters Using a Hybrid Intelligent System. *Water Resour. Manage*. doi:10.1007/s11269-022-03177-2

Data Analytics and Forecasting with Machine Learning Models

The data analytics and machine learning models chapter compares different models to select the best performing one and use it to forecast mine water parameters. This approach trained and tested the multivariate long short-term memory (LSTM), artificial neural network (ANN), deep neural network (DNN), random forest and gradient boosting regression tree algorithms. Exploratory data analysis techniques are also explored in this chapter with the use of robust statistics. Data supporting this study are available as an online resource (www.wolkersdorfer.info/disskgmore). This chapter is an extended version of the following research output:

- More, K. S. & Wolkersdorfer, Ch. (2022): Data Analytics behind Forecasting of Nonlinear Systems Using Machine Learning Models — An Example with Mining Influenced Water Data. *Water Resour. Ind*. Under Review.

The Internet of Mine Water — Python Machine Learning Graphical User Interface

This graphical user interface (GUI) chapter gives a detailed overview of the developed Python ML GUI for this thesis. In this GUI, data can be loaded and data pre-processing and visualisation can be performed on the loaded data. Several ML algorithms are imbedded in the GUI which can be used on the loaded data. An executable file for the GUI application is available as an online resource (www.wolkersdorfer.info/disskgmore). The GUI chapter forms part of the following research output:

- More, K. S. & Wolkersdorfer, Ch. (2022): Intelligent Mine Water Management Tools — *e*Metsi and Machine Learning GUI. *Mine Water Environ*. Under Review.

Conclusions and Recommendations

In the conclusions chapter, overall concluding remarks and recommendations according to the findings of the research are presented. The discoveries of the investigations are outlined and recommendations are put forward for further studies relating to this research study.

1.2 Background of the study

Effective treatment of mine water has often proven to be an illusion, with a number of mines contributing severely to water pollution. The current models used in treating mine water are not ideal as they treat mine water based on the compositions and volumes of water entering the plant (Figure 1.1 left). This means that the plant needs to react instantly when the volumes or chemistry of mine water changes. In most cases, there is no interaction between the precipitation, water inflow into the mine, technological changes within the mine, water analyses of the plant and the outflow of the treated water (Gao *et al.*, 2014; Golestanifar & Ahangari, 2012; More & Wolkersdorfer, 2019; Nalecki & Gowan, 2008; Usher *et al.*, 2010; Wolkersdorfer, 2008).

Technology used in the mines (e.g. data management technologies), especially South African mines, is aging and needs to be improved. Therefore, all sectors of the South African mines needs to be modernised to prevent premature mine closures, job losses and persisting water contamination by harmful contaminants from the mines. Consequently, technological improvement in the mines will lead to a safer and healthier environment, improved skills development and decrease in mine fatalities. The introduction of the Internet of Mine Water (IoMW), which was first mentioned and described by Wolkersdorfer (2013), will ensure that the technology in

the mines is improved, and all aspects of mine water management from sampling to data processing and eventually intelligent mine water management (iMineWa) are included (Figure 1.1 right).

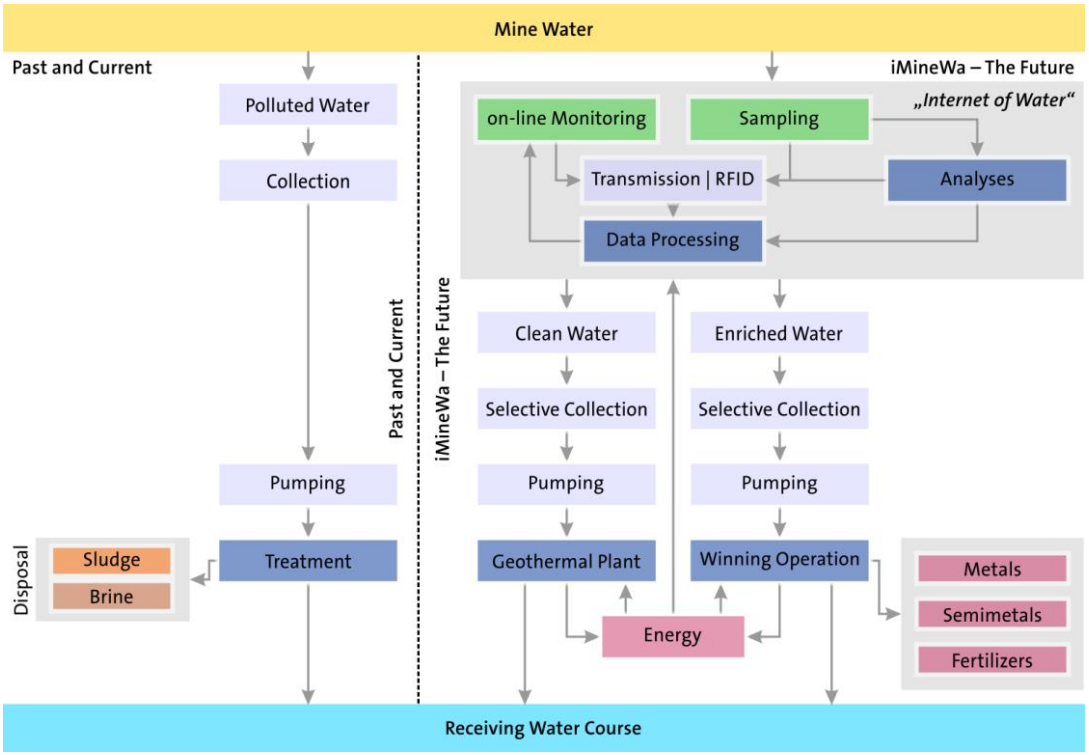


Figure 1.1: Comparison of past and current mine water treatment (left) with future mine water treatment plant using elements of iMineWa (right)

Developing such systems in the mining environment helps to thoroughly research, implement and improve existing technology in the mines (Zulkifli *et al.*, 2018). The research reported in this thesis introduces the IoMW and uses it to optimise mine water management. The IoMW is able to deal with large quantities of information from the mine water data and can also analyse it to ultimately make practical business decisions. In addition, the IoMW platform can be compared to systems in modern cars on which sensors are constantly collecting status information about the car's systems and performance. These data are transferred and stored in a device passing the data to a gateway, which integrates and processes this data. Based on a

set of algorithms, relevant information is transmitted to the manufacturer of the car. Yet, this data is only transmitted if, after a registration process, a secure communication channel between the car's platform, gateway and the manufacture is established. Should, for example, after a longer drive, the pressure in the break system deviates from the normal status, an error message will be logged. This, thereafter, triggers a "check engine" message for the driver, who then knows that the car has to be brought to a garage, where the mechanics can review the detailed error messages from a central server (Chen *et al.*, 2015; Lee & Lee, 2015; Wortmann & Flüchter, 2015). All these processes are based on the Internet of Things (IoT).

The IoT is a system of physical things embedded with sensors, electronics, software and connectivity to allow it to perform better by exchanging information with other connected devices (Lee & Lee, 2015). Therefore, the IoMW is derived from the idea of the IoT, and this technology was applied in this thesis for mine water management. The IoT, which very likely was first mentioned in the year 2000 in the United States of America (SRI Consulting Business Intelligence, 2008), is referred to all "things" that are somehow tangible and, theoretically, could be reached by an internet address. Yet, the IoMW is more than just referring to "things". It also includes data and the results of data processing coming from common statistical procedures or the procedures involved to analyse "big data". Big data is extremely large amounts of data that can only be analysed computationally to reveal patterns and trends (Millie *et al.*, 2013); it is not identical with "a lot of data".

Further comparison of the IoMW can be made with the industry 3.0 (e.g. Wolkersdorfer, 2013) or industry 4.0 (e.g. Gilchrist, 2016; Lasi *et al.*, 2014) because of the advanced technologies (e.g. artificial intelligence and IoT) integrated in its

system. Nanda (2020) describes the importance of integrating Industry 4.0 technologies in mines to improve production and increase profits. In order to survive in the competitive market, smart mining is imperative. This will see technologies such as Big data, IoT, Machine to Machine (M2M), data analytics, sensor networks, drones, and robotics increase efficiency in critical mining activities such as mine water management, drilling, surveying, processing and transportation (Nanda, 2020).

An additional advantage of Industry 4.0 applications in the mine water sector is its contribution to Integrated Mine Water Management Plans (IMWMP) as outlined by Chahbandour (2013). An IMWMP also requires that all mine departments cooperate: mine water management is incorporated into the overall mine plan, on-site and off-site risks relating to mine water are understood and internal and external commitments of the mine are supported at all times (Chahbandour, 2013). Water contamination by mining activities is a concerning issue and this was accurately emphasised by Hopwood and Deloitte Touche Tohmatsu Limited (2018): “With each passing year, water has become a more critical issue for the mining sector” and because of the many changes that the mining industry encounters, “mining companies must enhance their approach to water management”.

Organisations have begun presenting various IoT based items and services. It was about time that this kind of innovation is introduced in the mining industry, particularly in mine water management, as it is difficult to manage mine water. Only a few mines are currently utilising these new emerging technologies, and their experience showed that they are profiting substantially from these innovations. These mines include the Hull Rust (United States, Minnesota), Garzweiler (Germany, North Rhine-Westphalia), Escondida (Chile, Antofagasta Region), Bingham Canyon (United States, Utah), Berkeley Pit mines (United States, Montana) (Chadwick, 2016), Rio

Tinto Pilbara iron ore (Australia, Perth) (Nanda, 2020) or Newtrax Technologies (Canada), who also implemented artificial intelligence systems (International Mining, 2019). All of them have in common that they use state-of-the-art technologies and software, ensuring that they maximise their profit without having to cut off their employees (Athresh *et al.*, 2017; Mousavi & Sellers, 2019). Experts progressively recognize the IoT could develop into a market worth 7.1 billion US dollars by 2020 and beyond (Lee & Lee, 2015). Therefore, the IoMW has a great potential, in the wake of being created, of booming to such a market worth as IoT. Though the term IoT is currently extensively utilized; however, there is presently no basic definition of what the IoT really incorporates (Wortmann & Flüchter, 2015). With IoMW, it was diverse such that it was comprehended what it really includes, i.e. data and the results of data processing.

In the past years, the IoT started to change the way people are living, how they drive, how they purchase items and how the energy is distributed into their houses or factories. In a growing number of items, sensors or chips are implemented to help them communicate with each other and transferring data to allow a better understanding of these items performance or behaviour (Atmaja *et al.*, 2021; Song *et al.*, 2020). Key for these devices or applications to work together or exchange data is a common IoT platform and a common language. This allows these devices to share large quantities of data for, e.g., improving industrial processes or to monitor peoples' health conditions (e.g. Lee & Lee, 2015; SRI Consulting Business Intelligence, 2008; Wortmann & Flüchter, 2015). In this thesis, IoT technology was used to develop an application which share mine water data in real-time.

1.3 Problem statement and definitions

Many mining companies worldwide are reluctant or waiting to fully adopt Industry 3.0 or 4.0 technologies. However, efforts are being made at several mining companies that already use automation options and computational intelligence methods (Corke *et al.*, 1998; Duff *et al.*, 2003). This is a first step to qualify mines into the 4th industrial revolution, yet more efforts might be needed (Lööw *et al.*, 2019). The slow advancement of this technology sees the mines slowing down in production compared to those that implement state-of-the-art technologies, and others affecting nearby community water sources, which can be considered a severe concern in some areas.

The main limitation blocking most of the mines to smoothly transition to Industry 4.0 is the type of network they use. Highly recommended for this transition is the 5G network (Figure 1.2), because of its fast speed (Rodriguez, 2016; Xiang *et al.*, 2017). The 5G network will help fuel the rise of IoT technology and provide the infrastructure needed to carry out big data processing (Al-Dulaimi *et al.*, 2018; Millie *et al.*, 2013). This network will add vital contributions to the industry, such as transferring or moving big data with greater speed (Millie *et al.*, 2013), increase the responsiveness of connected devices such as wireless sensor networks and other smart devices (Agrawal *et al.*, 2019; Al-Dulaimi *et al.*, 2018; Asif, 2019; Farrugia, 2011).

Current mine water management usually consists of “blindly” end-of-the-pipe treatment. This treatment does not take into consideration mine water chemistry or volume changes prior to arriving at the plant (Wolkersdorfer, 2008). Neither does it take into consideration the effects of the weather or operational procedures. Yet, all these parameters can be monitored and processed regularly and potential changes

in volume or the water chemistry can be identified hours or weeks in advance, allowing the mine water management (e.g. pumping rates) and treatment to adjust to the predictable, future changes. In addition, this allows optimising the work flow at the mine water management site and ordering consumables or scheduling electricity needs before they are necessary.

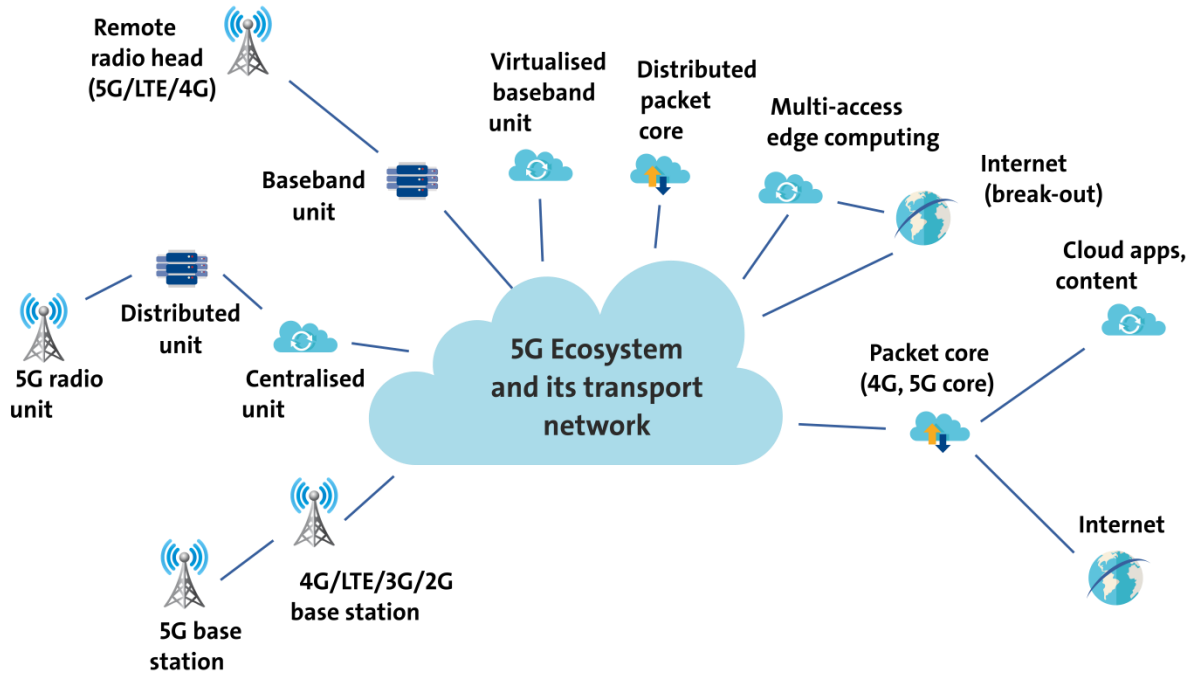


Figure 1.2: The 5G ecosystem

In this thesis, smart technologies were used to improve the current processes and describe and test a path forward to intelligent mine water management. Combination of IoT and artificial intelligence techniques were applied to build predictive and forecasting models for mine water quality. The computational intelligence techniques applied in this thesis used physico-chemical parameters (i.e. Mn, Fe, SO_4, temperature, electrical conductivity, turbidity, pH, TDS, TSS), RFID controlled sample management, results of chemical analyses of the mine water and data processing (e.g. statistics of various scales). Because of the large volumes of data

used, big data analyses were applied in the process and included ML algorithms such as random forest (Breiman, 2001), gradient boosting tree (e.g. Johnson *et al.*, 2017) and neural networks (e.g. Diamantopoulou, 2005; Singh *et al.*, 2009). Therefore, this eventually resulted in “smarter” ways of analysing big mine water datasets.

1.4 Hypothesis

The Internet of Mine Water provides an integrated control of all relevant water related processes within the mine and the plant to optimise mine water management and treatment.

1.5 Aims and Objectives of the study

The main aims of this research were to improve current mine water management technologies by:

- Improving existing technologies,
- Implementing existing technologies,
- Researching new technologies (e.g. artificial intelligence), and
- Setting up a case study for the RFID controlled mine water sampling processes.

The main goals for the thesis were as follows:

- Improve the technology in the mines by implementing the IoMW. This ensured that new technologies are researched, improved and implemented,
- Bring mine water management to the next stage by helping with accurate data collection and monitoring, monitor any small change in mine water quality and re-use process water,

- Build models that can be used to forecast the future chemistry of mine water, and
- Enable the mining industry to treat and manage mine water at a lower cost and eventually contribute positively in reducing receiving water sources pollution.

Having performed the above-mentioned activities, the main objective was to develop the “Internet of Mine Water” which can provide integrated control of all relevant water related processes within the mine, and achieve the technologically advanced treatment plant. The idea behind this was that large datasets are collected and be made available to build computational intelligence techniques to optimise mine water management.

1.6 Proposed solutions

This thesis, by means of ML or AI automation, was aimed to develop a model of a mine that uses the IoMW. It also developed a case study for the Radio Frequency Identification (RFID)-based mine water sampling. The term RFID refers to an identification system involving small radio frequency devices for easy identification and tracking purposes (Cheng *et al.*, 2008; Zhuiykov, 2012). The use of ML helped to visualise and dynamically simulate the large amounts of data collected on mine site.

The IoMW is advantageous because of real-time mine water data monitoring/sharing and accurate big data analysis. Additionally, it helps in detecting any changes in the mine water quality trend before it occurs, i.e. early detection of changes in the mine water chemistry and volumes. More future advantages include inflow and outflow monitoring that will allow for an accurate water balance and therefore take

precautions when there are risks of overflow in the plant, and reduced uncertainty in determining the potential risks posed by the mine on the environment, e.g. flooding will be detected before it occurs. Costs of treating mine water will be reduced as there might be no necessity for regular chemical monitoring of the full set of analyses, and all necessary steps can be done using the system discussed in this context. The technology will also help in improvement of mine water models, i.e. chemical and water balance models.

IoMW features are unique, and make it possible to achieve the true potential of state of the art technologies in the mines. These are the four unique features: 1) connect various things to the IoMW platform, 2) analyse the data collected and use it to build business intelligence, 3) integrate various models to improve user experience and 4) optimise mine water management and treatment.

The IoT and industry 4.0 describe a similar process, which is connecting things to the internet. However, the only difference is that IoT is related to a consumer while industry 4.0 is related to production (Gilchrist, 2016; Lasi *et al.*, 2014). Industry 4.0 was implemented for the rapid growth of economic and social digitisation. The main idea of the industrial internet is to combine the production methods, state of the art information and communication technology.

Based on the literature and information review for this thesis, the following summary can be given: all the reference lists of the included papers, information gathering from consultants, electronic database and grey literature searches, reveal that mines are widely using the traditional or “old style” of mine water treatment and management (e.g. Gao *et al.*, 2014; More & Wolkersdorfer, 2019; Nalecki & Gowan, 2008; Wolkersdorfer, 2008). By introducing the new technologies, e.g. IoT (Duff *et*

al., 2003; Lee & Lee, 2015; Sun *et al.*, 2012), WSN (Agrawal *et al.*, 2019; Farrugia, 2011; Molina *et al.*, 2011; Song *et al.*, 2010) or the 5G network (Al-Dulaimi *et al.*, 2018; Asif, 2019) at the mine sites would ensure that all parameters that need to be known for an optimised mine water process will be collected and used to predict and forecast mine water chemistry and treat mine water effectively, as this was the case in this thesis. This, for example, is also being practiced in Ghanaian mines whereby they integrate IoT technologies to maximise production (pers. comm. R. Bonner, Miwatek). In the next chapter, the mentioned technologies will be thoroughly explained and their linkage to this thesis will be shown.

CHAPTER 2

2 Literature Review

2.1 Explanation of Technologies

2.1.1 The Internet of Things (IoT)

Internet of Things (IoT) aims at growing the intensity of the internet past cellular phones and computers to an entire scope of different things, environments and processes. The “things” in the IoT are used to gather and send data or do both (Coetzee & Eksteen, 2011; Díaz *et al.*, 2016; Miorandi *et al.*, 2012). IoT involves the internet, sensors and machines, giving organisations and people better knowledge and control of almost all the environments including objects that are still not accessible through the internet. In this manner, individuals and businesses are enabled to be progressively associated with their general surroundings and theoretically to do more relevant work (Boyes *et al.*, 2018; Muthukumar *et al.*, 2019; Sarma & Girão, 2009; Shrouf *et al.*, 2014).

An internet connection is an advantageous thing as it gives everyone all sorts of new benefits that were unthinkable before now. For instance, a cellular phone in the early 80s could only text, yet now it is possible to watch a movie, listen to music or even read a book from a smart phone (Aloi *et al.*, 2017; Miorandi *et al.*, 2012; Muthukumar *et al.*, 2019). These new “phones” can even do many other incredible things – the point is that connecting devices or things to the internet yields numerous advantages. Therefore, the IoT essentially implies that taking all the things that can be imagined and connecting them to the internet (Chen *et al.*, 2015; Sarma & Girão, 2009; SRI Consulting Business Intelligence, 2008; Wortmann & Flüchter, 2015). When a device is connected to the internet it can only do one of two things or both,

and that is sending and/or receiving data. A straightforward case of this idea is with smartphones; people can listen to any song on the planet with the cellular phones – these songs are stored somewhere on the internet, and the phones can connect to the internet and ask for a particular song (Aloi *et al.*, 2017; Jeong *et al.*, 2018). In the IoT, all devices that are being connected to the internet can be grouped into three classes: devices that gather data and after that send it, devices that receive data and then act on it, or devices that do both (Civerchia *et al.*, 2017; Sadeghi *et al.*, 2015; ur Rehman *et al.*, 2019; Wortmann & Flüchter, 2015).

Devices that gather data and transfer it are commonly known as sensors. There are different types of sensors such as temperature, motion, air quality, moisture to name a few of them (Farrugia, 2011; Zhuiykov, 2012). These sensors, together with a connection, make it possible to collect data automatically on the ground which at that point allows making reliable business decisions. In mine water treatment plants, automatically receiving data about the mine water quality can precisely indicate how to treat mine water and what chemicals are needed to be injected to raise the pH and precipitate metals (Gilchrist, 2016; Losavio *et al.*, 2019; More & Wolkersdorfer, 2019; Sun *et al.*, 2012).

Devices that receive data and act on it are known as machines. There are endless examples of this concept such as a vehicle (machine) receiving data from the vehicle keys and the doors open, or a network printer receiving a document and printing it (Miorandi *et al.*, 2012; Shrouf *et al.*, 2014; Singh *et al.*, 2017b). However, the real power of the IoT is when devices can do both, i.e. they can collect and send data and also receive and act on it. For example, in farming, the sensors can collect soil moisture data, and instead of the farmer watering the crops, the irrigation system can turn on automatically with regards to the amount of moisture content in the soil

(Sarma & Girão, 2009; Tzounis *et al.*, 2017) until the system gets information to stop watering. Irrigation systems can also receive weather data from sensors; therefore, during a rainy period the irrigation system can automatically switch off as the rain will be watering the crops (Civerchia *et al.*, 2017; Díaz *et al.*, 2016; Gama *et al.*, 2012; Miorandi *et al.*, 2012; Sarma & Girão, 2009; Tzounis *et al.*, 2017).

2.1.2 Fourth Industrial Revolution (Industry 4.0)

Current industries are experiencing a vast change of technology; from the first industrial revolution which involved powering of machines through steam (Deane, 1979) to the second in which electricity played a major role (Yin *et al.*, 2018). The power of internet usage was then introduced in the third industrial revolution and also included advancement in computer usage and introduction of automation (Boyes *et al.*, 2018; Stock & Seliger, 2016; Yin *et al.*, 2018). Industry 4.0 takes over from the industry 3.0, adopting computer usage and automation and support it with smart and autonomous systems driven by big data communicating with each other (e.g. Iqbal *et al.*, 2017; Lee *et al.*, 2013; Millie *et al.*, 2013; Tekiner & Keane, 2013) and AI or ML (e.g. Bonabeau *et al.*, 1999; Corke *et al.*, 1998; Duff *et al.*, 2003). In simpler terms, industry 4.0 modifies the computer usage of industry 3.0 (Gilchrist, 2016; Lasi *et al.*, 2014; Sishi & Telukdarie, 2017; Stock & Seliger, 2016).

The industry 4.0 refers to the industrial internet of things (IIoT) (Gilchrist, 2016; Sadeghi *et al.*, 2015, Figure 2.1), while the IoT, as explained earlier, refers to a system of physical things and technologies which consist of software, electronics, sensors and connectivity to ensure that the performance is enhanced by sending data to other connected devices and vice versa (Lee & Lee, 2015; SRI Consulting Business Intelligence, 2008; Wortmann & Flüchter, 2015). However, the IoT is not well received in the industry world as compared to the consumer world. The IoT in

the industry comes with a number of disadvantages such as nature and levels of security and potential disruption of production due to failure of maintenance of the technologies (Gilchrist, 2016; Sadeghi *et al.*, 2015). Overcoming the barriers then gives birth to the IIoT. This system can be applied in a lot of sectors such as mining, water, energy and even the building environment. It comes with several advantages such as saving costs, increased production and flexibility in the workplace (Lee & Lee, 2015).

Introduction of computers with their technology during the third industrial revolution was a disruptive technology. With the introduction of industry 4.0, these computers became connected and communicate with one another to make human decisions (Boyes *et al.*, 2018; Gilchrist, 2016; Lasi *et al.*, 2014; Shrouf *et al.*, 2014; Sishi & Telukdarie, 2017; Stock & Seliger, 2016; Yin *et al.*, 2018). There is a rise in the number of smart factories (e.g. Shrouf *et al.*, 2014), and industry 4.0 is also spreading; this is being made possible through a combination of the IoT, cyber-physical systems and the internet of systems (Gilchrist, 2016; Iqbal *et al.*, 2017; Wortmann & Flüchter, 2015). The substantial increase in AI knowledge makes machine usage in industries more understandable, and brings positivity and production growth in factories. Digital connection of these machines in the industries makes it possible for data to be shared easily which is the heart of industry 4.0 (Civerchia *et al.*, 2017; Lasi *et al.*, 2014; Zuo, 2017).

A lot of factories are still hung up on industry 2.0 or 3.0 and are not aware that this could severely damage their business (Yin *et al.*, 2018). Industry 4.0 comes with several advantages such as data collection in previously impractical areas – machines are able to collect data that cannot be done by human beings, and can also collect this data in areas that are inaccessible by human beings. This data can

be used to build a platform which can help with performance, maintenance, and analysing the data to identify trends and patterns, and many other issues (Gilchrist, 2016; Millie *et al.*, 2013; Shrouf *et al.*, 2014; Sishi & Telukdarie, 2017; Tekiner & Keane, 2013). A lot of data from sensors saw one of the biggest African gold mines solve oxygen concentration problems during leaching. Other advantages of industry 4.0 include robotic systems which are mainly used for moving goods and eventually cut costs, autonomous equipment and vehicles which are now used in mines through self-driving trucks and also 3D printing (Boyes *et al.*, 2018; Feldner & Herber, 2018; Gilchrist, 2016; Lasi *et al.*, 2014; Shrouf *et al.*, 2014; Sishi & Telukdarie, 2017; Stock & Seliger, 2016; Yin *et al.*, 2018).

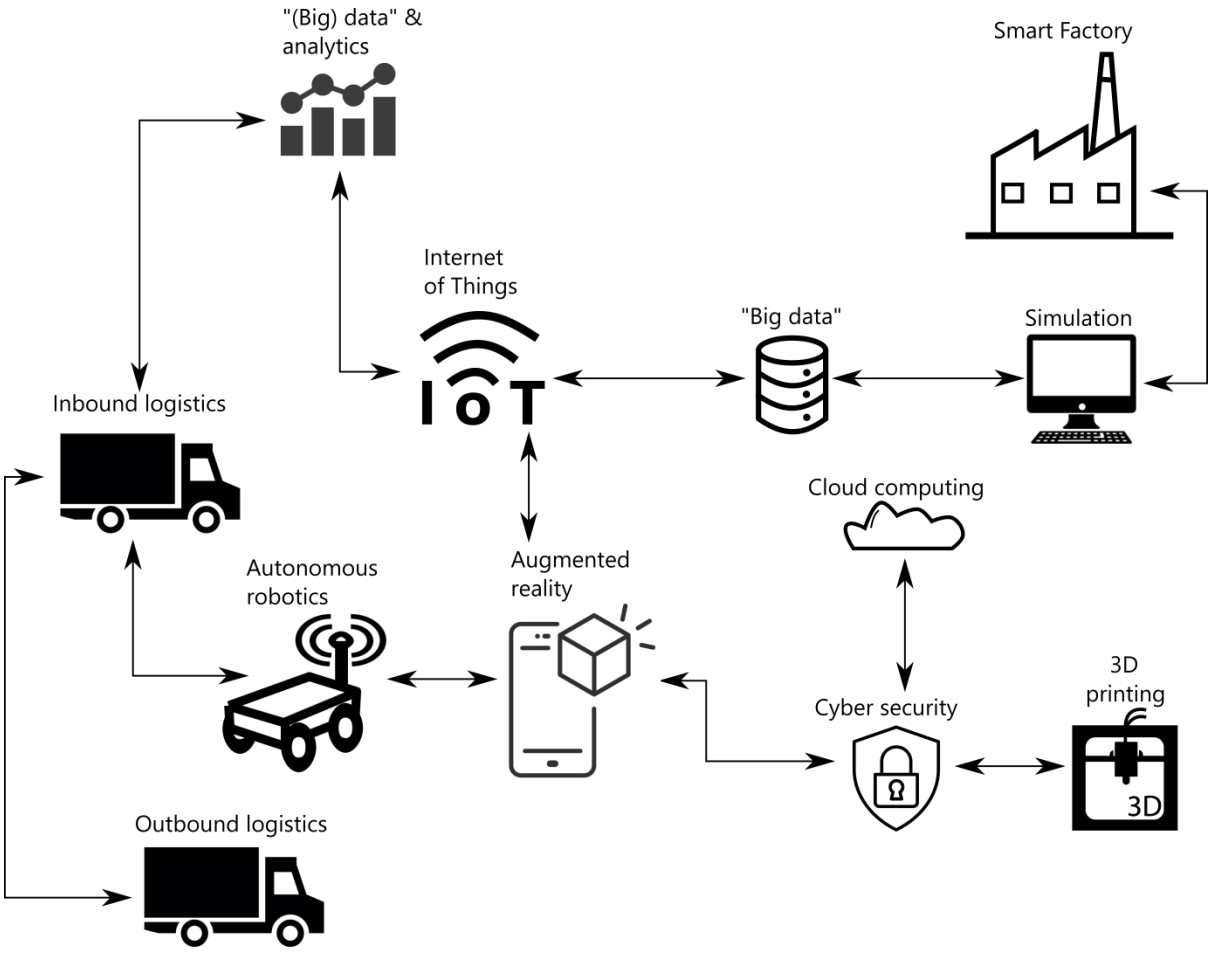


Figure 2.1: The fourth industrial revolution framework

2.1.3 Wireless Sensor Network (WSN)

Sensor network refers to a spatially distributed group of sensors that monitor and record physical or environmental conditions, such as vibrations, pollutants, temperature or sound (Figure 2.2). The idea of sensor networks was first introduced in the military and it is now widely used, including in healthcare applications, industrial process monitoring and applications, environmental monitoring and traffic control (Cheng *et al.*, 2008; Farrugia, 2011). They basically gather data for monitoring a wide range of parameters. This can be considered an automated data collection and control system, i.e. the sensor network consists of a large number of sensor nodes, and these nodes are able to gather, process information and communicate with other connected nodes in the sensor network (Li *et al.*, 2010).

The ability of sensor nodes to gather, process and communicate information with other connected sensor nodes results from its three functional units (Farrugia, 2011): the sensing unit helps in gathering information such as vibrations, motion, temperature and sounds, information which is processed through the computational unit, and lastly, the information is communicated to the neighbouring sensor node via a communication unit (Cheng *et al.*, 2008; Farrugia, 2011; Haenselmann & Müller, 2011; Li *et al.*, 2010). These nodes constantly gather data which eventually result in big data, i.e. extremely large amounts of data that can only be analysed computationally to reveal patterns and trends (Millie *et al.*, 2013).

Tuna *et al.* (2013) describes a theoretical approach where continuous water monitoring was modelled with MATLAB using autonomous buoys and boats on the Kirklareli Baraji Dam (Turkey), collecting data, which was transmitted via a wireless network. Their proposed autonomous mini boat measured temperature, electrical conductivity and nitrate concentrations. A real system was designed, installed and

tested by Sun *et al.* (2016) in a lake near Lamar University, Beaumont (Texas, USA). They monitored water temperature, dissolved oxygen and pH values in real-time using STORM 3 data loggers, and they used a wireless network to transmit the data from the loggers to the server. Based on their studies, they concluded that appropriately chosen locations for the network are imperative for a well-functioning system. Similar issues with choosing a good location were also eminent when a mine water monitoring station (temperature, electrical conductivity, pressure) at the Nikolaus-Bader-Schacht in Tyrol (Austria) was installed (pers. comm. Ch. Wolkersdorfer). Though the provider's maps showed a good cell phone connection, trees and a narrow valley made finding a good location for the antenna tedious.

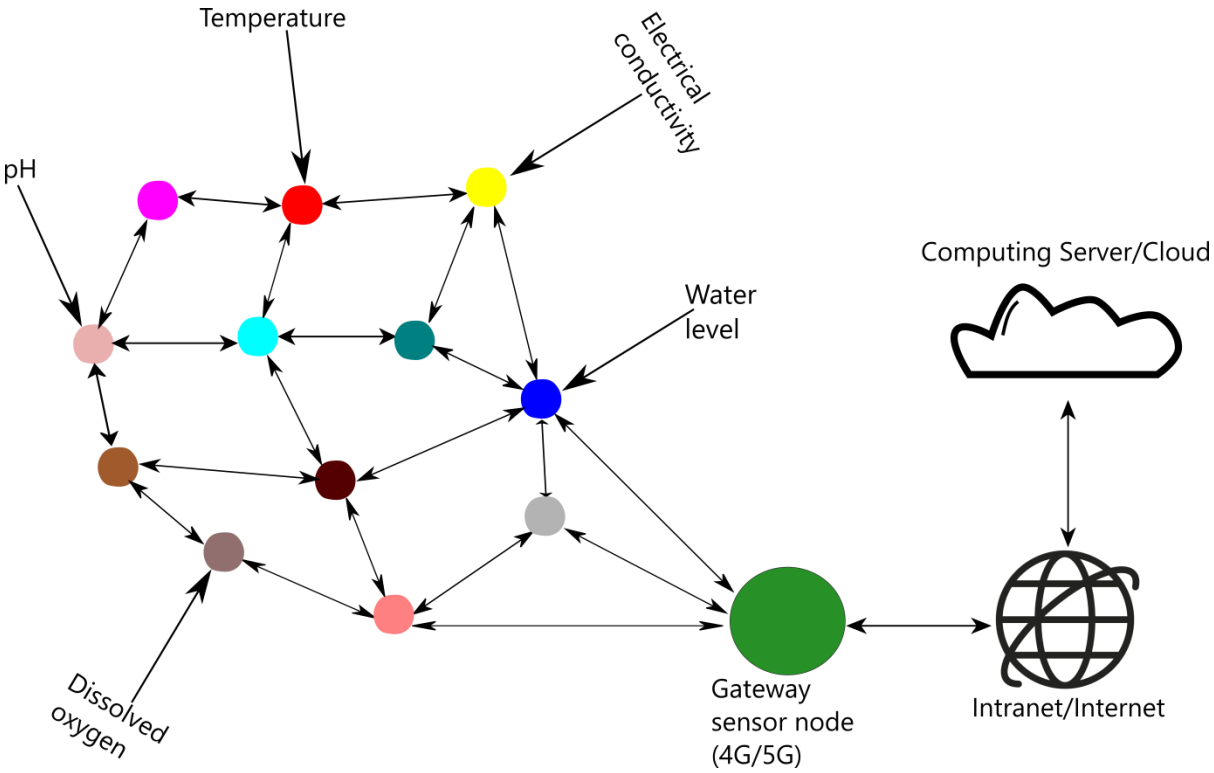


Figure 2.2: Example of a sensor network interconnected with the gateway sensor node and the server through the internet network. Sensor nodes are represented by different coloured circles

In the future, wireless sensors will be installed all over the mine to gather flow and physico-chemical parameters such as alkalinity, TDS, electrical conductivity, temperature, pH, acidity and TSS. This will be advantageous because of accurate and precise measurements the sensors will be gathering.

The core duties of a sensor network are efficient and effective data collection and management. There need to always be communication between the sensors and the server for data transfer. It is always the case that the data is stored in the sensor network and the server retrieves it randomly accessing a few sensor nodes. Sensor nodes have no knowledge of what data other nodes are recording; therefore, when retrieving data by the server this may result in data replication (Farrugia, 2011), and thus RFID tagging systems are important and need to be introduced, similar to the RFID project in this thesis. RFID tags can help in identifying and correlating data to its specific sensor by tagging the RFID chip and reveal the information.

2.1.4 Radio Frequency Identification (RFID)

Radio Frequency Identification (RFID) is a technology that identifies and tracks items by using radiated and reflected radio frequency power. Usually, an RFID system comprises of a reader and a tag (Glover & Bhatt, 2006; Sweeney, 2010; Wang *et al.*, 2010; Zhu *et al.*, 2012). An RFID reader is made up of at least one antenna, a radio frequency transmitter and a radio frequency receiver (Bhattacharyya *et al.*, 2010), while an RFID tag is an electronic tag consisting of an antenna and a microchip (Sweeney, 2010; Wang *et al.*, 2010; Zhu *et al.*, 2012). Both, the tag and reader communicate with each other via backscattering which means the reflection of radio frequency waves back to their original place (Chaix *et al.*, 2003; Sweeney, 2010; Wong *et al.*, 1998), and they utilise the ultra-high frequencies (UHF) from 860 MHz to 960 MHz (Dobkin, 2012).

RFID tags consists of a small microchip which, depending on the type, usually store not more than two KB of data (Kumar *et al.*, 2021). Information from these tags can be read from a wide range of distance, e.g. toll roads use an electronic toll collection process and the reading distance there is more than three meters. Toll roads or clothing stores use RFID technology incorporating cars or cloths embedded with a tag while the reading system is near or above the road or the store's exit location and data is exchanged through UHF radio waves (Dobkin, 2012; Glover & Bhatt, 2006; Gonçalves *et al.*, 2014; Sweeney, 2010; Zhu *et al.*, 2012). Mining work clothing can also be incorporated with RFID tags to track employee locations and ensure that they are safe (Nanda, 2020).

RFID systems have been successfully applied in a wide range of areas such as healthcare, logistics, transportation, agriculture, manufacturing, and many other services (e.g. Kumari *et al.*, 2015; Ngai *et al.*, 2007; Singh *et al.*, 2017b; Sweeney, 2010; Wang *et al.*, 2010; Zhu *et al.*, 2012), mainly for tracking and identification purposes. It can also be used to store important information for products or in biometric, electronic passports (Malčík & Drahanský, 2012). In agriculture, RFID tags have already been used for tagging vine rootstocks to identify various hybrids for scientific investigations (Luvisi *et al.*, 2014) and an RFID tag was developed for the tracking of a tracer injection probe (pers. comm. Ch. Wolkersdorfer). A future technology, which was developed in this thesis, relates to an application in mine water sampling. Currently, the sample bottles are labelled at the sampling location and recorded in the field book. However, the sample bottles could get wet during the process and it would be difficult to write on them. In addition, the process is time-consuming, and during transport the marking or a sticker may smear or scratch. With water sample bottles incorporated with RFID technology, the analysis of mine water

is faster and smoother. This thesis demonstrates a RFID-related project in which a microchip is incorporated in a plastic bottle using an RFID application to store relevant information such as sampling location, time and date.

2.1.5 Big data

Receiving extremely large amounts of data, with steadily increasing volume and very high speed is termed big data. Big data is more complex datasets, extremely large amounts of data, particularly from new data sources (Figure 2.3; Chen *et al.*, 2014; Jifa & Lingling, 2014). Using state of the art technology and big data helps to deal with business problems that were difficult to solve in the past. Big data often consists of unstructured data that needs to be processed, such as data from equipment that is embedded with sensors (e.g. Zhuiykov, 2012), social media feeds such as Twitter or Facebook data feeds (e.g. Yang *et al.*, 2015), or a mobile application (e.g. Aloi *et al.*, 2017). This is not just large data, but it can be hundreds of Exabyte of data (Millie *et al.*, 2013; Zikopoulos & Eaton, 2011). Furthermore, the increasing understanding of big data enables to process more unstructured data which can be in the form of video, text or audio. Additionally, the speed at which big data is received and processed is crucial. In this context, speed refers to the rate at which data is received and acted on (Gandomi & Haider, 2015; Lee *et al.*, 2013). With the rise of IoT, products operate in real time, and in most cases this requires real time action and evaluation (Iqbal *et al.*, 2017; Lee *et al.*, 2013; Singh *et al.*, 2017b; Zhuiykov, 2012; Zikopoulos & Eaton, 2011).

Data has essential value; however, it is useless until that value is found. It is also important to know how real data is and if processes would depend on it. Talk about big data means talking about money which is largely visible in the presently biggest technology companies (Gandomi & Haider, 2015; Jifa & Lingling, 2014; Lee *et al.*,

2013; ur Rehman *et al.*, 2019). These companies generate large amounts of money from their data as they continually analyse this data to increase productivity and create new products. Technological advancement has made the storage of data easier and cheaper. Therefore, this enables a lot of companies to expand their business by having big data at their disposal and effortlessly enables them to make more accurate business decisions (Jifa & Lingling, 2014; ur Rehman *et al.*, 2019; Zikopoulos & Eaton, 2011).

Twitter and Facebook use big data to possibly give them relevant information about its users which they use to improve their marketing techniques and eventually attract more users (Yang *et al.*, 2015; Zikopoulos & Eaton, 2011). Companies using big data are more advantageous and have the potential to outgrow the ones that do not use it because they have the ability to make quick and more informed business decisions. Besides, using big data enables organisations to focus on providing a positive customer experience.

Computer infrastructure that is currently used imposes serious difficulties in handling big data resulting in the rise of quantum computing (section 2.1.6) over the past years (Ikeda, 2018; Tabakin, 2017). Organisations find it difficult to reach the required speed when processing big data decreasing their productivity. In most cases, power required to process big data can easily result in server breakdown (Chen *et al.*, 2014; Gandomi & Haider, 2015; Tekiner & Keane, 2013; ur Rehman *et al.*, 2019; Zikopoulos & Eaton, 2011). Therefore, companies invest in quantum computing and cloud storage for big data tasks (Chen *et al.*, 2014; Gandomi & Haider, 2015; Ikeda, 2018; Iqbal *et al.*, 2017; Lee *et al.*, 2013; Millie *et al.*, 2013; Ruparelia, 2016; Tabakin, 2017).

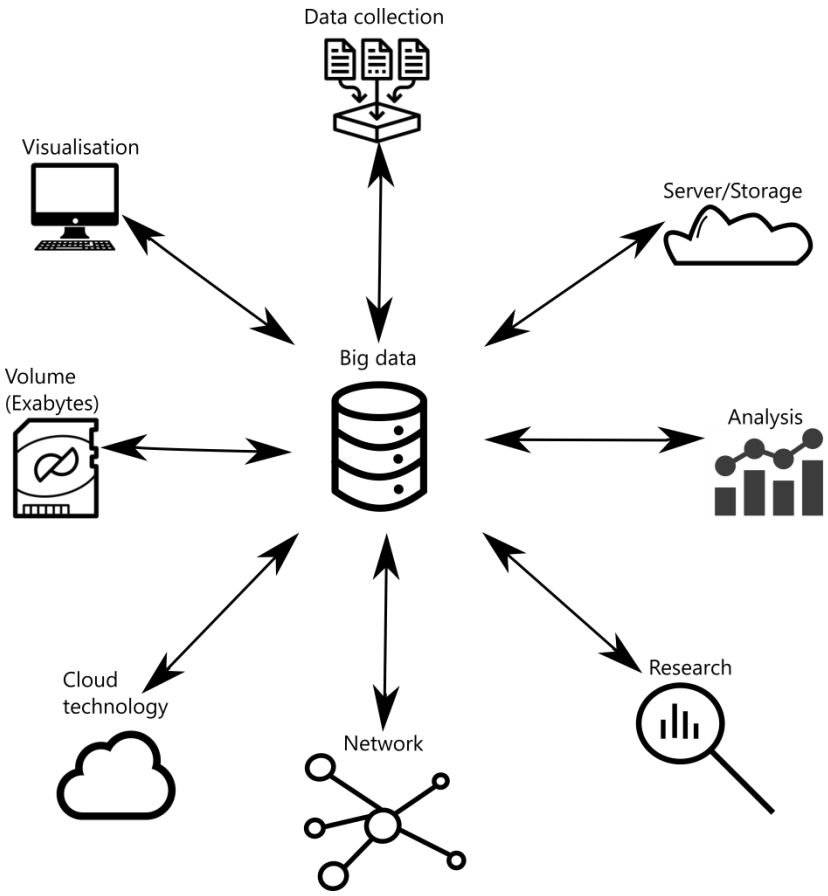


Figure 2.3: Components of big data

2.1.6 Quantum computing

As the technology advances, computers are drastically changing pushing for suitable hardware usage. In spite of the fact that more and more computers are becoming small and more powerful, most normal or classical computers are limited (Bandyopadhyay, 2005; McCaskey *et al.*, 2018). Quantum computers are seen as a solution to problems that classical computing cannot solve. In general, classical computers consist of various components such as control unit, arithmetic unit, main memory and many other parts (McCaskey *et al.*, 2018). They perform a lot of functions through these components, such as controlling, representation and processing of data. A computer contains small computers in it, referred to as logic gates (DiVincenzo & Loss, 1999; Tabakin, 2017). Their main responsibility is to produce an output by reading an input. Within a computer, other computer modules

consisting of groups of electronic circuits, such as transistors, act as an information transporter and either permit or deny data to go through it (Biham *et al.*, 2004; Häffner *et al.*, 2008).

Computer parts are getting smaller enforcing them to operate sometimes more slowly and thus hindering technological advancement. This normally occurs because transistors are getting too small not blocking all the data; therefore, electrons find a passage referred to as quantum tunnelling (Freedman *et al.*, 2002; McCaskey *et al.*, 2018; Tabakin, 2017). Computers are based on a binary base 2 numerical framework using 0 and 1 as bits. A bit is one unit of information in a computer (Biham *et al.*, 2004; DiVincenzo & Loss, 1999; Häffner *et al.*, 2008; McCaskey *et al.*, 2018; Tabakin, 2017). More complex information and situations are represented through a combination of these bits.

A quantum bit, also referred to as qubit (e.g. DiVincenzo & Loss, 1999; Häffner *et al.*, 2008; Tabakin, 2017), is used in quantum computing which are quantum systems with two different states existing in any superposition of bit values, 0 and 1 (Freedman *et al.*, 2002; Ikeda, 2018; McCaskey *et al.*, 2018; Tabakin, 2017). Due to this ability, they can store substantially more data than just 0 and 1. The superposition in quantum computing enables computers to analyse and interpret more complex data than a classical computer (e.g. De Wolf, 2017). Therefore, quantum computing implies the storing, analysing and processing of big data using photons, electrons or atoms independently which increases the probability of faster computers (Bandyopadhyay, 2005; Biham *et al.*, 2004; DiVincenzo & Loss, 1999; Freedman *et al.*, 2002; Häffner *et al.*, 2008; Ikeda, 2018; McCaskey *et al.*, 2018; Tabakin, 2017).

Big companies such as Google, IBM and Microsoft are currently running quantum computing tests expected to quickly sweep the world in the mid to late 2020s. By then, classical computers will be outperformed by quantum computers in many tasks such as financial, molecular and material modelling, cryptography, or other big data tasks (De Wolf, 2017). One of the reasons is that quantum computers use algorithms that enable them to perform tasks in seconds that would take classical computers years to perform (DiVincenzo & Loss, 1999; Häffner *et al.*, 2008; McCaskey *et al.*, 2018; Tabakin, 2017).

2.1.7 Swarm intelligence (SI)

Swarm intelligence (SI) is a branch of AI which provides possibilities of exploring collective or distributed ideas or problem solving without having to resort to a centralised control unit. It studies and focuses largely on the interactions of individuals and their environment, which can be natural or artificial systems (Bonabeau *et al.*, 1999; Karaboga & Akay, 2009; Walters, 2011). Examples of natural SI systems include swarms of bees, herds of land animals, flock of birds, school of fish, to name a few; and artificial SI systems can be seen in industries through multi-robotic systems and certain computer programs such as plagiarism checkers and modelling software (Bonabeau *et al.*, 1999; Huang & Liu, 2009; Karaboga & Akay, 2009; Krause *et al.*, 2010; Walters, 2011).

A number of mines already use swarming drones to gather large quantities of data. Swarming drones are a fleet of drones that operate together and can make decisions based on information gathered by one, many or all the drones contributing to the swarm (Tahir *et al.*, 2019). They originate from military research seeking to reduce the loss of soldiers and equipment in warfare (Lachow, 2017), using a technique allowing them to evaluate collective problem solving without having any centralised

control (Bonabeau *et al.*, 1999). Aerial photography, surveillance, site mapping and infrastructure inspection are some of the areas where swarming drones are already incorporated (Danilov *et al.*, 2015; Micklethwaite, 2018; Otto *et al.*, 2018).

This data collection method is highly preferred as compared to collecting data by using helicopters because these swarming drones are cheaper to use, faster, can collect data in large quantities and are highly reliable (Micklethwaite, 2018). Some areas on the mine site might not be accessible with self-driving trucks, and therefore, mining companies use drones in such areas to gather data. Another key point with drones is that they have cameras and they can take images and videos while collecting data. This camera feature enables specialised software to create 3D models of the mine site, their infrastructure and open pits (Danilov *et al.*, 2015; Micklethwaite, 2018; Otto *et al.*, 2018). Therefore, drones and swarming drones substantially contribute to time saving and costs cutting as opposed to creating 3D models using ground-based lasers or surveyors.

Some of the drones are even operating with embedded sensors which are utilised for mine water management, e.g. at Century Mine in northern Queensland, Australia, where sensor embedded drones were used to identify pyrite oxidation in subsurface rocks (Micklethwaite, 2018). At the Hannukainen mining development site, Northern Finland, Rautio *et al.* (2017) used UAVs and TIR to investigate groundwater-surface water interactions that might be relevant for the final mine design. The same technique was used to support the mine development of the Sakatti mine site, also in Northern Finland (pers. comm. V. Salonen).

2.1.8 Artificial intelligence (AI)

Artificial intelligence (AI) drives the decision-making in a lot of industries (Chau, 2006; Sakizadeh, 2015). They utilise smart data and ML to enhance the efficiency in operations, safety at the mine and workflow in production (Bonabeau *et al.*, 1999; Diamantopoulou, 2005; Singh *et al.*, 2009; Zuo, 2017). Using AI technology in the mines ensures that data are processed faster than the current data processing methods. As the mining industry changes and grows, AI and ML influence the future choices of today's mines. For example, AI is used in economic geology, to optimise the mineral exploration process (Zuo, 2017). It is also used in machine autonomous vehicles, e.g. in narrow mine tunnels where self-driving trucks require AI technology for easy navigation (Park & Choi, 2020). AI technology also helps to ensure the safety of the miners and improves the safety of mining workplaces and the environment in general.

Introducing this technology in mine water management should be in the form of computational modelling, the knowledge driven algorithms which are the computer programs that utilises expert systems and fuzzy concepts to make decisions (e.g. Chau, 2006; More & Wolkersdorfer, 2022; Rooki *et al.*, 2011; Sakala *et al.*, 2019; Sakizadeh, 2015). Main tools used in ML to predict and forecast mine water quality are the artificial neural networks (ANN) which are computational models consisting of processing elements which receive input data and immediately produce an output. For instance, Maier *et al.* (2004) used ANN to model aluminium dosage and eventually predict the residual aluminium concentrations and pH values in southern Australian surface waters.

2.2 Mining Technology and Mine Water Treatment Plants

2.2.1 Background

Currently, mining industry is confronted with many challenges such as low commodity prices, increasing cost of electricity and production or pressures from NGOs to provide fast and useful monitoring data. These are driving the mining sector to Industry 4.0, as it brings the industrial transformation. Implementation of these new technologies drives an increase in skills demand, meaning that more jobs are and still will be created (Olalekan *et al.*, 2016). As much as these technologies will not change basic mining principles, using electricity and mechanics, the mine workers will communicate with their equipment through the IoT (Lee & Lee, 2015; Sishi & Telukdarie, 2017; Stock & Seliger, 2016), and some might call themselves IT experts in the future. This thesis emphasises that upgrading mines to Industry 4.0 will not affect employment negatively, as this is the fear of almost everyone (Azevedo & Almeida, 2011; Garetti & Taisch, 2012; Shrouf *et al.*, 2014). On the contrary, as other industries have shown, it will create jobs, many of them needing higher qualifications, and requesting better education. Some of these adjustments can already be seen in mines in Ghana, Chile, USA, Finland and Australia.

Mine water treatment branches must start embracing the advantages that come with Industry 4.0. These technologies will allow “smooth” reactions to changing water qualities or quantities in the mine, and enables the mine to not only know that the pump has stopped, but to also know why, meaning that the pump, its control systems and motor will have to be connected to a network that will allow the operator to know all relevant data (Ghodrati *et al.*, 2015; Iarovyi *et al.*, 2016; Stock & Seliger, 2016; Zuehlke, 2010). Ideally, the system will even predict when the pump is going to stop. New skills that will be brought by this interruptive improvement, includes

configuring wireless devices, setting up networks or knowledge on internet protocols. Therefore, this technological advancement will benefit the mining sector as a whole. In this thesis, Industry 4.0 is defined as IIoT (Boyes *et al.*, 2018; Gilchrist, 2016; Sadeghi *et al.*, 2015) with the influence of cloud computing or cloud-based systems (Ruparelia, 2016; Sun *et al.*, 2012). It will increase the safety and security on the mine site, will enable the mines to reach full production (Sadeghi *et al.*, 2015) and optimise mine water management. With a data lake in the picture, communication between all the departments in the mine that deal with water will become faster, easier and more reliable (Lee *et al.*, 2013; Millie *et al.*, 2013; Tekiner & Keane, 2013).

In some areas of the mining sector, falling productivity is highly notable, partly due to lower commodity prices and lower grade ore deposits, which can be prevented with the adaptation of Industry 4.0. A combination of technologies, production, communication and information that already can be seen in some of the mines (e.g. Kansake *et al.*, 2019; Lacey *et al.*, 2019) is what the rest of the mines should be investing in, respectively investing in the future of their mines. It is important to have real-time data in modern mine water management to ensure that good and valuable business decision making becomes normality (Sishi & Telukdarie, 2017) in a world where water management becomes increasingly important.

2.2.2 The aging technology in the mines – Data Silos

Outdated technology in numerous mines is highly disadvantageous to their production. These old systems are time consuming, slow in data capturing and processing, they do not have user-friendly environments, are less accurate when compared to advanced computer processing, and have need for a lot of paper work. During field work, data is commonly written into a field book, which might result in

confusion for later, when information is assembled and data is not managed properly because of this old technology. Several industries, including the mining industry, still work through “data silos”. A data silo can be considered as a group of raw data that can only be administered or controlled by one department and is isolated from the rest of the organisation or other departments (Tekiner & Keane, 2013), often to keep secrets or exclude others from getting overall insights. Figuratively speaking, data silos are the electronic safes of modern times and are, therefore, preventing efficient development on site. For example, a mine site consists of various departments (e.g. the mine, processing plant, treatment plant, tailings facilities), which are not always communicating with each other in terms of data sharing, resulting in a data silo style of data management.

Data silos come with many disadvantages such as difficulties in analysing the data, i.e. data may sometimes be stored in formats that are inconsistent with one another resulting in time consuming standardisation of data and compilation into appropriate formats before it can be used (Gallego *et al.*, 2015; Tekiner & Keane, 2013). Varying levels of security and data duplication is also caused by data silos (Beesley *et al.*, 2009; Gallego *et al.*, 2015; Tekiner & Keane, 2013). Therefore, this data handling practice needs to be stopped and a free flow of data and communication across the mine site departments implemented instead.

With the emerging technologies, some of the mines are moving into more comprehensive solutions for data management, i.e. all the mine departments will transfer the data into a single data retrieving and management system (e.g. Jacobs & Webber-Youngman, 2017). This is known as either “data warehouse” or “data lake”, which is a data storage facility for the company on which multiple functions can be performed on such as real-time data management, on-premises data

management, data analytics and also goes as far as using the data for ML (Figure 2.4; Gorelik, 2019; Herman *et al.*, 2018). Simplified, a data warehouse stores, processes and analyses data in an organised structure, while a data lake holds raw or unstructured data of various types, and processes and analyses this data at the time of usage (Gorelik, 2019). One of the first integrated mine management systems having been invented was Al.Vis from the German company Wismut GmbH in 2003/2004 (pers. comm. M. Haase). This resulted in an easier data management on their mine sites and helped to improve the effectiveness of the water and site management processes.

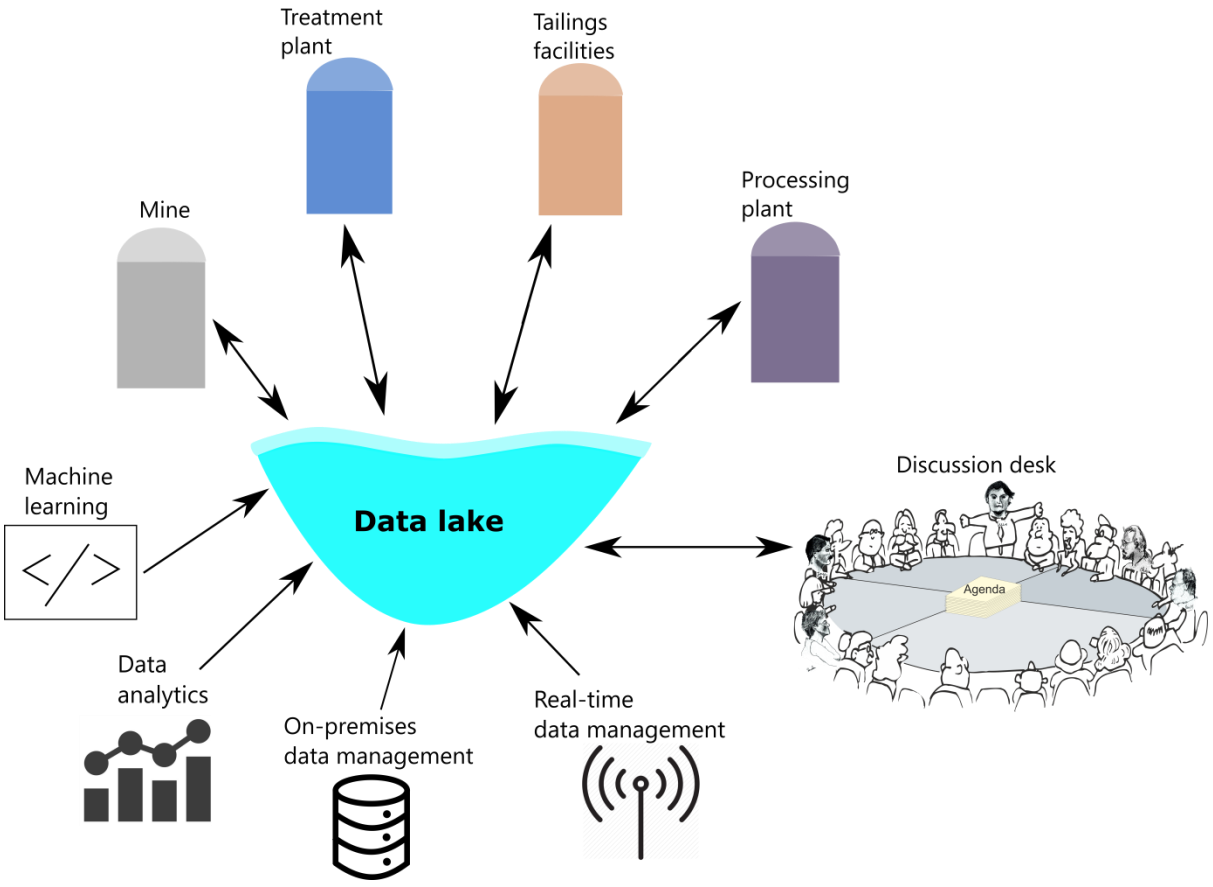


Figure 2.4: Data lake development on the mine site for free flow of data

One of the reasons why the industry still uses the old technology is because it is cheap; they do not want to pay high prices for the new and smart technology though

it will save costs in the long run as opposed to the cheaper technology that costs the companies millions every minute. With the current system, it is often difficult to get access to data at anytime and anywhere or conduct data verification or consistency testing. In this case, cloud storage can be used to store large amounts of data and for easy accessibility of this information, resulting in an optimised working environment. The old technology brings a lot of inactivity in the industry and does not spark the brain to reach the industry's full potential (Athresh *et al.*, 2017; Chadwick, 2016; Ghose, 2009; Jacobs & Webber-Youngman, 2017; Lacey *et al.*, 2019).

2.2.3 Why the Internet of Mine Water (IoMW)?

Internet of Mine Water (IoMW) as introduced by Wolkersdorfer (2013) is all about integrating disruptive technologies on the mine site. Recent technologies in mining show a convincing industry shift towards sustainability. These new technologies allow the mines to reach full production, and tackle problems such as acid mine drainage. There is little progress of implementation of these technologies for mine water management; however, other mining departments are experiencing good technology advancement through the implementation of AI (e.g. Bonabeau *et al.*, 1999; Corke *et al.*, 1998; Duff *et al.*, 2003; Grehl *et al.*, 2018; Sadeghi *et al.*, 2015; Zuo, 2017), SI (e.g. Bonabeau *et al.*, 1999; Danilov *et al.*, 2015; Micklethwaite, 2018; Otto *et al.*, 2018; Walters, 2011) and WSN (e.g. Agrawal *et al.*, 2019; Farrugia, 2011; Haenselmann & Müller, 2011; Li *et al.*, 2010; Losavio *et al.*, 2019; Zhuiykov, 2012). Additional technologies are still to be implemented and they will surely disrupt and slowly replace the existing technologies.

IoMW will link all the mining departments together to improve the techniques used to extract ore and treat mine water. It will help in innovative ways of storing and managing data: this will result in better business decision making due to AI and ML

which analyses this data and builds business patterns and trends. Consequently, the application of IoMW increases operational efficiency and reduces costs. Big data collected by WSN, AI and ML during exploration phase can vastly reduce operational costs by avoiding trial and errors that humans would cause.

Introducing these disruptive technologies in the mines will create smart mines, and thus the IoMW is the mining IIoT. IoMW incorporates the 4th industrial revolution to increase adaptability, visualisation and predictability. This is a connected chain of systems, machines and processes that enables autonomous control on the mine site. IoMW enables real time data monitoring and sharing, ensures that mine water is treated and managed efficiently, reduces energy consumption and deals with safety concerns especially in underground mines.

2.2.4 Failing mine water treatment and management plans

Quality and quantity of mine water on the mine site and the surroundings is highly influenced by mining activities which drastically changes the hydrogeological, hydrological and topographical conditions of the area. Water management plans are a challenging factor in the mines worldwide: in areas with excessive water, the challenging factors are too much water on the mine site and mine water quality and in drought stricken areas they have less water issues and mine water quality problems on the mine site (Hancock & Wolkersdorfer, 2012; Wolkersdorfer, 2008; Younger *et al.*, 2002). Too much or lack of water can both cause problems on the mining operations and the surrounding environment. Lack of proper mine water balance models might result in AMD issues (Bhakdisongkhram *et al.*, 2007; Côte *et al.*, 2010; Kalin, 2004; Kolhinen *et al.*, 2017). Therefore, water management and modelling skills must be improved in mines, continuous online data collection and monitoring for both water quality and quantity must be improved, and free-flow of

data must then be introduced, i.e. linking all mine site departments, thus the development of IoMW.

A proper planning and development of the mine water balance can be beneficial to the mine. This can cut costs of treating and managing mine water, e.g. a good water balance gives room for proper diversion of different water types and water might be re-used for other mining purposes (Bhakdisongkhram *et al.*, 2007; Côte *et al.*, 2010; Gao *et al.*, 2014; Kolhinen *et al.*, 2017). Therefore, the mine could then become socially acceptable when it has excellent environmental performance. Process of managing mine water is a continuous one and changes throughout the mine life cycle which commonly is referred to as longevity of mine water (Asr *et al.*, 2019). Data needs to be collected and updated continuously; therefore, increasing the need of smart technologies on the mine site to collect data and develop water quality and quantity predictive models (Bhakdisongkhram *et al.*, 2007; Côte *et al.*, 2010; Losavio *et al.*, 2019; More & Wolkersdorfer, 2019; Nalecki & Gowan, 2008; Younger *et al.*, 2002). This might result in an end of the stereotypical mine water treatment techniques and the use of old technologies.

Application of modelling software, especially AI techniques, must be made fashionable to replace old spreadsheet-based techniques. Such modelling software are useful and provide good predictive and forecasting models and they can lead to successful mine water management. Spreadsheet-based techniques, on the other hand, are disadvantageous as they cannot handle more complex modelling or be coupled with hydrogeochemical software tools to build accurate mine water models.

2.3 Developing Artificial Intelligence Systems (Algorithms used in this thesis)

2.3.1 Background

AI includes making use of computers to perform functions that normally requires human intervention which implies developing calculations or algorithms to characterise, analyse and make decisions from data (Russell & Norvig, 2002; Wolfgang, 2011). It additionally involves performing functions on given data, learning from new data and sometimes improve it over time (Negnevitsky, 2005). Unlike many other programs that characterise every conceivable situation and only operate within those characterised situations, AI trains a program for a particular task and allows it to explore and improves on its own (Negnevitsky, 2005; Russell & Norvig, 2002; Wolfgang, 2011). AI systems operate on large data so they are able to learn from it, improve scenarios and make better decisions (e.g. Khandelwal & Singh, 2005; Rooki *et al.*, 2011; Singh *et al.*, 2009; Wolfgang, 2011). AI has multiple branches which include ML, expert systems, speech, robotics, to name a few (Figure 2.5), whereupon this thesis will use ML (Figure 2.6).

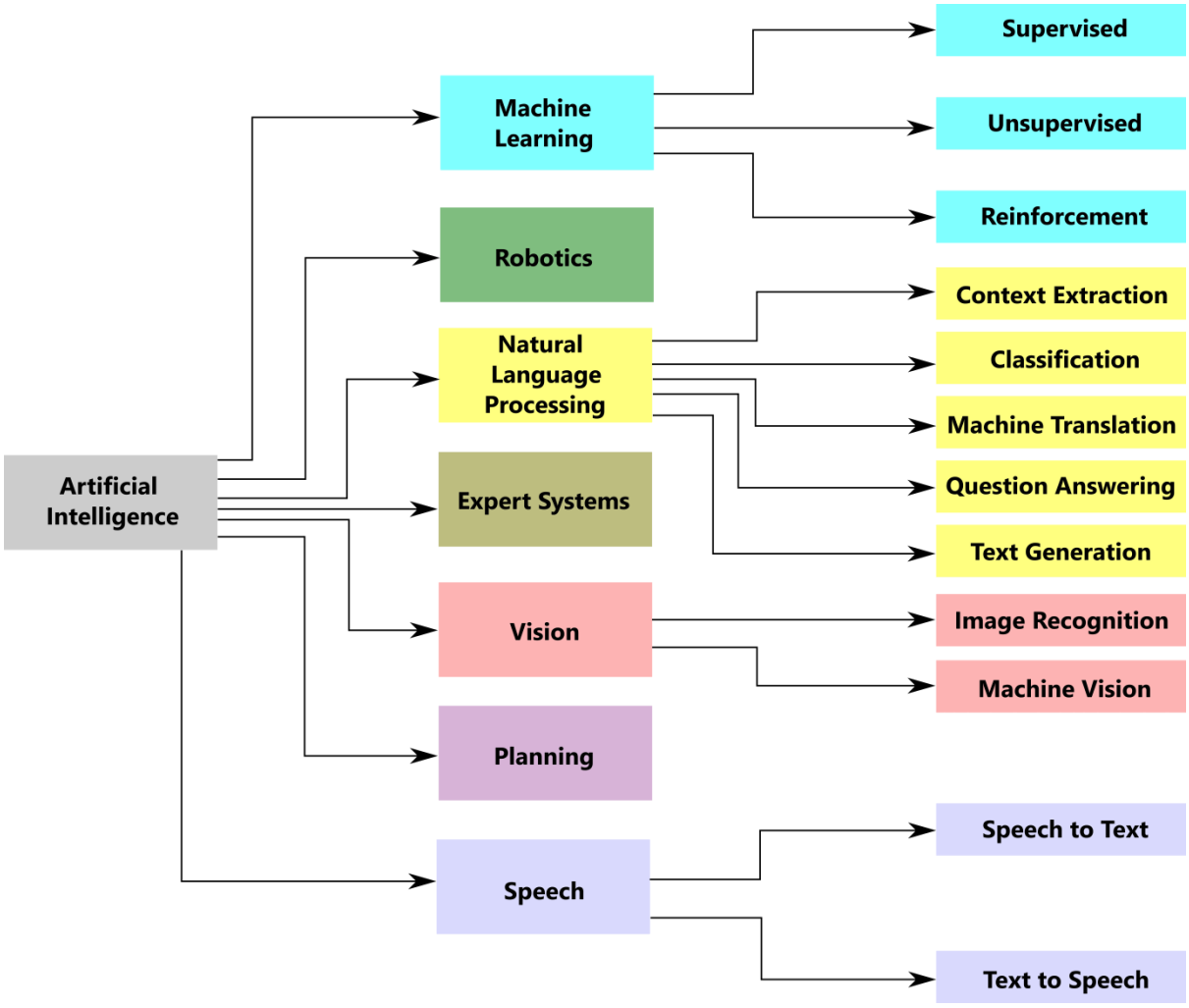


Figure 2.5: Simplified branches of Artificial Intelligence (modified after Sutton *et al.*, 2016)

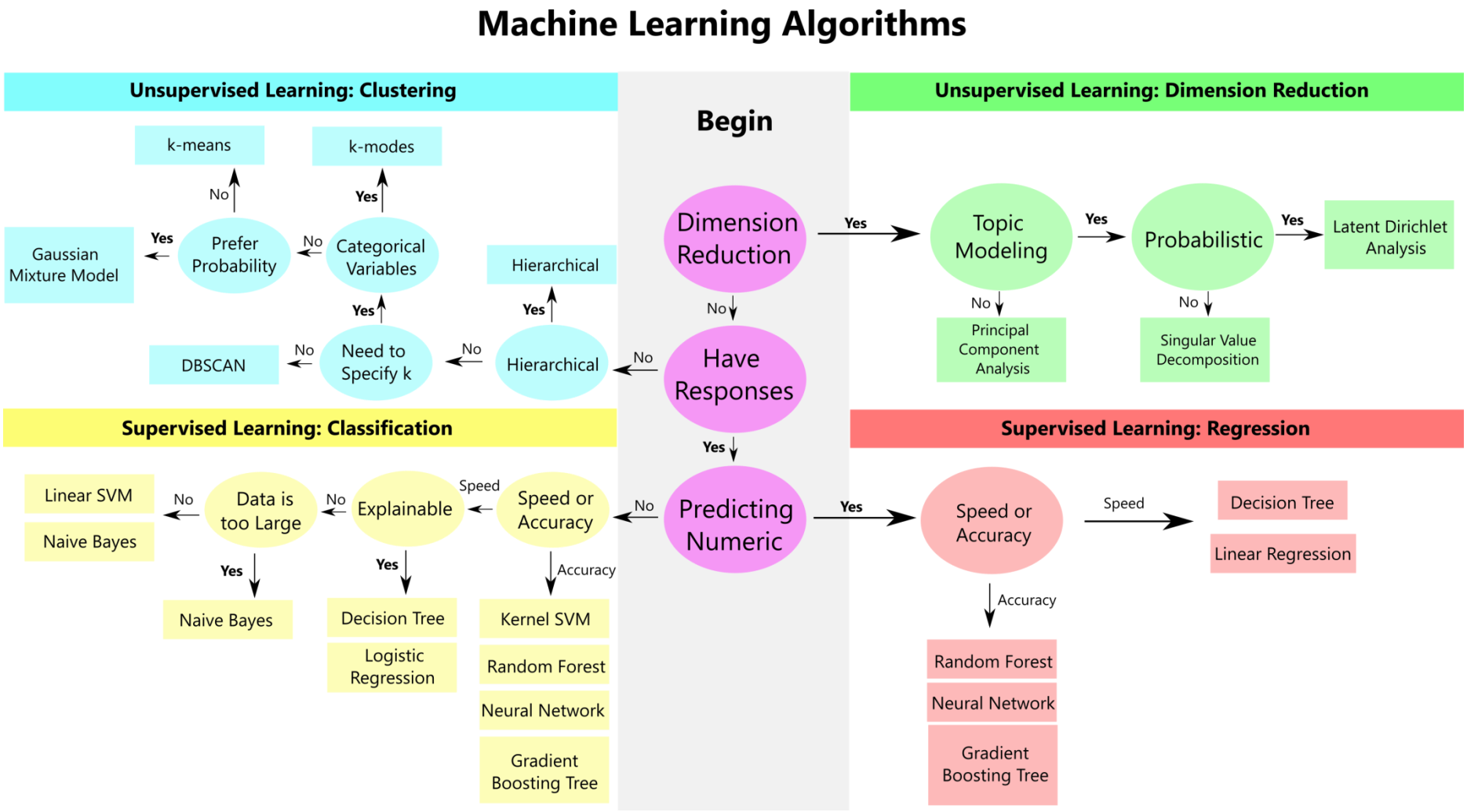


Figure 2.6: Machine learning algorithms; k: the number of clusters from the supplied data, SVM: support vector machine, DBSCAN: density-based spatial clustering of applications with noise

2.3.2 Random forest

Random forest, proposed by Breiman in the early 2000s (Breiman, 2001), is one of the supervised ML algorithms which makes use of ensemble learning to perform either classification or regression tasks. In this study, random forest regression will be explored. Random forest is built on the concept of decision trees algorithm (Figure 2.7; Boulesteix *et al.*, 2012). Decision trees are an ML algorithm where the data are repeatedly split according to a certain parameter. The trees consist of decision nodes where the data are split and leaves where the final outcomes and decisions are made. Since the focus is on random forest regressor, regression trees will be used to build random forest models (e.g. Belgiu & Drăguţ, 2016; Chen *et al.*, 2020; Singh *et al.*, 2017a). Random forests are operated by building multiple decision trees at training time and give a mean prediction of the individual trees. During training time, the trees are run in parallel and do not interact with each other (Biau & Scornet, 2016; Breiman, 2001).

Random forests applies a technique known as “bagging” (Sexton & Laake, 2009). This technique is used to reduce the variance for algorithms that mainly have high variance, such as decision trees. Bagging enables random forests to make decision trees run independently and ultimately aggregates the outputs to give the final output without preferring one model over the other. Two modifications are made when aggregating the results of multiple predictions from several decision trees; for random forest models to not depend largely on individual features, the number of features that can be split on at every single node is limited to a hyper-parameter, which is the total percentage of prediction. Another modification is the prevention of overfitting through random sample selection from the original dataset by the decision

tree when its splits are generated (Biau & Scornet, 2016; Breiman, 2001; Sexton & Laake, 2009). These modifications help in preventing over-correlation of the trees.

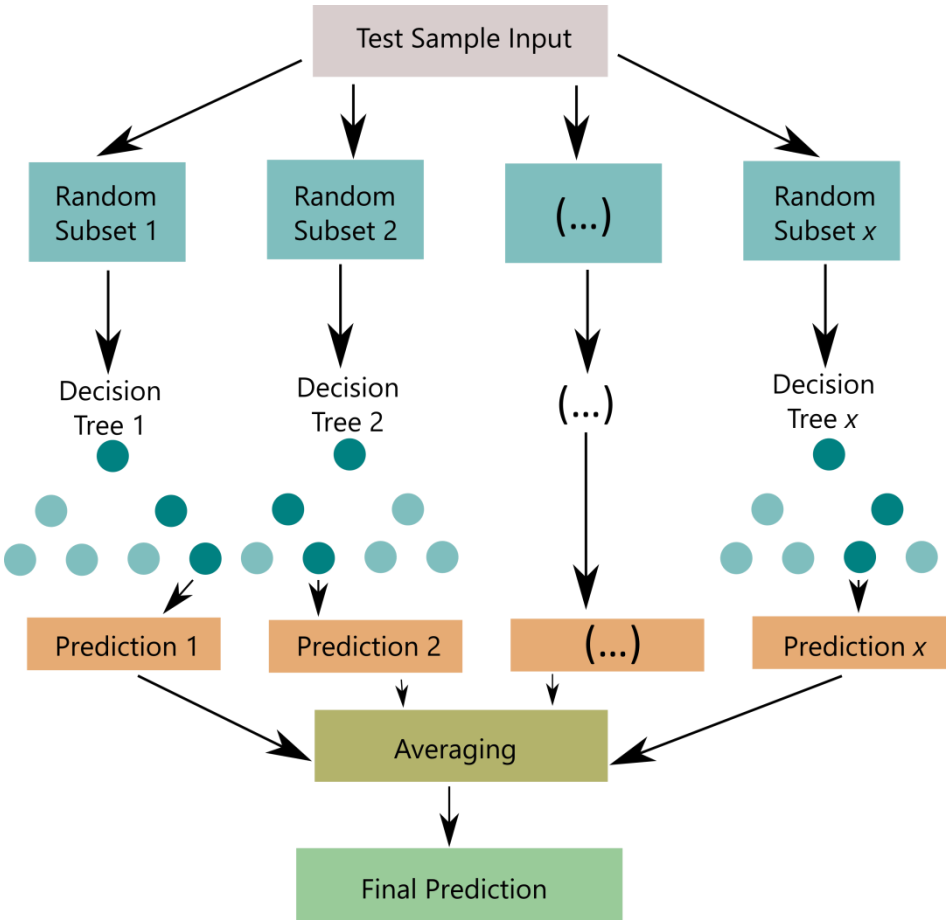


Figure 2.7: Random forest structure (modified after Araya *et al.*, 2017)

2.3.3 Artificial Neural Network System

Artificial neural network (ANN), which is part of the ML branch, is a data-driven system. Data-driven ANN systems work with known input data without any assumptions (Negnevitsky, 2005; Palani *et al.*, 2008; Singh *et al.*, 2009; Wolfgang, 2011). It can conclude on meaningful and workable data relationships that can be utilised to give output data when only input data are presented (e.g. Diamantopoulou, 2005; Khandelwal & Singh, 2005; Maier *et al.*, 2004; Palani *et al.*, 2008; Rooki *et al.*, 2011; Singh *et al.*, 2009). ANNs are one of the many tools applied in AI. The “neural” part of their name implies that they are brain-inspired systems

designed to perform what humans can do as elaborated by Russell and Norvig (2002). Neural networks are made up of (i) an input layer, where the data are initially presented to the model and computation is performed, (ii) a hidden layer, where the ANN model data are processed, and (iii) an output layer, where the results of the ANN model are produced (Figure 2.8; Russell & Norvig, 2002; Wolfgang, 2011).

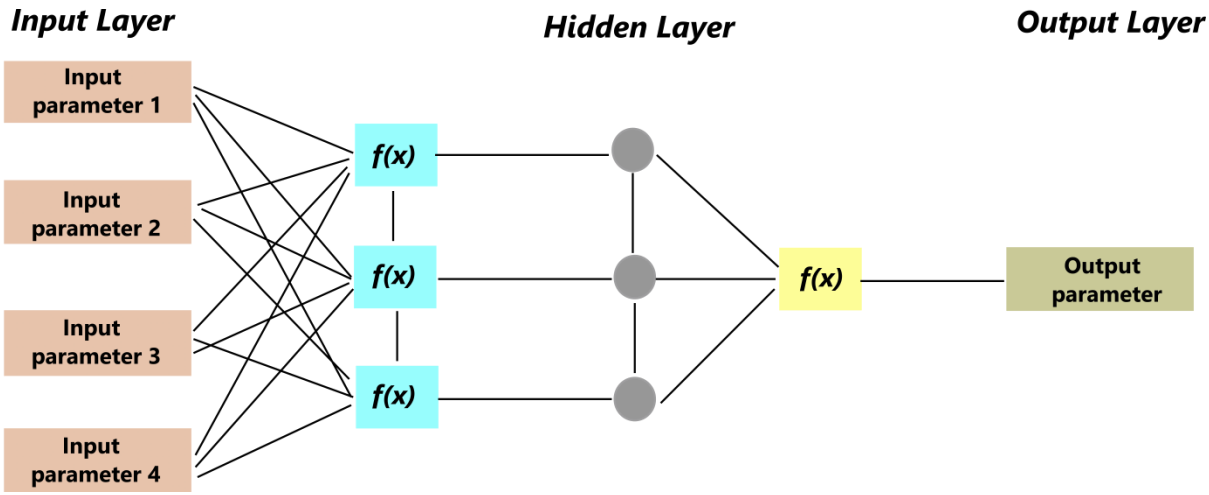


Figure 2.8: ANN structure; *f(x)*: equations

Each layer in this ANN structure consists of at least one basic element which can be referred to as a neuron. In this context, a neuron is defined as a non-linear algebraic function (Negnevitsky, 2005; Wolfgang, 2011). Selecting the number of neurons to use in the hidden layer is important as a larger number may result in model over-fitting while a smaller number can lead to data not being properly captured (Yuan *et al.*, 2003). ANN systems are widely utilised for finding complex patterns for humans to extract and apply the knowledge to a machine to recognise. Many modelling techniques are disadvantageous because of only dealing with linear relationships. ANN can be a useful computational technique, as it can also model complex non-linear relationships (e.g. Bui *et al.*, 2020a; Diamantopoulou, 2005; Maier *et al.*, 2004; Palani *et al.*, 2008; Sakizadeh, 2015).

2.3.4 Gradient boosting regression tree

Gradient boost algorithm is a technique, similar to random forest, which uses an ensemble or decision trees to predict a target label. With random forest focusing on reducing the variance on complex trees, gradient boosting's main aim is to decrease the bias of simple trees and make them more expressive (Johnson *et al.*, 2017; Zhang & Haghani, 2015). This technique optimises the predictive value of a model over multiple steps in the learning process. Each iteration process of the decision tree (Figure 2.9) aims in adjusting the values of the weights, coefficients or biases for each input variable used to build the model to predict the target value, with the main goal being to reduce the loss function, i.e. the difference between the predicted and actual target values (Cai *et al.*, 2020; Li & Bai, 2016). The incremental adjustment made in each step of the model is the gradient, while boosting can be referred to as the activity of speeding up the improvement in predictive accuracy to a desired value.

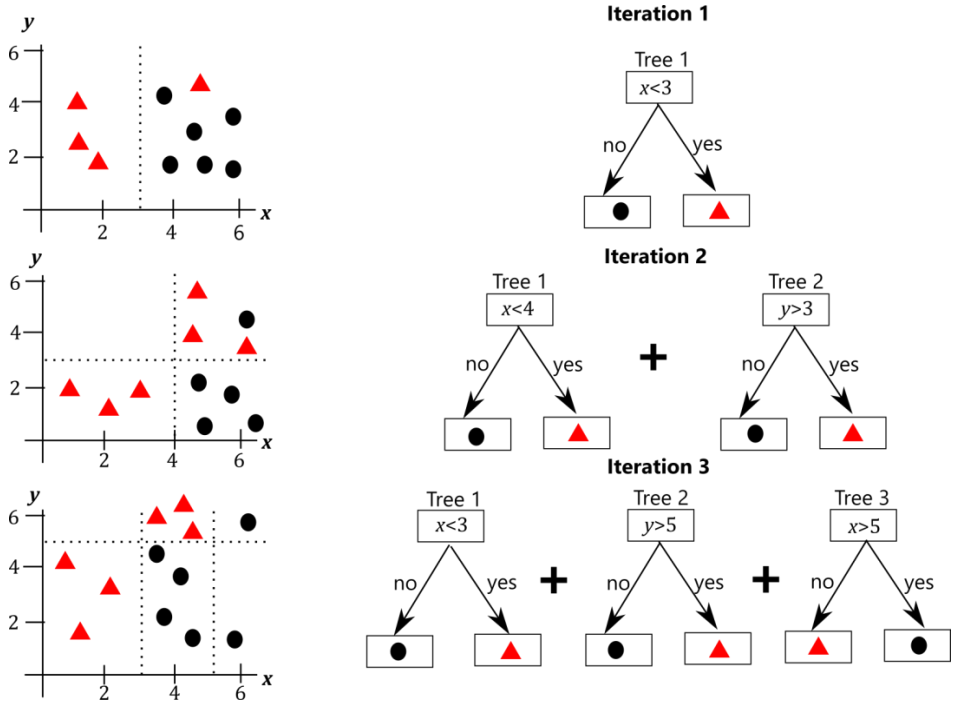


Figure 2.9: Gradient boosting tree structure (modified after Zhang *et al.*, 2018)

2.4 Other algorithms for predictive analysis

2.4.1 Decision tree regression

Decision tree is a supervised machine learning model which is used to predict a target variable by learning decision rules from associated features. This algorithm consists of a parent node which is divided into child nodes. Furthermore, the child nodes are divided into decision nodes to create branches, and the bottom nodes of the tree are known as terminal or leaf nodes (Figure 2.10).

Decision tree can be applied in both continuous and classification target variable prediction. Decision tree regression is used to predict the target variable whose values are continuous in nature. It is a flexible algorithm that can be considered for a wide range of applications. Hamoud (2016) used decision tree algorithm to classify and predict the willingness of students to enrol and complete higher education qualification after completing secondary school. This algorithm was also applied in the health sector by Romero *et al.* (2020) to help in controlling bovine tuberculosis disease. Decision trees are mostly favoured in regression problems because of their ability to exhaust all possible scenarios that can potentially influence decision making. Additionally, the output of decision tree algorithms is easy to read and interpret without having to use robust statistical methods, and thus it is the “go to” algorithm for regression problems. The mechanism of decision tree algorithm is about making decisions along the way to narrow down the possible values in order to tune the model until it gets confident enough to compute the final prediction (Figure 2.10). The decisions made are all in a True or False form.

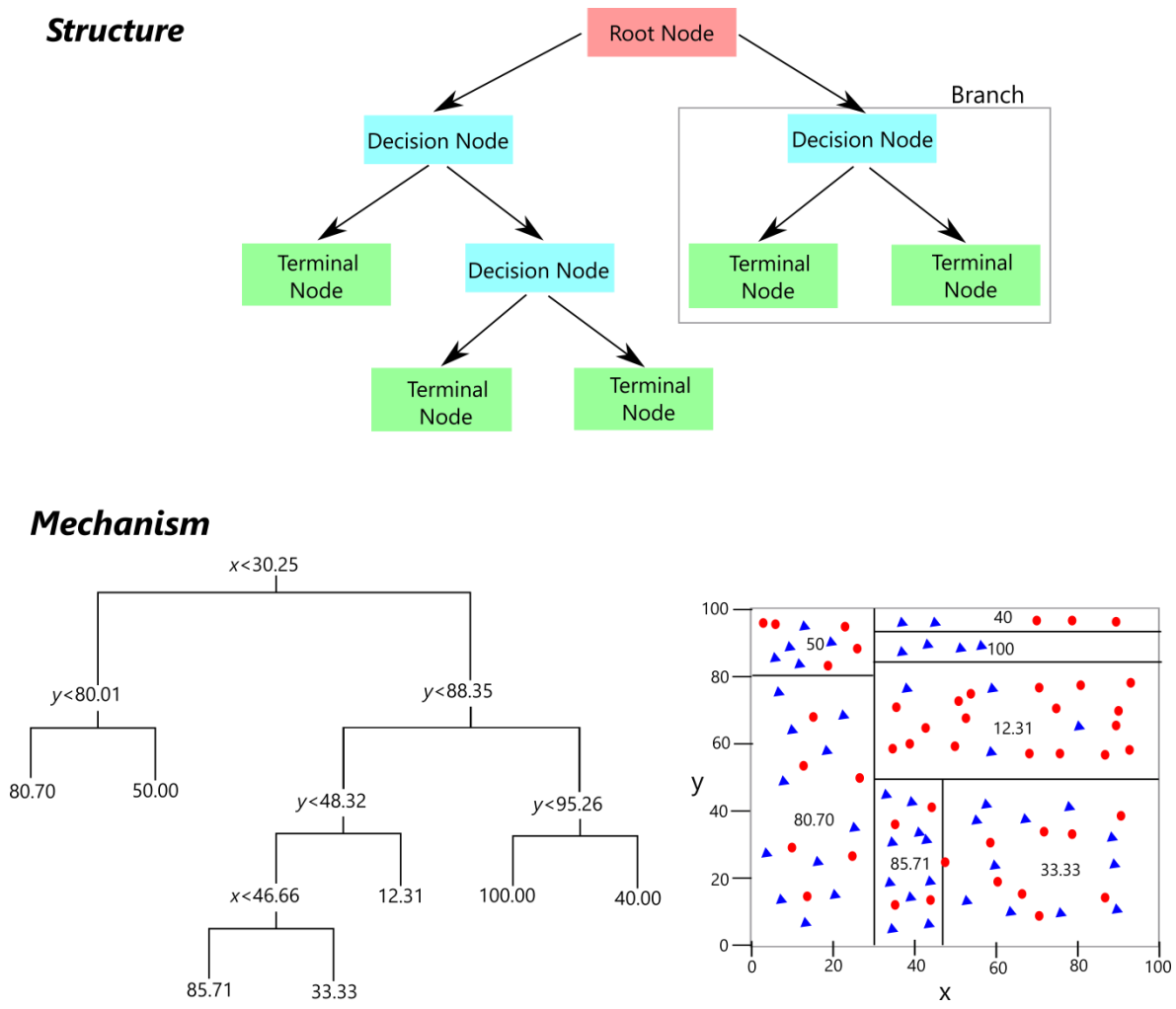


Figure 2.10: Decision tree algorithm structure and its mechanism (modified and supplemented after Wagner *et al.*, 2019)

2.4.2 Linear regression

Linear regression is s statistical model that aims to show the relationship between two variables through a linear equation, i.e. the algorithm aims to find the relationship between variables *x* and *y*. This means that every value of *x* has one corresponding value of *y* if the data are continuous. Linear regression algorithm can be used in several applications such as evaluating trends and sales estimates or analysing the effect of price changes in businesses (e.g. Ghosalkar & Dhage, 2018), forecasting the daily behaviour of stock markets (e.g. Bhuriya *et al.*, 2017) or in the energy sector (e.g. Ighalo *et al.*, 2020).

In linear regression algorithm, the relationship between the independent and dependent variables is shown by a line of regression, mathematically known as the '$y = mx + c$' line (Figure 2.11). This line can either show a positive or negative relationship. In case of a positive relationship, the independent and the dependent variable increases. When the independent variable increases and the dependent variable decrease, the resulting relationship is negative.

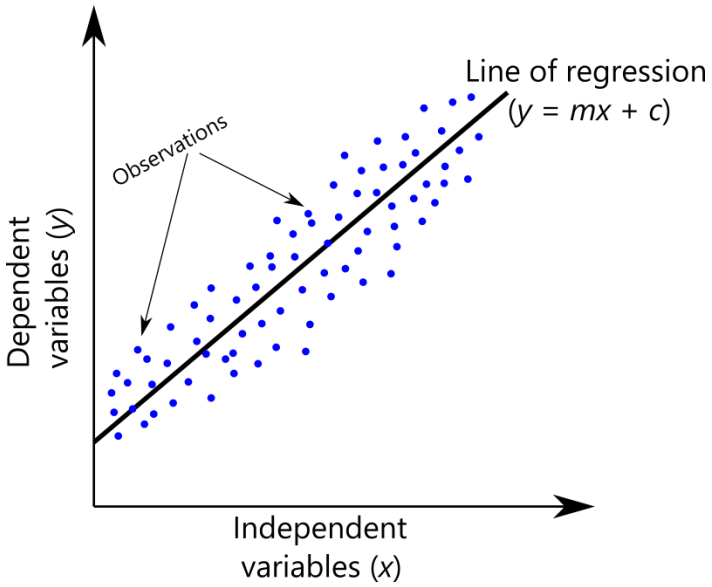


Figure 2.11: Linear regression visualisation

2.4.3 Fuzzy Expert System

Fuzzy optimisation is a technique on which reasoning that resembles human reasoning is involved. This method is much more similar to how experts perform decision making. It brings together all the intermediate possibilities between "yes" and "no", e.g. for the question of "is it hot?" a Boolean logic would provide "yes" or "no" as an answer, but fuzzy logic would give answers such as "very hot", "a little hot" or "very less hot". Fuzzy logic was made popular by Zadeh (1965), who coined the technique as one of the first approaches to AI. This technique is used for both commercial and practical purposes, from controlling machines to consumer products,

and its reasoning is not always accurate but acceptable. Fuzzy logic consists of a variable name, set of terms, universe of discourse, syntax rules and semantic rules (Figure 2.12; Buckley & Eslami, 2002).

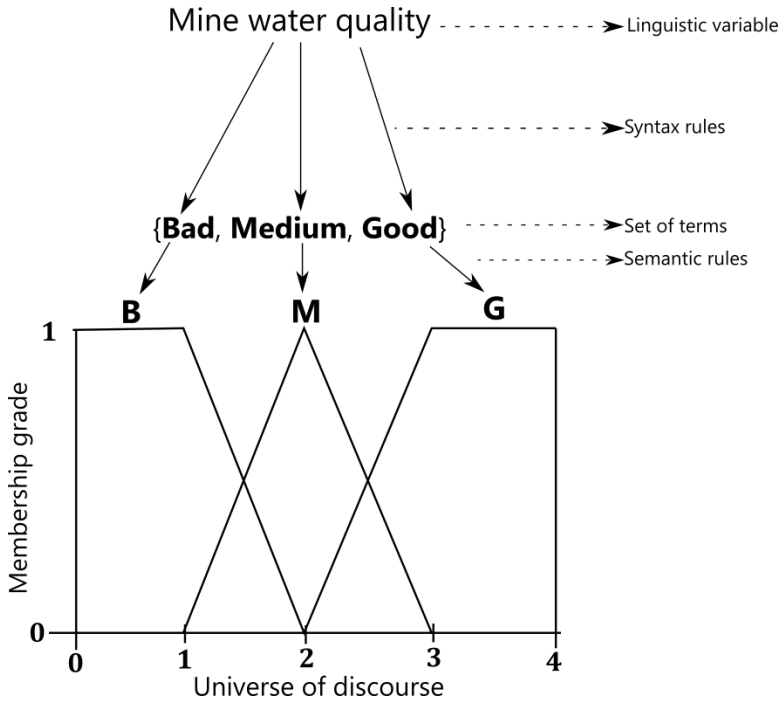


Figure 2.12: Fuzzy logic example for predicting mine water quality

In fuzzy systems, a linguistic variable is a classical set, *A*, which is the set of variable *x* such that *x* comes from the universe of discourse *X*:

$$A = [x|x \in X] \qquad 2.1$$

And a fuzzy set can be given by:

$$A_{fuzzy} = [x, \mu_A(x)|x \in X, \mu_A(x) \in (0;1)] \qquad 2.2$$

where μ_A refers to the fuzzy membership function (FMF). The FMF quantifies the degree of the state of belonging of *x* to *A*. Since *A* is a linguistic variable, its meaning can be modified.

The implementation of knowledge-driven expert systems involves solving problems that are normally solved by human expertise. It requires understanding, substantial knowledge base, a reliable inference engine and an efficient user interface to solve problems related to work it is designed for by interacting with users (Figure 2.13). This system acquires substantial knowledge from the human expertise through the knowledge acquisition process (Chen & Pham, 2001; Cornelius, 1998; Siler & Buckley, 2005).

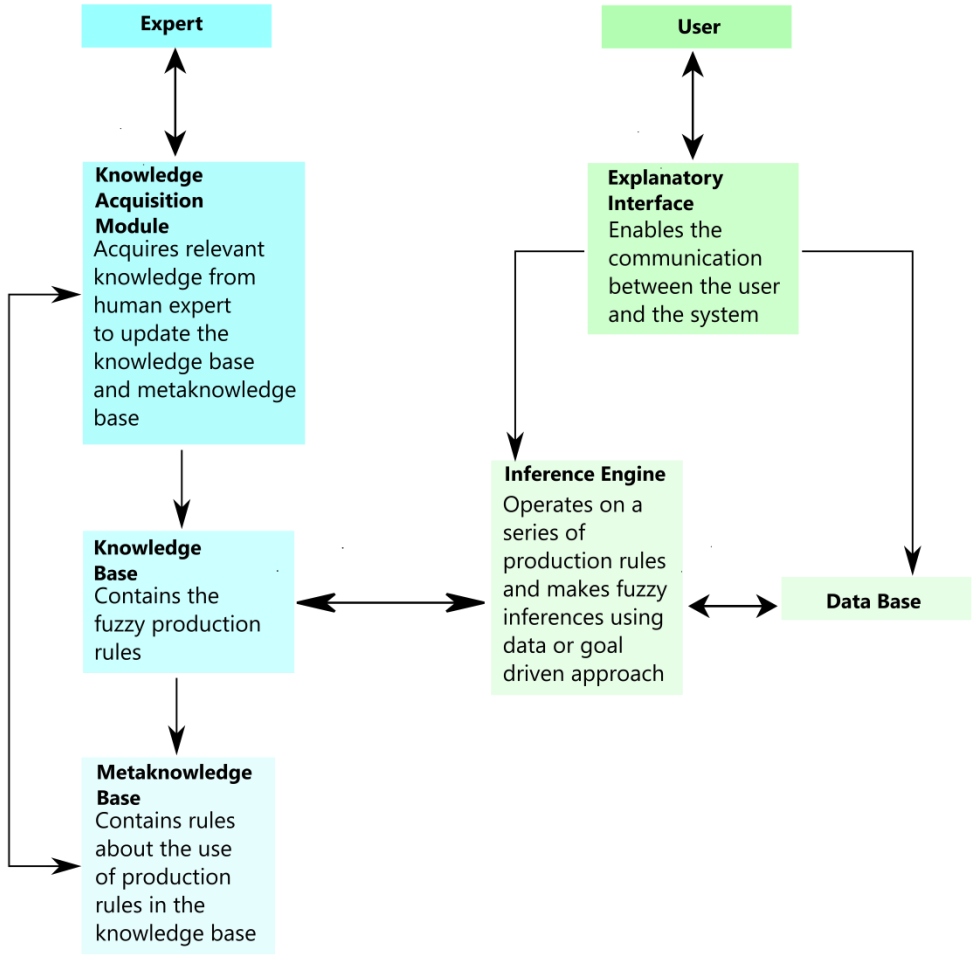


Figure 2.13: Overview of an expert system

A fuzzy expert system is a form of an AI that involves a set of fuzzified inference rules and formulating computer-based systems to solve the problems related to work

it is designed for (Siler & Buckley, 2005). The fuzziness came about because of the uncertainty and imprecisions that were encountered in certain situations and negatively affected decision making capability. Fuzzy reasoning technique based on the Mamdani fuzzy inference method is widely used in several applications (e.g. Deb *et al.*, 2008; Mahapatra *et al.*, 2011; Nasr *et al.*, 2012).

Mine water quality is commonly classified based on chemical, physical and biological indicators, and these include pH, redox potential, electrical conductivity, water temperature, air temperature, oxygen saturation, water level, biological oxygen, metal concentrations and total hardness (Wolkersdorfer, 2008). However, decisions on mine water quality cannot be reached by just using “crisp” datasets. Therefore, a fuzzy expert system makes it more understandable to consider mine water quality data as a fuzzy set (Figure 2.15), which makes it easy to express the term quality in a linguistic manner, e.g. poor, medium or good (e.g. Deb *et al.*, 2008; Liu *et al.*, 2019; Mahapatra *et al.*, 2011; Nasr *et al.*, 2012; Sahu *et al.*, 2011). Values ranging from 0 (worst water quality) to 100 (best water quality) are mainly used to express the deffuzified water quality (Chen & Pham, 2001; Cornelius, 1998; Siler & Buckley, 2005).

The goal with fuzzy expert modelling is to make a machine understand the developed algorithms so it can work as an autonomous system without an expert’s intervention and make decisions on its own. Fuzzy logic is a process in which a vague statement or data can be quantified so that the information can be well understood. So, the need to develop fuzzy logic models result from current computers being able to manipulate only precise facts defined numerically. Four components are used to build the model: fuzzifier, inference engine, defuzzifier, and fuzzy knowledge base which is connected to the inference engine (Figure 2.14).

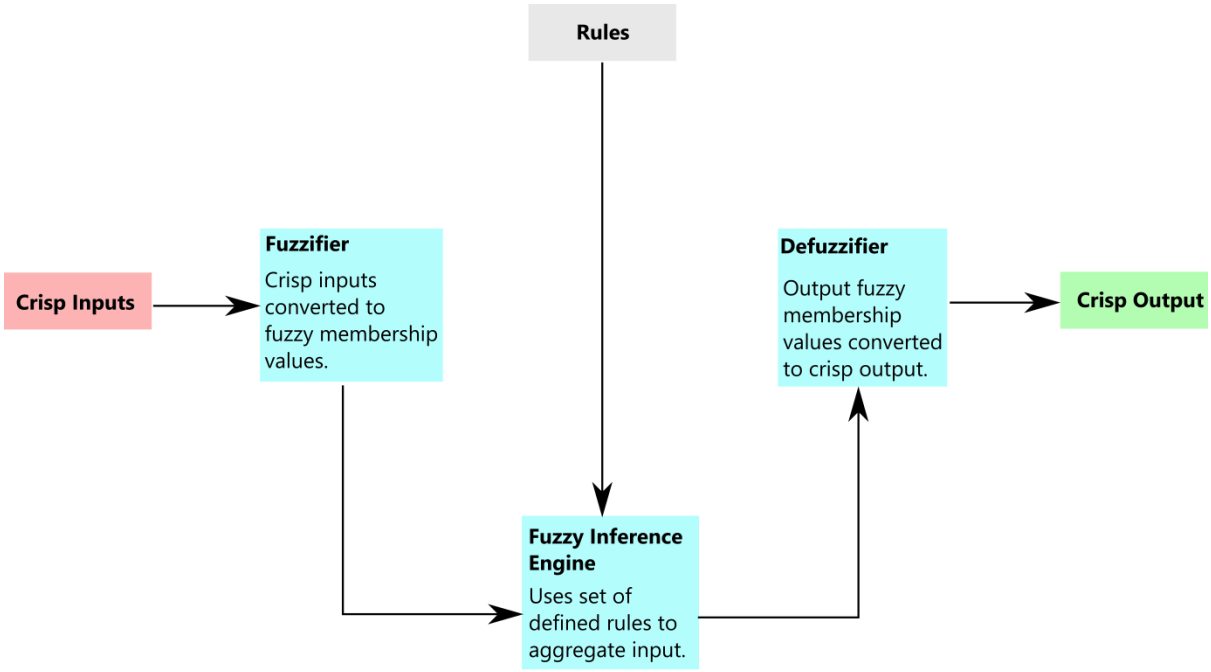


Figure 2.14: A more generalised fuzzy expert system model

Input parameters go through the fuzzifier, and the defuzzifier produces an output. In this case, the crisp inputs and output are non-fuzzy numbers limited to a specific range. Fuzzification process is performed in the fuzzifier, i.e. transforming a crisp set to fuzzy membership values (FMV) that form the fuzzy membership function (FMF), while the defuzzifier converts the fuzzy output of the inference engine to crisp value. Fuzzy inference engine, which can be regarded as the "brain" of an expert system, is the main part of a fuzzy logic system with decision making as its primary work. Fuzzy output is computed in the inference engine by combining all the FMV for all the input layers with fuzzy rules. It uses several rules and connectors such as the "*IF-THEN*" rules along with "*OR*", "*AND*", "*SUM*", "*PRODUCT*" or "*GAMMA*" connectors (Figure 2.15). Some of the common fuzzy operators are as follows:

The *AND* operator which is the minimum of the truth values is given by:

$$\mu_{AND} = \text{MIN}(\mu_A, \mu_B, \ldots, \mu_n) \quad 2.3$$

where the μ refers to the FMF.

The *OR* operator which is the maximum of the truth values is given by:

$$\mu_{OR} = \text{MAX}(\mu_A, \mu_B, \ldots, \mu_n) \quad 2.4$$

The *SUM* operator which combines the FMV to get a larger output result compared to the largest contributing FMV is given by:

$$\mu_{SUM} = 1 - \prod_{i=1}^{n} (1 - \mu_i) \quad 2.5$$

The *PRODUCT* operator which combines the FMV to get a smaller output result compared to the smallest contributing FMV is given by:

$$\mu_{PRODUCT} = \prod_{i=1}^{n} \mu_i \quad 2.6$$

The defuzzifier computes the crisp output variable by converting the fuzzy output membership values.

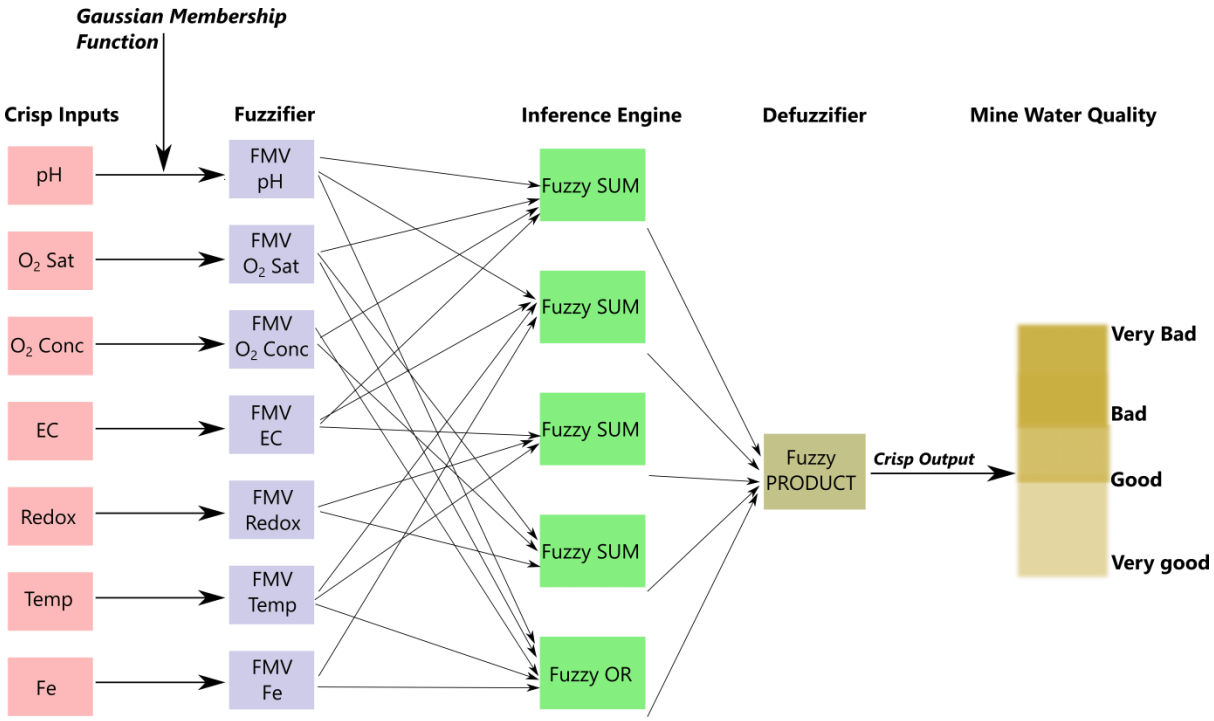


Figure 2.15: Example of a fuzzy expert system model for mine water management. O_2 Sat: Oxygen saturation, O_2 Conc: Oxygen concentration, Redox: Redox potential, Temp: Water temperature, Fe: Iron concentration, EC: Electrical conductivity, FMV: Fuzzy Membership Value

2.4.4 Hybrid intelligent system

A hybrid AI system combines at least two AI methods in order to produce a single outcome. These systems can combine methods such as, neural networks and fuzzy expert systems to name just one possible combination of them. This approach has proven to be effective in solving complex problems. Both the fuzzy expert system and ANN model have their flaws, e.g. the neural network models do not give explanations on how they reach their decisions, and fuzzy expert systems do not have the ability to automatically acquire sets of rules on reaching decisions. In summary, a fuzzy system is bad in learning ability, while neural networks are not good in knowledge representation and explanation ability (Kar *et al.*, 2014). Hybrid intelligent systems therefore combine the ability of both the ANN and fuzzy systems and can be able to overcome their limitations. On a hybrid system, if ANN can't

perform a certain function, it will be replaced by a fuzzy system to carry out that function.

Combining a neural network with a fuzzy system to produce a heterogeneous hybrid neuro-fuzzy system is one of the most common hybrid intelligent systems used in predictive analysis problems. Hybrid neuro-fuzzy systems come in different forms, however, only the adaptive neuro-fuzzy inference system (ANFIS), a method proposed by Jang (1993), produce the most successful results. Its inference system corresponds to a set of fuzzy *IF-THEN* rules and *AND* operators that have learning capability to approximate non-linear functions. It is generally regarded as a universal estimator, and uses the learning abilities of a neural network by automatically adjusting the membership function in the fuzzy inference system (Chung & Halim, 2014; Jang, 1993). This model has been applied for many years in the water sector, e.g. Nayak *et al.* (2005) proposed the ANFIS methodology for river flow forecasting, Galavi and Shui (2012) used hybrid learning algorithms for water resource forecasting, and Sudheer and Mathur (2010) used the methodology for groundwater flow detection.

In terms of the architecture for the ANFIS model, it is made up of five layers with two types of nodes, fixed and adaptable. Layer 1 can be referred to as the fuzzification layer, layer 2 as the rule layer, layer 3 as the normalisation layer, layer 4 as the deffuzification layer and layer 5 as the output layer (Figure 2.16).

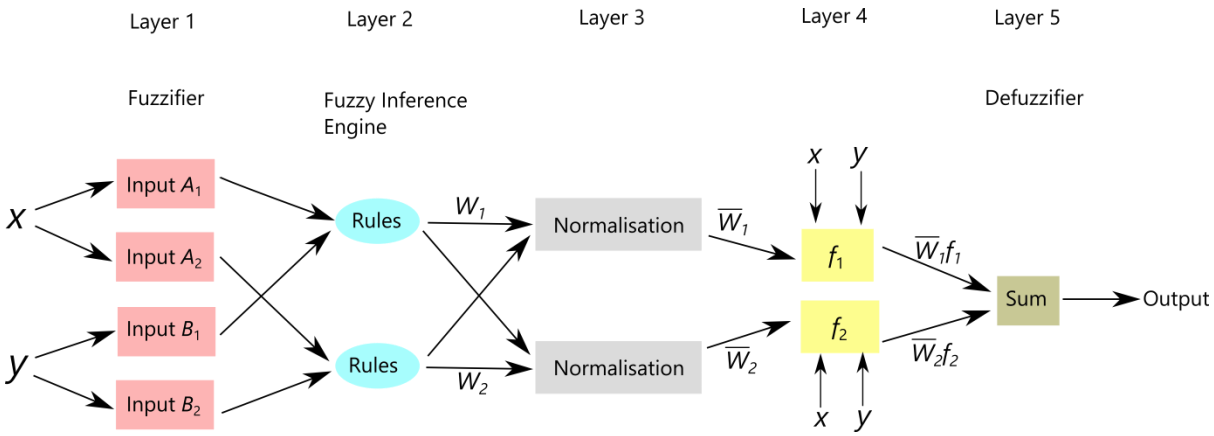


Figure 2.16: ANFIS architecture (modified after Jang, 1993)

Layer 1, which contains adaptive nodes, takes the input values and determines the membership functions belonging to them. Membership grades of each function are generalised by utilising the premise parameter set known as [a_i, b_i, c_i]. Therefore, the crisp inputs are converted into fuzzy inputs, and each crisp input has its own group of membership functions to which they are converted. Layer 1 is the linear transfer function and the outputs are the same with the corresponding inputs. Membership function applied in this model is the generalised bell function:

$$O_{1,i} = \mu_{Ai}(x) = \frac{1}{1+\left|\frac{x-c_i}{a_i}\right|^{2b_i}} \qquad 2.7$$

where $O_{1,i}$ is the output of the i^{th} node in layer 1, x is the input to node i and A_i is a linguistic label (vague terms, e.g. low, medium or high) from fuzzy set A.

Layer 2 is responsible of computing the firing strengths for the rules, e.g. rule 1: if x is A_1 and y is B_1 then f_1 will be $xp_1 + yq_1 + r_1$, and rule 2: if x is A_2 and y is B_2 then f_2 will be $xp_2 + yq_2 + r_2$, where p, q and r are the design parameters determined during the training process, and f refers to the outputs within the fuzzy region determined by the fuzzy rule. Every node in this layer is a fixed node, and the process in this layer

is to find out what should be the membership value. The output of every single node in this layer is fuzzy *AND* (product or min):

$$O_{2,i} = W_i = \mu_{Ai}(x) \times \mu_{Bi}(y);\ i \in (1,2) \qquad 2.8$$

Layer three's role is to normalise the computed firing strengths. Every node in this layer is a fixed node. Output of the i^{th} node is the ratio of the firing strength of the i^{th} rule to the sum of all the rules of the firing strength:

$$O_{3,i} = \overline{W_i} = \frac{W_i}{W_1 + W_2};\ i \in (1,2) \qquad 2.9$$

Layer 4 carries out the task of fuzzy inference. It takes the normalised values as inputs and the consequence parameter set. Every node in this layer is an adaptive node. The output of each rule is a linear combination of the input variables added by a constant value. Therefore, this layer returns the defuzzificated values:

$$O_{4,i} = \overline{W_i} f_i = \overline{W_i}\left(xp_i + yq_i + r_i\right);\ i \in (1,2) \qquad 2.10$$

Therefore, the computed values in layer 4 are passed to the last layer to return the final output. This layer contains fixed nodes. Lastly, the final output can be considered as the weighted average of each rule's output:

$$O_{5,i} = \sum_i \overline{W_i} f_i = \frac{\sum_i W_i f_i}{\sum_i W_i} \qquad 2.11$$

In the following chapter, methodologies applied in this thesis will be clearly explained. Procedures for the radio frequency identification based mine water sampling process and protocol, and the application of machine learning methodologies will be outlined. The following chapter will give full details of Python

programming language (used to develop the machine learning models for this thesis) and its libraries.

CHAPTER 3

3 Methodology

3.1 Introduction

This section contains a discussion of the different methods applied and sample collection procedures taken during the course of the thesis, including mine water quality data analysis and interpretation methods. In order to obtain a holistic understanding of the observed mine water quality and mine water management, different artificial intelligence (AI) methods were applied, including regression analysis and neural network algorithms. Furthermore, in order to include chemical and physical variation of mine water and the temporal and spatial variation thereof, historical data from monitoring networks and geochemical characterisation of the study areas were obtained from the mine. The project was implemented in a four-step case study approach on which different technologies were applied.

3.2 Method development and phases of the project

In order to get a proper mine water data analysis, water balance on the mine is required. A water balance for a mine site is an accounting for the water flowing into the mine from possible sources. After a mine was selected, their mine water management was stepwise converted and included in an IoMW system. However, some of the phases were not fully completed because of the unavailability of the advanced technologies that were supposed to be included in the IoMW system. Therefore, the completed phases included setting up an RFID controlled sample management and developing AI methodologies (Table 3.1). A project plan was set up in conjunction with the mine and all relevant water pathways into and out of the treatment plant. In this concept, all water pathways into, through and out of the mine

need to be monitored and the data can be used to develop machine learning models (Figure 3.1).

Table 3.1: Completed phases of the thesis

Case studies (thesis phases)	Location	Duration	Technologies involved	Expected outcomes
RFID controlled mine water sample management	Westrand Treatment Plant, Randfontein, South Africa	13 months	NFC (RFID) microchip attached to the mine water sampling bottle(s). IoT via the mobile and website applications	Simplified mine water sampling. Easy recording and storing of on-field water quality parameters. No interchange of data and samples during data extraction in the laboratory. Simplified sample identification and data extraction using NFC mobile application.
Random forest and Gradient boosting algorithms modelling	Eastrand Treatment Plant, Ekurhuleni, South Africa	8 months	Artificial intelligence (machine learning)	Forecasting (prediction) of mine water quality. Machine learning website application.
Neural networks (LSTM, ANN and DNN) and regression trees (random forest and gradient boosting) modelling	Westrand Treatment Plant, Randfontein, South Africa	10 months	Artificial intelligence (machine learning)	Forecasting (prediction) of mine water quality. Application of different exploratory data analysis techniques.
Internet of Mine Water (IoMW)	Westrand and Eastrand Treatment Plants, South Africa	10 months	Artificial intelligence (machine learning), IoT and GUI	Develop the IoMW GUI embedded with machine learning algorithms to perform forecasting analysis.

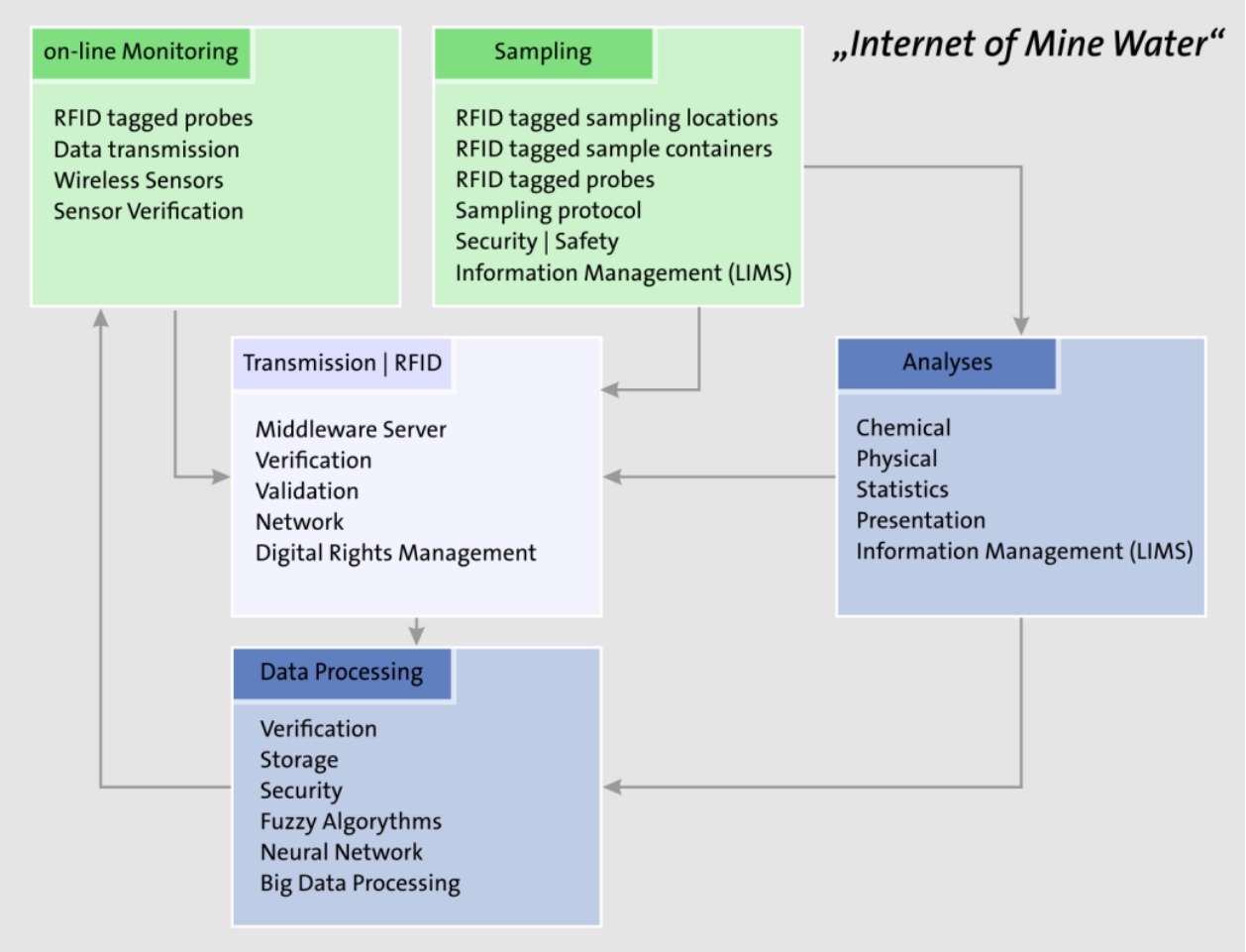


Figure 3.1: Relevant steps and their connection in the Internet of Mine Water (IoMW) system (from Wolkersdorfer, 2013)

3.3 Dynamic mine water management

A dynamic system approach for treating mine water is slowly becoming a norm with AI technology and modelling software such as GoldSim, MATLAB, Simulink, STELLA, Vensim and PHREEQC being favoured over spreadsheet-based approaches. Dynamic modelling is highly useful when the aim is to predict the mine water quality and quantity over time, as this was the case in this thesis. Nowadays, it is recommended to use dynamic and probabilistic methods when designing a mine water management system, which will require a dynamic system modelling package that can integrate all possible factors that might affect mine water management

(Awuah-Offei & Frimpong, 2011; Nalecki & Gowan, 2008). An AI modelling approach was practiced in this thesis and yielded positive results.

Applying the aforementioned water software and technologies will result in good mine water balance development for the whole mine site (George *et al.*, 2009). With the wide range of dynamic software, GoldSim is the most used in generating mine water balance models, mainly because of its flexibility. Though dynamic models are not new and have been around for a long time, they are not utilised in mine water management to a degree that would be possible. Dynamic water balance calculations include all the mine site departments; therefore, this approach provides a highly qualitative tool which can be used to track the system's performance. Dynamic system modelling is helpful because it includes all possible factors contributing to the mine water quality and quantity. Shifting to this approach will be vital for operating future mines and such consideration will also become important in the water sector as shown in this thesis and also elaborated by Kunz and Moran (2016).

3.4 Predictive analysis — Machine learning models and time series forecasting

Anaconda, a Python distribution platform for machine learning (ML), was used to develop predictive analysis techniques to predict the future mine water chemistry. This included the development of ML models: random forest (section 2.3.2), gradient boosting (section 2.3.4), neural networks (section 2.3.3) and long short-term memory (LSTM) time series analysis techniques. The aforementioned methods were used to build AI systems such as a web application, and applied as the predictive analysis method. Predictive analysis is a technique that consists of a variety of statistical methods from big data analysis, predictive modelling and ML models that analyse

the current and historical patterns and trends to make future predictions (Figure 3.2). Historical data is used to build predictive models that capture the important patterns and trends. Current data is therefore used on the built predictive models to predict the future events. Predictive analysis can be applied in several fields such as intelligence gathering and crime analysis (McCue, 2014), the healthcare (e.g. Bos *et al.*, 2014), education (e.g. Fernandes *et al.*, 2019) and water sectors (e.g. Arismendy *et al.*, 2020). In this thesis, predictive analysis techniques were used to predict future mine water chemistry.

For trained and tested ML models to make their predictions, they must be supplied with new data. ML and statistical forecasting techniques were applied to build a time series model for the input parameters. ML forecasting is most effective in capturing patterns and trends in either well-structured or unstructured dataset. Time series refers to the sequence of observations measured in constant time intervals, for example, in this study the measurements were taken on a daily basis. Time series analysis include building models used to explain the observed time series and understand the whole meaning behind its dataset, including the trends and patterns (Chatfield, 2000). Time series forecasting uses the best fitting model crucial in predicting the future observations using patterns and trends of previous and current data.

When choosing suitable models for time series prediction, it is always crucial to understand time-series data components such as: a) seasonality – which will note the repeating patterns of cycles of behaviour over time, b) cyclicity – identifying repetitive changes in the time series and explain their positioning in the cycle, c) trends – which is frequently observed as a linear model; it shows the decreasing and increasing behaviour of the time series, and d) anomalies – to detect observations

deviating from the time series model (Chatfield, 2000; Zhang, 2003). After several tests and data analysis techniques were conducted, LSTM (Figure 3.3) was used to forecast the input parameters.

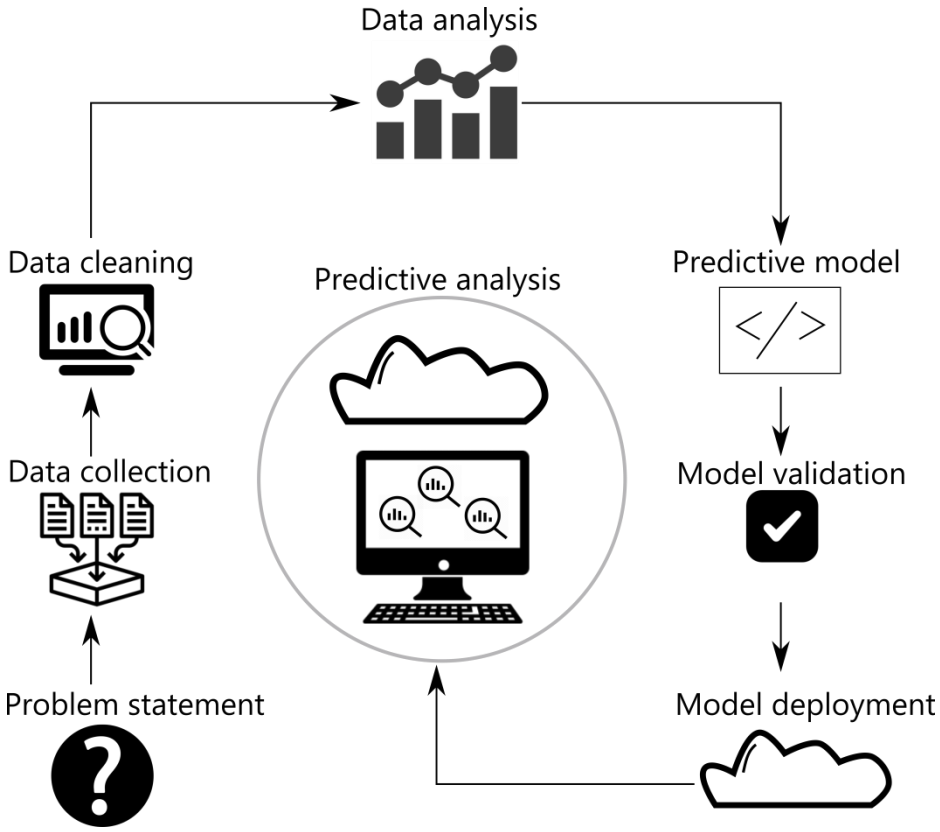


Figure 3.2: Predictive analysis mechanisms

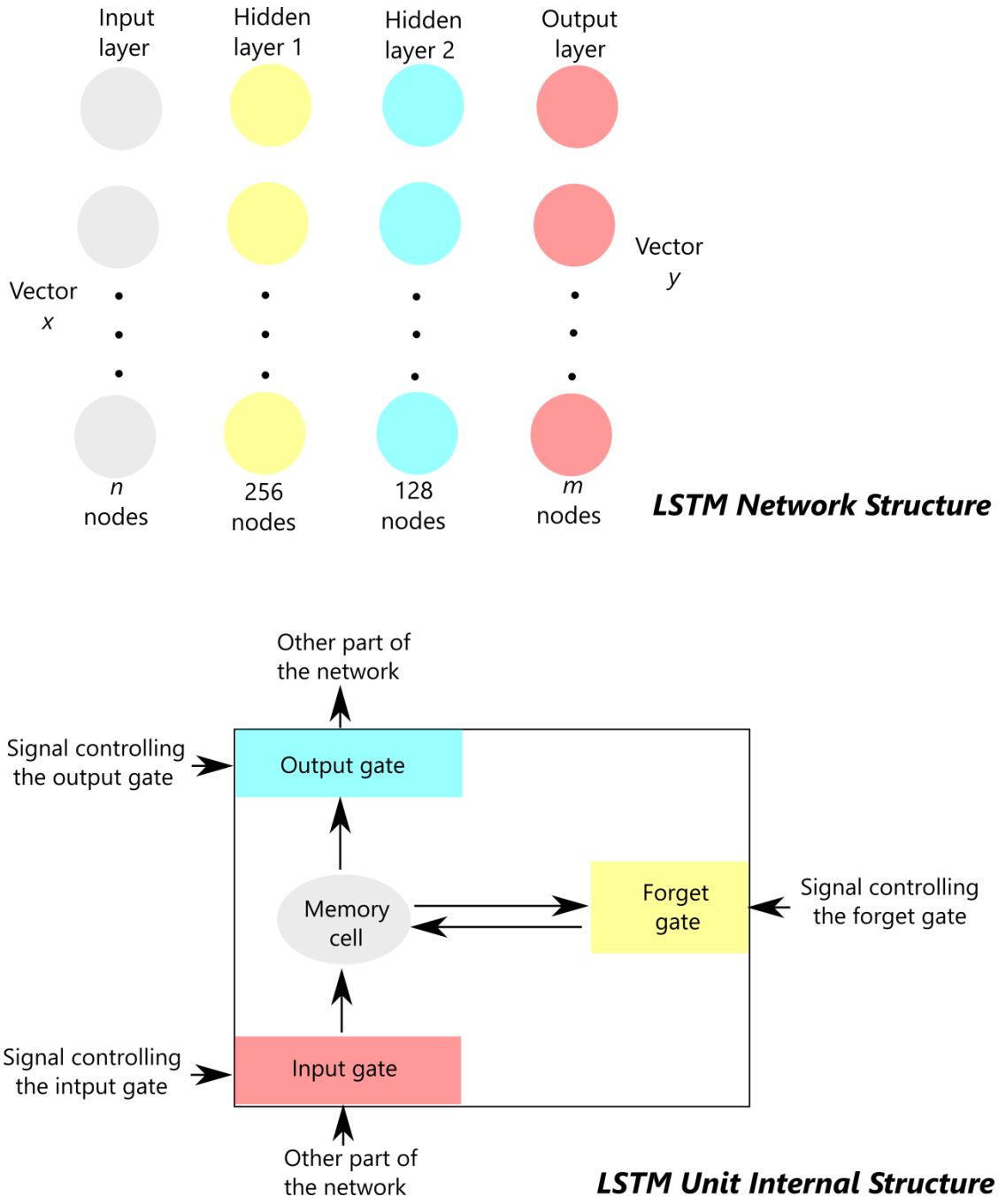


Figure 3.3: Long short-term memory network and unit structures (modified after Mei *et al.*, 2019)

3.5 Python Libraries used

Python 3.7.1 programming language was used to develop optimisation techniques or ML models (van Rossum, 1991). A more suitable Python distribution platform for ML is Anaconda 4.11 (Wang & Oliphant, 2012), which was utilised in this thesis. Libraries used to develop the models include Matplotlib 3.5.0, NumPy 1.19.2,

Pandas 1.3.1, TensorFlow 2.5.0, Keras 2.4.3, Scikit-learn 0.24.2, Seaborn 0.11.1, Statsmodels 0.12.2, Math 2021.2, Pickle 2.0.0, Flask 1.1.2, PyQt 5.9.2, and Pyinstaller 4.8. Spyder 5.1.5 was used as the integrated development environment software to write all the Python code of this thesis.

Matplotlib is a two-dimensional plotting library which produces quality figures in a variety of hardcopy format and interactive environments across platforms. NumPy provides a high performance multi-dimensional array object and tools for working with these arrays. Pandas is used for data structures and operations for manipulating numerical tables and time-series. TensorFlow is used as a deep learning application for fast numerical computing. Keras is a neural network library and acts as an interface for the TensorFlow library. Scikit-learn is used to build ML models and contains various ready-to-use algorithms such as classification, regression and clustering.

Seaborn is used for data visualisation and is based on the Matplotlib library. Statsmodels is a Python package that makes it possible for users to estimate statistical models, explore data and conduct statistical tests. The Math module enables users to make use of mathematical functions. Pickle is used in serialising and de-serialising a Python object structure, i.e. it converts a Python object into a byte stream to store it in a database. Flask is a Python module that enables users to develop web applications easily. PyQt is a Python binding for Qt designer, which consists of C++ libraries and development tools that include platform-independent abstractions for the GUI. Pyinstaller is a Python library that is used to convert Python files or projects into an executable file that can be used on machines without Python or Python’s set packages installed. Pyinstaller is used together with programmes that

convert applications to installers, e.g. in this thesis, the Python library was used together with the InstallForge programme.

The following chapter will explain the development and use of the *e*Metsi application. In this chapter, the developed application makes use of radio frequency identification technology (RFID) to manage mine water sampling results. *e*Metsi application utilises the information of technology (IoT) techniques to share sampling data in real-time with the use of near-field communication (NFC) mobile application and a website application acting as data cloud storage.

CHAPTER 4

4 *e*Metsi – RFID Controlled Mine Water Sample Management

4.1 Introduction

This chapter presents a framework for radio-frequency identification (RFID) technology in conjunction with mobile and website applications in mine water sampling (*e*Metsi – which is a Setswana Language meaning electronic water and is also an abbreviation for e-Tag based Mine Water Evaluation, Testing, Sampling and Identification Application). This application is the first of its kind in the mine water sector. RFID is the wireless contactless technology that uses radiated and reflected radio frequency waves to transfer data. Embedding items with RFID tags enables users to identify and track them (Finkenzeller, 2010; Shepard, 2005).

RFID is a disruptive technology in a sense that it replaced the two-dimensional barcodes (2D QR codes), and has substantially upgraded the automatic identification technology (Perret, 2014; Shepard, 2005). Data on the RFID tag can be read without line of sight, and some of them have a read range of more than 100 m depending on the RFID type (e.g. Qing & Chen, 2007; Sunny *et al.*, 2016; Xianming *et al.*, 2007). This technology improves every year and the cost of using it reduces drastically with massive improvements, thus gaining traction in different industries. Demiralp *et al.* (2012) shows that using data collection systems that are supported by RFID technology can reduce costs and save time.

RFID systems have several frequency bands, and the most used one are low frequency (LF), high frequency (HF) and ultra-high frequency (UHF). With the LF spectrum, the general frequency ranges from 30 – 300 kHz, the primary frequency

ranges from 125 – 134 kHz, and a read range is up to 10 cm (Shepard, 2005; Sunny *et al.*, 2016). It can be applied in sampling bottles, animal tracking, key fobs and access control (e.g. Chen *et al.*, 2011; Floyd, 2015; Hasanuzzaman *et al.*, 2013; Voulodimos *et al.*, 2010). The HF spectrum has a primary frequency range and a read range of 13.56 MHz and 0 – 30 cm respectively. It can be used in gaming chips, personal ID cards, library books and near field communication (NFC) applications (e.g. Chang *et al.*, 2010; Ching & Tai, 2009; Cho *et al.*, 2013; Silva-Pedroza *et al.*, 2017). UHF has two types of RFID, i.e. active and passive RFID. Active RFIDs are mainly applied in mining, vehicle tracking, construction, car manufacturing and asset tracking (e.g. Kelm *et al.*, 2013; Prasanna & Hemalatha, 2012). They have a read range of more than 100 m with a primary frequency range of 433 MHz (Goodrum *et al.*, 2006). Passive RFIDs are used in manufacturing, pharmaceuticals, electronic trolling, asset and supply chain tracking (e.g. Çakıcı *et al.*, 2011; Luvisi & Lorenzini, 2014; Mo *et al.*, 2009; Qing & Chen, 2007; Zhu *et al.*, 2012). They have a much lower reading range (0 – 16 m) when compared to active RFIDs, and their primary frequency range is between 860 and 960 MHz (Dobkin, 2012; Glover & Bhatt, 2006; Shepard, 2005; Sweeney, 2010).

This thesis presents the usage of NFC technology in mine water management. A typical NFC system is made up of an active NFC device, passive NFC device and a mobile application or a host computer (Figure 4.1; Coskun *et al.*, 2013; Silva-Pedroza *et al.*, 2017). A lot of industries have explored ways of embedding electronic identification tags on bottles for tracking, identification or both. The study explored the practice of incorporating NFC tags to sampling bottles, usage of a NFC mobile application for recording on-site parameters during sampling and display of this data on a website application. Merging of these technologies in mine water management

has never been used before, and this thesis presents the first of its kind. Parties that are involved in mine water sampling include the sampler, laboratory technician and end-user of the results. This development allows these three parties to interact with each other via interconnectivity and sharing of sampling data, time and results made possible via a mobile application that can be accessed anywhere by all of them.

Several programming languages were used to carry out this study including Extensible Mark-up Language (XML) and Java code for the Android mobile application, and Cascading Style Sheets 3 (CSS3) language, Hyper Text Mark-up Language 5 (HTML5) and JavaScript programming for the website application. The applications also required server provisioning and Structured Query Language (SQL) configuration to ensure maximal flexibility of the system. As this thesis describes the use of NFC technology, mobile and website applications in mine water sampling, it did not go into the details of the programming languages applied.

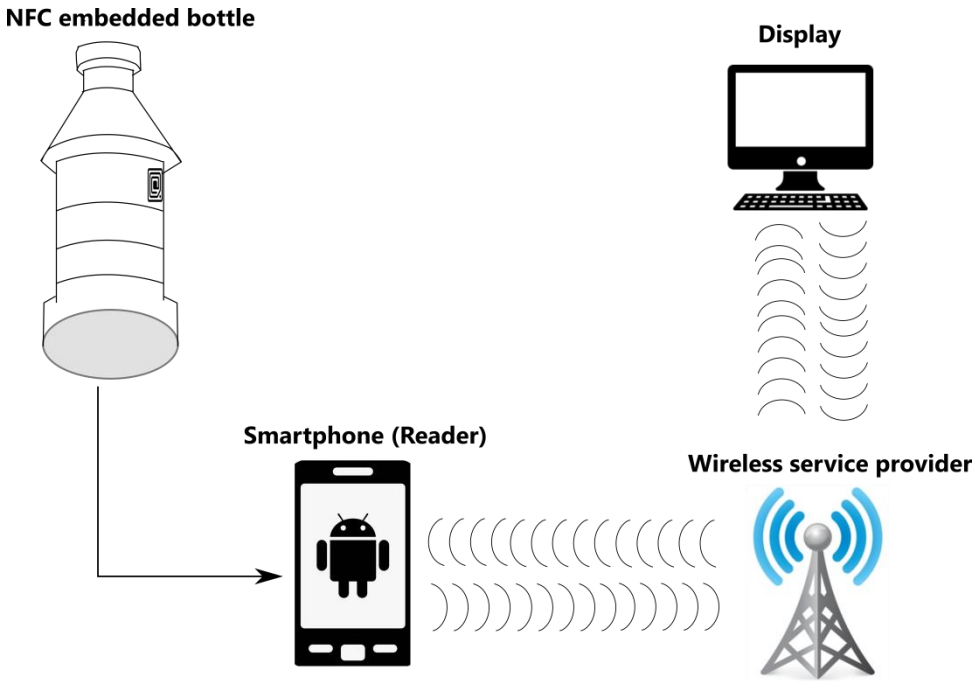


Figure 4.1: NFC application structure

4.2 From Barcodes to RFID-NFC tags

2D QR codes are widely used in many industries for tracking and identifying products. This item identifying and tracking style can be considered outdated technology and comes with its disadvantages. It is now commonly replaced by NFC which is part of the RFID community (Glover & Bhatt, 2006; Sweeney, 2010). NFC technology can be applied in anything that QR or bar codes or can; however, they are often more efficient. NFC can provide product information, track and also identify products faster than QR codes. During mine water sampling, 2D codes sometimes wear off on the bottle and it becomes difficult to scan or read them, whereas such occasions can rarely happen with NFC. QR or bar codes are sometimes difficult to scan as they require an unobstructed view of the code being scanned, a steady hand and sometimes it requires some time to scan them. This implies that codes cannot work in moving products such as vehicles, since it must be placed steady where scanning is practical. Therefore, QR or bar codes are associated with poor or very low scanning success and they are not always user friendly – it is now time to move forward with the NFC technology for these cases.

NFC technology is well received in the IoT and smart packaging industries. Nowadays, modern smartphones are embedded with the NFC readability feature, and they can read information of the products that have NFC microchips, which is a much faster and easier process than scanning the codes. NFC microchips also enable consumers to get more information about the product instantly and hassle free. Data generated during the food production phases are stored in the cloud storage and can be accessed by the public with the usage of NFC supported smartphones. This process is done to ensure that the public know the full details of the products they are buying from different food outlets (Pigini & Conti, 2017).

4.3 Evaluation of the study

Currently, when mine water samples are taken, there is no direct communication between the operator on the sampling site, the laboratory and the sampling institution. This slows down data exchange and can result in high response times to regulators or, in case of treatment plants, regulating the plant's parameters. In addition, samples can easily get mixed up when being taken from the sampling site to the laboratory and data could be misplaced or lost.

Mine water management in almost all operational mines consists of weaker optimisation techniques and does not generally take advantage of the power of digital technologies (More *et al.*, 2020). If applied accordingly, these technologies can reduce costs of treating mine water and save time. The technology of interest in this thesis is the RFID-NFC and was used as an identifier coupled with website application and storage location, and not solely for tracking purposes it is famously known for. Mine water samples that are taken to the laboratory are sometimes interchanged, resulting in confusion as the contents in the bottles are the same, i.e. they all contain mine water. Using NFC as an identifier helps to differentiate the samples: it helps in correlating the samples to its specific data. Specifically, *e*Metsi consists of identifying the samples electronically at the time of sampling, storing the on-site parameters and sample data, transferring the data to a cloud storage location, and allowing end-users to use the sample identifying data (Figure 4.2). The combination of these technologies closes a gap between the samplers, the laboratory and end-users of the results by communicating sampling times and locations, sampling parameters and results between the parties involved in the process.

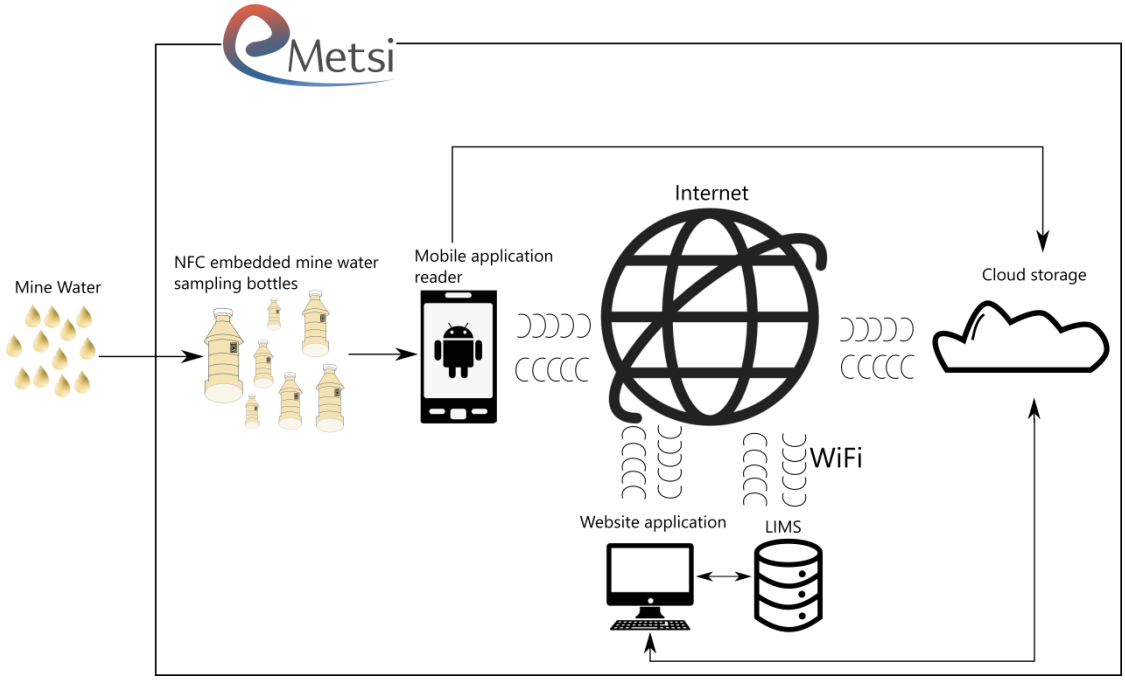


Figure 4.2: *e*Metsi – RFID controlled mine water sampling. LIMS: Laboratory Information Management Systems

4.4 Proposed solution and study design

4.4.1 NFC embedded sampling bottles

An NFC microchip was incorporated on mine water sampling bottles used during the testing sampling process. The NFC microchips used in this study are the HF spectrum chips with a primary frequency range of 13.56 MHz, communication distance of up to 30 mm, and have a circular shape with a diameter of 10 mm (Microsensys GmbH, Erfurt, Germany; Figure 4.3). This is advantageous because it is easy to differentiate the sample bottles, i.e. multiple sampling bottles of different volumes can be used and applying this technique will make it easy to identify them at a later stage.

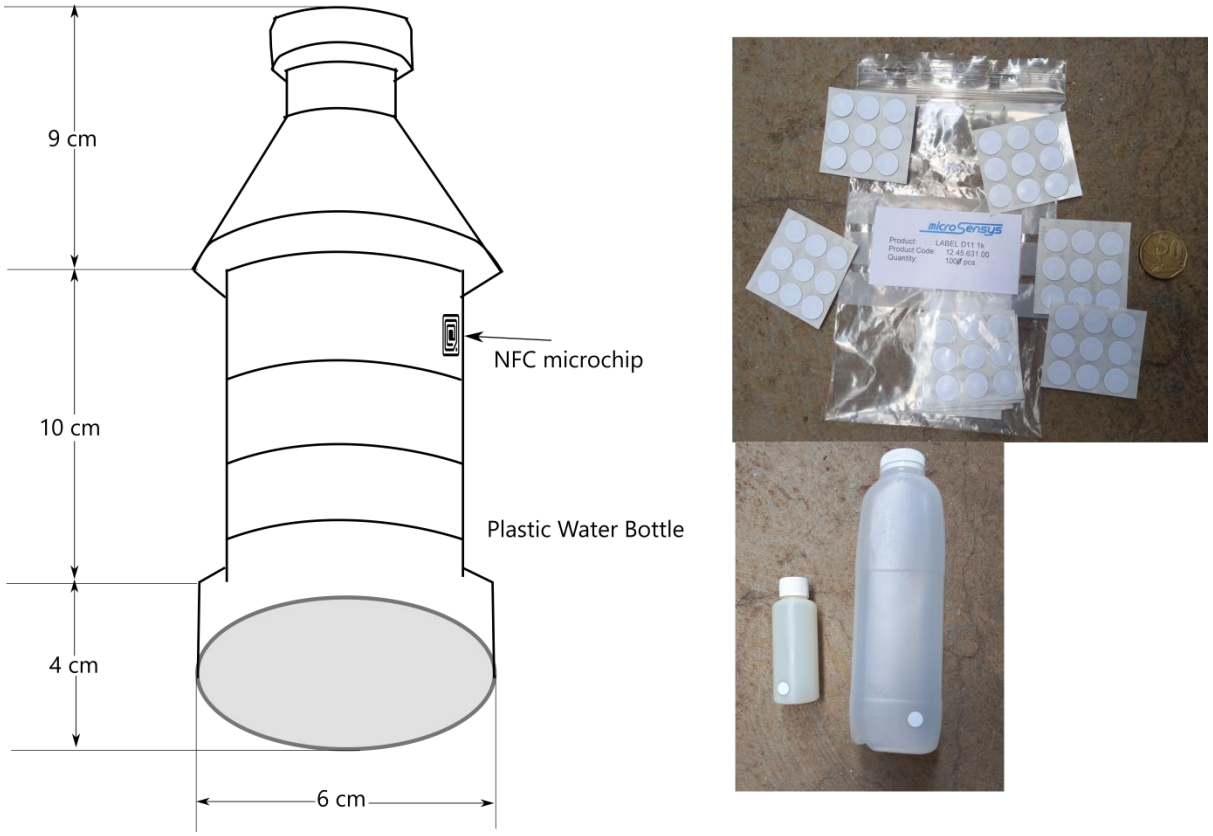


Figure 4.3: Water sampling bottle (not to scale) incorporated with a NFC microchip and NFC tags used in this study

4.4.2 What problem does *e*Metsi solve?

*e*Metsi is advantageous as it allows identifying samples at the sampling site and the laboratory, thus reducing errors in exchanging data relevant for the sample and laboratory results. As described earlier, using barcoded sample containers for sampling is challenging as samples and sampling locations might be mixed up or labels become unreadable. Additionally, it might be time consuming and inconvenient recording all sample data manually (e.g. on sampling sheets, field books, note books, field computers) or without using a database (e.g. cloud based). Therefore, *e*Metsi will ensure that sample parameters are recorded and accessed using mobile applications.

*e*Metsi is not solely restricted to mine water samples but can be expanded to general samples taken in various environments (e.g. manufacturing, extra-terrestrial expeditions, life science, agriculture) including the communication between the various users. Sample data can be exchanged electronically via wireless communication between the users and the electronic storage location. *e*Metsi also comes with the advantage such as the ability to store various types of data in different storage locations of the website application. The data include physico-chemical parameters (e.g. water temperature, oxygen saturation, pH, oxygen concentration, redox potential, electrical conductivity) written into the memory at various sampling location points, providing users with detailed and readily accessible sample data.

4.5 NFC mobile application and its screens

*e*Metsi mobile application works on NFC supported smartphones on the Android platform. It can be used by anyone involved in a sampling project – in this case, the sampler, laboratory technician and end-user of the results, and requires login details. The mobile application has several functionalities such as allowing the sampler and laboratory technician to write data back to the database. It also enables the end-user to only view the data that is on the database. Furthermore, the application can be used offline, but a user management system is required. It is also connected to a website application with data that is added using the mobile application being pushed to the server and it can be viewed using the website application.

*e*Metsi mobile application has a satellite based location feature (e.g. GPS) and an Open Street Maps functionality; therefore, the user is able to pin the sampling location for every sampling point. It is developed using XML for defining layouts and Java code to provide processing logic. Mixture of XML and Java is widely used for

Android application development. Layouts only declare the appearance of the application which is carried out by XML. To define what the application must do, Java code can be used. *e*Metsi mobile application enables the user to navigate through multiple screens while recording and storing sampling data. The following is a brief explanation of the mobile application screens:

Login

On the login screen (Figure 4.4 A), the user is able to login as one of the sampling project users. The login details can be generated by the administrator.

Start

After logging in, a new screen (Figure 4.4 B) shows up. Tapping on “start” allows the user to start adding or reading a tag. This can also be done offline and the application will save the data on the device as a CSV file and push it to the server as soon as the phone is connected to the internet.

Main Menu

The “main menu” screen (Figure 4.4 C) gives the user three options – “add tag”, “view all data” and “saved locations”. The “add tag” option allows the user to scan a tag and add information, but data will not be available until administrator user approves; “view all data” option shows the samples that have been saved (Figure 4.4 H), and full information of these samples can be viewed on the website; “saved locations” option allows the user to view (Figure 4.4 K) or add possible locations (Figure 4.4 J).

Ready to scan

The “ready to scan” screen (Figure 4.4 D) shows up after tapping “add tag”. This is where the user will scan a tag.

Add New

On the “add new” screens (Figure 4.4 E & F), after scanning a tag, the user can add on-site parameters during mine water sampling. When the user taps on “submit”, data will be saved and made available on the website application. The user can also select saved sampling locations on this screen (Figure 4.4 G).

Locations

After tapping “saved locations” on screen (C), “view locations” screen (Figure 4.4 I) will appear. On this screen, the user is able to see all the sampling locations, and can also add new locations by selecting the “+” option shown at the bottom right of screen (I). The “+” option will take the user to screen (J) on which they can select their desired location, i.e. sampling location by using the “drag and drop” option provided. Selecting the locations on screen (I) will take the user to screen (K) to view the selected location on a map.

Logout

Screen (C) has a logout option on the top right position, indicated by a small arrow. Tapping on this option will give the user a chance to logout of the application with the reading on the screen showing the writing “are you sure you want to logout?” (Figure 4.4 L).

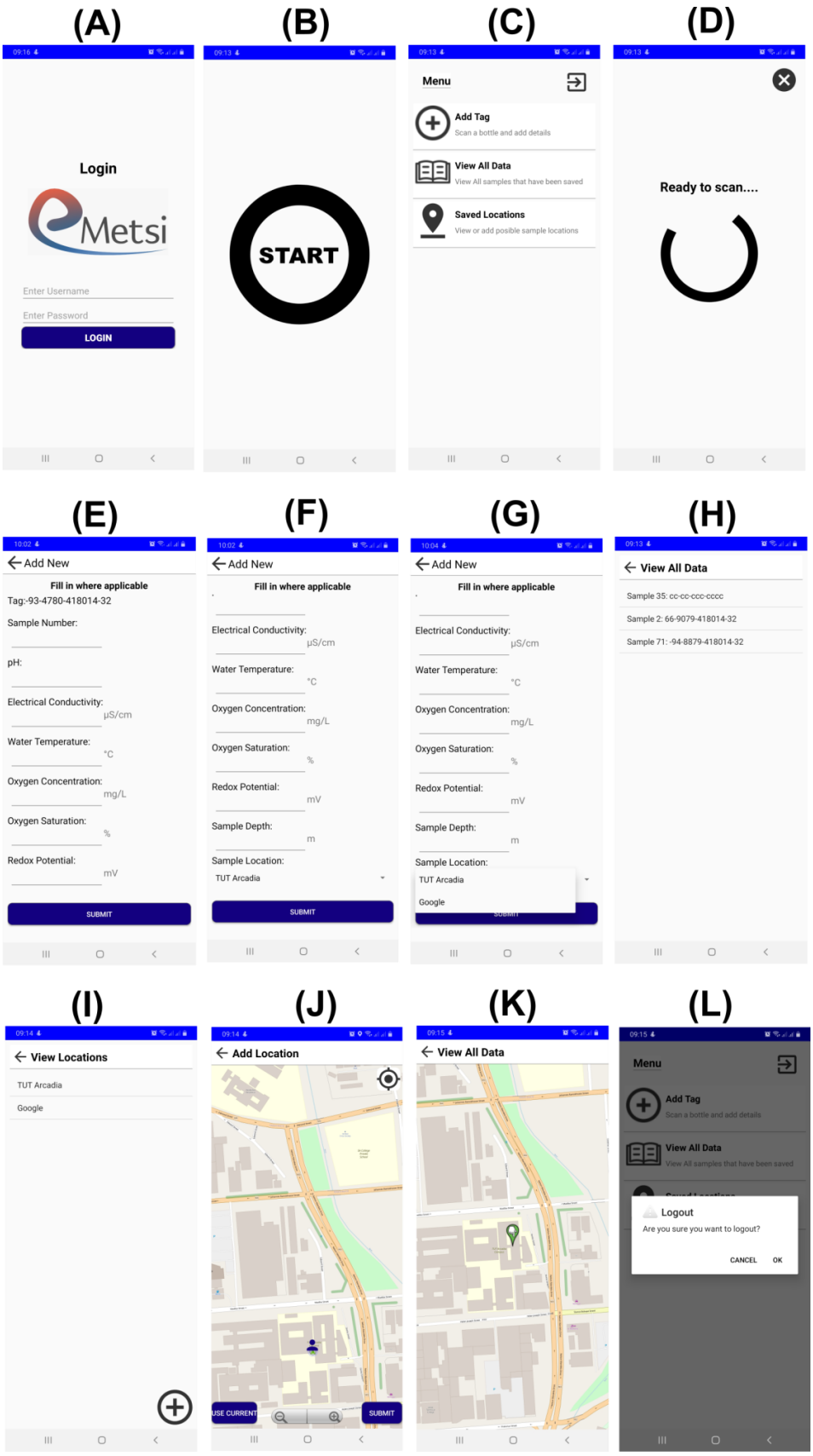


Figure 4.4: *e*Metsi NFC mobile application screen pages. Details in the text

4.6 Website application and its pages

The website application can be accessed via an internet browser using an electronic device (e.g. computer). It was created using CSS3 to add the look and style in webpages, HTML5 to structure the webpages and their contents, and JavaScript programming which enables users to interact with webpages. Data stored on the mobile application is pushed to the server and can be viewed on the website application. Applying a user management, the application can be used by five different user types, having access to different webpages with different functionalities (Table 4.1).

Table 4.1: The different webpages with their functionalities for each user

User Type	Webpage Types	Functionality
Administrator	Data	Read-only all data
	Available parameters	Read/write parameter used for analysis
	Add new request parameter	Read/write parameters to be analysed
	View parameters	Read-only parameters
	Users	Read/write users
	Add new user	Read/write users
	Locations	Read/write locations
	Add new location	Read/write locations
Leader	Data	Read-only all data
	Requested parameters	Read/write parameters to be analysed
	Available parameters	Read/write parameters used for analysis
	New parameter request	Read/write parameters to be analysed
	View requests	View parameters to be analysed
	View parameters	Read-only parameters
Sampler	Data	Read/write sample data
	Locations	Read/write locations
	View parameters	Read-only parameters
Worker	Data	Read-only sample data
	View parameters	Read/write parameters
Client	Data	Read-only all data
	View parameters	Read-only parameters

At the beginning, the website requires login details which can be created by the administrator for everyone using it. The administrator controls the data that goes into the database, the locations and other users (Figure 4.5). Project leaders can add or view parameters that need to be analysed (Figure 4.6). Sampling persons in the field can view or add on-site data and locations (Figure 4.7), while the project worker and

the client have limited functions with the worker only being able to view sample data and view or add parameters (Figure 4.8), and the client can only see each data point and where the data came from (Figure 4.9). Project workers and the client can only see the data that has been approved by the administrator.

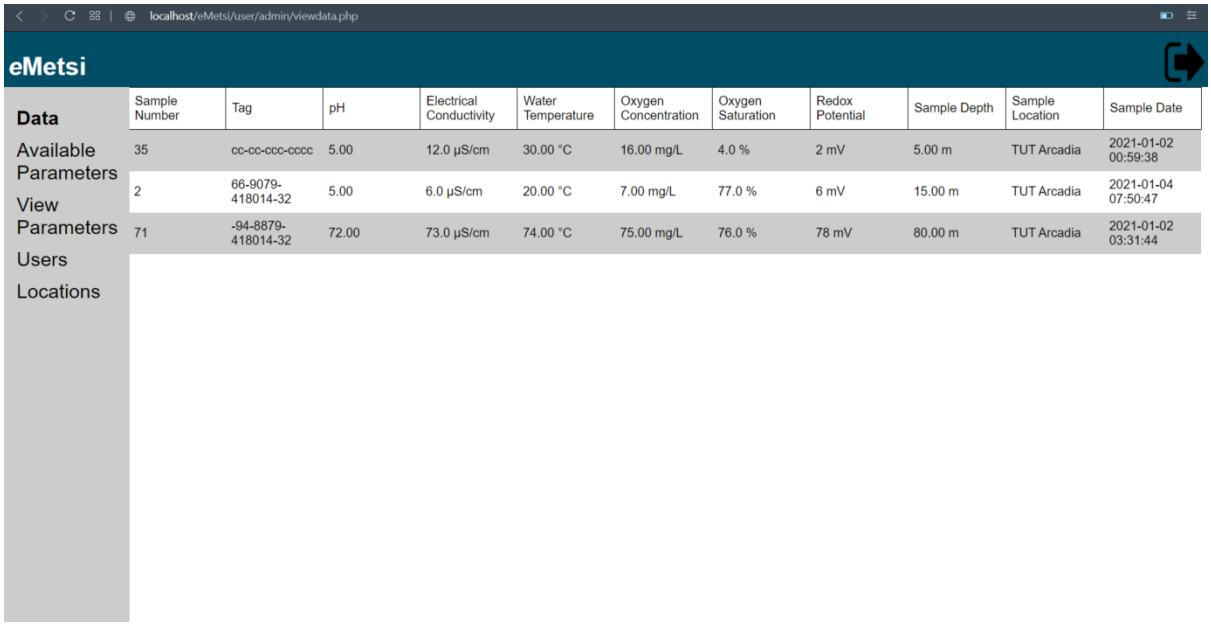


Figure 4.5: The administrator's webpages (example data was used)

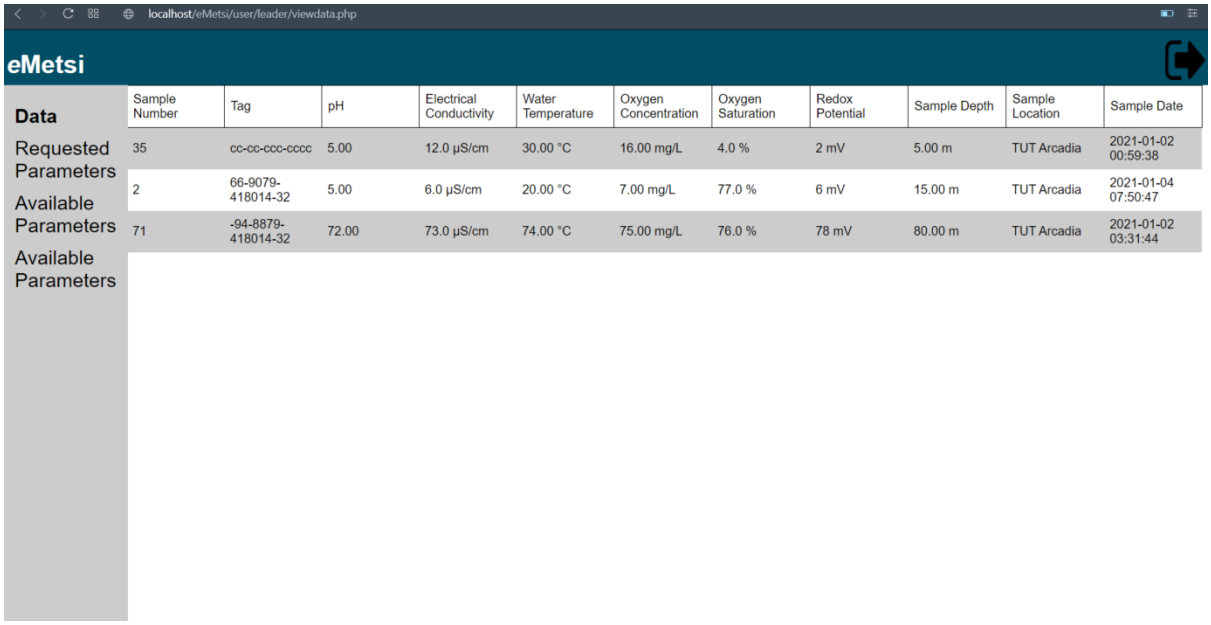


Figure 4.6: The project leader's webpages (example data was used)

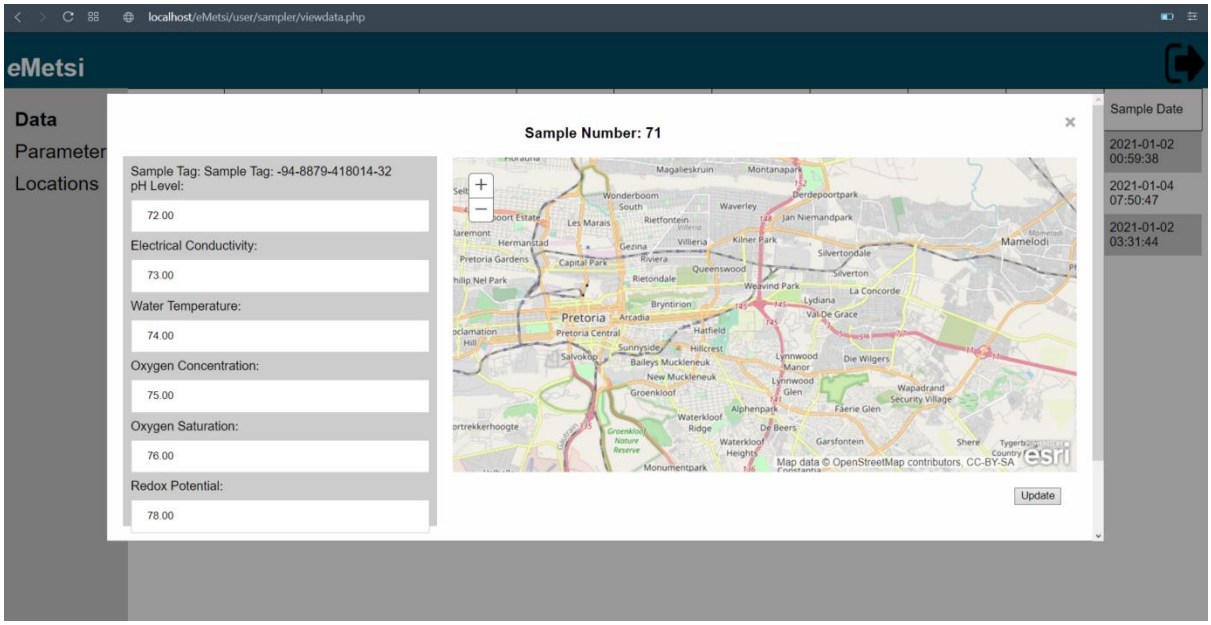
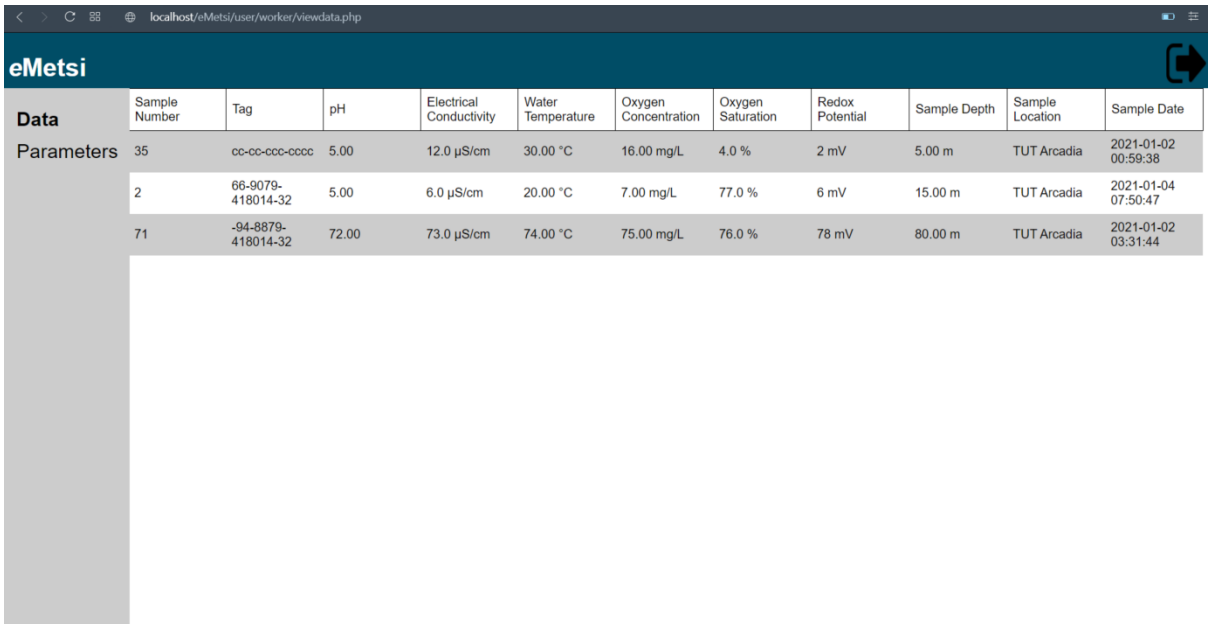


Figure 4.7: The sampler's webpages (example data was used)



Sample Number	Tag	pH	Electrical Conductivity	Water Temperature	Oxygen Concentration	Oxygen Saturation	Redox Potential	Sample Depth	Sample Location	Sample Date
35	cc-cc-ccc-cccc	5.00	12.0 µS/cm	30.00 °C	16.00 mg/L	4.0 %	2 mV	5.00 m	TUT Arcadia	2021-01-02 00:59:38
2	66-9079-418014-32	5.00	6.0 µS/cm	20.00 °C	7.00 mg/L	77.0 %	6 mV	15.00 m	TUT Arcadia	2021-01-04 07:50:47
71	-94-8879-418014-32	72.00	73.0 µS/cm	74.00 °C	75.00 mg/L	76.0 %	78 mV	80.00 m	TUT Arcadia	2021-01-02 03:31:44

Figure 4.8: The project worker's webpages (example data was used)

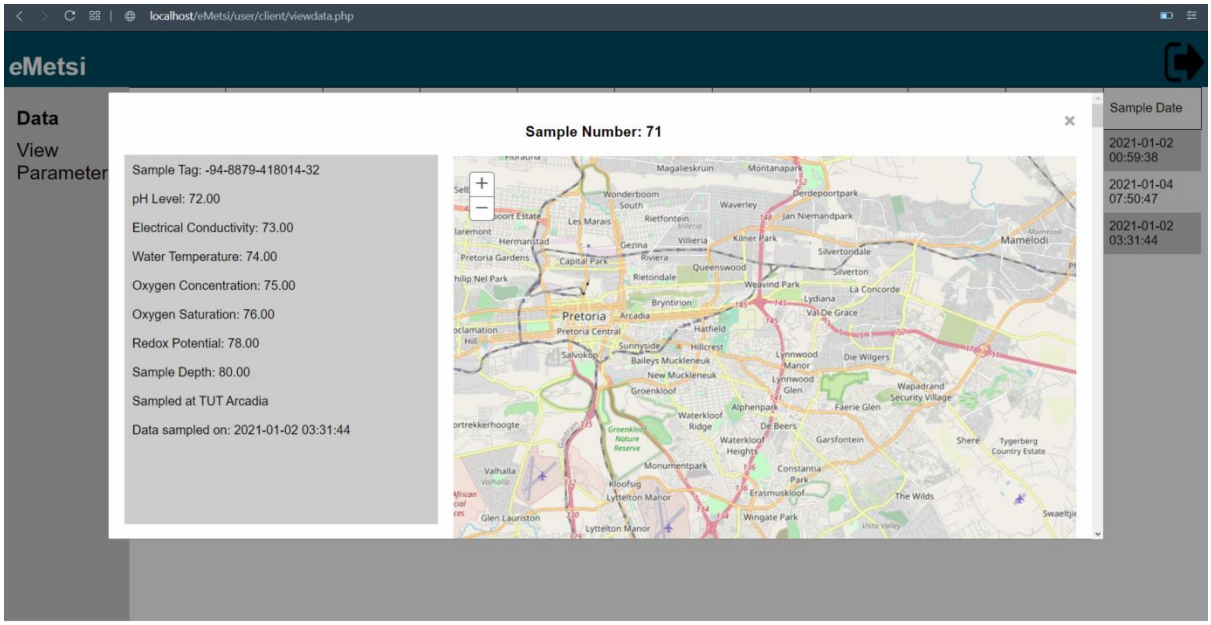


Figure 4.9: The client's webpages (example data was used)

4.7 Validation and discussion

Research and technology are two unique sections that support each other. In the context of using technology to simplify research, relevant technological inclusion can help make research investigation much easier. For example, the inclusion of NFC tags in the sampling bottles makes it easy for laboratory technicians to identify the bottles and simply correlate the on-site data to its specific bottle. Utilisation of this android-based application resulted in positive qualitative data, and was highly convenient as opposed to manually recording everything on the field book or using a handheld scanner. This application development is one solution to contribute to modernising mine water management.

The combination of NFC technology with mine water sampling will be advantageous for many reasons, where bulk sampling is at the centre stage. From the sampler's point of view, the technology allows for the optimisation of time. This can be observed in the field during sampling where the sampler uses the mobile application to store on-site data. Data communication between the sampler, laboratory

technician and the end-user is simplified as data is easily extracted from the server through the developed website application.

In the next chapter, techniques that can be applied in time series data for handling missing data and properly detect anomalies will be discussed. Some of the techniques that will be discussed in the next chapter are applied in data analysis processes in the following chapters of this thesis. The purpose of the next chapter is to introduce a new focus point of the thesis, i.e. data science, which forms a large part of the Internet of Mine Water.

CHAPTER 5

5 Review of Some of the Techniques for Missing Data Interpolation and Anomaly Detection in Time Series

5.1 Summary

Time series are a set of ordered data that are observed at successive points in time. In this type of dataset, time is the dependent variable, while the *y*-axis represents a dependent collection of observations (e.g. Chatfield, 2000; Chen *et al.*, 2005; Reich *et al.*, 2016). Time series data is useful as it helps in understanding the past and can be used to forecast the future. Yet, any missing observations and anomalies in time series data can negatively affect interpretations if not dealt with appropriately. Examples of time series can be found in a variety of disciplines, such as science or economics. Measurements or readings in time series are records taken through time and missing observations or anomalies are quite common. This may be due to equipment malfunctioning, a mistake, personnel not working, or halting of operations. When a time series contains multiple missing values and anomalies, it may be necessary to apply appropriate techniques to interpolate the missing values and detect anomalies. This thesis presents a mixture of useful old and new techniques that can be applied to time series for interpolating the missing values and detecting anomalies.

5.2 Missing Data

Time series measurements are conducted at various times under different conditions, and sometimes missing data occur due to problems that are known as the missingness mechanism (Fielding *et al.*, 2009; Little *et al.*, 2014; Newman, 2014). Missingness mechanism can be in three different forms: missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR)

(Figure 5.1). A variable is MCAR if the probability of missingness is the same for all units, i.e. there is no dependencies of the missingness probability related to the variable itself — *p*(missing) is unrelated to all variables, observed and unobserved (equation 5.1). For example, MCAR can occur due to equipment not working properly or due to sensor data recording failure (e.g. Bähr *et al.*, 2020). A variable is MAR if the probability of missingness is depending only on available information — *p*(missing) is only related to observed data (equation 5.2). For example, this can occur when certain days are skipped in a survey or during sampling. For the MNAR, *p*(missing) is only related to the unobserved missing data (equation 5.3). The missing data in MNAR are incomplete data that cannot be verified or predicted.

$$p(\text{missing}|\text{complete data}) = p(\text{missing}) \quad 5.1$$

$$p(\text{missing}|\text{complete data}) = p(\text{missing}|\text{observed data}) \quad 5.2$$

$$p(\text{missing}|\text{complete data}) \neq p(\text{missing}|\text{observed data}) \quad 5.3$$

Several methods have been introduced to replace the missing observations. It is common practice to either impute or interpolate the missing values. Imputation is the process of using observations from the entire dataset to fill in the missing values, while interpolation refers to using neighbouring data points to estimate the missing values. In this thesis, an interpolation process will be explored. This thesis looked into some of the deterministic and stochastic models to interpolate the missing data.

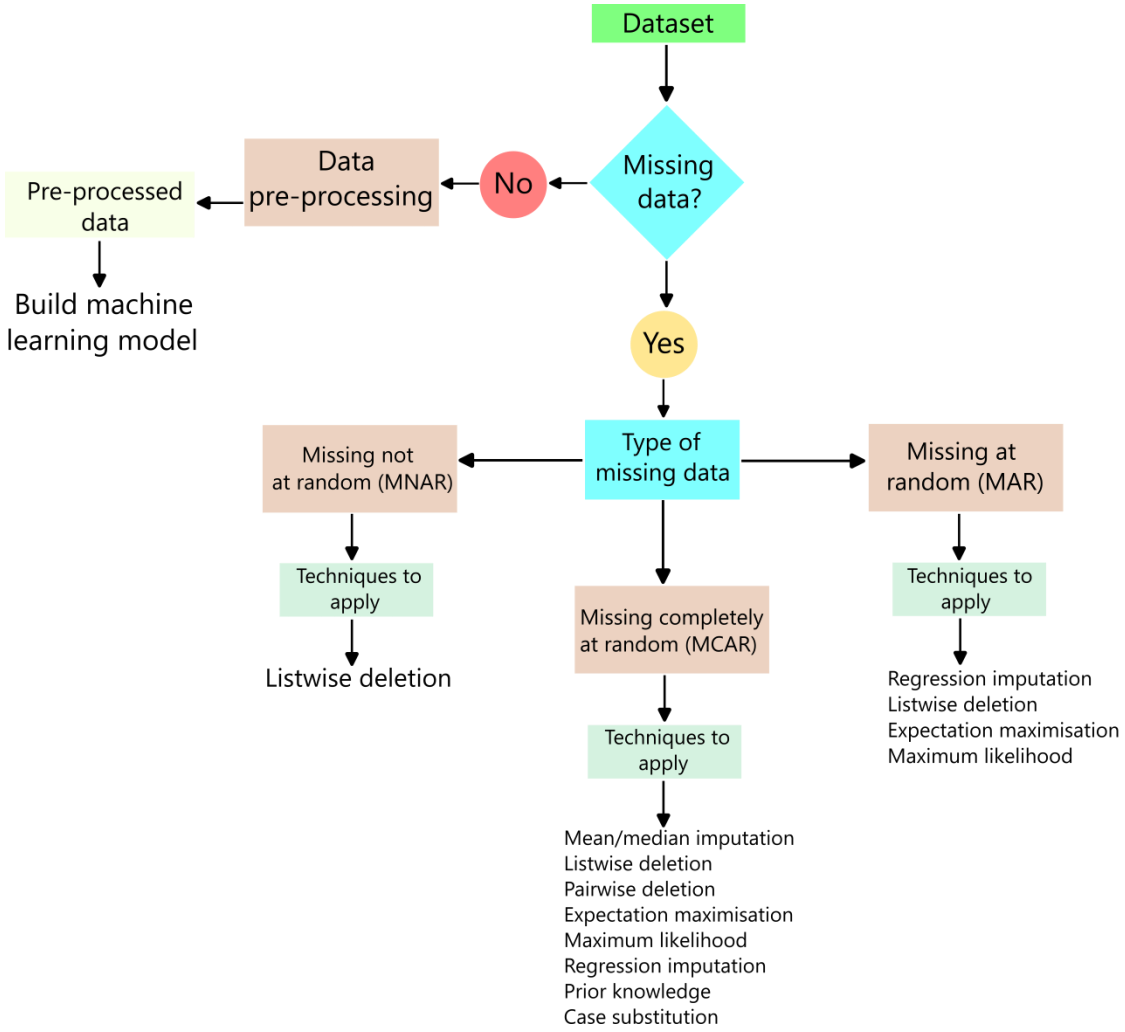


Figure 5.1: Recommended techniques for each missing data type (modified and supplemented after Houari *et al.*, 2014)

5.3 Deterministic Modelling

5.3.1 Background

Deterministic modelling, which can also be referred to as numerical analysis modelling, is an approach that assumes that the time series data corresponds to an unknown function and the main aim is to fit the function in the correct way (Wackerly *et al.*, 2014). The historical time series data is used to derive a best fit function, and the function is used to interpolate the missing values (Figure 5.2).

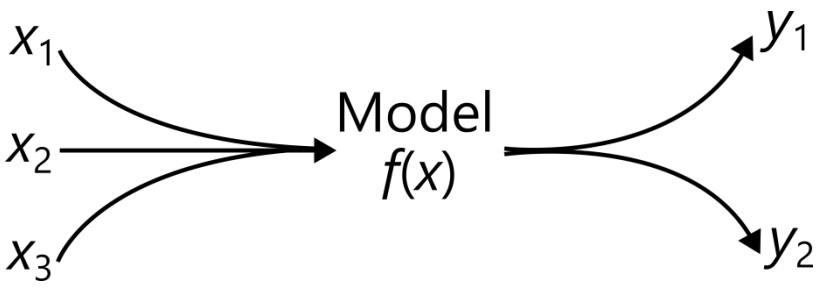


Figure 5.2: Function *f(x)* derived from the historical data x_n

In this approach, the outcomes are determined through known relationships among state and events, without room for random variation. Similar to a chemical reaction, a given input will always produce the same output. These models do not make room for error as they deal with systematic and definitive outcomes as opposed to random results. Through this approach, predictions can be computed and examine the relationship between variables. For example, if it is believed that *y* (mine water iron concentration) will be exactly two times *x* (mine water pH), then $y = 2x$. This means that the iron concentration can always be determined exactly when pH is known. The deterministic model prediction is a hypothetical "what-if" statement which helps in identifying the outcome using a particular *x*. This approach assumes certainty in its solution.

5.3.2 Basis spline (B-Spline)

Basis spline (B-Spline) is an interpolation technique in which the order chosen for the B-Spline curve is independent of the control points (Averbuch *et al.*, 2014; Kineri *et al.*, 2012). Control points are used to determine the shape of a spline curve. B-Splines are unique and more advantageous when compared to other interpolation or approximation techniques such as splines or Bezier curves (Baydas & Karakas, 2019). Splines follow the general shape of the curve, while the Bezier curve generally follows the shape of a defining polygon (Averbuch *et al.*, 2014; Baydas & Karakas, 2019; Han *et al.*, 2008). In the B-Spline curve, the independency of the

order of the curve over the local control points makes it a reliable interpolation technique. Thus, the B-Spline curve provides local control through control points in every segment of the curve (Hoffmann & Juhász, 2001; Kineri *et al.*, 2012). For example, a curve may have 40, 60 or 100 points and there will still be freedom to fix the curve to any shape of e.g. quadratic, cubic or higher order.

As can be seen, a B-Spline curve is not a single curve, but it is made up of a number of curve segments and all have the same continuity requirement depending on the order of the curve (Averbuch *et al.*, 2014; Elbanhawi *et al.*, 2015). B-Splines can be applied for the open and closed curves, and changing any of the control points changes only a specific segment of the curve (Figure 5.3), while in techniques like Bezier, the whole curve changes. Therefore, applying B-Spline interpolation in time series ensures that the missing data are interpolated within the population space and are not distorted or isolated.

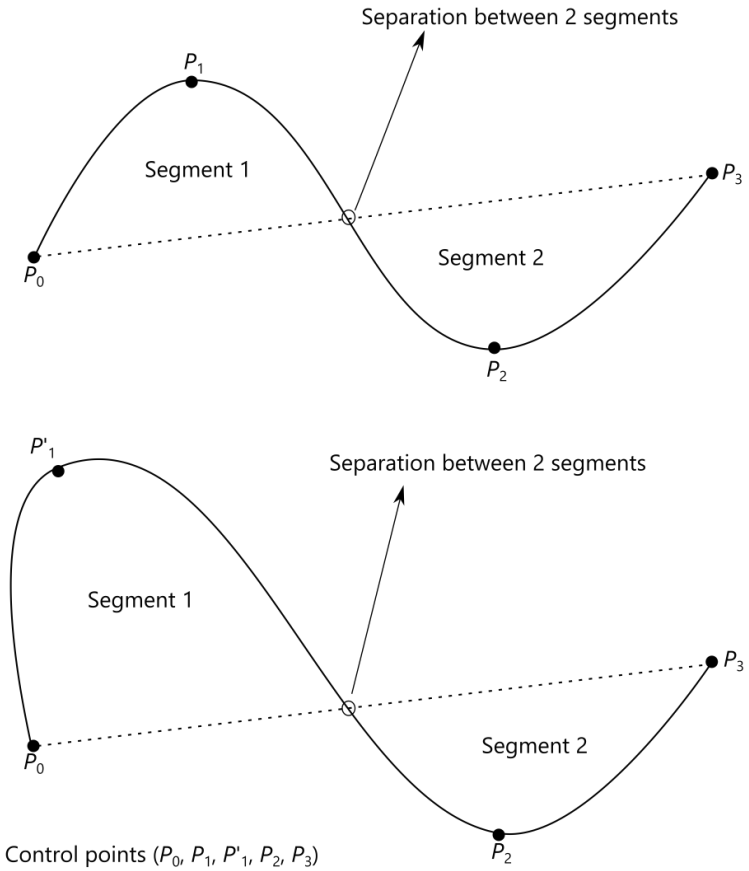


Figure 5.3: B-Spline curve behaviour when changing a control point position. Changing the position of control point P_1 only changed segment 1, while segment 2 remained intact

5.3.3 Non-uniform rational basis spline (NURBS)

Non-uniform rational basis spline (NURBS) curves are basically a modification or rational version of B-Spline curves. Their advantage is that they can create smoother surfaces with fewer control points (Shao & Xiao, 2011). NURBS are mathematical representations of complex structures, be it two- or three-dimensional objects, e.g. cars, buildings, cones or simple curves. They are the computer graphics design industry standard when it comes to creating or interpolating complex objects (Rogers, 2001; Sevilla *et al.*, 2008; Shao & Xiao, 2011). Generally, they work exceptionally well in three-dimensional modelling, enabling the designer to easily manipulate control points and the contours' smoothness.

NURBS are not necessarily dependent on a lot data; indeed it takes little data to define them. They are highly influenced by control points and weights (Rogers, 2001). The non-uniform in NURBS refers to the idea that some segments or sections of a defined shape can be easily manipulated relative to other sections of the overall shape with control points being associated with weights (positive numbers). When these control points all have the same weight, the curve is called non-rational (Boudjemaa *et al.*, 2003; Hoffmann & Juhász, 2001; Rogers, 2001) and the rationality of NURBS means that the curves have the possibility of being rational, i.e. NURBS have the ability to give more weight to the control points in the overall curve shape. Mathematically, NURBS can be defined as follows:

$$C(U)=\frac{1}{\sum_{i=0}^{n} N_{i,p}(U)w_i}\sum_{i=0}^{n} N_{i,k}(U)w_iP_i \qquad 5.4$$

This NURBS curve, $C(U)$, of the degree k is defined by control points P_0, P_1, …, P_n, knot vector $U = u_0, u_1, \ldots, u_m$, and weights w_0, w_1, …, w_n. The $N_{i,k}$ is the normalised B-Spline function of degree k defined by:

$$N_{i,k}(U)=\frac{u-t_i}{t_{i+k}-t_i}N_{i,k-1}(U)+\frac{t_{i+k+1}-u}{t_{i+k+1}-t_{i+1}}N_{i+1,k-1}(U) \qquad 5.5$$

and

$$N_{i,0}(U)=\begin{cases}1, \text{ if } t_i \le u < t_{i+1}\\ 0, \text{ else}\end{cases} \qquad 5.6$$

Where t_i are the knots forming a knot vector $u = t_0, t_1, \ldots, t_m$.

5.3.4 Wavelet transform

Wavelet transform is a technique derived from the Fourier Transform (FT) and Short-Time Fourier Transform (STFT). Limitations that arise from the FT and STFT techniques basically gave birth to the wavelet transform (Averbuch *et al.*, 2014;

Torrence & Compo, 1998). The FT provides frequency information of a signal that represents frequencies and their magnitude. However, it does not tell when in time these frequency components exist (Hansen, 2014; Serov, 2017). Therefore, it is ideal for signals that do not change with time, i.e. signals that have a constant frequency throughout. Consequently, FT's disadvantage is that it lacks capability to provide frequency information for a localised signal region in time. STFT was therefore developed to overcome the poor time resolution of the FT. Thus, STFT explains the time frequency representation of the signal and it assumes that a certain portion of the non-stationary signal is stationary (Baba, 2012; Veer & Agarwal, 2015).

The main limitation of STFT is that high frequency components appear as short bursts, thus needing higher time resolution (Boulet, 2006; Hon, 2013). So, the wavelet transform improves on this shortfall, i.e. it results in analysing a signal into different frequencies at different resolutions. Therefore, a wavelet is a rapidly decaying wave-like oscillation that has zero mean and exists for a finite duration (Averbuch *et al.*, 2014; Bolton *et al.*, 1995; Graps, 1995; Lord *et al.*, 2000; Qi *et al.*, 2018; Torrence & Compo, 1998; Veer & Agarwal, 2015). Wavelet interpolation is a key factor in time series data due to the data's stationarity properties. It is advantageous to apply this technique when the function is not smooth and is oscillating.

PyWavelets, an open source wavelet transform software for Python, can be used to interpolate the missing values for oscillating data (Lee *et al.*, 2019). PyWavelets combines a simple high level interface with low level C and Cython performance. Python consists of different types of wavelet families, e.g. Haar, Morlet, Gaussian or Frequency B-Spline wavelets. Therefore, for accurate results, a wavelet family which

fits best with the supplied data must be chosen. Each wavelet family is useful for a different purpose as they have different smoothness, shape and compactness.

5.4 Stochastic Modelling

5.4.1 Background

Stochastic modelling which can also be referred to as probabilistic modelling, is a mathematical representation of a random phenomenon (Olofsson & Andersson, 2012; Palma, 2016). This approach is based on statistical concepts and principles and is flexible in modelling historical time series data with its patterns and trends (Paolella, 2018). Stochastic modelling is also defined as statistical analysis tool that estimates, on the basis of historical data, the probability of an event occurring again. Stochastic models have some level of randomness, i.e. even with the same initial conditions, the model is bound to produce different results. The stochastic model includes both the deterministic component and a random error component. Thus, for example, *y* (iron concentration) will be two times *x* (pH) plus a random error, i.e. $y = 2x$ + random error. This thesis gives a brief overview of the probabilistic models that can be applied for data interpolation in time series. These models include autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA).

5.4.2 Autoregressive models (AR)

AR models forecast a series based solely on the past values of the series referred to as lags. A model that only depends on the previous lag is referred to as AR model of order one (Liu & Fomel, 2011; Oudre, 2018):

$$y_t = \omega + \phi y_{t-1} + e_t \qquad 5.7$$

where y_t is the target variable, y_{t-1} is the lagged target, e_t is the error of the model, ω and ϕ are the intercept and coefficient, respectively.

The lagged target of the AR model occurs for every single point in the series, i.e. the recursion in time goes back until the beginning of the series, and these can be referred to as long memory models (Oudre, 2018):

$$y_{t-1} = \omega + \phi y_{t-2} + e_{t-1} \quad 5.8$$

$$y_{t-2} = \omega + \phi y_{t-3} + e_{t-2}$$

$$\vdots \quad = \quad \vdots \quad + \quad \vdots + \quad \vdots$$

$$y_{t-n} = \omega + \phi y_{t-n} + e_{t-n}$$

An AR model can have an order of more than one, i.e. the lookback period can be two, three, four or more. A time series that is a linear function of n past values plus an error is referred to as an autoregressive process of order n:

$$y_t = \omega + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_n y_{t-n} + e_t \quad 5.9$$

AR model can be used to find a line of best fit for a series, and ultimately be used to interpolate data (e.g. Liu & Fomel, 2011; Oudre, 2018).

5.4.3 Moving average models (MA)

Moving average (MA) models are generally applied to stationary time series. They forecast a series based only on the past errors in a regression-like model which are referred to as error lags. MA models depend on the errors in the series that exist over time. A MA model that depends only on one lag of error in the past can be mathematically described as:

$$y_t = \omega + \phi e_{t-1} + e_t \quad 5.10$$

The error of the MA is computed by finding the difference between the predicted and actual observation (Figure 5.4; Saeipourdizaj *et al.*, 2021).

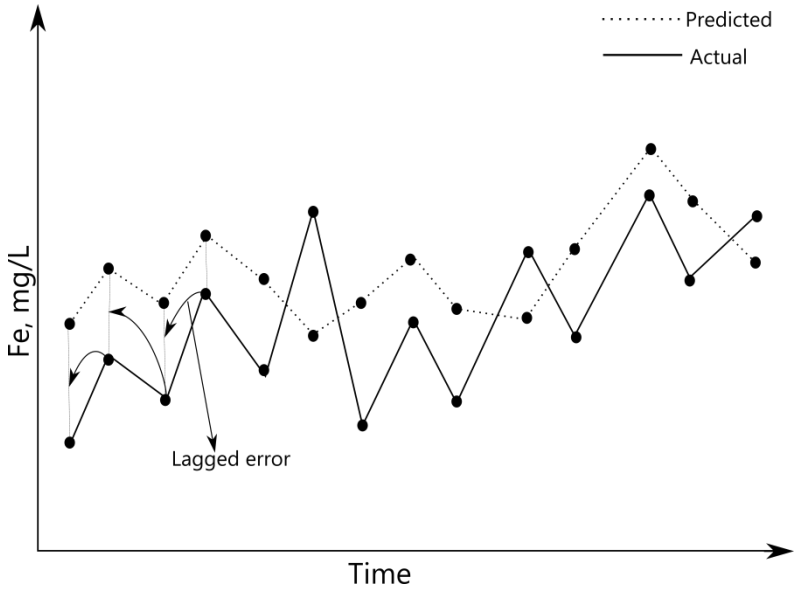


Figure 5.4: Moving average model mechanism

5.4.4 Autoregressive (integrated) moving average models (ARMA/ARIMA)

ARMA is a technique that combines AR and MA, and the model order must first be computed before a regression line can be fitted. The order of ARMA is known as ARMA(*p*, *q*), where *p* is related to the AR model, and it refers to the number of time lags, and *q* refers to the order of the MA model (Paolella, 2018). In the ARMA model, the influences of the previous lags together with the lag errors are considered when forecasting the future values in a series. Thus, ϕ_{AR} represents the coefficients of the AR model and ϕ_{MA} represents the MA model coefficients. For example, if an ARMA model is computed with only one lag for the AR model being used and only the first lag error is considered for MA model, then the order of the ARMA model will be ARMA(1,1). An ARMA model is a stationary model, and the AR enables it to make predictions using previous values of the dependent variables, while MA enables the ARMA model to make predictions using the series mean and previous errors (Olofsson & Andersson, 2012; Palma, 2016; Paolella, 2018; Wackerly *et al.*, 2014):

$$y_t = \omega + \phi_{AR1}y_{t-1} + \phi_{AR2}y_{t-2} + \ldots + \phi_{ARn}y_{t-n} + \phi_{MA1}e_{t-1} + \ldots + \phi_{MAn}e_{t-n} + e_t \quad 5.11$$

ARIMA approach is similar to ARMA except for the differencing part. In statistics, differencing refers to when data has one less data point than the original dataset (Paolella, 2018; Wackerly *et al.*, 2014). The ARIMA model is known with the order of ARIMA(*p*, *d*, *q*), where *p* and *q* are similar to ARMA order, and *d* is the number of differencing required to make the time series stationary. An ARIMA model with the order ARIMA(1, 1, 1) can be mathematically represented as:

$$y_t - y_{t-1} = W_t \quad 5.12$$

$$W_t = \omega + \phi_{AR1}W_{t-1} + \phi_{MA1}e_{t-1} + e_t$$

where W_t is the number of differencing required to reach stationarity in time series.

5.5 Anomalies

5.5.1 Background

Time series often contain unexpected observations in its dataset which are referred to as anomalies. During exploratory data analysis, missing data interpolation is normally carried out in parallel with anomaly detection. It is imperative to detect anomalies in a dataset before any predictive and forecasting models are built to avoid getting unreliable results (Malhotra *et al.*, 2015; Teng, 2010). Time series data can have seasonality and trend, and it is easy to be mistaken as an anomaly. Thus, this thesis presents appropriate techniques that are able to learn the seasonality and trend of a time series to eventually detect anomalies thereafter.

5.5.2 Statistical profiling approach

Statistical profiling is a simple and fast process that involves calculating measures of central tendency of the historical data and examining them (Wackerly *et al.*, 2014). This can be done by calculating the mean, median or the moving average of the

data. Standard deviation can also be calculated and be used to set up the uppermost and lower bounds of the dataset as anomaly thresholds. This technique is known as the Z-score analysis in which the anomaly threshold is set by the three-standard deviations above and below the mean (e.g. Chikodili *et al.*, 2020). Values that are outside the three-standard deviation thresholds are considered strong anomalies (Figure 5.5). Simple moving average can also be used for anomaly detection: it is applied to capture the pattern in time series. The difference between the actual and simple moving average can be computed to determine the tolerance band and identify anomalies.

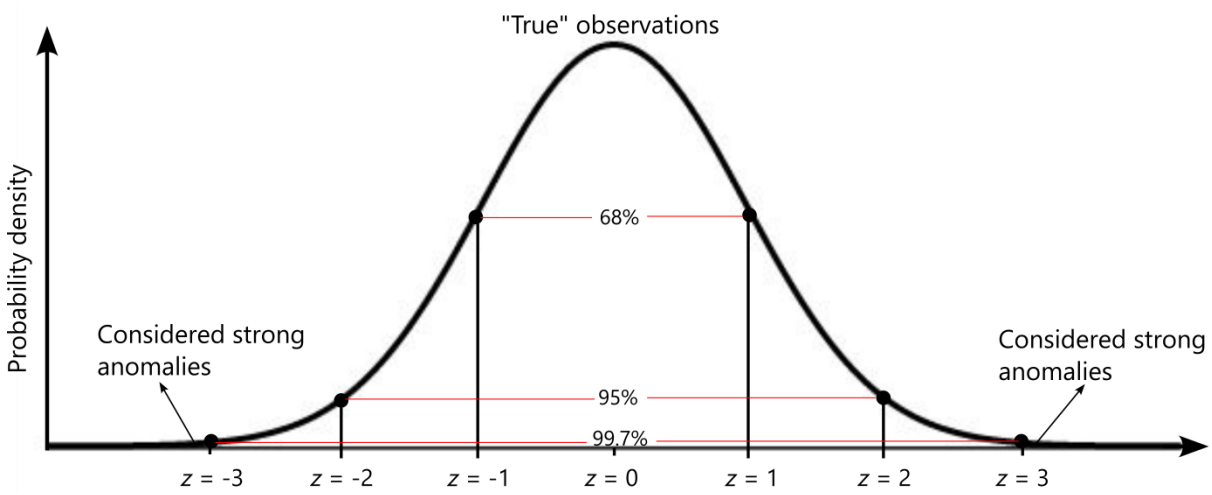


Figure 5.5: An example of anomaly detection using Z-score analysis

5.5.3 Predictive confidence level approach

Another way of detecting anomalies in time series data is by using the historical data to build a predictive model to get the overall trend, seasonality or cyclic pattern of the data. The model error can be analysed between the predicted and actual values, and use that to compute a confidence interval (Figure 5.6). The values falling beyond the confidence band can be regarded as anomalies. For example, an ARIMA model can be built and use the mean absolute percentage error (MAPE) to come up with a

confidence band (Kozitsin *et al.*, 2021; Yu *et al.*, 2016). Other ML or deep learning based algorithms such as long short-term memory (LSTM) or LSTM autoencoder produce accurate results for time series data, and thus can be used to find anomalies (Lindemann *et al.*, 2021). This technique is highly dependable on the accuracy and good performance of the predictive model. Therefore, the model must be tuned to ensure that it produces a low error.

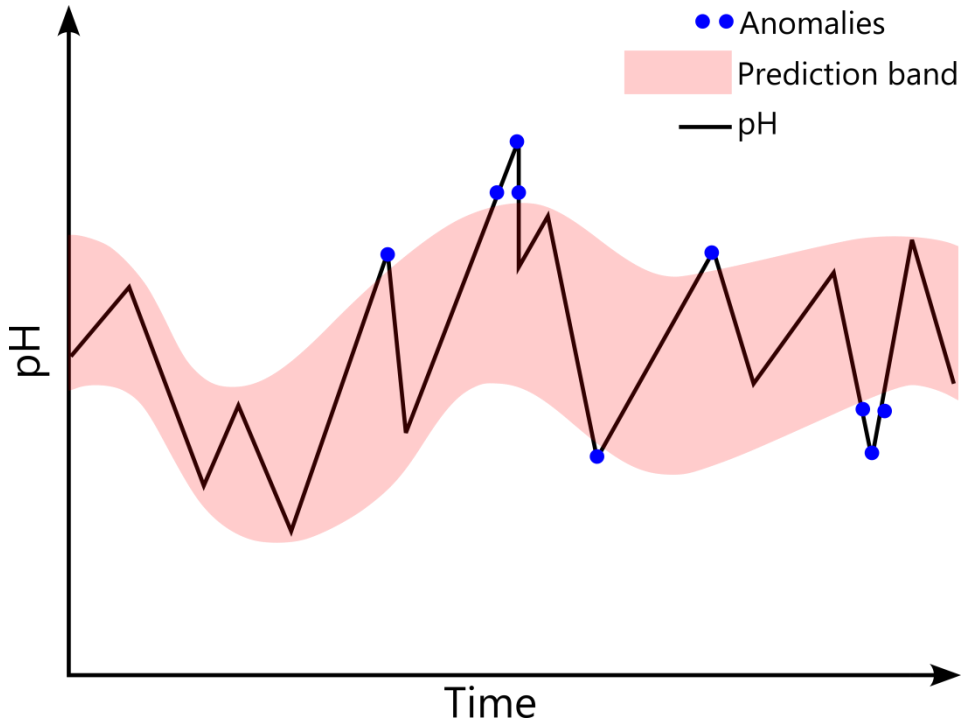


Figure 5.6: An example of applying predictive confidence level bands to detect anomalies

5.6 Conclusions

This review detailed different statistical approaches to model time series data with missing values and outliers. For the missing data, both numerical and probabilistic modelling approaches for time series were examined. Additionally, statistical profiling and predictive confidence level approaches were examined to understand anomaly detection in time series. When time series data contains missing values and anomalies, it is always necessary to compute estimation and anomaly detection

models. By computing these models and ultimately cleaning the data, it will be possible to understand the nature of the data and make accurate forecasting analysis. Different series may require different techniques to interpolate the missing values and detect anomalies. It is necessary to use the reviewed techniques effectively in order to obtain accurate results. Some of the techniques discussed in this review are applied in the coming chapters of this thesis.

The next chapter will introduce a hybrid intelligent model to forecast mine water chemistry, which applied a selection of the aforementioned techniques. This hybrid system is made up of three machine learning (ML) models: long short-term memory, gradient boosting and random forest models. Furthermore, ML models deployment is the main focus in this chapter, i.e. it goes into details of how to deploy trained and tested ML models using Flask application programming interface and Heroku cloud platforms. Finally, the best performing models are put in production as a web application which is used to make final predictions.

CHAPTER 6

6 Developing Artificial Intelligence Systems – Web Application for Predictive Analysis

6.1 Summary

Water treatment plants, including mine water treatment plants, need to stock chemicals, have reliable and enough energy and human resources to operate reliably. To avoid a process interruption, proper planning of these resources is important. Therefore, a scientifically based, practical tool to predict and forecast relevant water parameters will help plant operators to know in advance which chemicals and methods to use to treat and manage polluted water. Therefore, this chapter aims to develop a hybrid intelligent system to predict and forecast mine water parameters using electrical conductivity (EC) and pH of mining influenced water (MIW) from the mine water treatment plant in Ekurhuleni, South Africa as an example.

A hybrid intelligent system combines at least two artificial intelligence (AI) methods to predict a target label. Three machine learning (ML) algorithms: random forest regression, gradient boosting regression and a multilayer feed-forward artificial neural network (ANN) trained with backpropagation were compared to find the best learning model to be used for predictive analysis. These models were developed using historical data between the years 2016 and early 2021. The data contained a maximum of 14 parameters which were reduced to six after the dataset was cleaned and exploratory analyses were done. Therefore, the input variables of the models are turbidity, total dissolved solids (TDS), sulfate (SO_4) and iron (Fe), with EC and pH being the target outputs. Anaconda, a Python distribution platform for ML (Wang & Oliphant, 2012), was used to build the models.

Optimisation techniques and hyper-parameter tuning, including different numbers of hidden layers and neurons for ANN, optimisers, activation functions and test sizes were applied for all the models. Results of the models have been compared with the measured data on the basis of mean absolute error (MAE) and root mean square error (RMSE). It was found that random forest and gradient boosting regression models performed better than the neural network model. Therefore, the better performing ML models, i.e. random forest and gradient boosting, were deployed using the Flask application programming interface (API) and Heroku cloud platform for prediction of EC and pH. Time series forecasting technique, Long Short-Term Memory (LSTM), was used to predict turbidity, TDS, SO_4 and Fe values for 60 days, and these values were used to get the future values for EC and pH for the same period of time.

6.2 Background

Mining influenced water (MIW), especially acid mine or rock drainage (AMD/ARD) is a challenging problem encountered by many mining companies around the world (Verburg, 2011). Different techniques have been applied to predict the future mine water quality to better manage MIW (e.g. Côte *et al.*, 2010; Fox, 1999; Gao *et al.*, 2014; Khandelwal & Singh, 2005; Liu *et al.*, 2019; McCauley *et al.*, 2009; Rooki *et al.*, 2011). Modelling software such as GoldSim, MATLAB Simulink, Geochemists Workbench and PHREEQC have been previously favoured to predict future mine water quality (e.g. George *et al.*, 2009; Nalecki & Gowan, 2008; Usher *et al.*, 2010); however, researchers are trying to optimise these predictions by including artificial intelligence (AI) technology (More *et al.*, 2020; Vadapalli *et al.*, 2020). MIW occurs when iron-(di)sulfide minerals such as pyrite, marcasite or pyrrhotite react with water and oxygen (Blowes *et al.*, 2014; Singer & Stumm, 1970). In case no buffer minerals

such as carbonates are present, these reactions will result in AMD/ARD. This MIW may contain potentially toxic elements depending on the geological background (Wolkersdorfer, 2008).

AI being a disruptive technology presently drives the decision-making in many industries. They use smart data and machine learning to enhance the efficiency in operations, safety at the mines and workflow in production (e.g. Bui *et al.*, 2020a; Danilov *et al.*, 2015; Huang & Liu, 2009; Kansake *et al.*, 2019; Otto *et al.*, 2018). This chapter will present machine learning (ML) techniques in the form of a hybrid intelligent system to predict the indicator mine water parameters electrical conductivity (EC) and pH at a treatment plant in the east of Johannesburg, Ekurhuleni, South Africa. This chapter used historical data from the years 2014 to 2021, supplied by the Department of Water and Sanitation, Pretoria, South Africa. An AI system developed is a hybrid intelligent system as multiple methods are applied to predict the target parameters. The ML technique can be divided into different learning categories such as supervised, unsupervised and reinforcement learning (Alpaydin, 2010), whereas in this chapter, supervised learning is used.

Supervised learning is where a model has input variables and an output value and uses an algorithm to learn the mapping function from the input to the output (Burkov, 2019; Swamynathan, 2017). This method provides a computer system access to datasets that consists of input—output pairs, and the goal is to train the system to find a function that maps inputs to outputs. There are several tasks within supervised learning such as classification and regression. Osisanwo *et al.* (2017) describe classification as a supervised task to learn a function that maps an input point to a discrete category, and it classifies data using the nearest-neighbour or *k*-nearest-neighbour classification algorithms. These are algorithms that, when given an input,

choose the class of the nearest data point to that input or choose the most common class out of the *k*-nearest data point to that input, respectively (e.g. Ali *et al.*, 2019; Duivesteijn & Feelders, 2008). Regression is a supervised learning task of mapping an input point to a continuous value (Burkov, 2019). It is used to find the correlation between *x* and *y* variables, to know the strength of predictions and for forecasting an effect and trend (e.g. Betrie *et al.*, 2013; Ekemen Keskin *et al.*, 2020; Maier *et al.*, 2004; Singh *et al.*, 2017a). Therefore, supervised learning is a technique used in this chapter with the focus on data regression. In this paper, the main aim was on accuracy of the results, thus the ML algorithms explored include random forest, gradient boosting and neural networks, and the forecasting technique method used is the Long Short-Term Memory (LSTM).

Mine water quality can be evaluated using several parameters, e.g. EC, pH, major ions, turbidity, acidity and more. The amount of dissolved minerals in mine water is represented by total dissolved solids (TDS) and EC, which is crucial for mine water conditions. EC gives a measure of TDS as TDS = *f*(EC) (Hem, 1985; Hubert & Wolkersdorfer, 2015). Because accurate measurements of EC and TDS are important in determining mine water quality, EC is one of the target outputs in the developed ML models in this chapter. Another parameter used as a target output is pH, which is a measure of how acid or basic water is (Bates, 1964). Therefore, pH measurements are important in determining the amount of (semi-)metals that might be dissolved in mine water and the volume of clean water that can be discharged. Treated mine water is usually discharged into receiving water courses or for industrial and commercial use. Thus it is crucial to know the pH values of mine water entering the treatment plant beforehand so it can be increased or decreased accordingly to precipitate unwanted metals.

6.3 Machine Learning Algorithms Used

6.3.1 Random Forest Regression

Random forest regression builds a “forest”, which is an ensemble of decision trees trained using a “bagging” method. It works on the idea that a combination of models (decision trees) increases the accuracy of the results and yields better prediction. Bagging is a statistical technique that deals with high variance and bias problems. This algorithm involves random sampling of training observations in the process of building decision trees, and random subsets of features for splitting nodes. Random forest grows trees using binary partitioning, i.e. each parent node is split into two children. At each tree split, a random sample of m variables is selected out of all M possible variables, and only those m variables are taken for splitting. Normally, this is given by the expression: $m = \sqrt{M}$ (Biau & Scornet, 2016; Breiman, 2001).

In random forest, variables are selected using embedded algorithms containing their unique built-in variable selection methods. These algorithms are highly accurate, they make better generalisations and are interpretable (Genuer *et al.*, 2010; Nyongesa, 2020). Random forest consist of multiple decision trees, sometimes even more than one hundred trees, each of them constructed over a random selection of the observations from a dataset. The trees are de-correlated because they cannot “see” all the observations, and this guarantees less overfitting. At each node of the decision tree, the dataset is divided into two sets, each with observations much similar among themselves, yet different from the ones in the other set.

6.3.2 Gradient Boosting Tree Regression

For a gradient boosting algorithm to work, the loss function and additive model must be computed. Firstly, the boosting algorithm computes the first residual by finding the

average value of the variable that needs to be predicted. Therefore, the residual is computed by:

$$\text{residual} = \text{actual value} - \text{predicted value} \tag{6.1}$$

Therefore, a tree is built with the goal of predicting the residuals, meaning each leaf will contain a prediction as to the value of the residual and not the desired label. Some leaves may end up having more residuals, and when this happens, the average of the residuals is computed. Each sample will then pass through the decision node of the new tree until a given leaf is reached. Then the residual in this leaf is used to predict the actual output. A hyper-parameter known as the learning rate is then introduced to prevent overfitting. In the process of making a prediction, each residual is multiplied by the learning rate:

$$P_o = A_o + \alpha \times R_{DT} \tag{6.2}$$

where P_o is the predicted output, A_o refers to the average output, α is the learning rate parameter and R_{DT} refers to the residual predicted by decision tree.

Therefore, new residuals are calculated by subtracting the actual output from the predictions made in equation 6.2. These residuals will be used for the leaves of the next decision tree. After training, all trees in the ensemble are then used to make the final predictions, FP:

$$FP = A_o + \alpha_1 \times R_{DT1} + \alpha_2 \times R_{DT2} + \ldots + \alpha_n \times R_{DT\,n} \tag{6.3}$$

6.3.3 Neural Networks and structure

Artificial neural network (ANN) modelling is a technique in which the model is using known input and output parameters so it can learn from the historic data. The aim is to teach the model to produce output parameters by providing inputs. In other means, neural networks take in data, train themselves to recognise the patterns in this data, and finally predict the output for a new set of similar data (Alpaydin, 2010; Krenker *et al.*, 2011). ANN technique depends on the number of hidden layers to be used, the neurons in each hidden layer, activation function, learning rate and momentum, and the iterations and desired error level. Neural networks are made up of layers of neurons which are the core processing units of the network – the input layer neurons receive the input data, output layer neurons predicts the output parameter, and the hidden layer neurons perform most of the computations required by the network. ANN is a mathematical model that operates on three set of rules, i.e. multiplication, summation and activation as explained by Krenker *et al.* (2011) (Figure 6.1).

An input layer of the ANN gets input parameters, and its neurons are connected to the hidden layer neurons through “channels” that are assigned numerical values known as “weight”. Inputs are multiplied with the corresponding weights and the sum is transferred as inputs to the hidden layer neurons. The hidden layer neurons are associated with numerical values known as the “bias” and are added to the input sum. Furthermore, the resulting value is passed through a threshold function known as the activation function which determines whether a particular hidden layer neuron will get activated or not. Activated neurons transmit data to the neurons of the next hidden layer over the channels; this process is known as feed forward propagation (e.g. Svozil *et al.*, 1997; Yuan *et al.*, 2003). Therefore, in the output layer, the neuron

with the highest probability determines the output. Finally, the predicted outputs are then compared to the actual outputs, and an error can be deduced. This is therefore transferred back to the network in which weights can be adjusted according to the derived errors; this process is known as back propagation (e.g. Law, 2000).

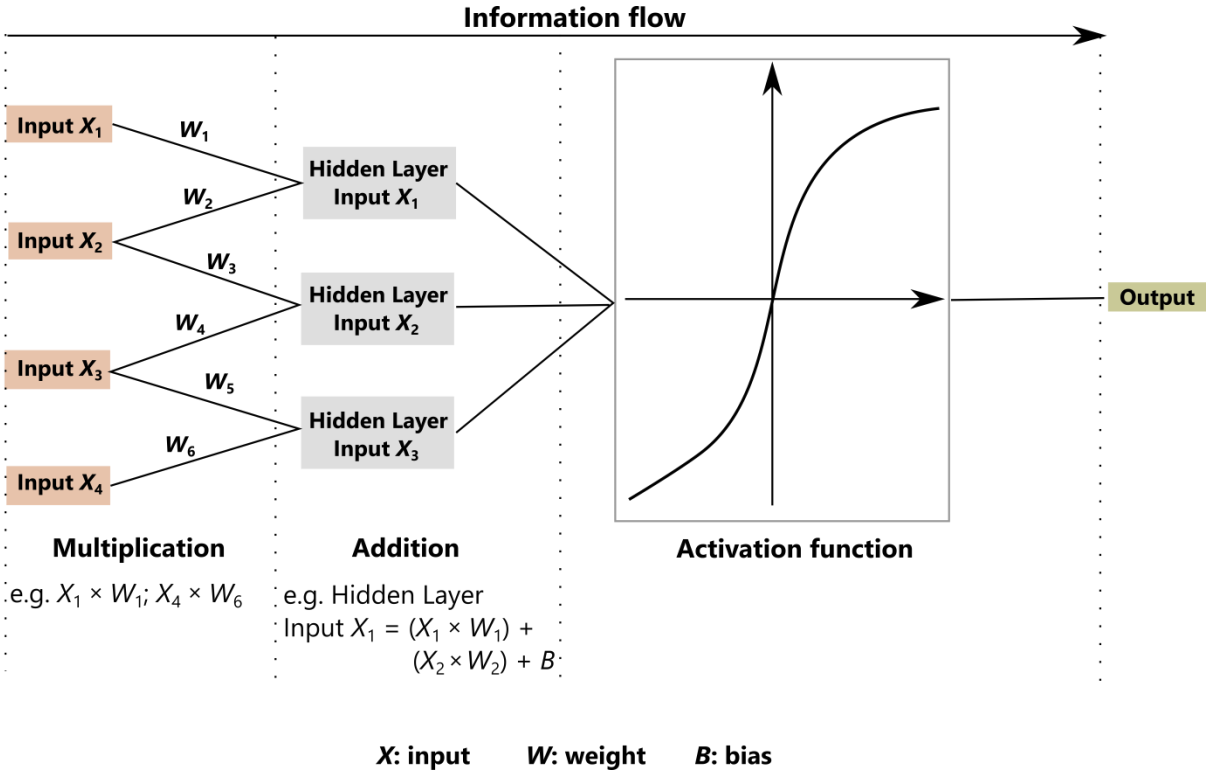


Figure 6.1: Artificial neuron's working concept (modified after Krenker *et al.*, 2011)

In this section, a supervised learning algorithm for neural network modelling was adopted, i.e. a back propagation algorithm. For this network, the architecture has one input layer with four neurons, two hidden layers all with six neurons and one output layer with two neurons (Figure 6.2). It should be noted that the two hidden layers automatically make this neural network a deep learning model or deep neural network (e.g. Montavon *et al.*, 2018).

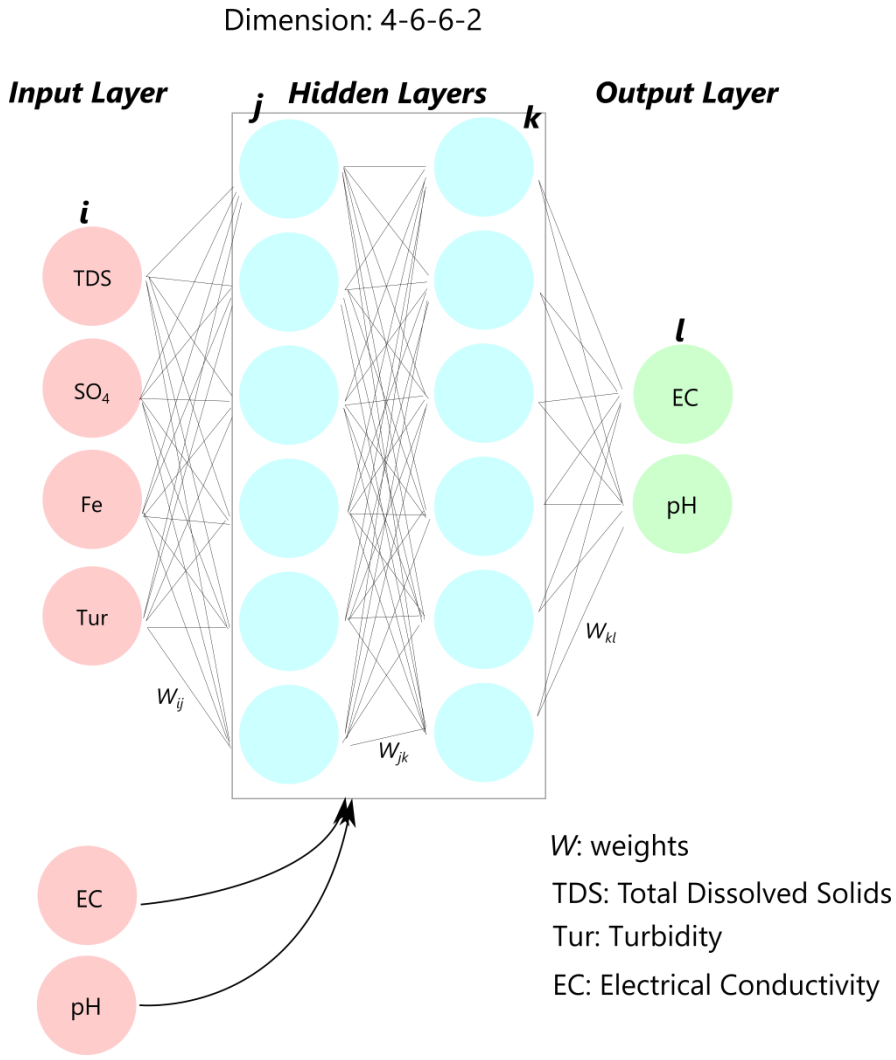


Figure 6.2: Neural Network architecture for this study

From the above structure (Figure 6.2), the j^{th} neuron of the hidden layer 1 is connected to several input neurons, x_i:

$$x_i = x_1 + x_2 + x_3 + x_4 \qquad 6.4$$

The total input values in the hidden layer 1 can then be given by:

$$Tot_j = \sum_{i=1}^{n} x_i W_{ij} + \beta_j \qquad 6.5$$

where n is the total number of input units, x_i is the input unit, W_{ij} refers to the weight connecting the i^{th} neuron to the j^{th} neuron and β_j is the bias neuron.

Therefore, the total hidden layer 1 output, y_i, is computed using the logarithmic sigmoid function:

$$y_j = \frac{1}{1 + e^{-(\mathrm{Tot}_j + \beta_j)}} \tag{6.6}$$

The total input in the output layer is given by:

$$\mathrm{Tot}_l = \sum_{l=1}^{n} y_k W_{kl} + \beta_l \tag{6.7}$$

where W_{kl} is the weight connecting the k^{th} neuron to the l^{th} neuron and β_l is the bias neuron.

Therefore, the total computed output, y_l, can be given by:

$$y_l = \frac{1}{1 + e^{-(\mathrm{Tot}_l + l)}} \tag{6.8}$$

With back propagation technique being applied, the computed output must be compared with the target output, T_l, and determine the error, e_l, at any neuron in the l^{th} layer:

$$e_l = T_l - y_l \tag{6.9}$$

The total error, E_l, which must be propagated from the output layer neurons to the hidden layer neurons, is given by:

$$E_l = \sum_{m=1}^{n} (T_l - y_l)^2 \tag{6.10}$$

Each weight must be updated using the calculated total error by applying partial derivatives:

$$\nabla W_{\mathrm{k\,l}} = \frac{\partial I}{\partial W_{\mathrm{kl}}} \tag{6.11}$$

with:

$$\frac{\partial E_{\mathrm{l}}}{\partial W_{\mathrm{kl}}} = \frac{\partial E_{\mathrm{l}}}{\partial y_{\mathrm{l}}} \times \frac{\partial I}{\partial \mathrm{Tot_l}} \times \frac{\partial \mathrm{Tot_l}}{\partial W_{\mathrm{kl}}} \tag{6.12}$$

and:

$$\frac{\partial E_{\mathrm{l}}}{\partial y_{\mathrm{l}}} = -(T_{\mathrm{l}} - y_{\mathrm{l}}) \tag{6.13}$$

$$\frac{\partial y_{\mathrm{l}}}{\partial \mathrm{Tot_l}} = y_{\mathrm{l}}(1 - y_{\mathrm{l}}) \tag{6.14}$$

$$\frac{\partial \mathrm{Tot_l}}{\partial W_{\mathrm{kl}}} = y_{\mathrm{k}} \tag{6.15}$$

Therefore, the new weights can be calculated by:

$$W_{\mathrm{kl}}^{\mathrm{new}} = W_{\mathrm{kl}} - \alpha \times \nabla W_{\mathrm{kl}} \tag{6.16}$$

where α is the learning rate parameter.

6.4 Mine Water Quality Dataset

The data used in this chaper was generated through the monitoring and managing of MIW, with nearly daily sampling being carried out during a period of six years (2016–2021). Parameters that were measured during this period are rainfall, temperature, EC, TDS, total suspended solids (TSS), acidity (as $CaCO_3$), pH, calcium (Ca as $CaCO_3$), magnesium (Mg as $CaCO_3$), sulfate (SO_4), aluminium (Al), iron (Fe) and manganese (Mn), and they were used in the units applied by the lab. All of these parameters had some of the measurements missing, leading to them not having the same number of observations. Fe, turbidity and pH had the most number of

observations, while rainfall and Mg had the very least of observations (Table 6.1). Therefore, the data had to be “cleaned” before it was used for training and testing ML models and for forecasting analysis.

Table 6.1: Mine water quality data supplied; *n*: number of measurements, $\overline{x}$: average, σ: standard deviation, min.: minimum observation, max.: maximum observation. True average for pH have been calculated as $-\log_{10}[(\sum C_i)/(n)]$, where *C* is the hydrogen ion (proton) activity and *n* is the number of measurements (www.wolkersdorfer.info/pH_en)

Parameter	*n*	$\overline{x}$	σ	Min.	Max.
Rainfall, mm	59	13.7	10.6	1	40
Temperature, °C	1387	23	2	17	29
EC, mS/cm	1387	3	0.1	2.4	3.2
TDS, mg/L	1381	2678	185	2014	3195
pH, —	1396	6.5	0.2	5.1	7.3
Turbidity, NTU	1398	54.2	59.9	0.5	450
TSS, mg/L	1386	74	75	0.0	378
Acidity, mg/L $CaCO_3$	898	310	41	166	442
Ca, mg/L $CaCO_3$	598	399	25	267	560
Mg, mg/L $CaCO_3$	591	101	35	19	389
SO_4, mg/L	1396	1474	204	658	1988
Al, mg/L	692	0.1	0.0	0	0.2
Fe, mg/L	1402	97.7	15.6	0.1	179.4
Mn, mg/L	1384	6.6	3.4	0.0	22.9

6.5 Data Cleaning and Exploratory Data Analysis

6.5.1 Data distribution

The graphical methods show how the data are distributed (Figure 6.3Figure 6.4) and help in visualising the spread, i.e. dispersion, variability and scatter. Distribution shape is important as it informs whether the data are normally distributed, skewed to

the right or left, or are distributed uniformly. Additionally, the Kolmogorov-Smirnov and Shapiro-Wilk tests were conducted to test the normality of the data (Table 6.2). In most cases, the Shapiro-Wilk test works better on small sample sizes ($n < 50$), while the Kolmogorov-Smirnov test is used on larger sample sizes ($n \geq 50$). Therefore, the Kolmogorov-Smirnov test was relied on to conclude on the normality of the data. From the constructed graphs and tests conducted, it shows that the data are not normally distributed, and all parameters have statistically significant outliers. The *p*-values for all the variables are less than 0.005, which implies that the data do not follow a normal distribution.

Table 6.2: Normality data test using Kolmogorov-Smirnov and Shapiro-Wilk methods; *n*: number of observations, *p*-value: probability value

		Kolmogorov-Smirnov		Shapiro-Wilk	
Parameter	*n*	Statistic	*p*-value	Statistic	*p*-value
Rainfall	59	0.143	0.004	0.912	0.000
Temperature	1387	0.089	0.000	0.970	0.000
Turbidity	1387	0.185	0.000	0.803	0.000
EC	1381	0.152	0.000	0.940	0.000
TDS	1396	0.176	0.000	0.918	0.000
pH	1398	0.111	0.000	0.947	0.000
TSS	1386	0.221	0.000	0.848	0.000
Acidity	898	0.197	0.000	0.912	0.000
Ca	598	0.341	0.000	0.469	0.000
Mg	591	0.319	0.000	0.338	0.000
SO_4	1396	0.141	0.000	0.917	0.000
Al	692	0.043	0.004	0.986	0.000
Fe	1402	0.139	0.000	0.788	0.000
Mn	1384	0.195	0.000	0.858	0.000

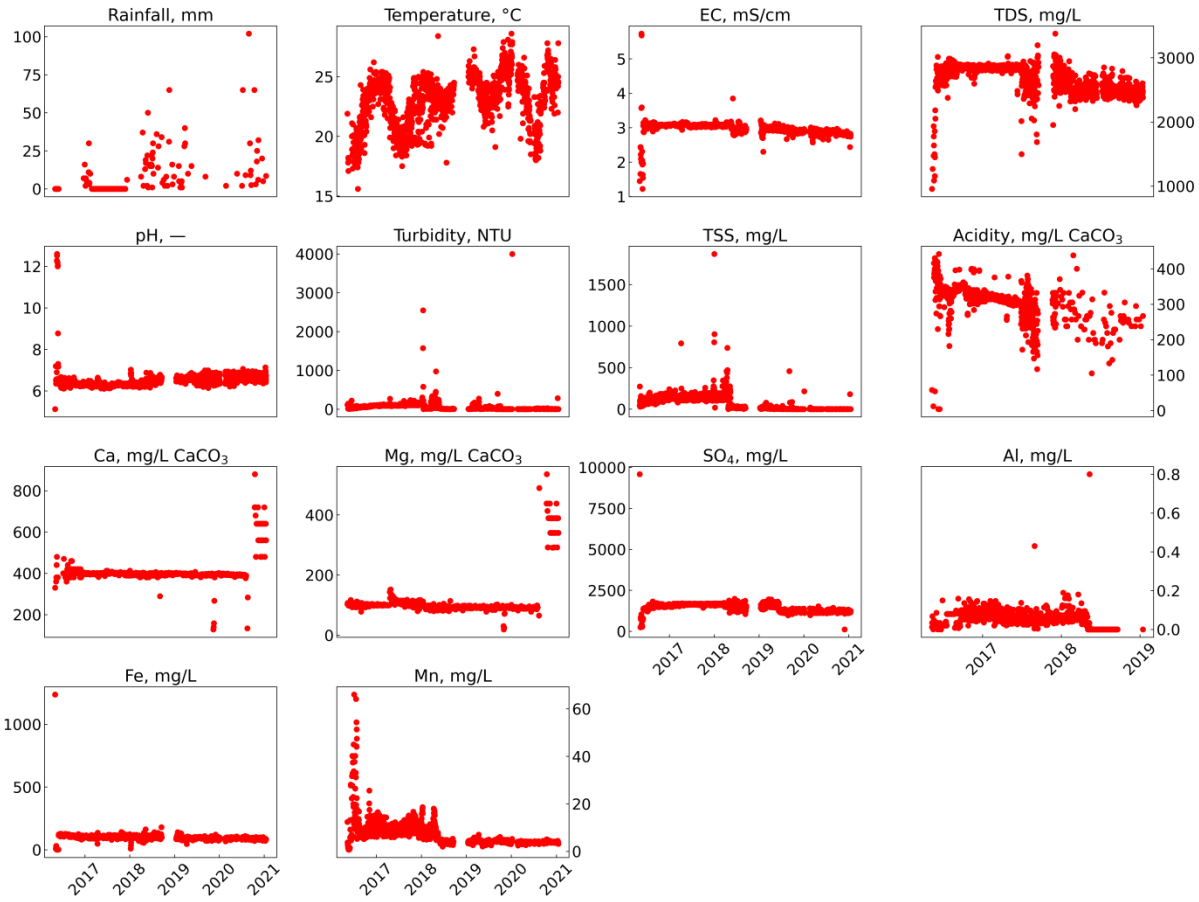


Figure 6.3: Time series plots of mine water data at the Eastrand mine water treatment plant over a period of six years (2016–2021)

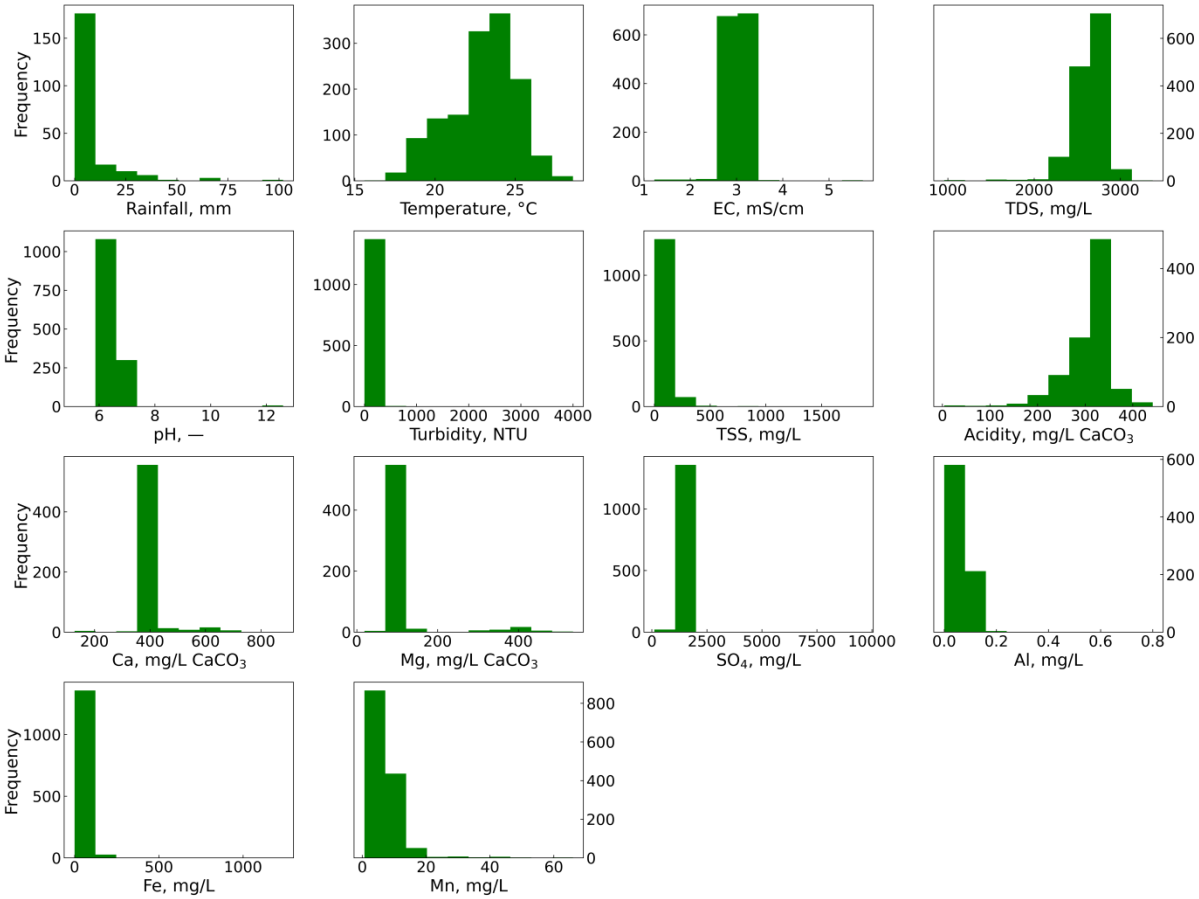


Figure 6.4: Histograms of Eastrand mine water treatment plant data (2016–2021)

Dependent variables and possible relationships between the parameters were visualised using the correlation matrix with statistical significance levels (Figure 6.5). Correlation coefficients (*r*) of the parameters vary from lowest to highest, with the notable strongest positive relationships of 0.7 and 0.69 between EC–TSS and turbidity–TSS, respectively. Even though most of the relevant parameters show a low correlation coefficient, they however display high statistical significance.

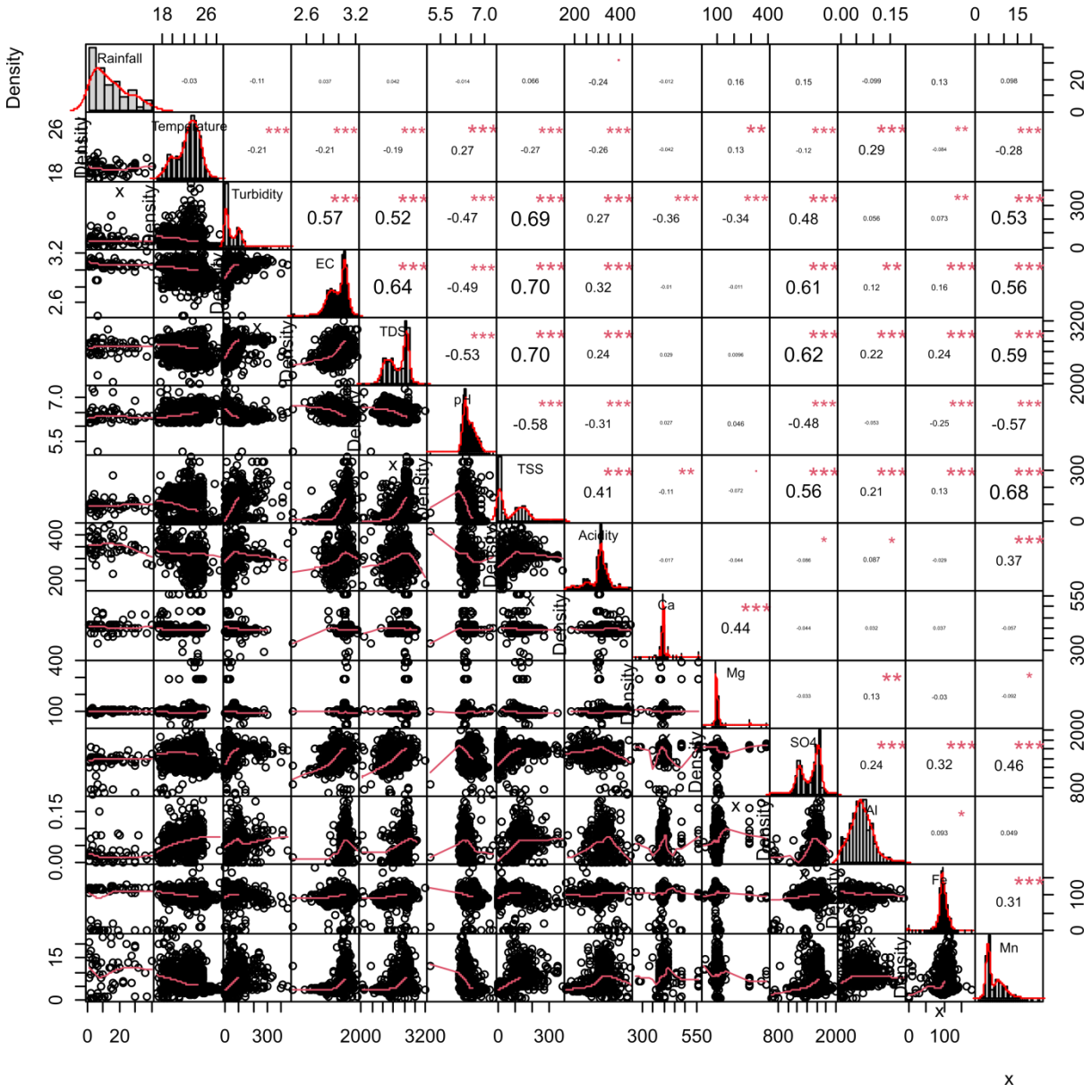


Figure 6.5: A correlation chart with the distribution of each parameter shown on the diagonal; on the bottom of the diagonal are the bivariate scatter plots with a fitted line; on the top of the diagonal are the values of the correlation with the significance levels shown as asterisks: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; correlation coefficient font size is related to the relationship between the parameters – the stronger the relationship the larger the font size and vice versa

6.5.2 Outlier visualisation, detection and removal

In statistical terms, an outlier is an observation that is different from the other observations (Wackerly *et al.*, 2014). This can happen due to a mistake during data collection, equipment malfunctioning or just an indication of variance in data

collected. Outliers can be identified using visual or statistical methods. In this chapter, the box plot visualisation methods were used, which graphically shows the groups of numerical data using their quartiles. Outliers are plotted as points while other data are displayed within boxes (Figure 6.6).

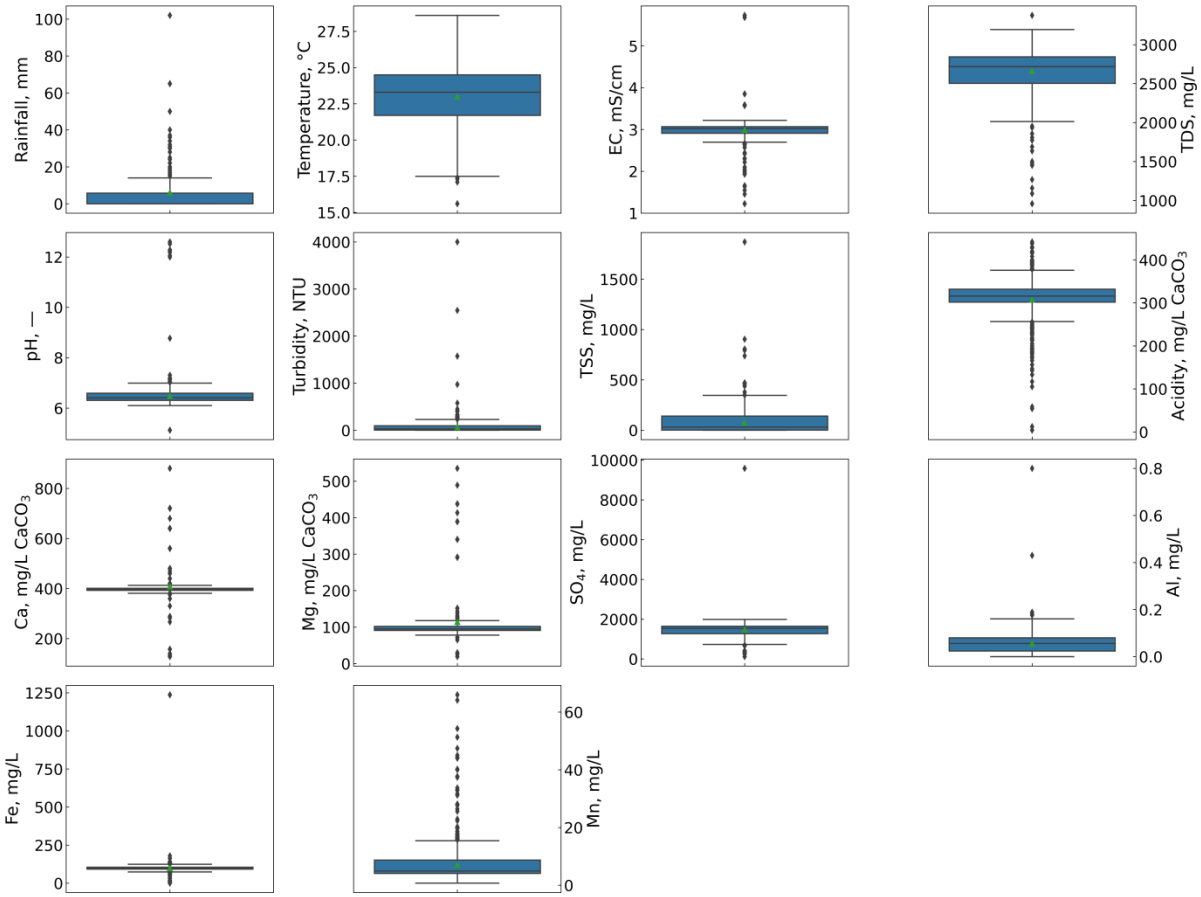


Figure 6.6: Box plots with outliers for the Eastrand AMD Treatment Plant data

Z-score analysis was additionally applied on the data for outlier detection and removal. *Z*-scores refer to the number of standard deviations above and below the mean, where a *Z*-score of ± 3 implies that an observation is three standard deviations above or below the mean. A *Z*-score for an observation is calculated by taking the observation (X), subtract the mean (μ) and divide by the standard deviation (σ):

$$Z = \frac{X - \mu}{\sigma} \qquad 6.17$$

The further away an observation's *Z*-score is from zero, the more chances of it being an anomaly. Standard cut-off values in determining outliers are *Z*-scores of ± 3 or even far away from zero (Figure 5.5).

6.5.3 "Clean" data

Data cleaning is the process of removing data that can potentially negatively affect the performance of the forecasting models. This process involves identifying and removing the outliers and also correctly replacing the missing values. After the data cleaning process, only nine parameters were statistically significant and some with good correlation coefficients. However, within these nine parameters, acidity, TSS and Mn were discarded. Acidity was dropped due to a low number of observations, i.e. 898 out of 1381 when compared to other observations, while TSS and Mn contains frequent observations with 0 mg/L or very low readings making them not effective for model training and testing. Therefore, parameters used for training and testing the ML models were turbidity, TDS, SO_4, Fe, EC and pH (Figure 6.7).

Some of the observations were missing for each parameter, although this was just a small proportion that might not necessarily negatively affect the quality of the results. In Python, it would not be possible to build ML models with dataset containing missing observations. There are different statistical techniques to impute missing observations, but mostly depend on the type of data available. Techniques such as sequential imputation for missing values (IMPSEQ), iterative robust model-based imputation (IRMI) and multiple imputations of incomplete multivariate data (AMELIA) have been put to test over the years to solve the missing values problem on historical data (Betrie *et al.*, 2016).

Various linear regression methods with single and multiple independent variables were applied in this study in an attempt to fill-in the missing values, but it was not possible due to the number of observations for parameters not matching. A well-known method of dealing with missing observations when building predictive models in Python is replacing missing values with large numbers, e.g. -9999 or 9999, so the algorithm can automatically detect it as an anomaly. However, this is not ideal when dealing with a uniformly distributed data. Therefore, this problem was solved by applying a robust, fast and simple method: getting the average of the three observations before and after the missing observation:

$$y_i = \frac{\sum_{m=i-3}^{i-1} y_m + \sum_{n=i+1}^{i+3} y_n}{6} \qquad 6.18$$

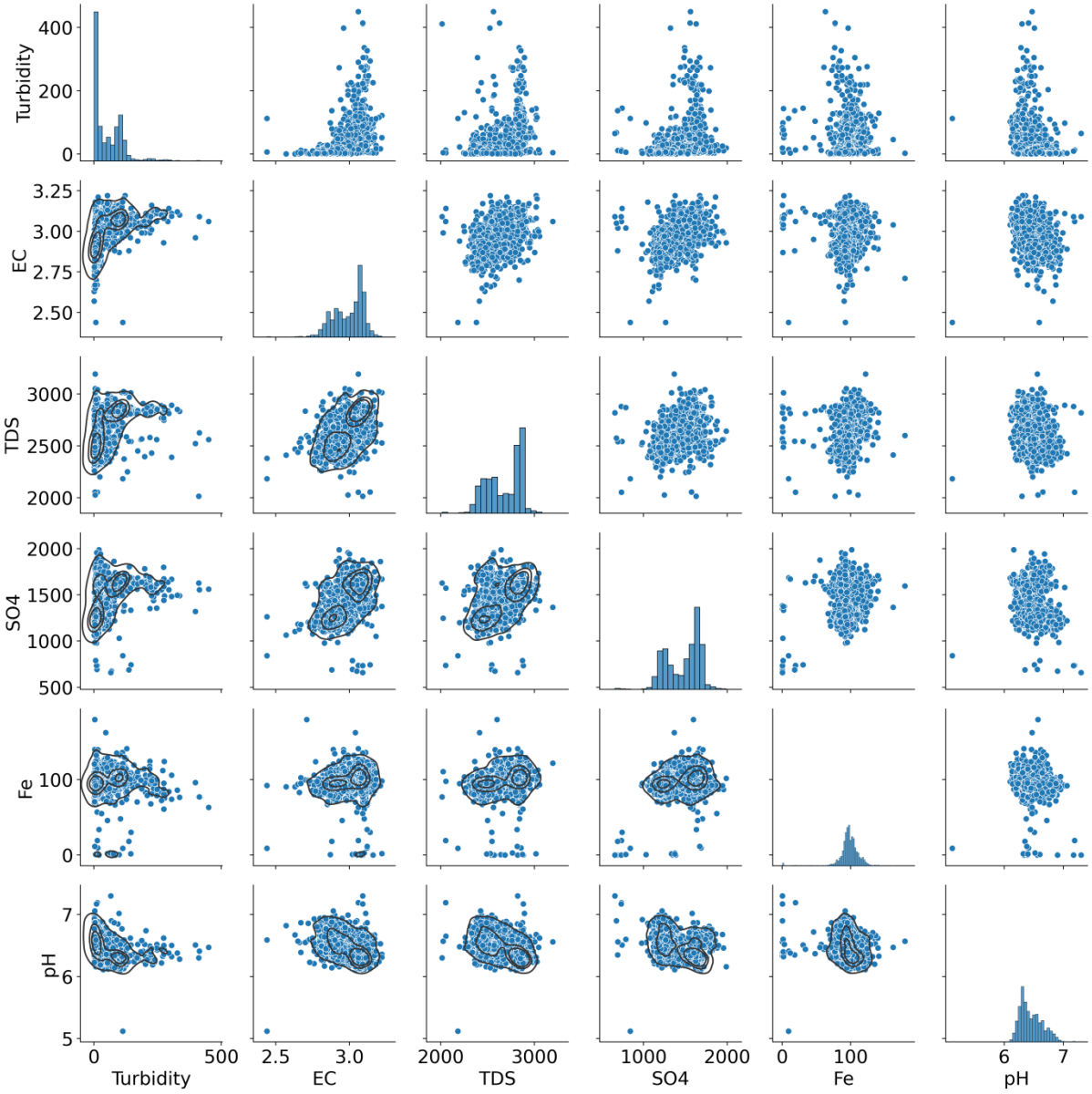


Figure 6.7: Density with cross plots (lower part of the graph), histograms (diagonal-middle part of the graph) and cross plots (upper part of the graph) of the parameters used to train and test the machine learning models

6.6 Results and Discussion

6.6.1 General results

This study used mine water data from a treatment plant in the east of Johannesburg, South Africa to predict EC and pH of mine water using random forest, gradient boosting and ANN algorithms on a Python distribution platform for ML known as Anaconda. Input data were the mine water parameters consisting of turbidity, TDS,

Fe and SO_4, and the output data were EC and pH values. In total, 1381 observations for each parameter were used for the model. In all the three algorithms, the data were split into two phases: training and testing. As recommended by Bui *et al.* (2020b), the dataset was split using a test size of 0.2; i.e. 20% of data (277 observations) were used to assess the models' performance and 80% (1104 observations) were used as training set.

Random forest regression algorithm has an option to select the number of decision trees to use, although the default number is 500 (Breiman, 2001). With the number of observations that were available for training and testing, a grid search optimisation method for selecting the number of trees was used. A grid search of 10, 50, 100, 150 and 200 decision trees was applied, and 10 trees gave the best results. Thus, 10 trees were used to ensure the richness of the forest. This model creates a number of bootstrap samples (smaller samples randomly selected from larger sample) and develops a regression for each. Each bootstrap iteration was used to predict EC and pH values and were averaged from all the decision trees.

With gradient boosting regression, the important parameters that determine the performance of the predictive model are: the number of trees, maximum tree depth and the learning rate (the value at which the weights are updated during model training). Using the same approach applied in the previous model, grid search optimisation technique was used to tune hyper-parameters, so the algorithm can give the best number of trees (100), maximum tree depth (2) and learning rate (0.05) to use.

ANN model takes into account the number of hidden layers and nodes, optimisers (learning rate, bias and momentum), epochs, batch size and activation functions. In

most cases, a deep neural network model with multiple hidden layers can make accurate predictions (e.g Easley *et al.*, 2018). The ANN structure was determined using the optimisation techniques and yielded a model with two hidden layers of six nodes and rectified linear activation function for each (Figure 6.2), a sigmoid activation function for the output layer, stochastic gradient descent optimiser with learning rate and momentum of 0.001 and 0.8 respectively, batch size of 16, and 50 epochs. MAE and RMSE were used to evaluate the performance of predicted EC and pH values for all the models.

The findings of this study suggest that random forest and gradient boosting analysis are powerful of predicting accurate EC and pH values due to their good model performance, while the ANN model performed poorly. Further hyper-parameter tuning to improve ANN algorithm performance was not conducted; therefore, this study relied on random forest and gradient boosting regression tree algorithms to perform predictive analysis.

All the models tested in this study were for predicting continuous data, and the evaluation metrics used are the mean absolute error (MAE) and root mean square error (RMSE) for both testing and training datasets (Table 6.3). The main aim of this evaluation was to predict the generalisation accuracy of a ML model on future data. Absolute error refers to the amount of error in measurements – it is the difference between the measured value and predicted value, while the MAE is the average of all absolute errors (Ahmed *et al.*, 2019; Reich *et al.*, 2016). Larger MAE implicates a larger error.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \qquad 6.19$$

where n is the number of data points, x_i refers to the observed values and x are the predicted values.

MSE shows how close a regression line is to a set of points. This is done by taking the errors (or distances) from the points to the regression line and squaring them to remove possible negative signs. With MSE, the average of a set of errors can be identified (Draper *et al.*, 2013; Gilroy *et al.*, 1990). The lower the MSE the better the model, the larger the MSE the larger the error, and zero MSE means the model is perfect:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2 \qquad 6.20$$

RMSE is the square root of MSE. It is the standard deviation of the prediction errors which is a measure of how spread out the prediction error is. RMSE shows how concentrated the data is around the regression line (Ahmed *et al.*, 2019; Draper *et al.*, 2013). The RMSE values between 0.2 and 0.5 show that the model can relatively predict data accurately.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x)^2}{n}} \qquad 6.21$$

Therefore, the smaller the values of MAE and RMSE imply that the model performed well. ANN is the only model with MAE and RMSE values greater than 1 (Figure 6.8), making it the only bad performing algorithm for this set of data.

Table 6.3: Performance on testing and training dataset for random forest, gradient boosting and neural network models

	Testing			Training		
Algorithm	MAE	RMSE	Quality	MAE	RMSE	Quality
Random Forest	0.074	0.109	Good	0.032	0.049	Good
Gradient Boosting	0.070	0.100	Good	0.069	0.100	Good
Artificial Neural network	3.822	4.114	Bad	3.815	4.110	Bad

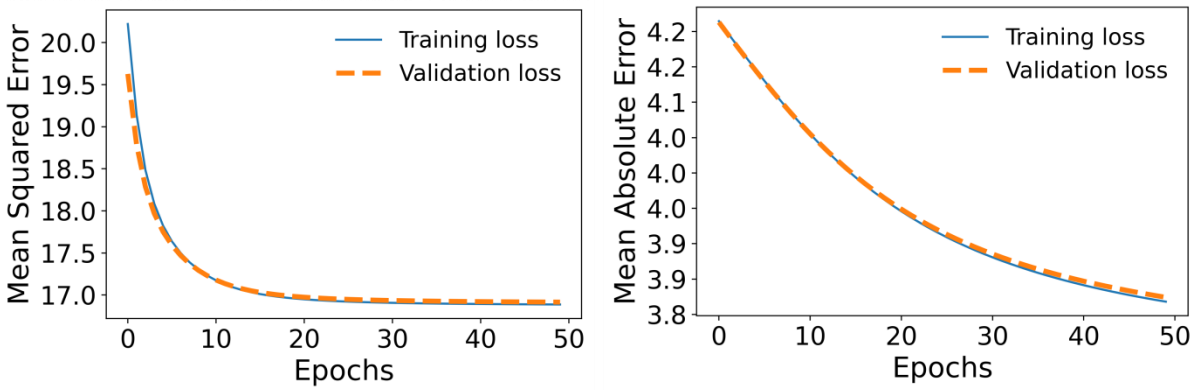


Figure 6.8: The change in training and validation loss throughout the training process using Mean Squared Error (the upper graph) and Mean Absolute Error (the lower graph)

A robust linear regression method was used to analyse the relationship between the predicted and measured values of EC and pH. From the graphs (Figure 6.9), the values show substantial and moderate correlation of EC and pH values, respectively. Predicted values of EC and pH produced correlation coefficients of 0.626 for EC and 0.432 for pH using random forest algorithm, while the gradient boosting algorithm produced correlation coefficients of 0.633 for EC and 0.433 for pH. The algorithms performed well with low values of MAE and RMSE, but prediction results show moderate correlation for pH and substantially good correlation for EC. Furthermore, significance levels were tested to determine which of the four relationships are more

significant than the others. Finally, the results show that gradient boosting's predicted data has a higher statistical significance than random forest's prediction (Figure 6.9).

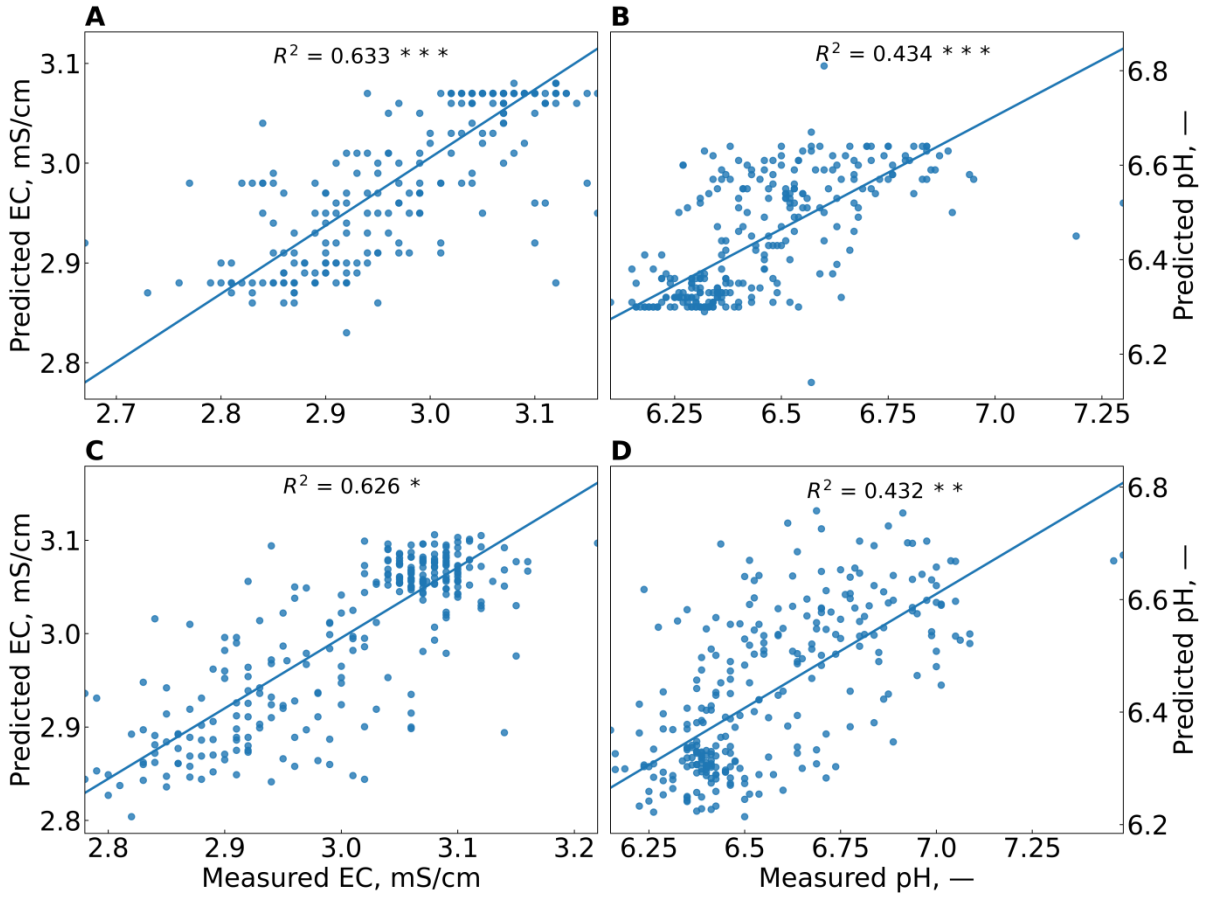


Figure 6.9: Comparing the measured and predicted EC and pH values; A: Gradient Boosting algorithm for EC, B: Gradient Boosting algorithm for pH, C: Random Forest algorithm for EC, D: Random Forest algorithm for pH; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

6.6.2 Model deployment

The main goal of developing ML models is to solve a problem, and a ML model can only achieve that when it is deployed and used in an AI system. Deployment is a process whereby the ML model is integrated into a suitable platform to make practical business decisions based on the data. The models have been built through training and testing techniques, with random forest and gradient boosting regression tree algorithms yielding good results, and therefore they are the only two out of three

tested algorithms deployed. Building models is as important as deploying them. A ML deployment process should be clear on how to get feedback from a model in production and how to set up continuous delivery. Actively tracking and monitoring the ML model state can be advantageous in knowing the model performance decay, major anomalies, or even data skewness.

In this study, an application programming interface (API) was created to deploy ML models using the Flask and Heroku platforms (Figure 6.10). Flask is a web framework for Python programming language, i.e. it provides functionality for building web applications. Heroku is a cloud platform on which applications can be deployed, managed and scaled (Lindenbaum *et al.*, 2007). Steps involved in this deployment process include a) training the model, b) creating a web application using Flask, c) committing programming code in GitHub (Preston-Werner *et al.*, 2008), d) linking GitHub to the Heroku cloud platform, e) and finally deploying the model as a web application (Figure 6.11).

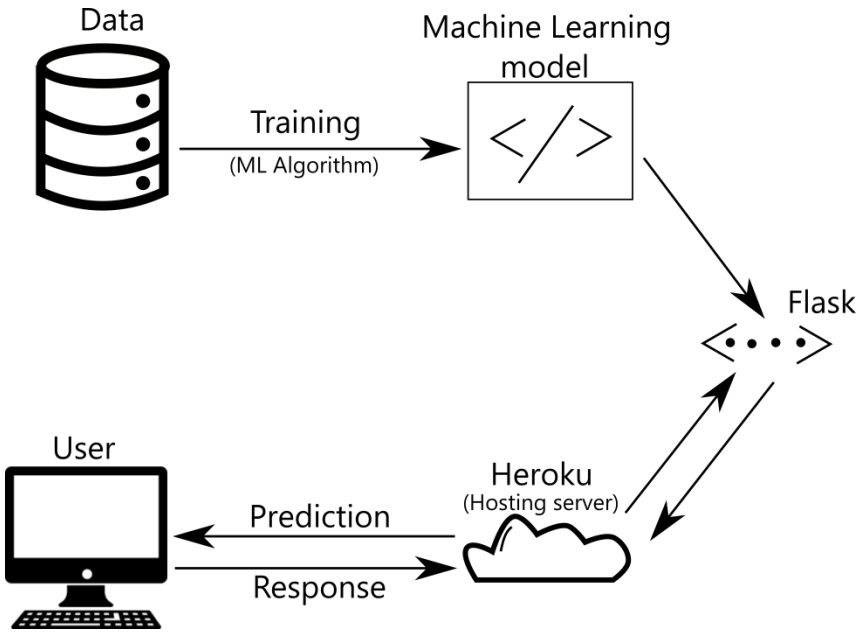


Figure 6.10: Deployment of machine learning models

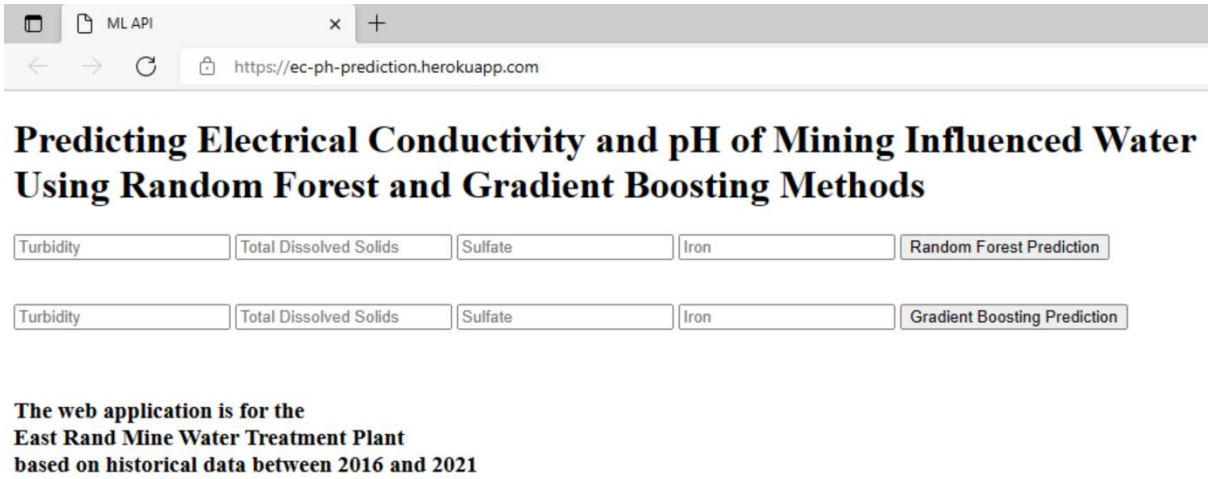


Figure 6.11: Screenshot of the web application used to predict EC and pH of mining influenced water for the Eastrand "Acid Mine Water" treatment plant in Springs, South Africa (https://ec-ph-prediction.herokuapp.com)

In this case, the web application was supplied with values of turbidity, TDS, SO_4 and Fe, and it gave predictions of EC and pH for both random forest and gradient boosting regression tree algorithms. These values were compared with the measured ones, showing little to no difference (Table 6.4). This implies that the developed web application can accurately predict the values of EC and pH for historical data when missing on the report.

Table 6.4: A comparison of measured EC and pH values versus the predicted values for historical mine water data using the web application. Prediction error is calculated as: |[(measured–predicted)/(measured)]×100%|

Gradient Boosting regression Algorithm						Random Forest Regression Algorithm					
EC, mS/cm			pH, —			EC, mS/cm			pH, —		
Measured	Predicted	|Prediction error, %|	Measured	Predicted	|Prediction error, %|	Measured	Predicted	|Prediction error, %|	Measured	Predicted	|Prediction error, %|
3.15	3.07	2.54	6.55	6.34	3.21	3.15	3.10	1.59	6.55	6.38	2.60
3.15	3.03	3.81	6.41	6.41	0.00	3.15	3.09	1.90	6.41	6.39	0.31
3.08	3.07	0.32	6.31	6.35	0.63	3.08	2.99	2.92	6.31	6.50	3.01
3.07	3.00	2.28	6.28	6.37	1.43	3.07	3.03	1.30	6.28	6.33	0.80
3.09	3.06	0.97	6.24	6.37	2.08	3.09	3.07	0.65	6.24	6.33	1.44
2.99	3.03	1.34	6.48	6.39	1.39	2.99	3.01	0.67	6.48	6.44	0.62
3.04	2.97	2.30	6.67	6.47	3.00	3.04	3.02	0.66	6.67	6.64	0.45
3.00	2.99	0.33	6.32	6.43	1.74	3.00	3.01	0.33	6.32	6.43	1.74
3.05	3.06	0.33	6.15	6.34	3.09	3.05	3.06	0.33	6.15	6.24	1.46
3.04	2.99	1.64	6.27	6.42	2.39	3.04	3.01	0.99	6.27	6.33	0.96

6.6.3 Forecasting and Web Application Prediction Results

Long short-term memory (LSTM) was used to forecast the observations for the input parameters (turbidity, TDS, SO_4 and Fe) for 60 days. These predicted input parameter observations were used to forecast EC and pH for the same period. Recurrent neural networks (RNNs) are often used in time series analysis. However, they have limitations such as failing to process longer sequences and the vanishing gradient. LSTM is a type of RNN that is developed to overcome the problems a basic RNN would encounter, and were thus used in this study.

In the LSTM model, the number of past days used to predict the future was set to be 100 days. A single hidden layer LSTM having 32 memory units with a rectified linear (ReLU) activation function was used in this study. The model was fitted over 50 epochs with a batch size of 32 and validation split of 0.2, and was further compiled using the adaptive moment estimation (Adam) optimiser and mean squared error (MSE) loss, with the lower MSE implying that the model performed well (Figure 6.12). Forecasting for turbidity, SO_4, TDS and Fe were therefore computed (Figure 6.13). The forecasted values of the input parameters were then fed to the trained random forest and gradient boosting regression tree models (Figure 6.14) to forecast the values of EC and pH for 60 days (Figure 6.15).

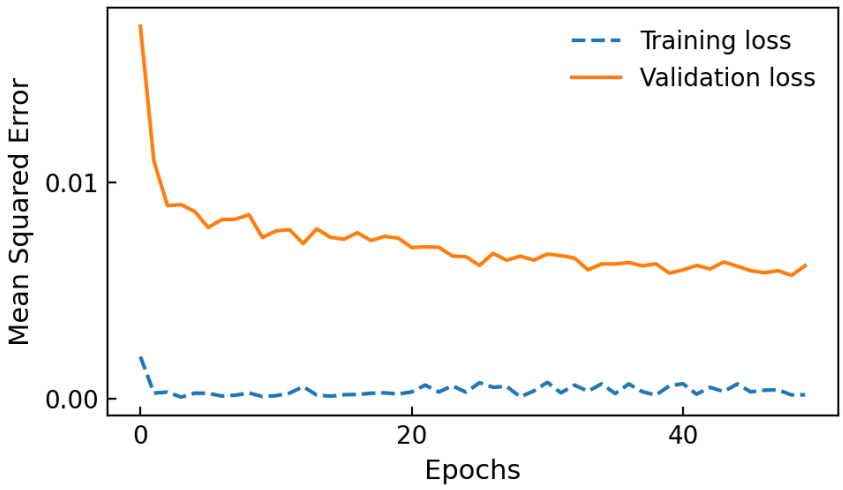


Figure 6.12: Performance of the LSTM model

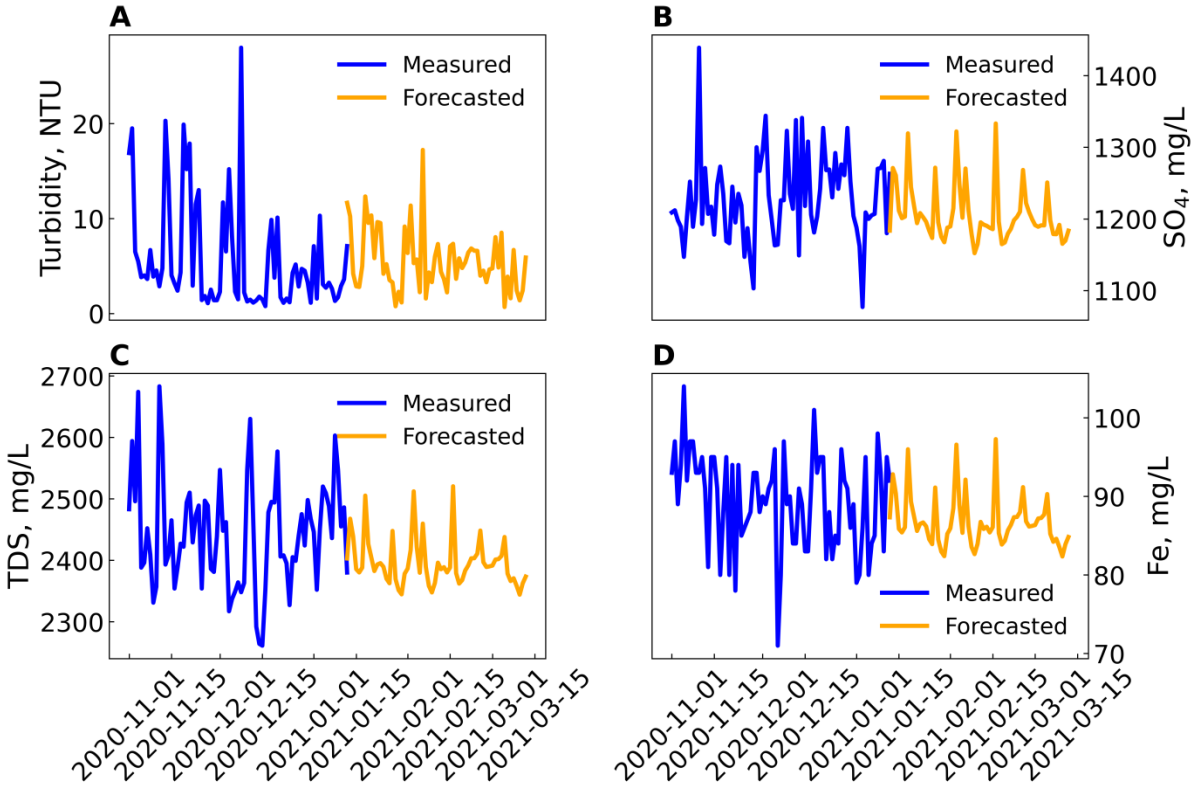


Figure 6.13: Forecasted values for turbidity, SO_4, TDS and Fe for 60 days using the LSTM model. Visualisation starts on 1 November 2020 due to a tiny forecast of 60 days and a large historical data of 1381 observations (fitting the whole 1381 observations would make it difficult to visualise the results); A: Turbidity forecast, B: SO_4^{2-} forecast, C: TDS forecast, D: Fe forecast

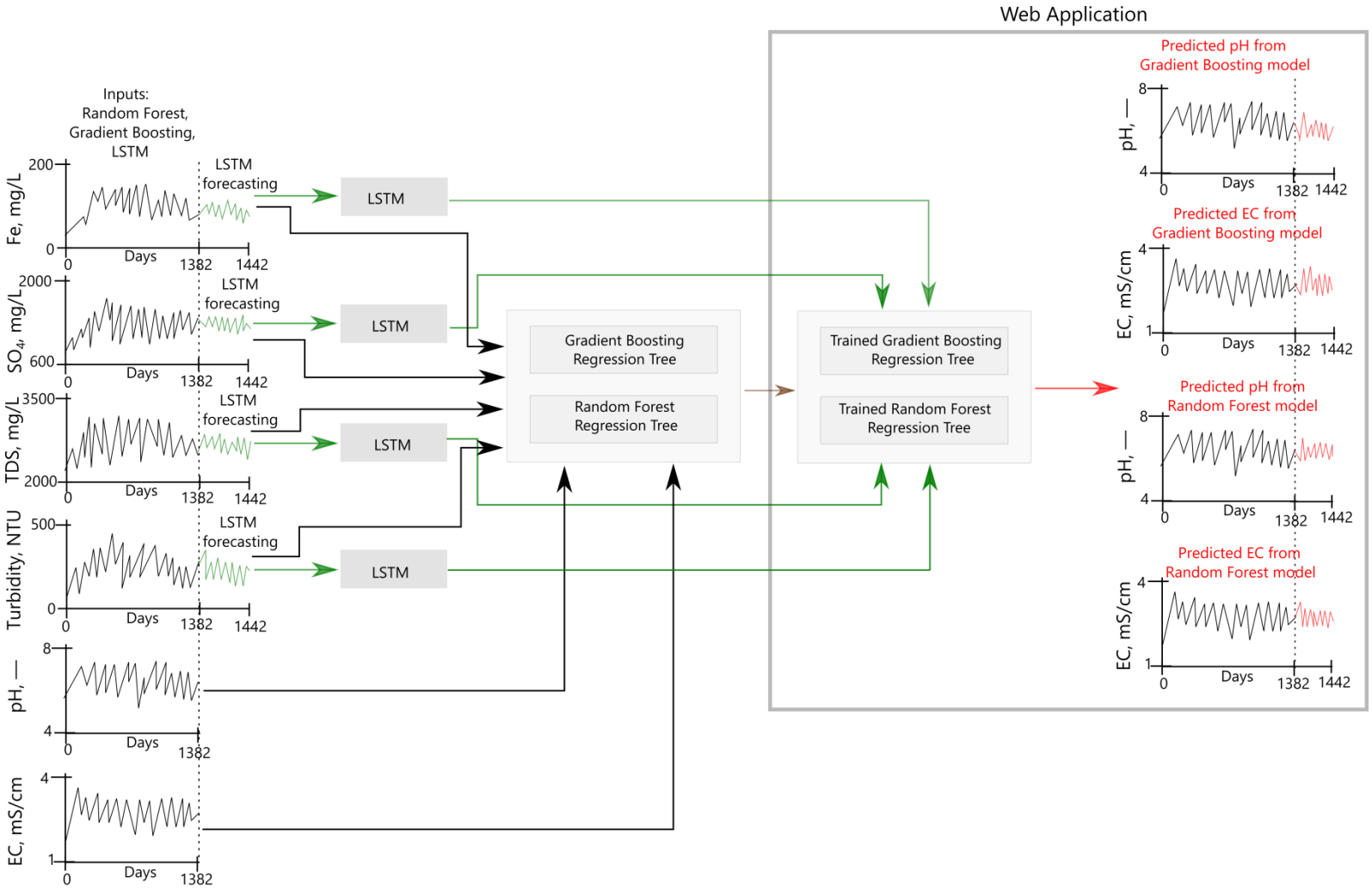


Figure 6.14: Electrical conductivity and pH prediction structure showing the connection between different algorithms used (modified and supplemented after Vadapalli *et al.*, 2020)

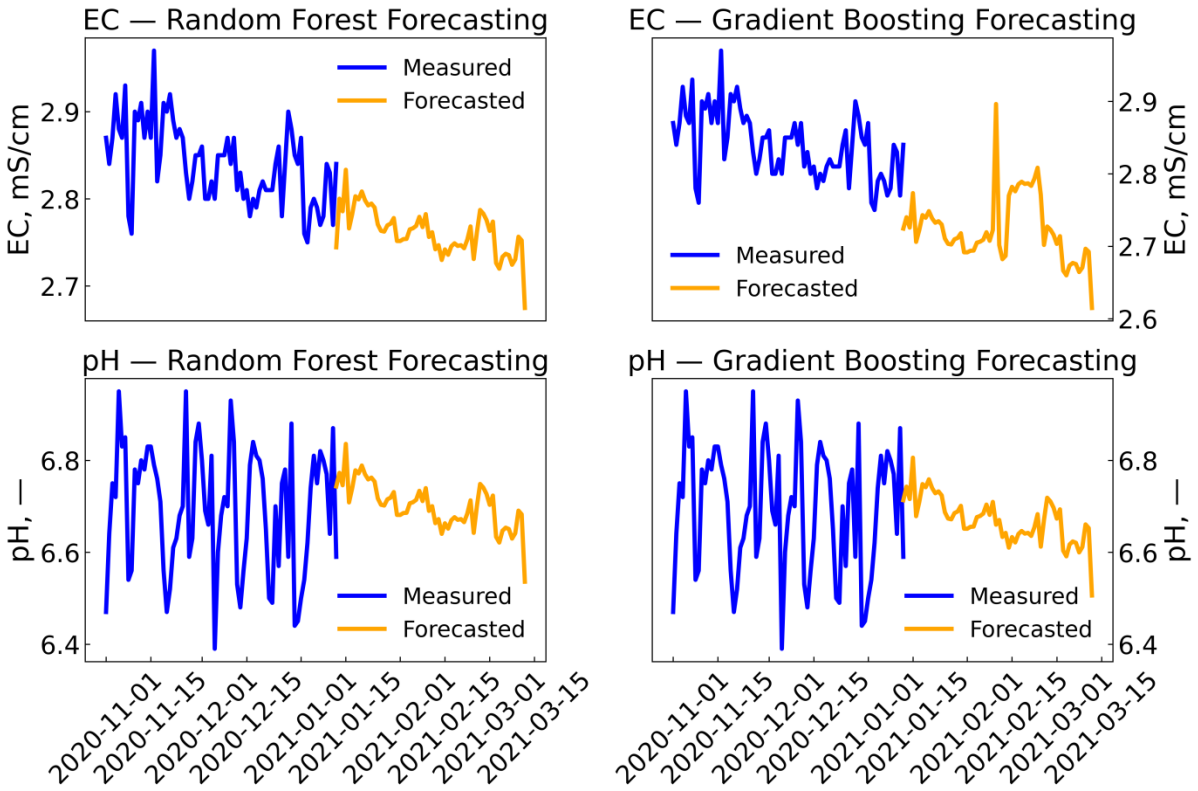


Figure 6.15: EC and pH forecasting visualisation using the web application

6.7 Conclusions

Random forest and gradient boosting algorithms yielded low error measures in forecasting EC and pH values using historical data. Therefore, the approach of using ML models is a scientifically important contribution in knowing the future quality of mine water. ANN's error measures were too high, making the algorithm not suitable for the MIW data. It should be noted that further hyper-parameter tuning for the ANN algorithm was not done to improve its performance, and therefore this study relied on the forecasting ability of random forest and gradient boosting algorithms. Predicted values of EC and pH with random forest and gradient boosting algorithms were not much different from the measured ones, giving much needed confidence and reliability in the prediction of mine water chemistry. In addition, the ANN model did not produce the best results, however, it can still be considered in other instances with extended datasets.

Incorporating ML models in mines can improve the current treatment practices, give a good mine water balance, and increase the effectiveness at which MIW is treated. Predicting the chemistry of MIW will ensure that the treatment plant operator knows in advance the quantity and type of chemicals and methods to use to treat and manage mine water. In mine water treatment, neutralising agents such as lime or quick lime are added to increase the pH for precipitating iron and other metals or semimetals. For mine water treatment plants it is therefore important to know the future development of the water parameters in order to stock for chemicals or to control the mine water level in a pumped system. This research proved that computational-intelligence techniques are much more effective than traditional dynamic modelling approaches. Therefore, the proposed approach can be an efficient tool and useful alternative for forecasting mine water quality parameters. Future research will focus on forecasting additional parameters such as Fe or SO_4 from other plants.

6.8 Remarks

It is crucial to understand that a machine learning model is only as good as the data it is fed. Researchers often publish articles without accurate data collection and not having performed thorough exploratory data analysis before deciding on suitable input parameters for the outputs to be predicted using ML models. In this study, data cleaning had to be performed before deciding on the input parameters to use. Some models performed better with good accuracies, others did not, and the forecasting trend did not take the entire shape of the historical data, but the results proved that ML models can be considered the future of mine water quality predictions. This implies that if the data were collected accurately from start to finish, without

equipment malfunctioning or missing observations, model performance would have been substantially better.

A mistake commonly done is to use accuracy to evaluate regression models (e.g. Khandelwal & Singh, 2005; Maier *et al.*, 2004; Rooki *et al.*, 2011). In ML, if the task is to predict a numerical value, then regression models can be used, and if the problem is dealing with discrete labels as a result then classification models can be applied. ML models need to be evaluated before they can be put in production. Researchers find it easy to use accuracy to evaluate ML model performance. However, accuracy only works when an observation is similar to a prediction; it focuses on whether the prediction is correct or not. In regression models, many predictions are not similar to the observations, and sometimes have low errors. Therefore, evaluation metrics recommended for regression models are RMSE and MAE as they can indicate how spread out the prediction error is, and can also identify the amount of errors in measurements.

The main aim to build ML models is to deploy them and make practical business decisions. In this study, a free cloud platform, Heroku, was used to show how ML model can be deployed or put in production. Several scientific articles only show how good the algorithms can perform, and never educate the readers on the deployment of ML models. Platforms such as Microsoft Azure, Amazon Web Services (AWS) Lambda, Google Cloud, and Algorithmia can be used to deploy ML models.

In the next chapter, another method for mine water quality forecasting at a different South African mine water treatment plant will be described. In this next chapter, neural networks and regression tree algorithms are compared, and the best performing model is used to perform final forecasting analysis. Furthermore, robust

statistical data analytics techniques (already discussed in CHAPTER 5) to handle the missing values and outliers are applied in this next chapter. Finally, the main aim of the next chapter is to address the important area of data science which is data analytics. In practice, data analytics makes up ≈80% of a machine learning related project, and thus proper statistical data analysis methods need to be understood.

CHAPTER 7

7 Data Analytics and Forecasting with Machine Learning Models

7.1 Introduction

Mining influenced water (MIW), especially acid mine drainage (AMD), but also circum-neutral and alkaline drainage, is a threat to the former mining areas in South Africa and worldwide (Wolkersdorfer *et al.*, 2022). It is therefore imperative that any mine water is treated before it can enter local water courses. Due to the temporal changes and the longevity of the mine water quality (Paul *et al.*, 2011; Younger, 1997), mine or treatment plant operators have to make provisions for the variability in the discharge qualities. If these changes can be quantified before they are about to occur, chemical stock, electricity or employee planning could be optimised to avoid over or under stocking or having too many personnel on site. This chapter, therefore, applied machine learning (ML) techniques to forecast MIW parameters for 60 days using Fe and acidity exemplified by the former Randfontein Estates gold mine located in Randfontein, South Africa (Westrand mine pool). Algorithms tested include the multi-layer perceptron (artificial) neural network (ANN), deep neural network (DNN), random forest and gradient boosting tree. Multivariate long short-term memory (LSTM) was used to generate new data for the best performing algorithm to forecast acidity and Fe of MIW. In the Westrand mine pool, MIW is pumped from shafts № 8 and № 9 to the treatment plant, and the main aim here is to forecast the mine water chemistry so that the plant operators can be prepared for changing water qualities ahead of time for optimal chemical dosing.

Usually, the traditional models fail to use all available parameters to forecast other parameters. These traditional models, such as auto regressive integrated moving

average (ARIMA) or Box-Jenkins, assume that time series data are linear processes (Khashei & Bijari, 2010; Zhang, 2003). Additionally, they forecast data of an individual time series by analysing the underlying data structure and using its patterns and trends. In many cases, real world scenarios are nonlinear (Zhang *et al.*, 1998), and thus, relying only on the traditional time series forecasting techniques is highly disadvantageous and would be inappropriate for time series datasets of MIW. Mine water parameters produce a nonlinear dataset, thus ML models were applied in this study. ML models such as the neural networks have gained overwhelming attention over the past years in nonlinear time series forecasting (e.g. Chen *et al.*, 2005; Jain & Kumar, 2007) and have yielded positive results. These techniques, including regression tree models, use the whole dataset's structure and analyse the relationships between the data of the parameters in the whole dataset to forecast the future patterns and trends.

In addition to applying ML models, thorough exploratory data analysis needs to be practiced to produce models that can forecast the data with accuracy and precision. Missing data and anomalies are frequently encountered while collecting MIW data, as this was the case with the data used for this study. Missing data compromise the statistical power of the study, while anomalies result in overfitting or underfitting of the models. Therefore, it is advantageous to apply suitable data interpolation and anomaly detection techniques on time series data before attempting to build forecasting models. In this chapter, numerical modelling was investigated and suitable techniques used to interpolate the missing data. Approaches investigated include the basis-spline (B-Spline) curves, non-uniform rational basis spline (NURBS) curves and wavelet transform as described earlier in this thesis (

CHAPTER 5). In case of anomalies, statistical profiling and predictive confidence level approaches were investigated.

7.2 Study design

7.2.1 Introduction

Machine learning models in this chapter were developed in a three-step process, i.e. initial forecasting using the multivariate LSTM model, choosing the best performing model by comparing the ANN, DNN, random forest and gradient boosting tree models and finally performing the final forecasting analysis (Figure 7.1). An LSTM model was used to forecast the values of alkalinity, pH and SO_4, and these forecasted values were supplied to the trained and tested best performing model to give the final concentrations of Fe and acidity for 60 days.

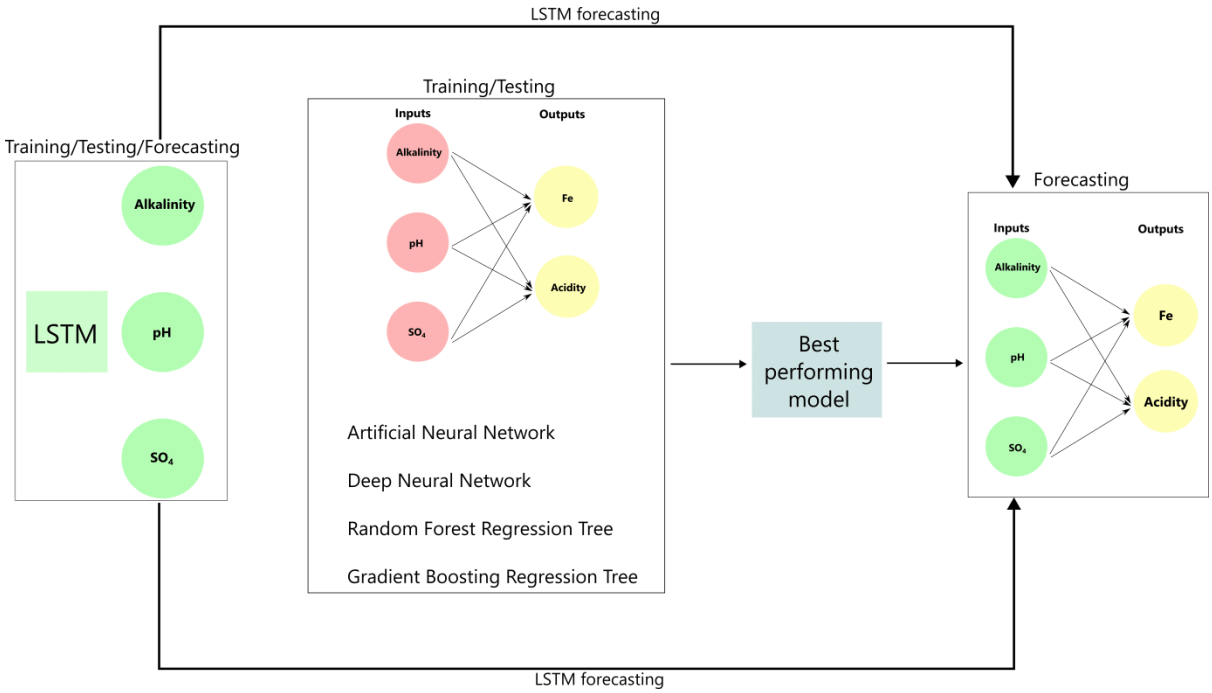


Figure 7.1: Machine learning mechanism for the Westrand mine pool water treatment plant data

7.2.2 Multivariate Long Short-Term Memory (LSTM)

Multivariate Long Short-Term Memory systems (LSTMs) are a special type of recurrent neural networks (RNN) and are mostly favoured because of the disadvantages that normal RNNs have. RNNs have no long-term memory, cannot use information from distant past, and cannot learn patterns with long dependencies (Manaswi, 2018). A way to overcome these issues is by introducing an LSTM which has memory cells that enable them to learn long-term patterns (Hochreiter & Schmidhuber, 1997; Manaswi, 2018). LSTM's default behaviour is remembering patterns and trends for a long period of time. They have a chain-like structure, similar to RNNs. However, the structure of the repeating module differs: for RNN, the repeating module is made up of a simple structure, such as a single tanh (hyperbolic tangent) layer, while LSTM's repeating module consist of four neural network layers interacting in a unique way (Figure 7.2).

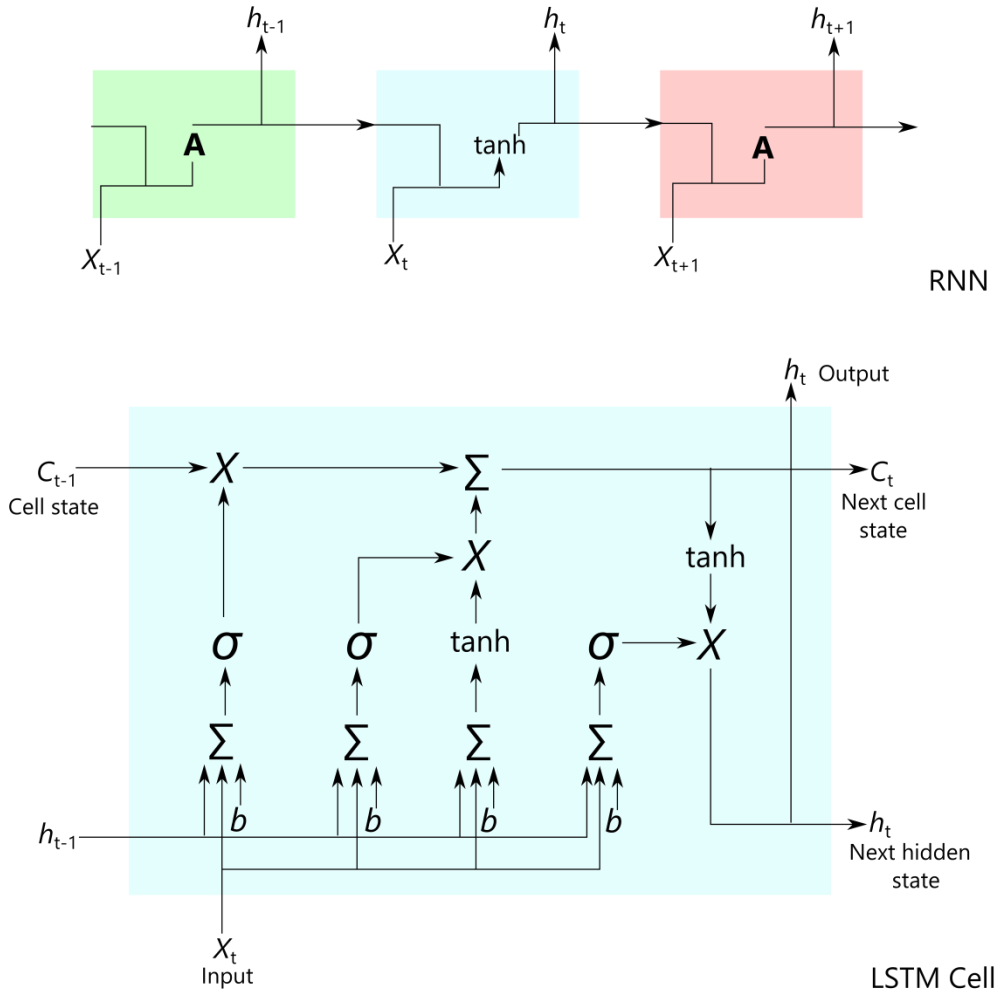


Figure 7.2: Mechanism structure for RNN (above) and LSTM cells (below). Inputs are labelled as X_t: current input, C_{t-1}: memory from last LSTM unit and h_{t-1}: output of last LSTM unit. Outputs are labelled as C_t: new updated memory and h_t: current output. Nonlinearities are labelled as σ: sigmoid layer and tanh: tanh layer. Vector operations are labelled as X: scaling of information and $\sum$: adding information; b: bias

7.2.3 Artificial Neural Network (ANN) and Deep Neural Network (DNN)

Artificial Neural Networks (ANN) and Deep Neural Networks (DNN) are from the same group of neural networks, but only differ by the number of hidden layers. A typical neural network is made up of the input layer, hidden layer and output layer, and if the network has one hidden layer then it is an ANN structure. A neural network that consists of two or more hidden layers is referred to as a DNN model. A fully connected layer in the neural network structure is practically composed of the

weights and the bias of each neuron, and the input size controls the number of weights. Each neuron has its own activation function (Chen *et al.*, 2005; Ekemen Keskin *et al.*, 2020; Khashei & Bijari, 2010; Maier *et al.*, 2004; Øyen, 2018). An input layer introduces values into the network, and it has no activation function. Hidden layer(s) perform the network's computations. Furthermore, the number of neurons in the input layer depends on the parameters that will be used in the network as inputs. An output layer makes final prediction for the network, and its neurons depend on the parameters that need to be forecasted, while hidden layers can have any number of neurons stacked together. Hidden and output layer neurons have an activation function such as the sigmoid, rectified linear unit (ReLU) or softmax (normalised exponential function).

7.2.4 Regression tree algorithms

Regression trees are one of the key algorithms used in complex structures such as mine water dataset. They are useful when the data has no obvious linear relationship between the input and output parameters. In this study, random forest and gradient boosting regression trees were used. Decision trees make the foundation of both the algorithms. Regression tree algorithms have a tree-like structure which consists of root node, branches and leaf nodes (Figure 7.3; Biau & Scornet, 2016; Zhang & Haghani, 2015). A random forest algorithm aims to reduce the variance in complex trees while gradient boosting aims to decrease the bias.

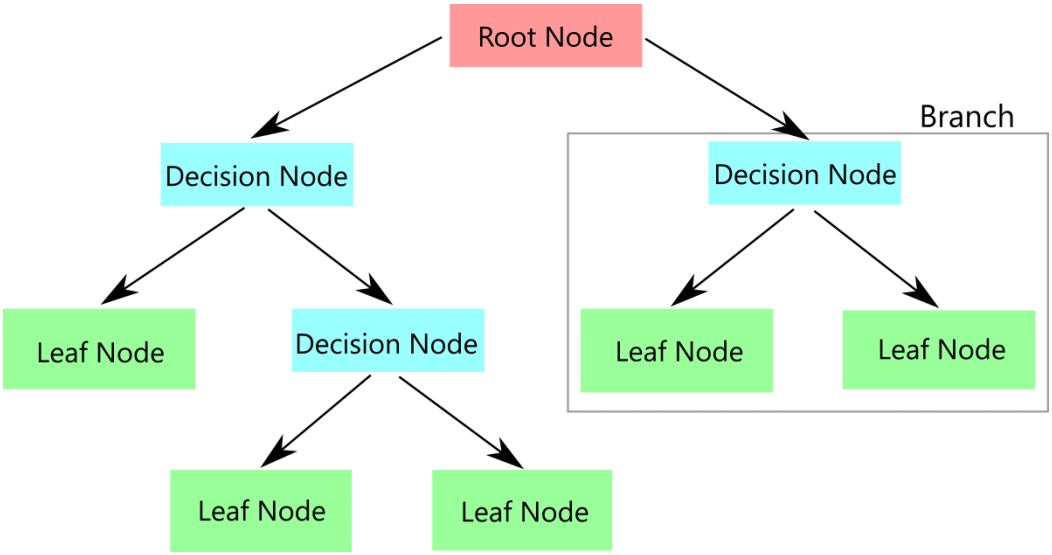


Figure 7.3: Components of regression tree structure

7.3 Dataset

7.3.1 Background

Sampling and monitoring of mine water is commonly done on a regular basis to carefully examine its chemistry. This study uses the South African Westrand mine water treatment plant's data gathered between 2016-03-07 and 2021-07-03. This data contains nine parameters, i.e. acidity, alkalinity, electrical conductivity (EC), Fe, Mn, pH, SO_4, temperature and turbidity, which were used in the units applied by the plant operators. These parameters do not have equal numbers of observations i.e. some of the measurements are missing. The highest number of observations of an individual parameter is 1123 (Table 7.1). Exploratory data analysis which included data interpolation and anomaly detection was conducted to prepare the data to be used in the machine learning models.

Table 7.1: Mine water quality dataset from shaft № 9 of the gold mine in Randfontein; *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

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	28.4	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
Temperature, °C	1123	19.7	2.1	9.5	26.2
Turbidity, NTU	1116	23.0	36.2	0.7	275.0

7.3.2 Stationarity test

Stationarity testing is critical as several statistical applications and models are computed based on its results. For example, numerical models are often applied to time series data when it is non-stationary, and probabilistic models are sometimes useful and accurate when the dataset is stationary (Wackerly *et al.*, 2014). Data interpolation and anomaly detection approaches are guided by stationarity tests. Therefore, the Augmented Dickey Fuller (ADF) test was used on the dataset to test the stationarity of the time series (Table 7.2). A stationary dataset will have a *p*-value that is highly significant (<0.05). Additionally, scatter plots for the dataset were drawn to visualise the patterns, trends and seasonality, and identify any stationarity or non-stationarity properties of the data (Figure 7.4).

From the statistical results, the *p*-value obtained for temperature is less than 0.05. Therefore, the null hypothesis is rejected, i.e. temperature time series is stationary.

However, the graphs show that temperature data is oscillating, a prominent seasonality can be observed in series. For the other parameters, the *p*-value for the time series is greater than 0.05. Therefore, the null hypothesis is accepted, i.e. the dataset for the parameters is non-stationary. Finally, the statistical results show that numerical modelling can be used to interpolate missing measurements and detect anomalies.

Table 7.2: Augmented Dickey Fuller test application on dataset to test for stationarity. № lags: number of lags, *n*: number of observations used for ADF regression and critical values calculation. Critical values at α = 0.01, 0.05 and 0.10 (–3.437, –2.864 and –2.568)

Parameter	ADF Statistic	*p*-value	№ lags	*n*
Acidity	–1.474	0.546	6	981
Alkalinity	–1.572	0.498	4	983
EC	–1.953	0.307	18	969
Fe	–1.580	0.493	11	976
Mn	–1.548	0.510	6	981
pH	–2.210	0.202	7	980
SO_4	–1.440	0.563	8	979
Temperature	–3.238	0.018	16	971
Turbidity	–2.739	0.068	22	965

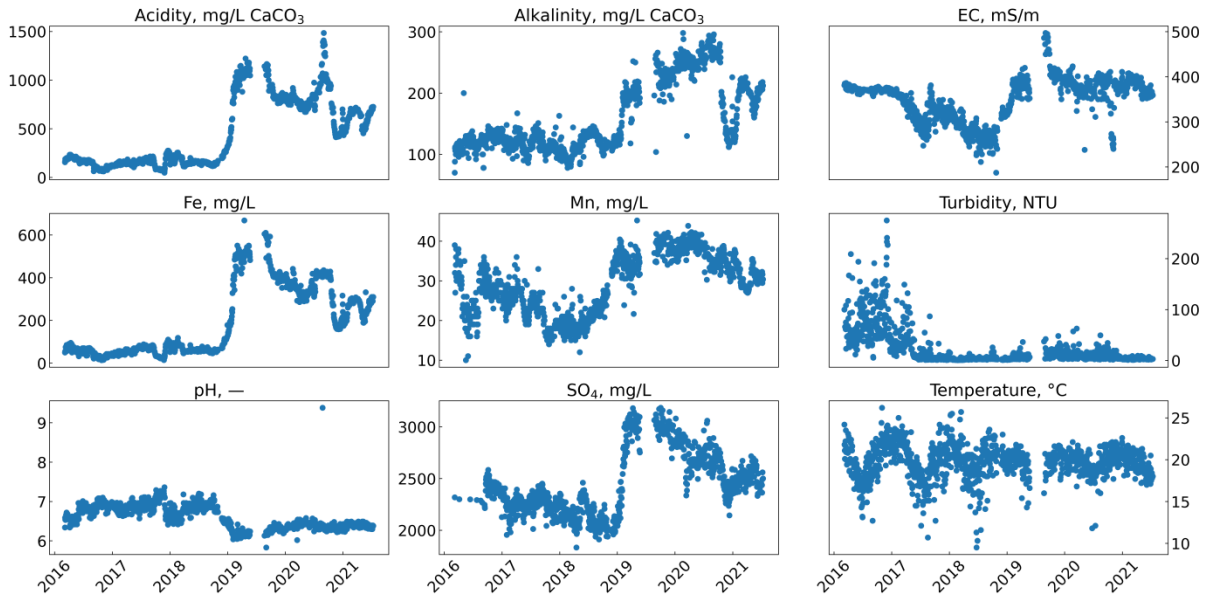


Figure 7.4: Temporal mine water data development of Shaft № 9 in the Western Basin of the Witwatersrand mines from 2016 to 2021

7.3.3 Normality test

Normality tests identify if the data has been sampled from a normal distribution. When data are plotted on a frequency distribution, the normal distribution can be explained by a bell-curve shape with majority of the observations being around the mean value. From the graphs plotted, the data are log-normally, bi-modally and multi-modally distributed, and only temperature shows a close to Gaussian distribution (Figure 7.5). Mining influenced water data are continuous data, thus the normality test is a crucial process for deciding statistical methods and measures of central tendency to perform data analysis. Apart from graphical methods, there are several statistical techniques applied to test for normality of data.

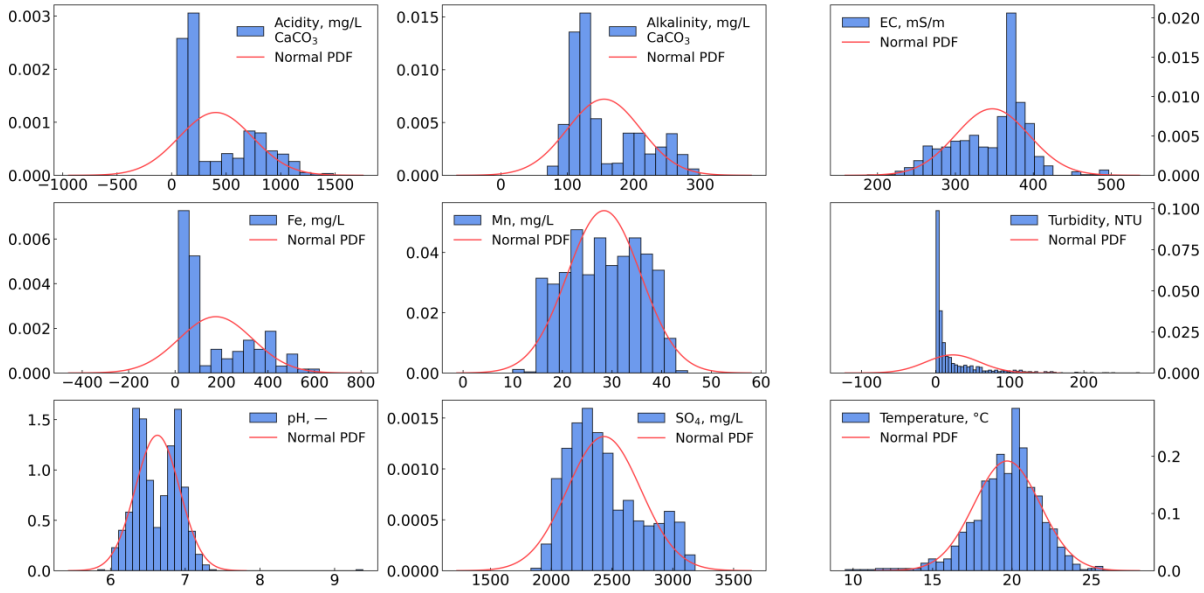


Figure 7.5: Histograms with fitted normal probability distribution function (PDF) curves. Normal PDF curves were drawn using the SciPy 1.7.0 module by utilising the mean and standard deviations of the data

In this study, SciPy module version 1.7.0 was used to compute the Kolmogorov-Smirnov, Shapiro-Wilk, and Anderson-Darling normality tests (Table 7.3). Statistical tests conducted using the Kolmogorov-Smirnov and Shapiro-Wilk methods show *p*-values for the parameter to be below the 5% significance level, meaning that the data do not follow a normal distribution. Using the Anderson-Darling test, the test statistics are well above the critical values at $\alpha = 0.15$, 0.10, 0.05, 0.025 and 0.01 (0.574, 0.654, 0.784, 0.915 and 1.088), which also implies that the data do not follow a normal distribution.

Table 7.3: Normality tests using the Kolmogorov-Smirnov, Shapiro-Wilk and Anderson-Darling tests; *n*: number of observations. Critical values for α = 0.15, 0.10, 0.05, 0.025 and 0.01 (0.574, 0.654, 0.0784, 0.915 and 1.088) (Anderson-Darling)

		Kolmogorov-Smirnov		**Shapiro-Wilk**		**Anderson-Darling**
Parameter	*n*	Test statistic	*p*-value	Test statistic	*p*-value	Test statistic
Acidity	1123	1.00	0.000	0.819	0.000	89.1
Alkalinity	1123	1.00	0.000	0.861	0.000	66.9
EC	1123	1.00	0.000	0.946	0.000	27.9
Fe	1111	1.00	0.000	0.815	0.000	89.2
Mn	1111	1.00	0.000	0.967	0.000	10.4
pH	1123	0.99	0.000	0.935	0.000	16.0
SO_4	995	1.00	0.000	0.942	0.000	18.5
Temperature	1123	1.00	0.000	0.972	0.000	5.4
Turbidity	1118	0.90	0.000	0.626	0.000	147.5

7.3.4 Data transformation

As has been shown in the previous section, the dataset is not normally distributed. Several statistical methods, especially time-series forecasting techniques, are based on the assumption that the data are normally distributed. Thus, building forecasting models with un-transformed data often results in inaccurate forecasting results. Therefore, data transformation is taking data that are not normally distributed and transforming it to a close-to normal distribution (Azzalini & Capitanio, 1999; Manikandan, 2010; Xie *et al.*, 2000; Zhang *et al.*, 2017). Data transformation does not change the relationship of the variables for mathematical and statistical

purposes. So, the procedure is a statistically necessary step towards building models that can forecast data with accuracy and precision.

From the results, it can be seen that alkalinity, EC and Mn have a multi-modal distribution, and acidity, Fe, pH and SO_4 show a bi-modal distribution. Temperature shows a close-to normal distribution and turbidity has a log-normal distribution; therefore, only turbidity was transformed. There are several ways to transform the data, such as log-normal, square-root, reciprocal, or Box-Cox transformation. In this study, a natural log-normal transformation was used: each variable of *x* was replaced by ln(*x*). Finally, quantile-quantile (Q–Q) and histogram graphs were plotted to test the normality of the transformed turbidity data (Figure 7.6). In a Q–Q plot, for a normally distributed data, observations lie approximately on a straight line. Therefore, the graphs show that turbidity, after being transformed, is close-to a normal distribution, slightly showing a bi-modal distribution.

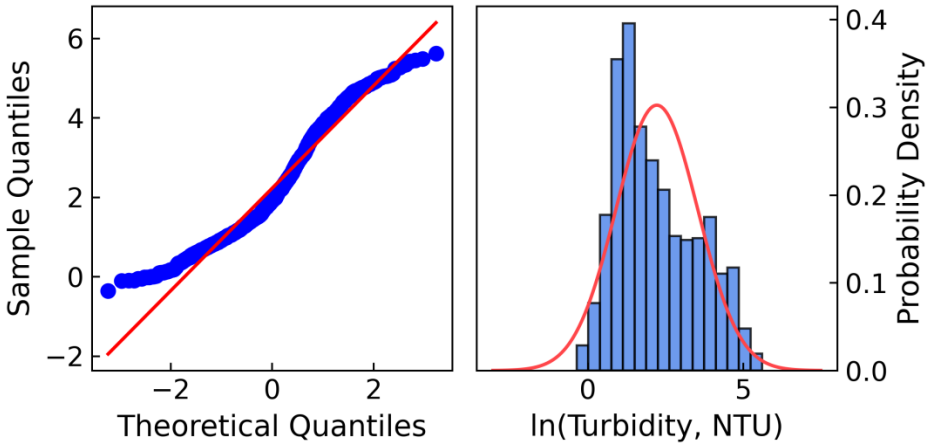


Figure 7.6: Q–Q plot (left) and histogram with a normal PDF curve (right) for the transformed turbidity data

7.4 Anomaly Detection and Data Interpolation

The dataset used contains sample times with missing measurements resulting from equipment malfunctioning or because no sample was taken, and has outliers

according to the statistical analysis conducted in the previous sections of this chapter. There are several statistical methods to interpolate missing values and detect anomalies for time series data, some of which have been discussed in this thesis. Predictive confidence level approach fitting nonlinear and linear regressors was used to detect anomalies for this dataset. It should be noted that only nonlinear regressors (gradient boosting regression models) were relied on to detect anomalies. Linear regressors (robust linear regression models) and an additional predictive confidence level approach (long short-term memory autoencoder) were used only to show the readers that there are other methods to detect anomalies.

In the approach of fitting a nonlinear regressor, Python's Scikit-learn 0.24.2 library was used to build the gradient boosting regression model. This model was fitted on the data with three prediction bands of α = 0.10, 0.50 and 0.90 (Figure 7.7, Annexure A). The $\alpha = 0.10$ prediction band represents the lower bound of the data (real observations plotting below this band are considered possible anomalies), while the α = 0.90 prediction band represents the upper bound of the data (real observations plotting above this band are considered possible anomalies). For the mid-prediction, alpha was set to 0.50 and this predicts the median of the original data. Therefore, values plotting on the α = 0.50 prediction band were used to replace the possible anomalies.

For the build-up of gradient boosting regression model, for each target output all other parameters were used as input variables, e.g. when acidity was set as the target output, alkalinity, EC, Fe, Mn, pH, SO_4, temperature and turbidity were set as input variables. This model used 100 trees, a maximum tree depth of 2 and learning rate of 0.05 to perform predictions, and the data were split into training (80%) and testing (20%) sets. Gradient boosting model's objective was to predict the values of

the parameters that will possibly be used to build the forecasting models. There are several hidden features, some of which are not included in the data, which affect each parameter. Therefore, the uncertainty in the estimates was shown by predicting the lower ($\alpha = 0.10$), middle ($\alpha = 0.50$) and upper ($\alpha = 0.90$) bands of the observations. The loss function of the gradient boosting model was changed to quantiles with selected prediction intervals (alphas). This configuration ensures that the model performs predictions which correspond to percentiles.

In addition to showcase more methods that can be applied to detect anomalies, robust linear regression was used. It should be noted that linear regressors are only suitable for linear data, thus the results from this approach were not relied on to make final conclusions. In the linear regressor approach, robust linear regression with upper and lower 95% confidence regions and prediction bands were computed. Observations that lie outside the 95% prediction bands were regarded as possible strong anomalies (Figure 7.8, Annexure B).

In addition to support the predictive confidence level approaches, deep learning techniques, LSTMs and Autoencoders, were used to detect anomalies on the dataset. Again, this is an additional method to educate the readers about robust ways to detect anomalies. The results from this approach were not used for this study.

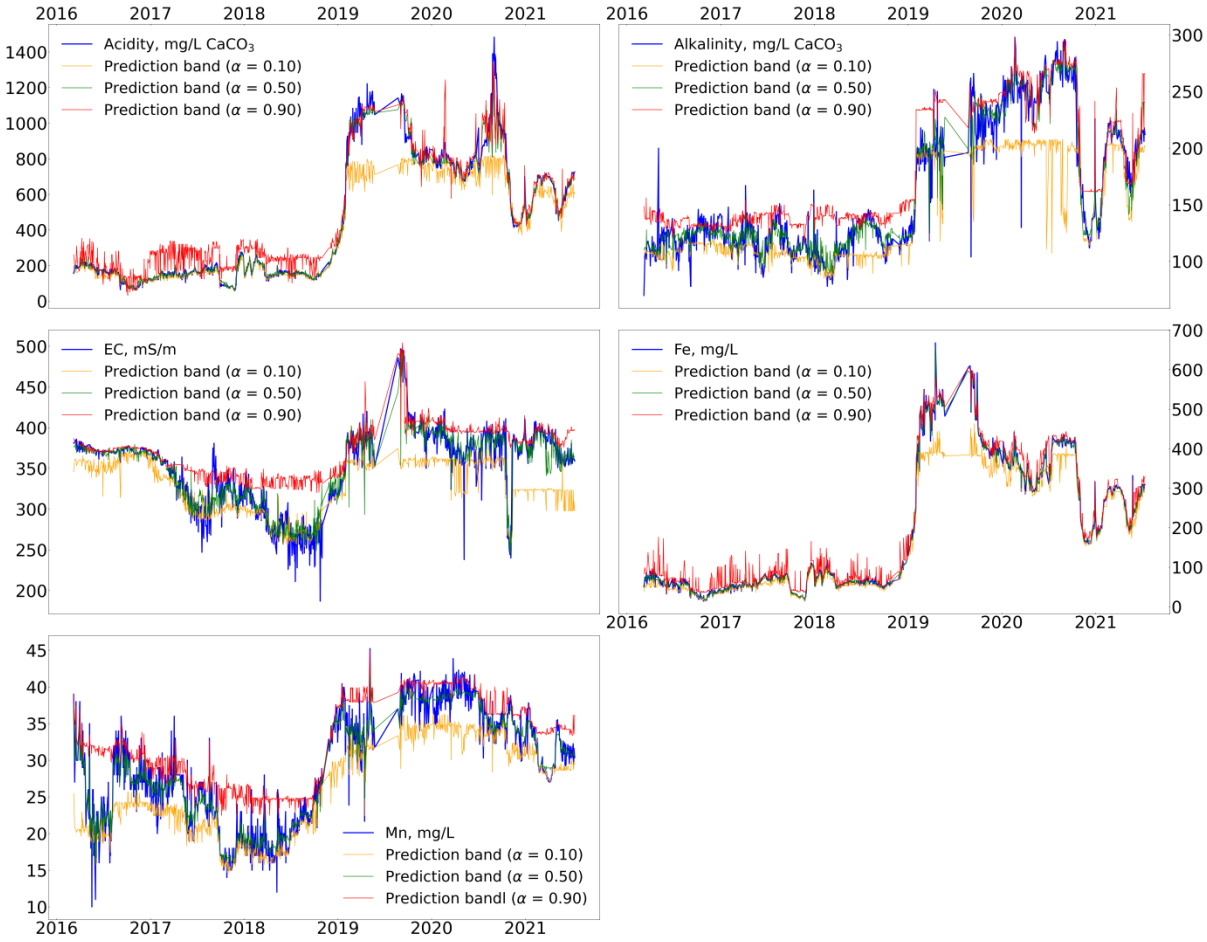


Figure 7.7: Gradient boosting regressors with prediction intervals of α = 0.10, 0.50 and 0.90 to detect anomalies in mining influenced water data. Only showing graphs for acidity, alkalinity, EC, Fe and Mn due to the data size

Autoencoder is a type of neural network that consists of an encoder and a decoder sub-model (Skansi, 2018). It is used to study hidden patterns and trends of raw data. In the LSTM Autoencoder model, input values are downscaled, captured as a latent vector, and decoded. LSTM Autoencoder is made up of five layers: input sequence, encoder, latent vector, decoder and reconstructed sequence (Manaswi, 2018; Michelucci, 2018; Skansi, 2018). The models for each parameter were built using Python version 3.7.1 within the Anaconda platform, with Tensorflow version 2.5.0 used as the backend and Keras version 2.4.3 as the core model development module.

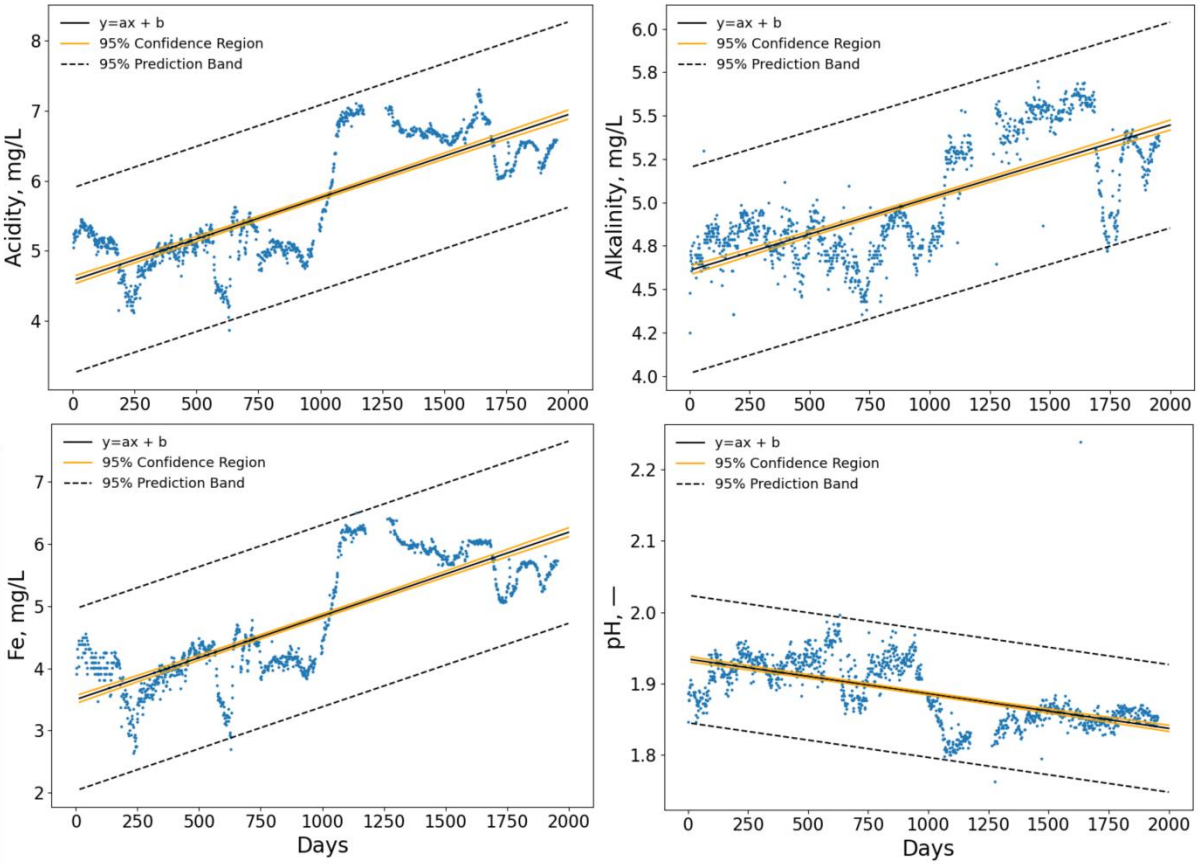


Figure 7.8: Robust linear regression with upper and lower bands of 95% confidence and prediction levels. Day 1: 2016-03-07 and day 1955: 2021-07-13

For each parameter, 90% of the observations were used for the training set while the remaining 10% were used for testing the model performance. The LSTM Autoencoder structure used in this chapter contained 16 memory units. 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 the 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. LSTM Autoencoder model was fitted over 30 epochs with a batch size of 32 and validation split of 10%, and was further compiled using the adaptive moment estimation (Adam) optimiser and mean absolute error (MAE) for calculating the loss

function. Finally, the models produced a small MAE value which implies that they performed relatively well (Figure 7.9, Annexure C).

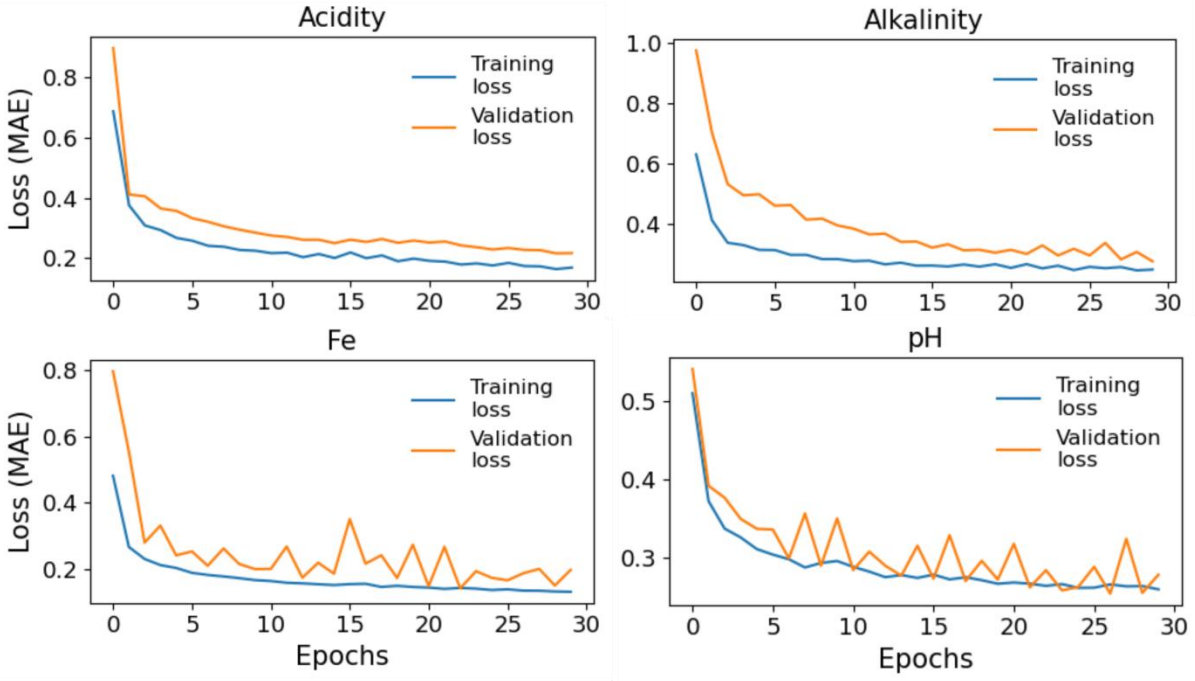


Figure 7.9: Performance of the LSTM Autoencoder model for each MIW data parameter to detect anomalies

After examining the loss function in the training and testing sets, a suitable threshold value for identifying anomalies was computed. Therefore, the reconstruction loss in the training and testing sets (Annexure D) were computed to determine when the observations of the parameters cross the anomaly threshold (Annexure E). Overall, the models show that the data has anomalies, same as the nonlinear gradient boosting regression and robust linear regression methods conducted.

Numerical analysis modelling was applied to interpolate the missing measurements. Numerical analysis assumes that the time series data trend and pattern represent an unknown function (Wackerly *et al.*, 2014). The main task when applying this technique is to find a suitable function for the data so it can be used to interpolate the missing values. B-Splines interpolation technique, using Python's SciPy library, was

used in this study (Annexure F). B-Splines interpolation is a form of interpolation where a continuous curve has various piecewise polynomials whose gradients match up at the measured data (section 5.3.2). Interpolation in this form takes place between two points that have missing values, i.e. a function is derived between the points and interpolation takes place. The polynomials are continuous up to their second derivative, and this process occurs for every paired-point that has missing values between them (Averbuch *et al.*, 2014; Habermann & Kindermann, 2007). Finally, the interpolated values fitted well into the known population, and these interpolated values did not change the trends and patterns of the original data (Figure 7.10). This implies that B-Spline interpolation is an accurate technique for non-stationary time series data.

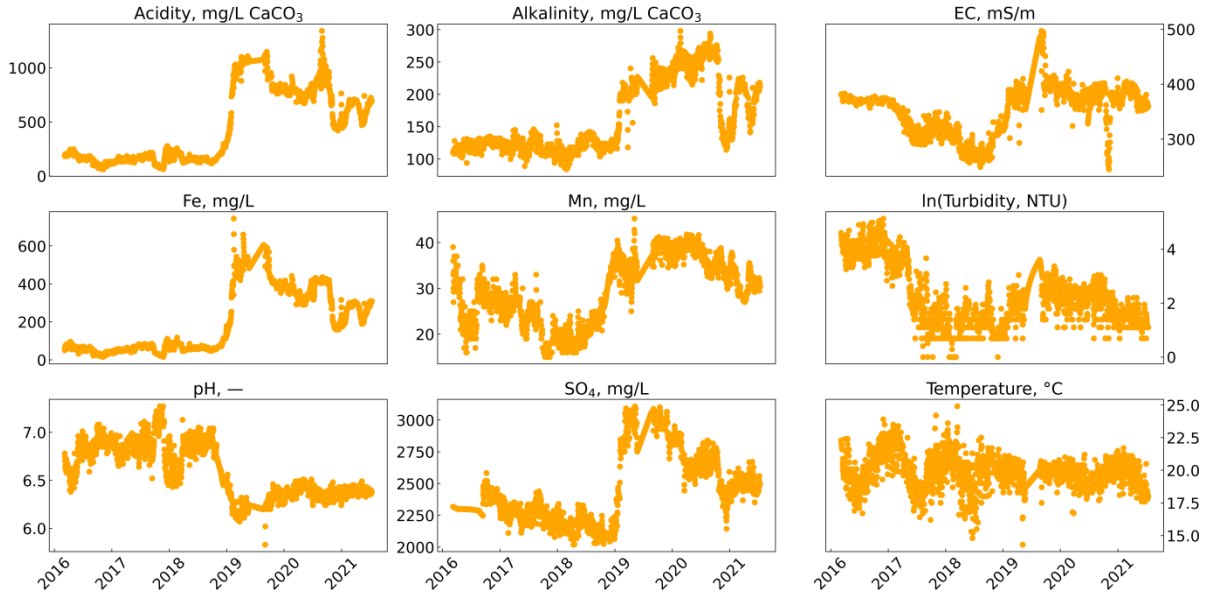


Figure 7.10: Temporal mine water data development of Shaft № 9 in the Western Basin of the Witwatersrand mines from 2016 to 2021 including interpolated data

7.5 Modelling Data

Data analytics approaches were conducted to detect anomalies and interpolate the missing observations by robust statistical techniques described in the previous

sections, which provided a “clean” dataset that could be used for modelling. A correlation chart of nine parameters, all with 1955 observations, was computed using R 4.1.2 on RStudio 1.1.456 (Figure 7.11). Parameters of concern at the Westrand mine water treatment plant are acidity and Fe. Therefore, their relationships with other parameters were thoroughly examined. The statistical analyses show that both acidity and Fe have high correlations with alkalinity, EC, Mn, pH and SO_4. Additionally, alkalinity, pH and SO_4 have good relationships with acidity and Fe with correlation coefficients (*r*) above 0.8. Therefore, alkalinity, pH and SO_4 were used as input parameters for final forecasting, with acidity and Fe being the target outputs. Turbidity and temperature were not used to build the forecasting models because they have a poor relationship with the rest of the parameters, as displayed by the correlation coefficients.

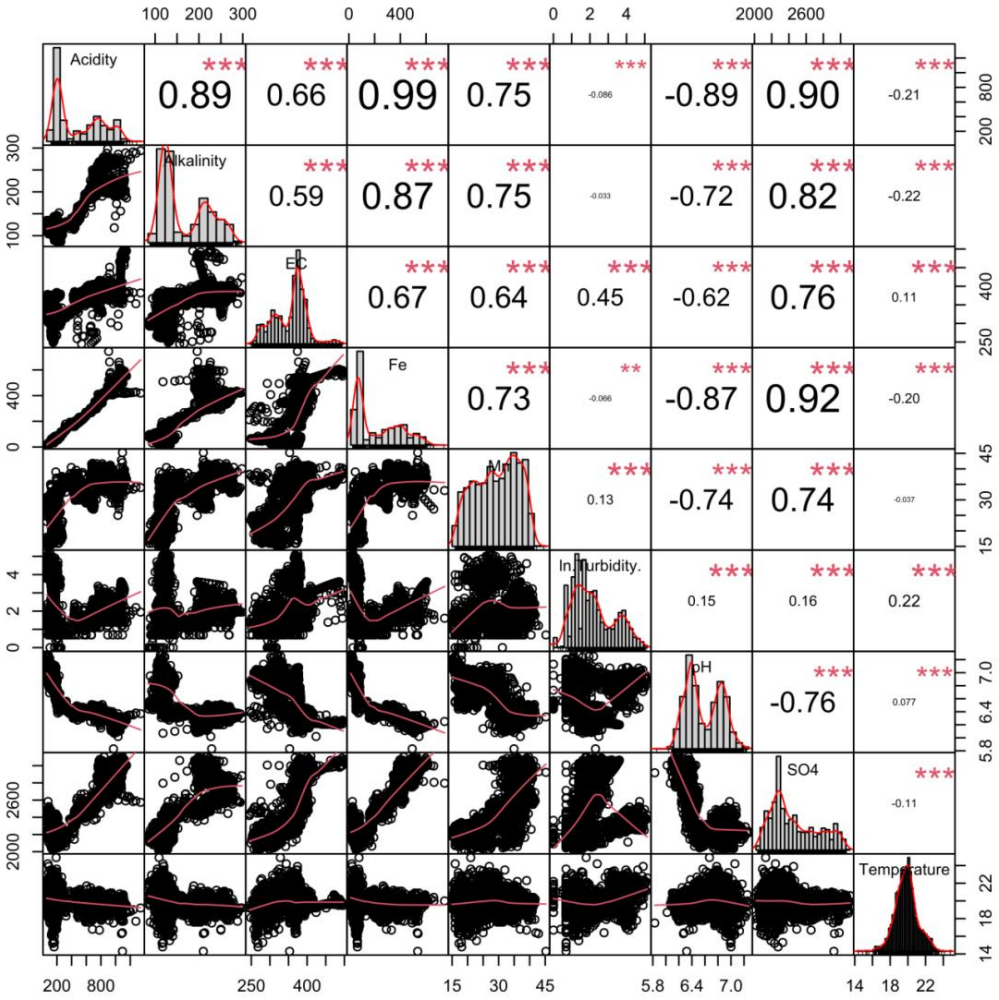


Figure 7.11: A correlation chart with the distribution of each parameter shown on the diagonal; on the bottom of the diagonal are the bivariate scatter plots with a fitted line; on the top of the diagonal are the values of the correlation with the significance levels shown as asterisks: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$; correlation coefficient font size is related to the relationship between the parameters – the stronger the relationship the larger the font size and vice versa

7.6 Model Development and Evaluation

All the models were trained and tested independently and only the best performing model was used in the project to create the final forecasting. The forecasted concentrations and values of alkalinity, pH and SO_4 by the LSTM model were supplied to the best performing model to forecast Fe and acidity concentrations. For the LSTM model, three structures were developed with different numbers of the

“lookback period” which explains the number of previous time-steps the model needs to use to predict the subsequent time-step, epochs, and validation split (Table 7.4). A single hidden layer multivariate LSTM model with 32 memory units based on the structure variations explained was compiled, and a ReLU activation function was used throughout. For all the structure variations, the models were fitted with a batch size of 32. The adaptive moment estimation (Adam) optimiser and MSE loss function were used in this model, and the low validation and training loss implies good model fitting on the new and training data (Table 7.5).

Table 7.4: Structure variations of the multivariate LSTM model

Model type	Lookback period	Epochs	Test size
Structure variation 1	150 days	30	15%
Structure variation 2	250 days	20	20%
Structure variation 3	300 days	50	25%

Different techniques were applied to tune the hyper-parameters for the neural network models, including grid search optimisation technique, keras tuner and “trial and error” method. For the ANN model, the configuration that yielded better results was a model with input layers that consisted of three neurons of alkalinity, pH and SO_4, while Fe and acidity were used in the output layer with a ReLU activation function. The model had one hidden layer of four neurons and a ReLU activation function. The DNN architecture consisted of two hidden layers, the first one with four neurons and another with two neurons, both with a ReLU activation function. Like ANN, input parameters were alkalinity, pH and SO_4, with target outputs being Fe and acidity. Both the neural network models used a test size of 20%, and the models were compiled using the Adam optimiser. They were then fitted for 500 epochs with a batch size of 64. The model performances were tested using the MSE and MAE

accuracy metrics (Table 7.5). A random forest regression tree model was built using 150 trees and a maximum depth of 8, with a test split of 20%. A gradient boosting regression tree model was compiled using 100 trees, maximum depth of 5 and learning rate of 0.05, with a test size of 20%. Regression tree model performances were also evaluated using MSE and MAE (Table 7.5).

Table 7.5: Performances (MSEs and MAEs) for the multivariate LSTM, ANN, DNN, random forest and gradient boosting models; RF: random forest, GB: gradient boosting

	LSTM1	LSTM2	LSTM3	ANN		DNN		RF		GB	
	MSE	MSE	MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Training	0.0532	0.0485	0.0481	0.0337	0.1308	0.0342	0.1313	0.0094	0.0620	0.0236	0.1039
Testing	0.0851	0.0790	0.0592	0.0258	0.1190	0.0279	0.1234	0.0192	0.0798	0.0254	0.1050
Comment	Good performance.			Data leaking. Not a good generalisation model.		Data leaking. Not a good generalisation model.		Good performance.		Acceptable performance.	
Decision				Do not use.		Do not use.		Use.		Do not use.	

7.7 Results and Discussions

Machine learning models developed in this study consisted of a multivariate LSTM model which was used to forecast alkalinity, pH and SO_4 for 60 days using three different structure variations (Figure 7.12). Furthermore, ANN, DNN, random forest and gradient boosting models were trained and tested using historical data. Random forest displayed an overall best performance and was used to forecast Fe and acidity concentrations for the same period as the LSTM model (Figure 7.13). The multivariate LSTM forecasting has three different forecasted concentrations for alkalinity and SO_4 as well as the pH values, which then resulted in the random forest model also forecasting three different concentrations for Fe and acidity. With random forest using alkalinity, pH and SO_4 as input parameters, the forecasted

concentrations and values by the LSTM model were fed to the trained random forest model. Therefore, the random forest model showed good performance and forecasted Fe and acidity concentrations with accuracy as the values fall within the population and follows the trend. In addition, the results suggest that ML models can be widely applied in mine water time series forecasting analysis.

The models were evaluated using MSE and MAE, and in all occasions, the error was substantially low. However, the neural network models' training loss was always greater than the validation loss, which meant that there is data leaking in the training process. Hyper-parameters were rigorously tuned using grid search, keras tuner and "trial and error" approach, but training loss continued to be slightly greater than the validation loss. Therefore, the final decision was that the models cannot be used for final forecasting analysis. Regression tree models displayed better performance, with random forest performing better than the gradient boosting. Thus, random forest was used to perform final forecasting analysis.

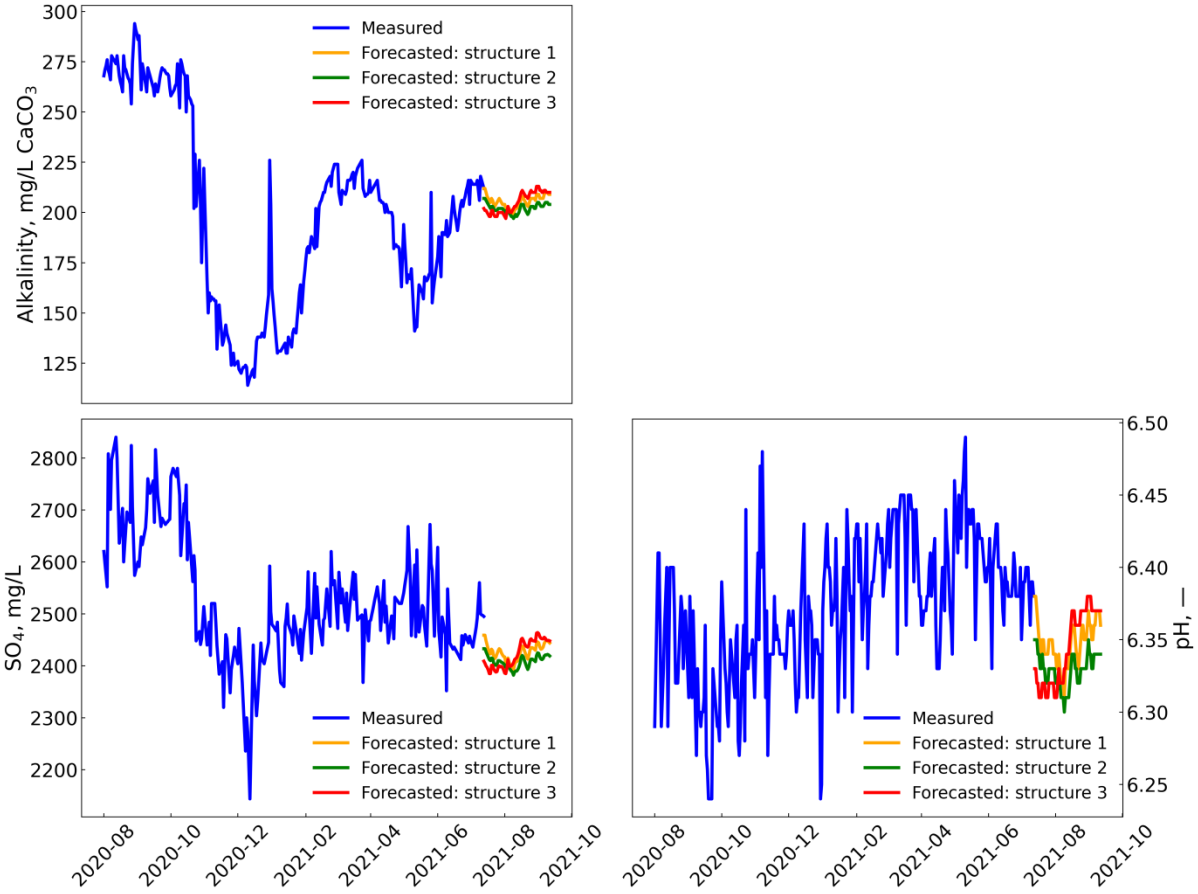


Figure 7.12: Forecasted concentrations of alkalinity (top), SO_4 (bottom left) and pH values (bottom right) using a multivariate LSTM model. Historical data was only used from 2020-08-01 to allow better visualisation of the forecasting results

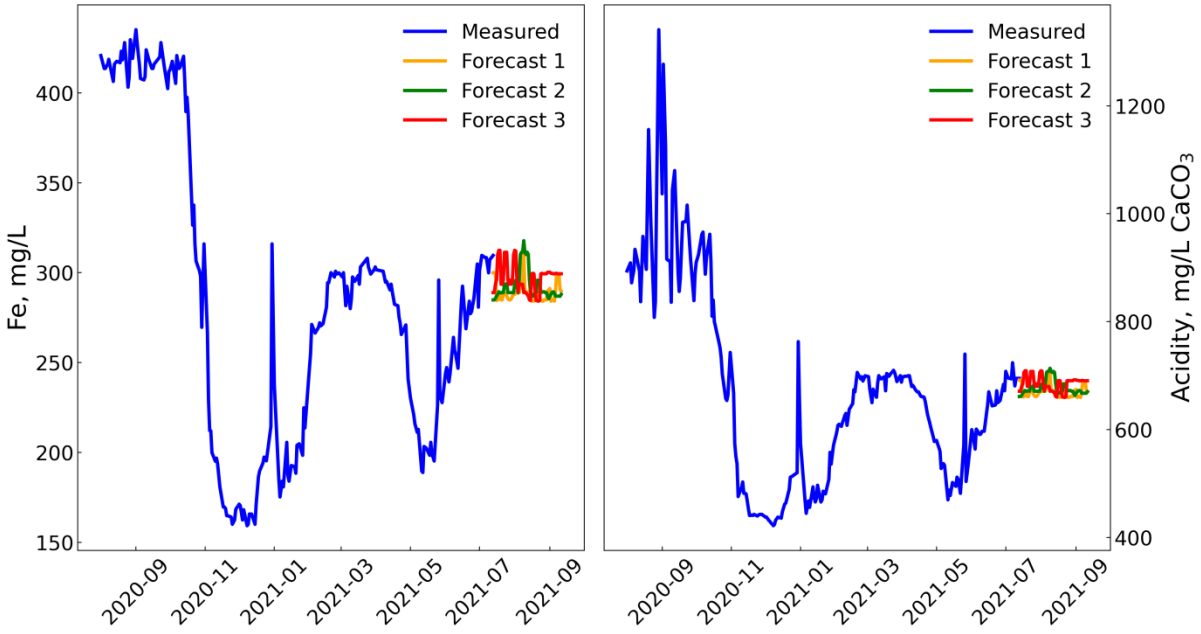


Figure 7.13: Forecasted concentrations of Fe and acidity using the random forest model. Historical data was only used from 2020-08-01 to allow better visualisation of the forecasting results

7.8 Comparing the Measured and Forecasted Data

Forecasted concentrations of Fe and acidity using random forest model were compared with the measured data by calculating the forecasting error (Table 7.6). This procedure was also done for the forecasted input parameter values by the multivariate LSTM model (Annexure G). Measured data only contains 23 observations while the forecasting period was for 60 days. This is because sampling was not carried out daily for this period, thus the forecasting error was only calculated for the available measured data. Furthermore, cross plots of measured and forecasted data with a robust regression lines were computed (Figure 7.14). Computed plots and calculations show both the low coefficients of determinations (r^2) and statistical significances. However, the calculated forecasting errors are relatively low with only four notable higher errors. The reason for such differences may be because of the sampling that was not conducted daily at the treatment plant.

Table 7.6: Forecasting error analysis for the random forest model. Error is calculated as: |[(measured–forecasted)/(measured)]×100%|

Measured Fe, mg/L	Forecast 1	|Error, %|	Forecast 2	|Error, %|	Forecast 3	|Error, %|	Measured Acidity, mg/L $CaCO_3$	Forecast 1	|Error, %|	Forecast 2	|Error, %|	Forecast 3	|Error, %|
309.6	300.0	3.1	284.9	8.0	288.9	6.7	726	692	4.7	661	8.9	671	7.6
307.2	299.4	2.5	285.6	7.0	293.5	4.4	724	691	4.6	663	8.4	680	6.1
306.4	286.9	6.4	288.9	5.7	312.4	2.0	724	667	7.8	671	7.3	709	2.1
301.6	284.9	5.6	289.1	4.1	293.5	2.7	724	661	8.7	672	7.2	680	6.1
304.0	288.7	5.0	293.5	3.4	311.4	2.4	728	671	7.8	680	6.6	708	2.8
296.8	285.6	3.8	288.9	2.7	295.8	0.3	702	663	5.5	671	4.4	685	2.4
263.2	284.9	8.2	288.9	9.8	293.5	11.5	662	661	0.1	671	1.4	680	2.7
307.8	284.9	7.5	288.9	6.1	295.8	3.9	742	661	10.9	671	9.5	685	7.7
304.0	286.9	5.6	288.9	5.0	295.8	2.7	708	667	5.7	671	5.2	685	3.3
289.6	288.9	0.2	295.8	2.2	312.0	7.7	700	671	4.1	685	2.2	708	1.2
280.8	288.9	2.9	293.5	4.5	295.8	5.4	700	671	4.1	680	2.8	685	2.2
284.0	288.9	1.7	295.8	4.2	293.5	3.4	710	671	5.5	685	3.5	680	4.2
287.2	293.5	2.2	310.2	8.0	293.5	2.2	710	680	4.2	706	0.6	680	4.2
276.8	288.9	4.4	311.2	12.4	289.4	4.5	680	671	1.3	707	4.0	673	1.1
278.4	288.9	3.8	310.1	11.4	288.7	3.7	684	671	1.9	706	3.2	671	1.8
264.0	289.4	9.6	310.2	17.5	286.9	8.7	648	673	3.8	706	9.0	667	3.0
256.0	284.2	11.0	287.9	12.5	299.4	16.9	610	660	8.1	670	9.8	690	13.2
255.2	284.5	11.5	286.9	12.4	299.6	17.4	614	660	7.6	667	8.7	691	12.5
248.0	284.2	14.6	287.9	16.1	299.4	20.7	600	660	10.0	670	11.7	690	15.1
197.6	289.7	46.6	286.9	45.2	299.6	51.6	518	671	29.6	667	28.9	691	33.4
198.4	299.4	50.9	286.9	44.6	299.6	51.0	530	691	30.3	667	25.9	691	30.4
208.8	299.4	43.4	286.9	37.4	299.4	43.4	540	690	27.9	667	23.6	690	27.9
208.8	290.4	39.1	286.9	37.4	299.4	43.4	536	673	25.5	667	24.5	690	28.8

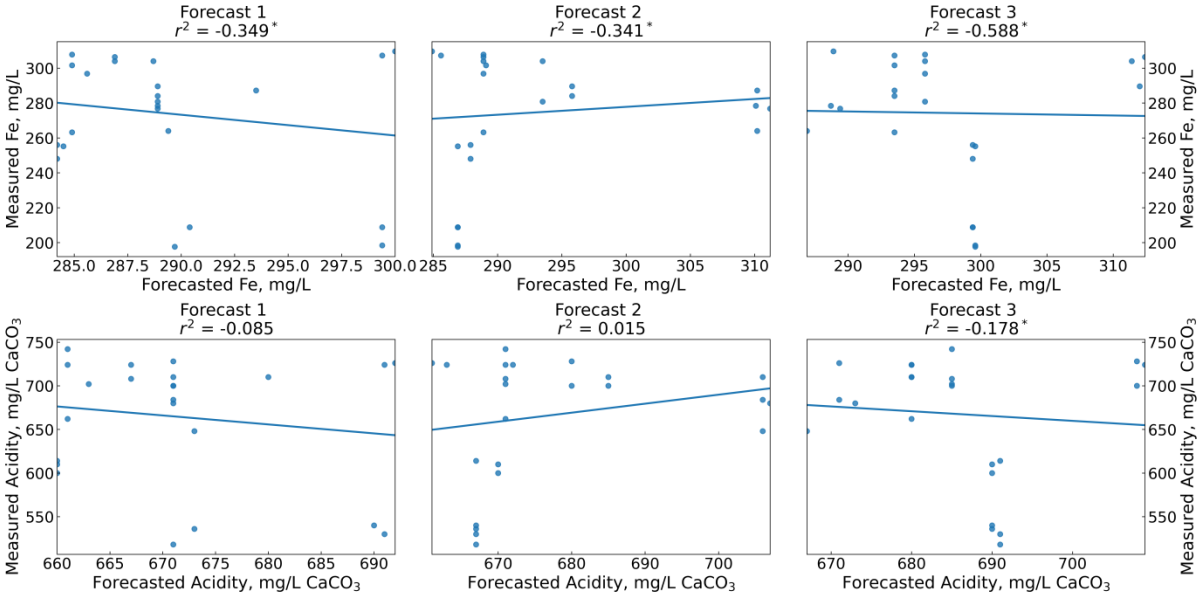


Figure 7.14: Cross plots with robust regression lines comparing the measured and forecasted Fe and acidity concentrations using the random forest model results; $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

7.9 Conclusions

Forecasted concentrations of Fe and acidity by the random forest model fall within the historical data population and follow its recent trend and pattern. Therefore, the proposed methodology can be applied with certainty and confidence in forecasting mine water chemistry. Machine learning forecasting approach proved that the application can use data from several parameters to forecast other parameters, i.e. the model was developed in a way that the computer learns the trends, patterns and seasonality of input data to forecast the target outputs. Parameters in a time series are related to each other and influence the outcomes in each parameter's dataset. Therefore, focusing only on one parameter to perform forecasting analysis would be inaccurate. Traditional statistical forecasting techniques such as ARIMA or Box-Jenkins, which forecast data of a parameter by learning its structure without relating it to other parameters, should be avoided in future applications. Thus, the forecasting technique proposed here will be a useful tool for water treatment plants because it

will help in understanding changes in the mine water chemistry and volumes in advance.

It can be concluded that forecasting mine water chemistry by applying ML models is a relevant contribution in and addition to mine water treatment plants. Comparing the neural network and regression tree models, the results show that random forest regression tree model performed better than the other models. Finally, the results obtained in this study indicate that regression tree algorithms are powerful and important mechanisms to model and forecast the complex mine water time series data or nonlinear systems. These approaches were able to analyse the hidden patterns, trends and seasonality among the historical mine water dataset in a much better and accurate approach compared to traditional time series analysis and statistical techniques. Lastly, the findings of this chapter have revealed that transforming time series data before using it for modelling is sometimes necessary to achieve more accurate forecasting results.

The next chapter will introduce a machine learning (ML) graphical user interface (GUI). This GUI is embedded with several ML algorithms and can perform several functions such as loading data, data pre-processing and visualisation. Most importantly, a ML algorithm can be trained and tested in this GUI and the built model can be downloaded as a pickle file to perform predictive analysis in Python.

CHAPTER 8

8 The Internet of Mine Water — Python Machine Learning Graphical User Interface

8.1 Introduction

After building several predictive machine learning (ML) models, it was necessary to bind them in a user friendly graphical user interface (GUI). In this chapter, a GUI was developed using PyQt 5.9.2 software and Qt designer. This kind of a GUI application is the first of its kind in the mine water sector. PyQt is a framework written in Python to develop desktop applications, and it allows the Qt framework to be used in Python code. Qt is a cross-platform application framework created for the use with C++ (Harwani, 2011; Rempt, 2001) and was later modified to accommodate Python language. For this thesis, the GUI was designed using Qt designer, and the GUI pages were linked and programmed using Python language via PyQt package. In addition, the Pyinstaller 4.8 library was used to convert the Python files to an executable file. Furthermore, InstallForge 1.4.2 was used to combine the files Pyinstaller created into a single file installer. This enabled the developed GUI to be able to be used in any machine with or without Python or specific Python libraries installed. Thus, this might be a necessity for data analysis in several organisations.

Not every worker in the treatment plant (or any other working environment) can programme. Thus, the purpose for the development of the GUI is to ensure that data is easily processed and analysed without having to write programming codes. The GUI can perform several functions such as loading data in CSV format, pre-processing the data such as normalising it using different scaling options (e.g. robust scaler) and visualising the data, e.g. scatter plots. Additionally, the GUI is embedded with several ML algorithms, e.g. gradient boosting regression tree. In a few clicks,

data can be loaded, visualised, pre-processed and ML models quickly built without having to write the programming code.

8.2 Overview of the IoMW GUI

8.2.1 Overview

The GUI was developed using Python programming language and Qt designer within an Anaconda platform, and is entitled “The Internet of Mine Water” (IoMW) (Figure 8.1). Several Python libraries are part of the GUI: Matplotlib, Seaborn, Numpy, Pandas, Scikit-learn, Pickle, and PyQt5 were used to write Python scripts in the Spyder integrated development environment (IDE) software, while Pyinstaller was used to build the Python files into an executable file. Qt designer was used to design the GUI pages, and CSS language was applied for styling of these pages. The files created by the Pyinstaller library were compressed into a single installer file using InstallForge. Therefore, this ensures that the IoMW can be installed and used in any machine that runs on the Windows system (Figure 8.2). IoMW GUI is currently on version 1.5 and is being used in this thesis. This software was built to be compatible for any Windows operating system, has End-User license agreement, and requires space of 1.14 GB to be installed.

Figure 8.1: The Internet of Mine Water GUI welcome page

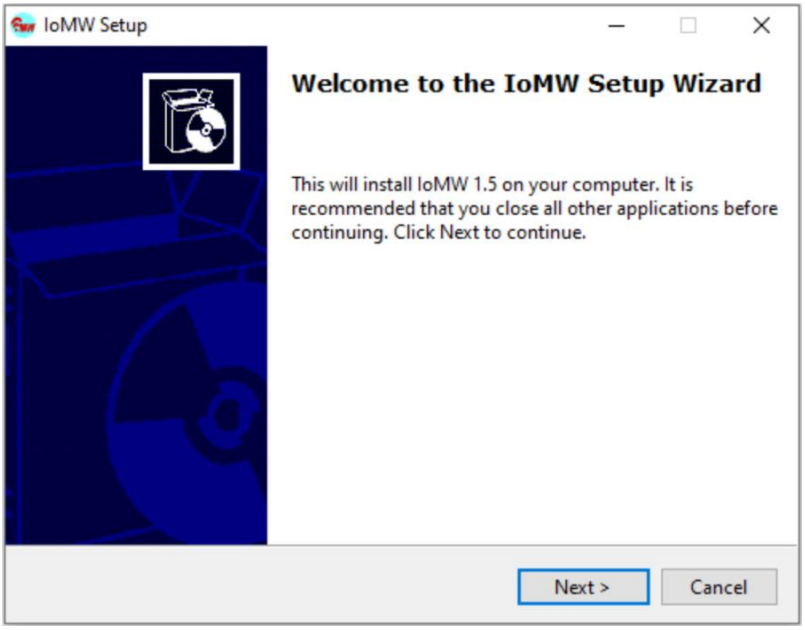


Figure 8.2: The windows setup wizard for the IoMW 1.5 GUI

The program allows the user to load CSV files and has three regression ML algorithms embedded in it (gradient boosting regression tree, random forest regression tree and linear regression). These algorithms can be used to build the models with the loaded data, and the models can be downloaded as pickle files to be used for predictive analysis. To use the software, the user must first load the CSV

file, select the target variable, drop columns (optional), visualise the data (optional) by plotting scatter/cross plots, histograms and Pearson correlation matrix, and normalise the data (optional) using different normalising options such as robust scaler, min/max scaler and standard scaler. Therefore, the user can select the ML algorithm to build the model (Figure 8.3).

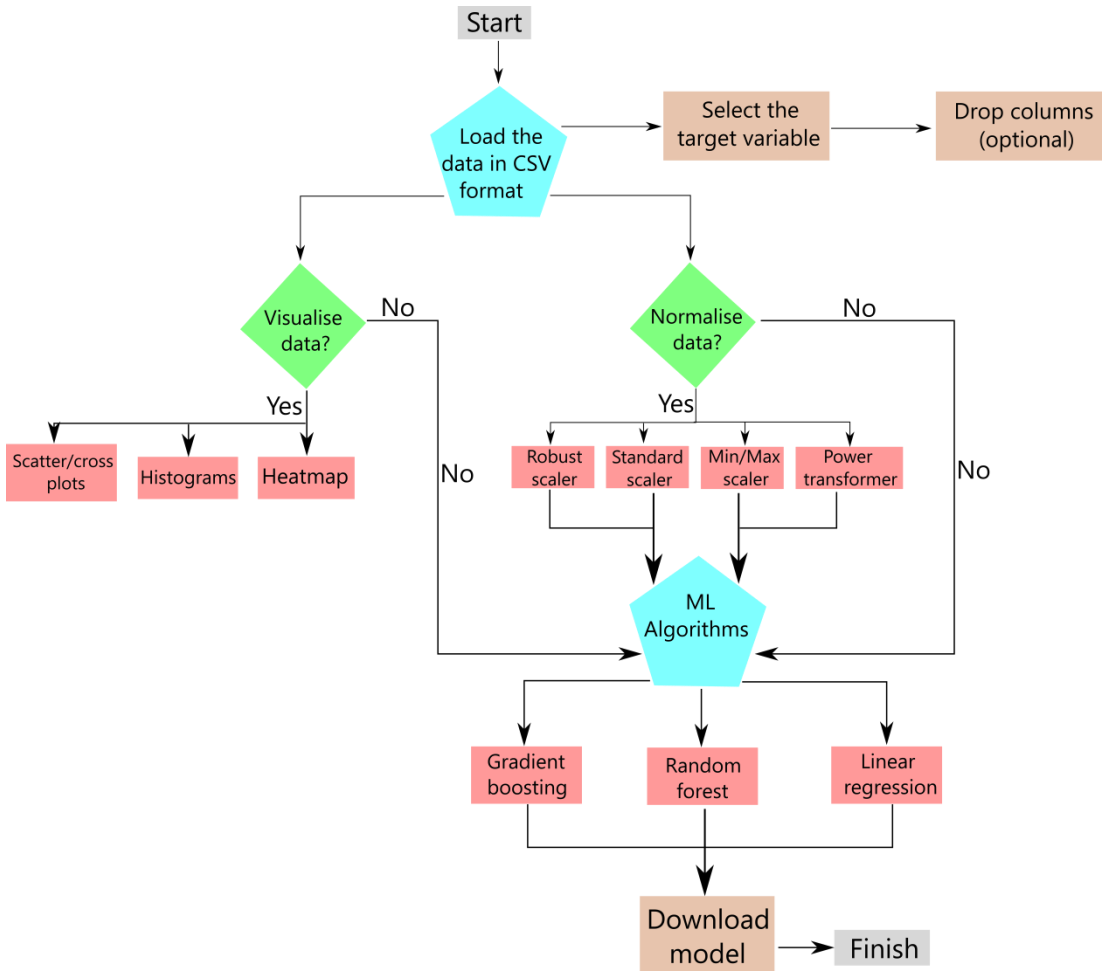


Figure 8.3: The IoMW user guide flow chart

8.2.2 Main window — the user interface

8.2.2.1 Data pre-processing

Data pre-processing is the process in which data are transformed into an "understandable" format so it can be used for further statistical analysis. This process involves several steps; however, the IoMW application's data pre-

processing consists of normalisation and dropping of columns. For the normalising option, the user has a choice to normalise the data using the robust scaler, standard scaler, min/max scaler, or power transformation (Figure 8.4).

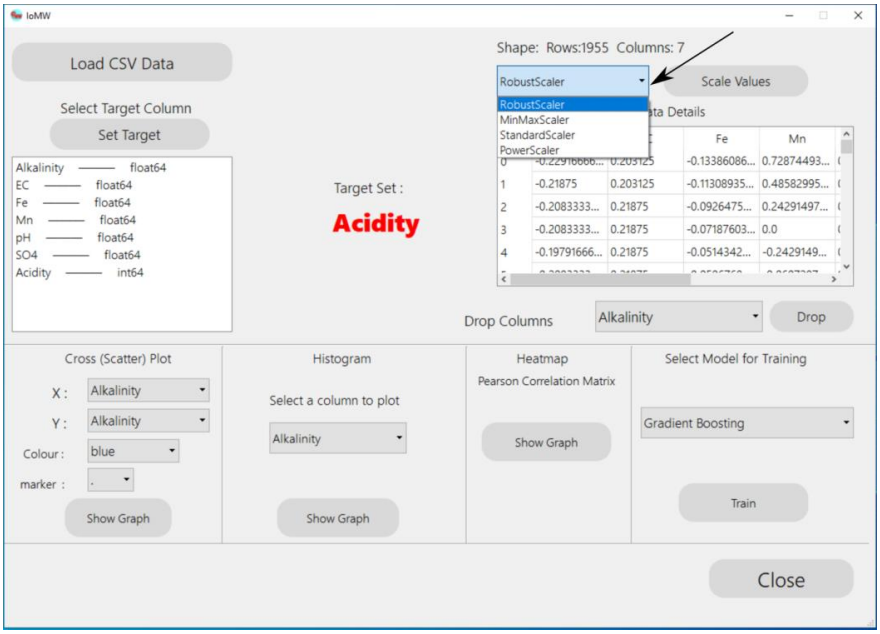


Figure 8.4: The data normalisation options for the IoMW application

8.2.2.2 Data visualisation

The IoMW helps the user to visualise the data before building ML models. In this part, the user can plot scatter/cross plots (Figure 8.5), histograms, or heatmap. Cross plots show one variable on the x-axis and the other on the y-axis, the histogram shows the probability on the y-axis, and the heatmap creates a Pearson Correlation Matrix with coefficient correlations (*r*).

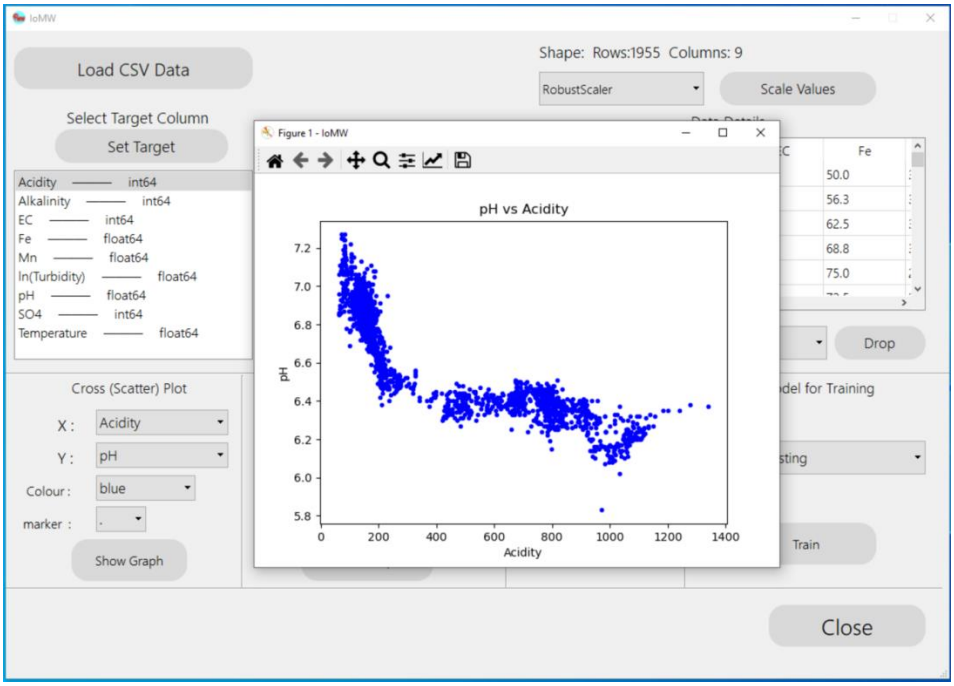


Figure 8.5: Cross plots on the IoMW application with acidity on the x-axis and pH on the y-axis

8.2.2.3 Machine learning algorithms

The application is embedded with three regression ML algorithms: gradient boosting regression tree, random forest regression tree and linear regression (Figure 8.6). When the user clicks on one of the algorithms and chooses to train the data by clicking “Train”, a new window will open. In the new window, the user is able to tune the ML algorithm hyper-parameters such as number of trees and learning rate, and the user also has an option to set the test size. Once the user is done training the algorithm, the model performance (mean absolute error, mean squared error and root mean squared error) will show. If the user is satisfied with the performance, the model can be downloaded as a pickle file. Once the user has new data, the pickle file model can be used in Python to perform predictive analysis (Figure 8.7). In future versions of this software, the user might be able to even perform predictive analysis without having to download a pickle model.

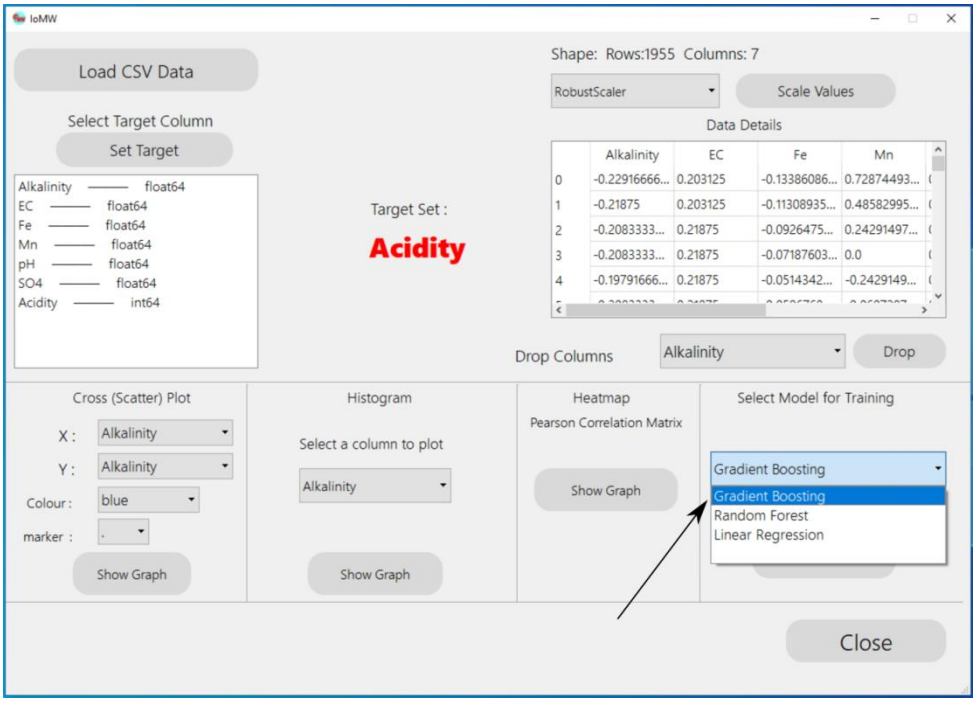


Figure 8.6: IoMW's embedded ML algorithms

```
#import necessary modules
import  pickle
import pandas as pd

#loading the pickle file (model)
model = pickle.load(open("model.pkl", "rb"))

#loading new data
newdata = pd.read_csv('NewData.csv')

#making predictions
predictions = model.predict(newdata)
```

Figure 8.7: An example Python code of loading a pickle file model and make predictions on new data

In the next chapter, overall concluding remarks and recommendations according to the findings of the research conducted are presented. The discoveries of the investigations are outlined and recommendations are put forward for further studies relating to this thesis.

CHAPTER 9

9 Conclusions and Recommendations

Improved mine water management techniques are a necessity especially in countries encountering polluted water crisis like in South Africa. Even though the mining industry contributes a low percentage to the polluted water crisis globally as compared to other industries like the agriculture, it still needs to be given undivided attention. Clean and safe water is essential to the mining communities, thus properly treating and managing mining influenced water (MIW) is important. For mine water treatment plants to operate efficiently, they need to have enough chemicals supply, energy and human resources. Therefore, the research reported in this thesis presented scientifically based, practical tools to predict and forecast relevant mine water parameters which will help plant operators to know beforehand which chemicals and methods to use to treat and manage polluted MIW (details below). An advanced method which integrates information of technology (IoT) to be able to identify different mine water samples taken, especially during bulk sampling, was also introduced. The presented methods will increase the speed and accuracy of sample identification and data processing, and ultimately optimise production and mine water management.

Incorporating IoT and artificial intelligence (AI) techniques in the mines improves the current treatment practices and improves the way at which MIW is treated. IoT methodologies presented in this thesis ensure that the sampling information and results are shared in real-time between the sampler, client and sampling institution. AI techniques in this thesis, which involved building predictive machine learning (ML) models, allows for the prediction and forecasting of the chemistry of MIW, which will

ensure that the treatment plant operators know in advance the quantity and type of chemicals and methods to use to treat and manage mine water, and when to start or close pumps.

In this thesis, the advantages of using IoT were shown by making use of the Near-Field Communication (NFC) technology, i.e. developing the *e*Metsi application. Sampling bottles were incorporated with NFC tags, a NFC mobile application for recording on-site parameters during sampling was developed, and a website application related to this NFC mobile application was developed to be used for data display. This approach proved to be the answer to closing a gap between the sampler, laboratory and the sampling institution, i.e. it enables the sharing of sampling parameters and results, sampling times and locations in real-time and reduces errors encountered when sampling.

Several programming languages were used to develop the *e*Metsi application, including Extensible Mark-up Language (XML) and Java code for the Android mobile application, and Cascading Style Sheets 3 (CSS3) language, Hyper Text Mark-up language 5 (HTML5) and JavaScript programming for the website application. The applications also required server provisioning and Structured Query Language (SQL) configuration to ensure maximal system flexibility. Specifically, *e*Metsi consists of identifying the samples electronically at the time of sampling, storing the on-site parameters and sample data, transferring the data to a cloud storage location, and allowing the end-users to use the sample identifying data.

Before ML models can be built, historical data needs to be examined. This thesis, based on historical time series data, presented different data cleaning and exploratory data analysis approaches which resulted in some being used for

interpolating missing data and detecting anomalies. Numerical models such as the B-Spline, NURBS and wavelet transform interpolate the missing data more precisely and accurately than the probabilistic models. Anomaly detection proved to be more effective when it is conducted by using predictive confidence level methods than statistical profiling approach. The approaches presented in this thesis proved that nonlinear systems require different exploratory data analysis techniques.

This thesis tapped in the power of AI by applying different hybrid intelligent systems in its studies. Computer models such as long short-term memory (LSTM), random forest regression tree, gradient boosting regression tree, artificial neural network (ANN) and deep neural network (DNN) were used to predict and forecast the chemistry of MIW. The models were trained and tested using Python within an Anaconda platform. For the predictive analysis study in the Ekurhuleni mine water treatment plant, the performance of the models were evaluated using the mean absolute error (MAE) and root mean squared error (RMSE). The algorithms produced RMSEs of 0.112 (for LSTM), 0.100 (for gradient boosting) and 0.109 (for random forest) for the testing set, and 0.270 (for LSTM), 0.100 (for gradient boosting) and 0.049 (for random forest) for the training set.

Regression tree models were further evaluated using the MAE and produced 0.070 (for gradient boosting) and 0.074 (for random forest) for the testing set, and 0.069 (for gradient boosting) and 0.032 (for random forest) for the training set. These numbers are indicative for a good fit of the model to the data. For the gradient boosting and random forest algorithms, a web application using Flask was created and the programming code committed in Github. Github was then linked to the Heroku cloud platform, and finally the hybrid model was deployed as a web application.

Another approach was introduced in this thesis which robustly analysed the data and compared the neural networks and regression tree algorithms. This technique was applied on the Randfontein mine water treatment plant data and it proved to be effective in dealing with nonlinear systems. The models applied in this study include the multivariate LSTM, ANN, DNN, random forest and gradient boosting, and were evaluated using the mean squared error (MSE) and mean absolute error (MAE). Regression tree algorithms performed better than the neural networks, with the random forest algorithm yielding better results. Therefore, this model was used to finally forecast Fe and acidity for 60 days.

It can be concluded that incorporating computational intelligence techniques in the mining environment can improve current treatment and management practices as it allows optimising resource management at the plant. The proposed approaches will be an efficient tool to forecasting mine water quality parameters assisting in this task. Additionally, the approaches in this thesis are important because they show what many studies lack to elaborate, which is how to put ML models into production in a real world application. Deployment of ML models is often overlooked, especially in academia. This thesis will assist people in industries seeking to maximise production and optimise management practices by informing them of the importance of AI and how to put good performing models into production. In addition, the explained techniques in this thesis will be an important booster to the mining industry, especially mine water treatment plant departments. Prediction and forecasting of relevant mine water parameters in mine water treatment plants is an important practical tool to optimise mine water management.

Finally, the IoMW graphical user interface (GUI) was developed to ensure that people from the industry find it easy to implement predictive ML models by doing

most of the work in the GUI and only applying less than five lines of Python code to reach conclusive results. The IoMW GUI application helps the user to perform important functions without having to write any programming code. With the IoMW application, the user can analyse the data and build predictive ML models.

This thesis presented a certain portion of the internet of mine water. More work still needs to be conducted to develop a fully functional IoMW; however, this thesis can be used as a “stepping stone”. Most of the presented industry 4.0 technologies in some parts of this thesis could not be explored due to time constraints, financial matters, and unavailability of these technologies in the mines. A fully functional IoMW would have technologies such as AI, IoT, wireless sensor networks, big data, quantum computing and 5G network; and this thesis only managed to explore AI and IoT. Additionally, the developed technological tools (*e*Metsi and IoMW GUI) in this thesis will be further improved. For example, the *e*Metsi application can be upgraded in future to be able to integrate big data systems. More ML algorithms, data pre-processing and visualisation options can be included in the IoMW GUI to allow the application to be more versatile. In conclusion, this thesis showed that AI and IoT techniques are applicable to mine water treatment plants and that it is possible to reliably forecast MIW parameters for a certain period.

References

AGRAWAL, S., DAS, M.L. & LOPEZ, J. 2019. Detection of Node Capture Attack in Wireless Sensor Networks. *IEEE Syst. J.,* 13(1): 238-247. https://doi.org/10.1109/JSYST.2018.2863229

AHMED, U., MUMTAZ, R., ANWAR, H., SHAH, A.A., IRFAN, R. & GARCÍA-NIETO, J. 2019. Efficient water quality prediction using supervised machine learning. *Water,* 11(11): 14. https://doi.org/10.3390/w11112210

AL-DULAIMI, A., WANG, X. & CHIH-LIN, I. 2018. *5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management.* Chichester: Wiley.

ALI, N., NEAGU, D. & TRUNDLE, P. 2019. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Appl. Sci.,* 1(12): 1559. https://doi.org/10.1007/s42452-019-1356-9

ALOI, G., CALICIURI, G., FORTINO, G., GRAVINA, R., PACE, P., RUSSO, W. & SAVAGLIO, C. 2017. Enabling IoT interoperability through opportunistic smartphone-based mobile gateways. *J. Netw. Comput. Appl.,* 81: 74-84. https://doi.org/10.1016/j.jnca.2016.10.013

ALPAYDIN, E. 2010. *Introduction to Machine Learning.* Cambridge: MIT Press. The Adaptive Computation and Machine Learning.

ARAYA, D.B., GROLINGER, K., ELYAMANY, H.F., CAPRETZ, M.A.M. & BITSUAMLAK, G. 2017. An ensemble learning framework for anomaly detection in building energy consumption. *Energy Build.,* 144: 191-206. https://doi.org/10.1016/j.enbuild.2017.02.058

ARISMENDY, L., CÁRDENAS, C., GÓMEZ, D., MATURANA, A., MEJÍA, R. & QUINTERO M, C.G. 2020. Intelligent system for the predictive analysis of an industrial wastewater treatment process. *Sustainability,* 12(16): 6348. https://doi.org/10.3390/su12166348

ASIF, S. 2019. *5G mobile communications: Concepts and technologies.* London: CRC Press.

ASR, E.T., KAKAIE, R., ATAEI, M. & TAVAKOLI MOHAMMADI, M.R. 2019. A review of studies on sustainable development in mining life cycle. *J. Clean. Prod.,* 229: 213-231. https://doi.org/10.1016/j.jclepro.2019.05.029

ATHRESH, A.P., AL-HABAIBEH, A. & PARKER, K. 2017. An Innovative and Integrated Approach for Using Energy from the Flooded Coal Mines for Pre-

warming of a Gas Engine in Standby Mode Using GSHP. *Energy Procedia,* 105: 2531-2538. https://doi.org/10.1016/j.egypro.2017.03.726

ATMAJA, A.P., EL HAKIM, A., WIBOWO, A.P.A. & PRATAMA, L.A. 2021. Communication systems of smart agriculture based on Wireless Sensor Networks in IoT. *J. Robot. Control,* 2(4): 297-301. https://doi.org/10.18196/jrc.2495

AVERBUCH, A.Z., NEITTAANMÄKI, P. & ZHELUDEV, V.A. 2014. *Spline and spline wavelet methods with applications to signal and image processing.* Cham: Springer. Non-Periodic Splines. 2. https://doi.org/10.1007/978-3-319-22303-2

AWUAH-OFFEI, K. & FRIMPONG, S. 2011. Efficient Cable Shovel Excavation in Surface Mines. *Geotech. Geol. Eng.,* 29(1): 19-26. https://doi.org/10.1007/s10706-010-9366-9

AZEVEDO, A. & ALMEIDA, A. 2011. Factory templates for digital factories framework. *Robot. Comput. Integr. Manuf.,* 27(4): 755-771

AZZALINI, A. & CAPITANIO, A. 1999. Statistical applications of the multivariate skew normal distribution. *J. R. Statist. Soc. B,* 61(3): 579-602. https://doi.org/10.1111/1467-9868.00194

BABA, T. 2012. Time-frequency analysis using short time Fourier transform. *Open J. Acoust.,* 5(1): 32-38. https://doi.org/10.2174/1874837601205010032

BÄHR, S., HAAS, G.-C., KEUSCH, F., KREUTER, F. & TRAPPMANN, M. 2020. Missing data and other measurement quality issues in mobile geolocation sensor data. *Soc. Sci. Comput. Rev.*: 1-24. https://doi.org/10.1177/0894439320944118

BANDYOPADHYAY, S. 2005. Computing with spins: from classical to quantum computing. *Superlattices Microstruct.,* 37(2): 77-86. https://doi.org/10.1016/j.spmi.2004.09.043

BATES, R.G. 1964. *Determination of pH: theory and practice.* Sydney: Wiley.

BAYDAS, S. & KARAKAS, B. 2019. Defining a curve as a Bezier curve. *J. Taibah Univ. Sci.,* 13(1): 522-528. https://doi.org/10.1080/16583655.2019.1601913

BEESLEY, C., FROST, A. & ZAJACZKOWSKI, J. 2009. A comparison of the BAWAP and SILO spatially interpolated daily rainfall datasets. In: Anderssen, R.S., Braddock, R.D. & Newham, L.T.H. (eds.). *18th World IMACS/MODSIM Congress*. Cairns, Australia: Citeseer: 13-17

BELGIU, M. & DRĂGUŢ, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.,* 114: 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

BETRIE, G.D., SADIQ, R., TESFAMARIAM, S. & MORIN, K.A. 2016. On the issue of incomplete and missing water-quality data in mine site databases: comparing three imputation methods. *Mine Water Environ.,* 35(1): 3-9. https://doi.org/10.1007/s10230-014-0322-4

BETRIE, G.D., TESFAMARIAM, S., MORIN, K.A. & SADIQ, R. 2013. Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques. *Environ. Monit. Assess.,* 185(5): 4171-4182. https://doi.org/10.1007/s10661-012-2859-7

BHAKDISONGKHRAM, T., KOOTTATEP, S. & TOWPRAYOON, S. 2007. A Water Model for Water and Environmental Management at Mae Moh Mine Area in Thailand. *Water Resour. Manage.,* 21(9): 1535-1552. https://doi.org/10.1007/s11269-006-9103-6

BHATTACHARYYA, R., FLOERKEMEIER, C. & SARMA, S. 2010. RFID tag antenna based sensing: Does your beverage glass need a refill? In: Kerr, T., Deavours, D., Engels, D.W. & Floerkemeier, C. (eds.). *IEEE International Conference on RFID (IEEE RFID 2010)*. IEEE: 126-133

BHURIYA, D., KAUSHAL, G., SHARMA, A. & SINGH, U. 2017. Stock market predication using a linear regression. In: Qadeer, A., Alexander, M. & Dimitri, P. (eds.). *International conference of electronics, communication and aerospace technology (ICECA).* IEEE: 510-513

BIAU, G. & SCORNET, E. 2016. A random forest guided tour. *TEST,* 25(2): 197-227. https://doi.org/10.1007/s11749-016-0481-7

BIHAM, E., BRASSARD, G., KENIGSBERG, D. & MOR, T. 2004. Quantum computing without entanglement. *Theor. Comput. Sci.,* 320(1): 15-33. https://doi.org/10.1016/j.tcs.2004.03.041

BLOWES, D.W., PTACEK, C.J., JAMBOR, J.L., WEISENER, C.G., PAKTUNC, D., GOULD, W.D. & JOHNSON, D.B. 2014. The Geochemistry of Acid Mine Drainage. In: Holland, H.D. & Turekian, K.K. (eds.). *Treatise on Geochemistry*. 2nd ed. Oxford: Elsevier, 131-190. https://doi.org/10.1016/B978-0-08-095975-7.00905-0

BOLTON, E.W., MAASCH, K.A. & LILLY, J.M. 1995. A wavelet analysis of Plio-Pleistocene climate indicators: A new view of periodicity evolution. *Geoph. Res. Letters,* 22(20): 2753-2756. https://doi.org/10.1029/95GL02799

BONABEAU, E., DORIGO, M. & THERAULAZ, G. 1999. *Swarm Intelligence: From Natural to Artificial Systems.* Oxford: Oxford University Press.

BOS, J.W., LAUTER, K. & NAEHRIG, M. 2014. Private predictive analysis on encrypted medical data. *J. Biomed. Inform.,* 50: 234-243. https://doi.org/10.1016/j.jbi.2014.04.003

BOUDJEMAA, R., FORBES, A., HARRIS, P. & LANGDELL, S. 2003. *Multivariate empirical models and their use in metrology* (No. CMSC 32/03). Middlesex: Software Support for Metrology Programme.

BOULESTEIX, A.L., JANITZA, S., KRUPPA, J. & KÖNIG, I.R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs. Data Mining Knowl. Discov.,* 2(6): 493-507. https://doi.org/10.1002/widm.1072

BOULET, B. 2006. *Fundamentals of Signals and Systems.* Boston: Thomson Learning.

BOYES, H., HALLAQ, B., CUNNINGHAM, J. & WATSON, T. 2018. The industrial internet of things (IIoT): An analysis framework. *Comp. Ind.,* 101: 1-12. https://doi.org/10.1016/j.compind.2018.04.015

BREIMAN, L. 2001. Random Forests. *Mach. Learn.,* 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

BUCKLEY, J.J. & ESLAMI, E. 2002. *An introduction to fuzzy logic and fuzzy sets.* Heidelberg: Springer. Advances in Soft Computing. https://doi.org/10.1007/978-3-7908-1799-7

BUI, X.-N., NGUYEN, H., CHOI, Y., NGUYEN-THOI, T., ZHOU, J. & DOU, J. 2020a. Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. *Sci. Rep.,* 10(1): 9939. https://doi.org/10.1038/s41598-020-66904-y

BUI, X.-N., NGUYEN, H., LE, H.-A., BUI, H.-B. & DO, N.-H. 2020b. Prediction of Blast-induced Air Over-pressure in Open-Pit Mine: Assessment of Different Artificial Intelligence Techniques. *Nat. Resour. Res.,* 29(2): 571-591. https://doi.org/10.1007/s11053-019-09461-0

BURKOV, A. 2019. *The hundred-page machine learning book.* Canada: Andriy Burkov.

CAI, J., XU, K., ZHU, Y., HU, F. & LI, L. 2020. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Appl. Energy,* 262: 114566. https://doi.org/10.1016/j.apenergy.2020.114566

ÇAKıCı, Ö.E., GROENEVELT, H. & SEIDMANN, A. 2011. Using RFID for the management of pharmaceutical inventory — system optimization and shrinkage control. *Decis. Support Syst.,* 51(4): 842-852. https://doi.org/10.1016/j.dss.2011.02.003

CHADWICK, J. 2016. Mine design and scheduling. *Mining Technology.* Available from: www.promine.com. Accessed: 2019-09-01

CHAHBANDOUR, J. 2013. Reliable Mine Water Management; Connecting the Drops to Operate "Water Smart" Mines. In: Brown, A., Figueroa, L. & Wolkersdorfer, C. (eds.). *Reliable Mine Water Technology*. Golden: International Mine Water Association, 813-817.

CHAIX, J.-F., GARNIER, V. & CORNELOUP, G. 2003. Concrete damage evolution analysis by backscattered ultrasonic waves. *NDT. E. Int.,* 36(7): 461-469. https://doi.org/10.1016/S0963-8695(03)00066-5

CHANG, Y.-S., CHANG, C.-L., HUNG, Y.-S. & TSAI, C.-T. 2010. NCASH: NFC Phone-Enabled Personalized Context Awareness Smart-Home Environment. *Cybernetics Syst.,* 41(2): 123-145. https://doi.org/10.1080/01969720903584241

CHATFIELD, C. 2000. *Time-series forecasting.* London: Chapman and Hall/CRC.

CHAU, K.-W. 2006. A review on integration of artificial intelligence into water quality modelling. *Mar. Pollut. Bull.,* 52(7): 726-733. https://doi.org/10.1016/j.marpolbul.2006.04.003

CHEN, F., DENG, P., WAN, J., ZHANG, D., VASILAKOS, A.V. & RONG, X. 2015. Data mining for the internet of things – literature review and challenges. *Int. J. Distrib. Sens. Netw.,* 11(8): 1-14. https://doi.org/10.1155/2015/431047

CHEN, G. & PHAM, T.T. 2001. *Introduction to fuzzy sets, fuzzy logic, and fuzzy control systems.* New York City: CRC Press.

CHEN, K., CHEN, H., ZHOU, C., HUANG, Y., QI, X., SHEN, R., LIU, F., ZUO, M., ZOU, X., WANG, J., ZHANG, Y., CHEN, D., CHEN, X., DENG, Y. & REN, H. 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.,* 171: 115454. https://doi.org/10.1016/j.watres.2019.115454

CHEN, M., MAO, S. & LIU, Y. 2014. Big data: A survey. *Mobile Netw. Appl.,* 19(2): 171-209. https://doi.org/10.1007/s11036-013-0489-0

CHEN, Y., YANG, B., DONG, J. & ABRAHAM, A. 2005. Time-series forecasting using flexible neural tree model. *Inform. Sci.,* 174(3): 219-235. https://doi.org/10.1016/j.ins.2004.10.005

CHEN, Y.Y., TSAI, M.L. & JAN, J.K. 2011. The design of RFID access control protocol using the strategy of indefinite-index and challenge-response. *Comput. Commun.,* 34(3): 250-256. https://doi.org/10.1016/j.comcom.2010.05.012

CHENG, X., DU, D.Z., WANG, L. & XU, B. 2008. Relay sensor placement in wireless sensor networks. *Wireless Netw.,* 14(3): 347-355

CHIKODILI, N.B., ABDULMALIK, M.D., ABISOYE, O.A. & BASHIR, S.A. 2020. Outlier Detection in Multivariate Time Series Data Using a Fusion of K-Medoid, Standardized Euclidean Distance and Z-Score. In: Misra, S. & Muhammad-Bello, B. (eds.). *International Conference on Information and Communication Technology and Applications*. Springer: 259-271. https://doi.org/10.1007/978-3-030-69143-1_21

CHING, S.H. & TAI, A. 2009. HF RFID versus UHF RFID — Technology for Library Service Transformation at City University of Hong Kong. *J. Acad. Librariansh.,* 35(4): 347-359. https://doi.org/10.1016/j.acalib.2009.04.005

CHO, J., SON, H., CHOI, W. & PARK, C. 2013, 7-13 July. HF RFID reader antenna generating horizontal magnetic fields for casino applications. In: Erricolo, D. (ed.). *2013 IEEE Antennas and Propagation Society International Symposium (APSURSI).* IEEE: 1118-1119. https://doi.org/10.1109/APS.2013.6711219

CHUNG, Y.M. & HALIM, Z.A. 2014. Adaptive Neuro-fuzzy Inference System as Cache Memory Replacement Policy. *Adv. Electr. Comput. Eng.,* 14(1): 15-25. https://doi.org/10.4316/AECE.2014.01003

CIVERCHIA, F., BOCCHINO, S., SALVADORI, C., ROSSI, E., MAGGIANI, L. & PETRACCA, M. 2017. Industrial Internet of Things monitoring solution for advanced predictive maintenance applications. *J. Ind. Inf. Integr.,* 7: 4-12. https://doi.org/10.1016/j.jii.2017.02.003

COETZEE, L. & EKSTEEN, J. 2011. Internet of things–promise for the future? An Introduction. In: Cunningham, P. & Cunningham, M. (eds.). *International Information Management Corporation.* 1-9

CORKE, P., ROBERTS, J. & WINSTANLEY, G. 1998. Vision-based control for mining automation. *IEEE Robot. Autom. Mag.,* 5(4): 44-49

CORNELIUS, T.L. 1998. *Fuzzy logic and expert systems applications.* California: Academic Press.

COSKUN, V., OZDENIZCI, B. & OK, K. 2013. A Survey on Near Field Communication (NFC) Technology. *Wireless Pers. Commun.,* 71(3): 2259-2294. https://doi.org/10.1007/s11277-012-0935-5

CÔTE, C.M., MORAN, C.J., HEDEMANN, C.J. & KOCH, C. 2010. Systems modelling for effective mine water management. *Environ. Modell. Softw.,* 25(12): 1664-1671. https://doi.org/10.1016/j.envsoft.2010.06.012

DANILOV, A., SMIRNOV, Y., PETROVA, T. & PASHKEVICH, M. 2015. Using drones of preconstruction monitoring conducting in mining enterprise. *Int. J. Ecol. Dev.,* 30: 36-42

DE WOLF, R. 2017. The potential impact of quantum computers on society. *Ethics Inf. Technol.,* 19(4): 271-276. https://doi.org/10.1007/s10676-017-9439-z

DEANE, P.M. 1979. *The first industrial revolution.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511622090

DEB, D., DESHPANDE, V.N. & DAS, K.C. 2008. Assessment of Water Quality around Surface Coal Mines using Principal Component Analysis and Fuzzy Reasoning Techniques. *Mine Water Environ.,* 27(3): 183-193. https://doi.org/10.1007/s10230-008-0030-z

DEMIRALP, G., GUVEN, G. & ERGEN, E. 2012. Analyzing the benefits of RFID technology for cost sharing in construction supply chains: A case study on prefabricated precast components. *Autmat. Constr.,* 24: 120-129. https://doi.org/10.1016/j.autcon.2012.02.005

DIAMANTOPOULOU, M.J. 2005. Artificial neural networks as an alternative tool in pine bark volume estimation. *Comput. Electron. Agr.,* 48(3): 235-244. https://doi.org/10.1016/j.compag.2005.04.002

DÍAZ, M., MARTÍN, C. & RUBIO, B. 2016. State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing. *J. Netw. Comput. Appl.,* 67: 99-117. https://doi.org/10.1016/j.jnca.2016.01.010

DIVINCENZO, D.P. & LOSS, D. 1999. Quantum computers and quantum coherence. *J. Magn. Magn. Mater.,* 200(1-3): 202-218. https://doi.org/10.1016/S0304-8853(99)00315-7

DOBKIN, D.M. 2012. *The RF in RFID: UHF RFID in practice.* Oxford: Newnes. https://doi.org/10.1016/B978-0-7506-8209-1.X5001-3

DRAPER, C., REICHLE, R., DE JEU, R., NAEIMI, V., PARINUSSA, R. & WAGNER, W. 2013. Estimating root mean square errors in remotely sensed soil moisture over continental scale domains. *Remote Sens. Environ.,* 137: 288-298. https://doi.org/10.1016/j.rse.2013.06.013

DUFF, E.S., ROBERTS, J.M. & CORKE, P.I. 2003. Automation of an underground mining vehicle using reactive navigation and opportunistic localization. In: Abbott, D. (ed.). *International Conference on Intelligent Robots and Systems (IROS 2003).* Las Vegas: IEEE: 3775-3780. https://doi.org/10.1109/IROS.2003.1249742

DUIVESTEIJN, W. & FEELDERS, A. 2008. Nearest Neighbour Classification with Monotonicity Constraints. In: Daelemans, W., Goethals, B. & Morik, K. (eds.). *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer: 301-316. https://doi.org/10.1007/978-3-540-87479-9_38

EASLEY, M., HANEY, L., PAUL, J., FOWLER, K. & WU, H. 2018. Deep neural networks for short-term load forecasting in ERCOT system. In: Kerr, T., Deavours, D., Engels, D.W. & Floerkemeier, C. (eds.). *2018 IEEE Texas Power and Energy Conference (TPEC)*. IEEE: 1-6. https://doi.org/10.1109/TPEC.2018.8312073

EKEMEN KESKIN, T., ÖZLER, E., ŞANDER, E., DÜĞENCI, M. & AHMED, M.Y. 2020. Prediction of electrical conductivity using ANN and MLR: a case study from Turkey. *Acta Geophys.,* 68(3): 811-820. https://doi.org/10.1007/s11600-020-00424-1

ELBANHAWI, M., SIMIC, M. & JAZAR, R.N. 2015. Continuous path smoothing for car-like robots using B-spline curves. *J. Intell. Robot. Syst.,* 80(1): 23-56. https://doi.org/10.1007/s10846-014-0172-0

FARRUGIA, L.I. 2011. *Wireless Sensor Networks.* New York: Nova Science. Computer Science, Technology and Applications.

FELDNER, B. & HERBER, P. 2018. A Qualitative Evaluation of IPv6 for the Industrial Internet of Things. *Procedia Comp. Sci.,* 134: 377-384. https://doi.org/10.1016/j.procs.2018.07.195

FERNANDES, E., HOLANDA, M., VICTORINO, M., BORGES, V., CARVALHO, R. & VAN ERVEN, G. 2019. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *J. Bus. Res.,* 94: 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

FIELDING, S., FAYERS, P.M. & RAMSAY, C.R. 2009. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health Qual. Life Outcomes,* 7(1): 57. https://doi.org/10.1186/1477-7525-7-57

FINKENZELLER, K. 2010. *RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication.* 3rd ed. London: Wiley.

FLOYD, R.E. 2015. RFID in Animal-Tracking Applications. *IEEE Potentials,* 34(5): 32-33. https://doi.org/10.1109/MPOT.2015.2410308

FOX, D.L. 1999. *Prediction of acid rock drainage (ARD) risk from sulphidic slates using GIS analysis of mineralogical, geochemical, magnetic and geological parameters – a test case in southern Nova Scotia.* Halifax: unpubl. PhD Thesis Dalhousie University.

FREEDMAN, M.H., KITAEV, A. & WANG, Z. 2002. Simulation of topological field theories by quantum computers. *Commun. Math. Phys.,* 227(3): 587-603. https://doi.org/10.1007/s002200200635

GALAVI, H. & SHUI, L.T. 2012. Neuro-fuzzy modeling and forecasting in water resources. *Sci. Res. Essays,* 7(24): 2112-2121. https://doi.org/10.5897/SRE11. 2164

GALLEGO, E., RUIZ, A. & AGUADO, P.J. 2015. Simulation of silo filling and discharge using ANSYS and comparison with experimental data. *Comp. Electron. Agr.,* 118: 281-289

GAMA, K., TOUSEAU, L. & DONSEZ, D. 2012. Combining heterogeneous service technologies for building an Internet of Things middleware. *Comput. Commun.,* 35(4): 405-417. https://doi.org/10.1016/j.comcom.2011.11.003

GANDOMI, A. & HAIDER, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manage.,* 35(2): 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

GAO, L., BARRETT, D., CHEN, Y., ZHOU, M., CUDDY, S., PAYDAR, Z. & RENZULLO, L. 2014. A systems model combining process-based simulation and multi-objective optimisation for strategic management of mine water. *Environ. Modell. Softw.,* 60: 250-264. https://doi.org/10.1016/j.envsoft.2014.06.020

GARETTI, M. & TAISCH, M. 2012. Sustainable manufacturing: trends and research challenges. *Prod. Plan. Control,* 23(3): 83-104. https://doi.org/10.1080/09537287.2011.591619

GENUER, R., POGGI, J.-M. & TULEAU-MALOT, C. 2010. Variable selection using random forests. *Pattern Recognit. Lett.,* 31(14): 2225-2236. https://doi.org/10.1016/j.patrec.2010.03.014

GEORGE, L., LUDWICK, W. & CHAHBANDOUR, J. 2009. Case study: Site-wide water balance of the Pierina Gold Mine, Peru. In: *Tailings and Mine Waste*. 1st ed. Boca Raton: CRC Press, 369-380.

GHODRATI, B., HADI HOSEINIE, S. & GARMABAKI, A.H.S. 2015. Reliability considerations in automated mining systems. *Int. J. Min. Reclam. Environ.,* 29(5): 404-418. https://doi.org/10.1080/17480930.2015.1091617

GHOSALKAR, N.N. & DHAGE, S.N. 2018. Real estate value prediction using linear regression. In: Dimitri, P., Massimo, C. & Alessandro, P. (eds.). *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE: 1-5. https://doi.org/10.1109/ICCUBEA.2018.8697366

GHOSE, A.K. 2009. Technology vision 2050 for sustainable mining. *Procedia Earth Planet. Sci.,* 1(1): 2-6. https://doi.org/10.1016/j.proeps.2009.09.003

GILCHRIST, A. 2016. *Industry 4.0: the industrial internet of things.* New York: Apress.

GILROY, E., HIRSCH, R. & COHN, T. 1990. Mean square error of regression-based constituent transport estimates. *Water Resour. Res.,* 26(9): 2069-2077. https://doi.org/10.1029/WR026i009p02069

GLOVER, B. & BHATT, H. 2006. *RFID essentials.* Massachusetts: O'Reilly.

GOLESTANIFAR, M. & AHANGARI, K. 2012. Choosing an optimal groundwater lowering technique for open pit mines. *Mine Water Environ.,* 31(3): 192-198

GONÇALVES, R., DUARTE, A., MAGUETA, R., CARVALHO, N.B. & PINHO, P. 2014. RFID tags on paper substrate for bottle labelling. *Procedia Tech.,* 17: 65-72. https://doi.org/10.1016/j.protcy.2014.10.217

GOODRUM, P.M., MCLAREN, M.A. & DURFEE, A. 2006. The application of active radio frequency identification technology for tool tracking on construction job sites. *Autmat. Constr.,* 15(3): 292-302. https://doi.org/10.1016/j.autcon.2005.06.004

GORELIK, A. 2019. *The Enterprise Big Data Lake – Delivering the Promise of Big Data and Data Science.* California: O'Reilly.

GRAPS, A. 1995. An introduction to wavelets. *IEEE Comput. Sci. Eng.,* 2(2): 50-61. https://doi.org/10.1109/99.388960

GREHL, S., LÖSCH, R. & JUNG, B. 2018, 2019-07-06. Perfect match: IoT and robotics in underground mining. *World Mining Frontiers.* Available from: www.tu-freiberg.de/en/aridua/article-in-world-mining-frontiers. Accessed: 2019-07-06.

HABERMANN, C. & KINDERMANN, F. 2007. Multidimensional spline interpolation: Theory and applications. *Comput. Econ.,* 30(2): 153-169. https://doi.org/10.1007/s10614-007-9092-4

HAENSELMANN, T. & MÜLLER, C. 2011. *Wireless Sensor Networks: Design Principles for Scattered Systems.* München: De Gruyter Oldenbourg.

HÄFFNER, H., ROOS, C.F. & BLATT, R. 2008. Quantum computing with trapped ions. *Phys. Rep.,* 469(4): 155-203. https://doi.org/10.1016/j.physrep.2008.09.003

HAMOUD, A. 2016. Selection of best decision tree algorithm for prediction and classification of students' action. *Am. Int. J. Res. Sci. Technol. Eng. Math.,* 16(1): 26-32

HAN, X.-A., MA, Y. & HUANG, X. 2008. A novel generalization of Bézier curve and surface. *J. Comput. Appl. Math.,* 217(1): 180-193. https://doi.org/10.1016/j.cam.2007.06.027

HANCOCK, S. & WOLKERSDORFER, C. 2012. Renewed Demands for Mine Water Management. *Mine Water Environ.,* 31(2): 147-158. https://doi.org/10.1007/s10230-012-0176-6

HANSEN, E.W. 2014. *Fourier transforms: principles and applications.* Hoboken: Wiley.

HARWANI, B. 2011. *Introduction to Python programming and developing GUI applications with PyQT.* Boston: Cengage Learning.

HASANUZZAMAN, F.M., YANG, X., TIAN, Y., LIU, Q. & CAPEZUTI, E. 2013. Monitoring activity of taking medicine by incorporating RFID and video analysis. *Netw. Model. Anal. Health Inform. Bioinforma.,* 2(2): 61-70. https://doi.org/10.1007/s13721-013-0025-y

HEM, J.D. 1985. *Study and interpretation of the chemical characteristics of natural water.* 3rd ed. Alexandria, VA: Department of the Interior, US Geological Survey.

HERMAN, J., HERMAN, H., MATHEWS, M.J. & VOSLOO, J.C. 2018. Using big data for insights into sustainable energy consumption in industrial and mining sectors. *J. Clean Prod.,* 197: 1352-1364. https://doi.org/10.1016/j.jclepro.2018.06.290

HOCHREITER, S. & SCHMIDHUBER, J. 1997. Long short-term memory. *Neural Comput.,* 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

HOFFMANN, M. & JUHÁSZ, I. 2001. Shape control of cubic B-spline and NURBS curves by knotmodifications. In: Alessandro, P., Lijun, C. & Dario, P. (eds.). *14th International Conference Information Visualisation*. IEEE: 63-68. https://doi.org/10.1109/IV.2001.942040

HON, T.K. 2013. Time-Frequency Analysis and Filtering based on the Short-Time Fourier Transform. Doctor of Philosophy, London, King's College London

HOPWOOD, P. & DELOITTE TOUCHE TOHMATSU LIMITED. 2018. *Tracking the trends 2018 – The top 10 issues shaping mining in the year ahead.* London: Deloitte Design Studio.

HOUARI, R., BOUNCEUR, A., TARI, A.K. & KECHA, M.T. 2014. Handling missing data problems with sampling methods. In: Mueller, P., Lorenz, P. & Challal, Y. (eds.). *2014 International conference on advanced networking distributed systems and applications*. IEEE: 99-104. https://doi.org/10.1109/INDS.2014.25

HUANG, Y.-M. & LIU, C.-H. 2009. Applying adaptive swarm intelligence technology with structuration in web-based collaborative learning. *Comput. Educ.,* 52(4): 789-799. https://doi.org/10.1016/j.compedu.2008.12.002

HUBERT, E. & WOLKERSDORFER, C. 2015. Establishing a conversion factor between electrical conductivity and total dissolved solids in South African mine waters. *Water SA.,* 41(4): 490-500. https://doi.org/10.4314/wsa.v41i4.08

IAROVYI, S., MOHAMMED, W.M., LOBOV, A., FERRER, B.R. & LASTRA, J.L.M. 2016. Cyber-physical systems for open-knowledge-driven manufacturing execution systems. *Proceedings of the IEEE,* 104(5): 1142-1154

IGHALO, J.O., ADENIYI, A.G. & MARQUES, G. 2020. Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. *Biofuels Bioprod. Bioref.,* 14(6): 1286-1295. https://doi.org/10.1002/bbb.2140

IKEDA, K. 2018. Security and Privacy of Blockchain and Quantum Computatuion. *Advances Comp.,* 111: 199-228

INTERNATIONAL MINING. 2019. *Newtrax tackles data silo issues with launch of IoT hub*. Available from: www.im-mining.com/2019/12/03/newtrax-tackles-data-silo-issues-launch-iot-hub/. Accessed: 2020-09-03.

IQBAL, R., DOCTOR, F., MORE, B., MAHMUD, S. & YOUSUF, U. 2017. Big Data analytics and Computational Intelligence for Cyber-Physical Systems: Recent trends and state of the art applications. *Future Gener. Comp. Syst.* https://doi.org/10.1016/j.future.2017.10.021

JACOBS, J. & WEBBER-YOUNGMAN, R.C.W. 2017. A technology map to facilitate the process of mine modernization throughout the mining cycle. *J. S. Afr. Inst. Min. Metall.,* 117: 636-648. https://doi.org/10.17159/2411-9717/2017/v117n7a5

JAIN, A. & KUMAR, A.M. 2007. Hybrid neural network models for hydrologic time series forecasting. *Appl. Soft Comput.,* 7(2): 585-592. https://doi.org/10.1016/j.asoc.2006.03.002

JANG, J.S.R. 1993. ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Trans. Syst. Man Cybern. Syst.,* 23(3): 665-685. https://doi.org/10.1109/21.256541

JEONG, J.P., YEON, S., KIM, T., LEE, H., KIM, S.M. & KIM, S.-C. 2018. SALA: Smartphone-assisted localization algorithm for positioning indoor IoT devices. *Wirel. Netw.,* 24(1): 27-47. https://doi.org/10.1007/s11276-016-1309-9

JIFA, G. & LINGLING, Z. 2014. Data, DIKW, Big data and Data science. *Procedia Comput. Sci.,* 31: 814-821. https://doi.org/10.1016/j.procs.2014.05.332

JOHNSON, N.E., IANIUK, O., CAZAP, D., LIU, L., STAROBIN, D., DOBLER, G. & GHANDEHARI, M. 2017. Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. *Waste Manage.,* 62: 3-11. https://doi.org/10.1016/j.wasman.2017.01.037

KALIN, M. 2004. Passive mine water treatment: the correct approach? *Ecol. Eng.,* 22(4-5): 299-304. https://doi.org/10.1016/j.ecoleng.2004.06.008

KANSAKE, B.A., KABA, F.A., DUMAKOR-DUPEY, N.K. & ARTHUR, C.K. 2019. The future of mining in Ghana: Are stakeholders prepared for the adoption of autonomous mining systems? *Resour. Policy,* 63: 11. https://doi.org/10.1016/j.resourpol.2019.101411

KAR, S., DAS, S. & GHOSH, P.K. 2014. Applications of neuro fuzzy systems: A brief review and future outline. *Appl. Soft Comput.,* 15: 243-259. https://doi.org/10.1016/j.asoc.2013.10.014

KARABOGA, D. & AKAY, B. 2009. A survey: algorithms simulating bee swarm intelligence. *Artif. Intell. Rev.,* 31(1-4): 61-85. https://doi.org/10.1007/s10462-009-9127-4

KELM, A., LAUSSAT, L., MEINS-BECKER, A., PLATZ, D., KHAZAEE, M.J., COSTIN, A.M., HELMUS, M. & TEIZER, J. 2013. Mobile passive Radio Frequency Identification (RFID) portal for automated and rapid control of Personal Protective Equipment (PPE) on construction sites. *Automat. Constr.,* 36: 38-52. https://doi.org/10.1016/j.autcon.2013.08.009

KHANDELWAL, M. & SINGH, T. 2005. Prediction of mine water quality by physical parameters. *J. Sci. Ind. Res.,* 64: 564-570

KHASHEI, M. & BIJARI, M. 2010. An artificial neural network (*p*,*d*,*q*) model for timeseries forecasting. *Expert Syst. Appl.,* 37(1): 479-489. https://doi.org/10.1016/j.eswa.2009.05.044

KINERI, Y., WANG, M., LIN, H. & MAEKAWA, T. 2012. B-spline surface fitting by iterative geometric interpolation/approximation algorithms. *Comput. Aided Des.,* 44(7): 697-708. https://doi.org/10.1016/j.cad.2012.02.011

KOLHINEN, V., VENTO, T., JAKKILA, J., HUTTUNEN, M., KORPPOO, M. & VEHVILÄINEN, B. 2017. Operational Water Balance Model for Siilinjärvi Mine Site. In: Wolkersdorfer, C., Sartz, L., Sillanpää, M. & Häkkinen, A. (eds.). *IMWA 2017 – Mine Water & Circular Economy*. Vol. I. Lappeenranta: Lappeenranta University of Technology, 73-77.

KOZITSIN, V., KATSER, I. & LAKONTSEV, D. 2021. Online Forecasting and Anomaly Detection Based on the ARIMA Model. *Appl. Sci.,* 11(7): 3194. https://doi.org/10.3390/app11073194

KRAUSE, J., RUXTON, G.D. & KRAUSE, S. 2010. Swarm intelligence in animals and humans. *Trends Ecol. Evol.,* 25(1): 28-34. https://doi.org/10.1016/j.tree.2009.06.016

KRENKER, A., BESTER, J. & KOS, A. 2011. Introduction to the artificial neural networks. In: Suzuki, K. (ed.). *Artificial neural networks — methodological advances and biomedical applications*. Rijeka: InTech. https://doi.org/10.5772/15751

KUMAR, A., JAIN, A.K. & DUA, M. 2021. A comprehensive taxonomy of security and privacy issues in RFID. *Complex Intell. Syst.,* 7(3): 1327-1347. https://doi.org/10.1007/s40747-021-00280-6

KUMARI, L., NARSAIAH, K., GREWAL, M. & ANURAG, R. 2015. Application of RFID in agri-food sector. *Trends Food Sci. Tech.,* 43(2): 144-161. https://doi.org/10.1016/j.tifs.2015.02.005

KUNZ, N.C. & MORAN, C.J. 2016. The utility of a systems approach for managing strategic water risks at a mine site level. *Water Resour. Ind.,* 13: 1-6. https://doi.org/10.1016/j.wri.2016.02.001

LACEY, J., MALAKAR, Y., MCCREA, R. & MOFFAT, K. 2019. Public perceptions of established and emerging mining technologies in Australia. *Resour. Policy,* 62: 125-135. https://doi.org/10.1016/j.resourpol.2019.03.018

LACHOW, I. 2017. The upside and downside of swarming drones. *Bull. Atom. Scient.,* 73(2): 96-101. https://doi.org/10.1080/00963402.2017.1290879

LASI, H., FETTKE, P., KEMPER, H.G., FELD, T. & HOFFMANN, M. 2014. Industry 4.0. *Bus. Info. Syst. Eng.,* 6(4): 239-242. https://doi.org/10.1007/s12599-014-0334-4

LAW, R. 2000. Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tour. Manag.,* 21(4): 331-340. https://doi.org/10.1016/S0261-5177(99)00067-9

LEE, G., GOMMERS, R., WASELEWSKI, F., WOHLFAHRT, K. & O'LEARY, A. 2019. PyWavelets: A Python package for wavelet analysis. *J. Open Source Softw.,* 4(36): 1237. https://doi.org/10.21105/joss.01237

LEE, I. & LEE, K. 2015. The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Bus. Horizons,* 58(4): 431-440

LEE, J., LAPIRA, E., BAGHERI, B. & KAO, H.-A. 2013. Recent advances and trends in predictive manufacturing systems in big data environment. *Manuf. Lett.,* 1(1): 38-41. https://doi.org/10.1016/j.mfglet.2013.09.005

LI, F.H., CHEN, H. & XIAO, Y. 2010. *Handbook On Sensor Networks.* Hackensack, New Jersey: World Scientific.

LI, X. & BAI, R. 2016. Freight vehicle travel time prediction using gradient boosting regression tree. In: Dario, P., Dimitri, P. & Massimo, C. (eds.). *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE: 1010-1015. https://doi.org/10.1109/ICMLA.2016.101

LINDEMANN, B., MASCHLER, B., SAHLAB, N. & WEYRICH, M. 2021. A survey on anomaly detection for technical systems using LSTM networks. *Comput. Ind.,* 131: 103498. https://doi.org/10.1016/j.compind.2021.103498

LINDENBAUM, J., HENRY, O. & WIGGINS, A. 2007. *Heroku*. Available from: www.heroku.com. Accessed: 2021-03-12.

LITTLE, T.D., JORGENSEN, T.D., LANG, K.M. & MOORE, E.W.G. 2014. On the joys of missing data. *J. Pediatr. Psychol.,* 39(2): 151-162. https://doi.org/10.1093/jpepsy/jst048

LIU, Y. & FOMEL, S. 2011. Seismic data interpolation beyond aliasing using regularized nonstationary autoregression. *Geophysics,* 76(5): V69-V77. https://doi.org/10.1190/GEO2010-0231.1

LIU, Y., WANG, T. & YANG, J. 2019. Evaluating the Quality of Mine Water Using Hierarchical Fuzzy Theory and Fluorescence Regional Integration. *Mine Water Environ.,* 38(2): 243-251. https://doi.org/10.1007/s10230-018-0567-4

LÖÖW, J., ABRAHAMSSON, L. & JOHANSSON, J. 2019. Mining 4.0—the Impact of New Technology from a Work Place Perspective. *Min. Metall. Explor.,* 36(4): 701-707. https://doi.org/10.1007/s42461-019-00104-9

LORD, G., PARDO-IGÚZQUIZA, E. & SMITH, I. 2000. *A practical guide to wavelets for metrology* (No. CMSC 02/00). Middlesex: Software Support for Metrology Programme.

LOSAVIO, M., LAUF, A. & ELMAGHRABY, E. 2019. The internet of things and issues for mine water management. In: Wolkersdorfer, C., Khayrulina, E. & Bogush, A. (eds.). *International Mine Association Conference.* Perm: 678-683

LUVISI, A. & LORENZINI, G. 2014. RFID-plants in the smart city: Applications and outlook for urban green management. *Urban For. Urban Gree.,* 13(4): 630-637. https://doi.org/10.1016/j.ufug.2014.07.003

LUVISI, A., PANATTONI, A., RINALDELLI, E., PAGANO, M., MANNINI, F., GRIBAUDO, I. & BANDINELLI, R. 2014. Application of tracking implants in grape hybrids: Adjustments to production practices and new health-compliant methodologies. *Comp. Elec. Agr.,* 108: 130-134. https://doi.org/10.1016/j.compag.2014.07.013

MAHAPATRA, S.S., NANDA, S.K. & PANIGRAHY, B.K. 2011. A Cascaded Fuzzy Inference System for Indian river water quality prediction. *Adv. Eng. Softw.,* 42(10): 787-796. https://doi.org/10.1016/j.advengsoft.2011.05.018

MAIER, H.R., MORGAN, N. & CHOW, C.W.K. 2004. Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environ. Modell. Softw.,* 19(5): 485-494. https://doi.org/10.1016/S1364-8152(03)00163-4

MALČÍK, D. & DRAHANSKÝ, M. 2012. Anatomy of Biometric Passports. *J. Biomed. Biotechnol.,* 2012: Article ID 490362. https://doi.org/10.1155/2012/490362

MALHOTRA, P., VIG, L., SHROFF, G. & AGARWAL, P. 2015. Long short term memory networks for anomaly detection in time series. In: Verleysen, M. (ed.). *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges: ESANN: 89-94

MANASWI, N.K. 2018. *Deep learning with applications using python: Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras.* Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-3516-4

MANIKANDAN, S. 2010. Data transformation. *J. Pharmacol. Pharmacother.,* 1(2): 126. https://doi.org/10.4103/0976-500X.72373

MCCASKEY, A.J., DUMITRESCU, E.F., LIAKH, D., CHEN, M., FENG, W.-C. & HUMBLE, T.S. 2018. A language and hardware independent approach to quantum-classical computing. *SoftwareX,* 7: 245-254. https://doi.org/10.1016/j.softx.2018.07.007

MCCAULEY, C.A., O'SULLIVAN, A.D., MILKE, M.W., WEBER, P.A. & TRUMM, D.A. 2009. Sulfate and metal removal in bioreactors treating acid mine drainage dominated with iron and aluminum. *Water Res.,* 43(4): 961-970. https://doi.org/10.1016/j.watres.2008.11.029

MCCUE, C. 2014. *Data mining and predictive analysis: Intelligence gathering and crime analysis.* 2nd ed. Oxford: Elsevier.

MEI, L., HU, R., CAO, H., LIU, Y., HAN, Z., LI, F. & LI, J. 2019. Realtime Mobile Bandwidth Prediction Using LSTM Neural Network. In: Choffnes, D. & Barcellos, M. (eds.). *International Conference on Passive and Active Network Measurement. PAM 2019.*: Springer. https://doi.org/10.1007/978-3-030-15986-3_3

MICHELUCCI, U. 2018. *Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks.* New York: Springer. https://doi.org/10.1007/978-1-4842-3790-8

MICKLETHWAITE, S. 2018. Drones in mining — the new possible. *AusIMM Bulletin*(Oct 2018): 32

MILLIE, D.F., WECKMAN, G.R., YOUNG II, W.A., IVEY, J.E., FRIES, D.P., ARDJMAND, E. & FAHNENSTIEL, G.L. 2013. Coastal 'Big Data' and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric. *Estuar. Coast. Shelf Sci.,* 125: 57-67

MIORANDI, D., SICARI, S., DE PELLEGRINI, F. & CHLAMTAC, I. 2012. Internet of things: Vision, applications and research challenges. *Ad hoc networks,* 10(7): 1497-1516. https://doi.org/10.1016/j.adhoc.2012.02.016

MO, L.-F., ZHANG, H.-J. & ZHOU, H.-L. 2009. Analysis of dipole-like ultra high frequency RFID tags close to metallic surfaces. *J. Zhejiang Univ. Sci. A,* 10(8): 1217-1222. https://doi.org/10.1631/jzus.A0820495

MOLINA, S., SOTO, I. & CARRASCO, R. 2011. Detection of gases and collapses in underground mines using WSN. In: Abbott, D. (ed.). *2011 IEEE International Conference on Industrial Technology.* Auburn: IEEE: 219-225. https://doi.org/10.1109/ICIT.2011.5754376

MONTAVON, G., SAMEK, W. & MÜLLER, K.-R. 2018. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.,* 73: 1-15. https://doi.org/10.1016/j.dsp.2017.10.011

MORE, K.S. & WOLKERSDORFER, C. 2019. Disruptive Technologies in Mine Water Management – The Future. In: Wolkersdorfer, C., Khayrulina, E., Polyakova, S. & Bogush, A. (eds.). *Mine Water – Technological and Ecological Challenges.* Perm: 597-602

MORE, K.S. & WOLKERSDORFER, C. 2022. Predicting and Forecasting Mine Water Parameters Using a Hybrid Intelligent System. *Water Resour. Manage.* https://doi.org/10.1007/s11269-022-03177-2

MORE, K.S., WOLKERSDORFER, C., KANG, N. & ELMAGHRABY, A.E. 2020. Automated measurement systems in mine water management and mine workings — A review of potential methods. *Water Resour. Ind.,* 24: 1-12. https://doi.org/10.1016/j.wri.2020.100136

MOUSAVI, A. & SELLERS, E. 2019. Optimisation of production planning for an innovative hybrid underground mining method. *Resour. Policy,* 62: 184-192. https://doi.org/10.1016/j.resourpol.2019.03.002

MUTHUKUMAR, N., SRINIVASAN, S., RAMKUMAR, K., PAL, D., VAIN, J. & RAMASWAMY, S. 2019. A model-based approach for design and verification of Industrial Internet of Things. *Future Gener. Comp. Syst.,* 95: 354-363. https://doi.org/10.1016/j.future.2018.12.012

NALECKI, P. & GOWAN, M. 2008. Mine Water Management–Dynamic, Probabilistic Modelling Approach. In: Rapantova, N. (ed.). *10th International Mine Water Association Congress.* Karlsbad, Czech Republic: 533-536

NANDA, N.K. 2020. Intelligent Enterprise with Industry 4.0 for Mining Industry. In: Topal, E. (ed.). *Proceedings of the 28th International Symposium on Mine Planning and Equipment Selection MPES 2019*. Cham: Springer, 213-218. https://doi.org/10.1007/978-3-030-33954-8_27

NASR, A.S., REZAEI, M. & BARMAKI, M.D. 2012. Analysis of groundwater quality using mamdani fuzzy inference system (MFIS) in Yazd Province, Iran. *Int. J. Comp. App.,* 59(7):

NAYAK, P., SUDHEER, K., RANGAN, D. & RAMASASTRI, K. 2005. Short-term flood forecasting with a neurofuzzy model. *Water Resour. Res.,* 41(4). https://doi.org/10.1029/2004WR003562

NEGNEVITSKY, M. 2005. *Artificial intelligence: a guide to intelligent systems.* Edinburgh: Pearson Education.

NEWMAN, D.A. 2014. Missing data: Five practical guidelines. *Organ. Res. Methods,* 17(4): 372-411. https://doi.org/10.1177/1094428114548590

NGAI, E.W., CHENG, T.E., AU, S. & LAI, K.-H. 2007. Mobile commerce integrated with RFID technology in a container depot. *Decision Support Syst.,* 43(1): 62-76. https://doi.org/10.1016/j.dss.2005.05.006

NYONGESA, D. 2020. Variable Selection Using Random Forests in SAS®. In: Duhon, D. (ed.). *SAS Global Forum*. SASGF: 1-11

OLALEKAN, O., AFEES, O. & AYODELE, S. 2016. An Empirical Analysis of the Contribution of Mining Sector to Economic Development in Nigeria. *Khazar J. Human. Social Sci.,* 19(1): 88-104

OLOFSSON, P. & ANDERSSON, M. 2012. *Probability, statistics, and stochastic processes.* 2nd ed. Hoboken: Wiley.

OSISANWO, F., AKINSOLA, J., AWODELE, O., HINMIKAIYE, J., OLAKANMI, O. & AKINJOBI, J. 2017. Supervised machine learning algorithms: classification and comparison. *Int. J. Comput. Trends Technol.,* 48(3): 128-138. https://doi.org/10.14445/22312803/IJCTT-V48P126

OTTO, A., AGATZ, N., CAMPBELL, J., GOLDEN, B. & PESCH, E. 2018. Optimization approaches for civil applications of unmanned aerial vehicles (UAVs) or aerial drones: A survey. *Networks,* 72(4): 411-458. https://doi.org/10.1002/net.21818

OUDRE, L. 2018. Interpolation of missing samples in sound signals based on autoregressive modeling. *Image Process. Line,* 8: 329-344. https://doi.org/10.5201/ipol.2018.23

ØYEN, S. 2018. Forecasting multivariate time series data using neural networks. MSc in Cybernetics and Robotics, Trondheim, Norwegian University of Science and Technology

PALANI, S., LIONG, S.-Y. & TKALICH, P. 2008. An ANN application for water quality forecasting. *Mar. Pollut. Bull.,* 56(9): 1586-1597. https://doi.org/10.1016/j.marpolbul.2008.05.021

PALMA, W. 2016. *Time series analysis.* Hoboken: Wiley. Probability and Statistics.

PAOLELLA, M.S. 2018. *Linear models and time-series analysis: regression, ANOVA, ARMA and GARCH.* Hoboken: Wiley. Probability and Statistics.

PARK, S. & CHOI, Y. 2020. Applications of unmanned aerial vehicles in mining from exploration to reclamation: A review. *Minerals,* 10(8): 663. https://doi.org/10.3390/min10080663

PAUL, M., METSCHIES, T., FRENZEL, M. & MEYER, J. 2011. The Mean Hydraulic Residence Time and its Use for Assessing the Longevity of Mine Water Pollution from Flooded Underground Mines. In: Merkel, B. & Schipek, M. (eds.). *The New Uranium Mining Boom. Springer Geology*. Heidelberg: Springer, 689-699. https://doi.org/10.1007/978-3-642-22122-4_79

PERRET, E. 2014. *Radio Frequency Identification and Sensors: From RFID to Chipless RFID.* Hoboken: Wiley.

PIGINI, D. & CONTI, M. 2017. NFC-based traceability in the food chain. *Sustainability,* 9(10): 1910. https://doi.org/10.3390/su9101910

PRASANNA, K.R. & HEMALATHA, M. 2012. RFID GPS and GSM based logistics vehicle load balancing and tracking mechanism. *Procedia Eng.,* 30: 726-729. https://doi.org/10.1016/j.proeng.2012.01.920

PRESTON-WERNER, T., HYETT, P.J., WANSTRATH, C. & CHACON, S. 2008. *GitHub*. Available from: www.github.com. Accessed: 2021-02-11.

QI, P., ZHANG, G., XU, Y.J., WANG, L., DING, C. & CHENG, C. 2018. Assessing the influence of precipitation on shallow groundwater table response using a combination of singular value decomposition and cross-wavelet approaches. *Water,* 10(5): 598. https://doi.org/10.3390/w10050598

QING, X. & CHEN, Z.N. 2007. Proximity Effects of Metallic Environments on High Frequency RFID Reader Antenna: Study and Applications. *IEEE Trans. Antennas Propag.,* 55(11): 3105-3111. https://doi.org/10.1109/TAP.2007.908575

RAUTIO, A.B., KORKKA-NIEMI, K. & SALONEN, V.-P. 2017. Thermal infrared remote sensing in assessing ground/surface water resources related to Hannukainen mining development site, Northern Finland. In: Wolkersdorfer, C., Sartz, L., Sillanpää, M. & Häkkinen, A. (eds.). *IMWA 2017 – Mine Water & Circular Economy*. Vol. II. Lappeenranta: Lappeenranta University of Technology, 1290-1296.

REICH, N.G., LESSLER, J., SAKREJDA, K., LAUER, S.A., IAMSIRITHAWORN, S. & CUMMINGS, D.A. 2016. Case study in evaluating time series prediction models using the relative mean absolute error. *Am. Stats.,* 70(3): 285-292. https://doi.org/10.1080/00031305.2016.1148631

REMPT, B. 2001. *GUI programming with Python: Qt edition.* Los Angeles: CommandPrompt.

RODRIGUEZ, J. 2016. *Fundamentals of 5G mobile networks.* Jersey City: Wiley.

ROGERS, D.F. 2001. *An introduction to NURBS: with historical perspective.* Amsterdam: Elsevier. https://doi.org/10.1016/B978-1-55860-669-2.X5000-3

ROMERO, M.P., CHANG, Y.-M., BRUNTON, L.A., PARRY, J., PROSSER, A., UPTON, P., REES, E., TEARNE, O., ARNOLD, M. & STEVENS, K. 2020. Decision tree machine learning applied to bovine tuberculosis risk factors to aid disease control decision making. *Prev. Vet. Med.,* 175: 104860. https://doi.org/10.1016/j.prevetmed.2019.104860

ROOKI, R., DOULATI ARDEJANI, F., ARYAFAR, A. & BANI ASADI, A. 2011. 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.,* 64(5): 1303-1316. https://doi.org/10.1007/s12665-011-0948-5

RUPARELIA, N. 2016. *Cloud Computing.* Cambridge: The MIT Press.

RUSSELL, S. & NORVIG, P. 2002. *Artificial intelligence: a modern approach.* New Jersey: Prentice Hall.

SADEGHI, A.-R., WACHSMANN, C. & WAIDNER, M. 2015. Security and privacy challenges in industrial internet of things. In: Li, X. (ed.). *52nd ACM/EDAC/IEEE Design Automation Conference (DAC).* San Francisco: IEEE: 1-6. https://doi.org/10.1145/2744769.2747942.

SAEIPOURDIZAJ, P., SARBAKHSH, P. & GHOLAMPOUR, A. 2021. Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbor methods. *Environ. Health Eng. Manag. J.,* 8(3): 215-226. https://doi.org/10.34172/EHEM.2021.25

SAHU, M., MAHAPATRA, S.S., SAHU, H. & PATEL, R. 2011. Prediction of water quality index using neuro fuzzy inference system. *Water Qual. Expos. Hea.,* 3: 175-191. https://doi.org/10.1007/s12403-011-0054-7

SAKALA, E., NOVHE, O. & VADAPALLI, V.R.K. 2019. Application of Artificial Intelligence (AI) to predict mine water quality, a case study in South Africa. In: Wolkersdorfer, C., Khayrulina, E., Polyakova, S. & Bogush, A. (eds.). *International Mine Water Association Congress.* Perm: 140-145

SAKIZADEH, M. 2015. Artificial intelligence for the prediction of water quality index in groundwater systems. *Modell. Earth Syst. Environ.,* 2(1): 8. https://doi.org/10.1007/s40808-015-0063-9

SARMA, A.C. & GIRÃO, J. 2009. Identities in the future internet of things. *Wirel. Pers. Commun.,* 49(3): 353-363. https://doi.org/10.1007/s11277-009-9697-0

SEROV, V. 2017. *Fourier series, Fourier transform and their applications to mathematical physics.* New York: Springer. Applied Mathematical Sciences. https://doi.org/10.1007/978-3-319-65262-7

SEVILLA, R., FERNÁNDEZ-MÉNDEZ, S. & HUERTA, A. 2008. NURBS-enhanced finite element method (NEFEM). *Int. J. Numer. Methods Eng.,* 76(1): 56-83. https://doi.org/10.1002/nme.2311

SEXTON, J. & LAAKE, P. 2009. Standard errors for bagged and random forest estimators. *Comput. Stat. Data Anal.,* 53(3): 801-811. https://doi.org/10.1016/j.csda.2008.08.007

SHAO, C. & XIAO, L. 2011. Nurbs model for chaotic time series. In: Massimo, C., Alessandro, B. & Dimitri, P. (eds.). *3rd International Conference on Computer Research and Development*. IEEE: 135-138

SHEPARD, S. 2005. *RFID: radio frequency identification.* New York: McGraw Hill.

SHROUF, F., ORDIERES, J. & MIRAGLIOTTA, G. 2014. Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm. In: Abbott, D. (ed.). *IEEE international conference on industrial engineering and engineering management.* Selangor, Malaysia: IEEE: 697-701. https://doi.org/10.1109/IEEM.2014.7058728

SILER, W. & BUCKLEY, J.J. 2005. *Fuzzy expert systems and fuzzy reasoning.* Hoboken: Wiley.

SILVA-PEDROZA, D., MARIN-CALERO, R. & RAMIREZ-GONZALEZ, G. 2017. NFC Evaluation in the Development of Mobile Applications for MICE in Tourism. *Sustainability,* 9(11): 20. https://doi.org/10.3390/su9111937

SINGER, P.C. & STUMM, W. 1970. Acidic mine drainage: the rate-determining step. *Science,* 167(3921): 1121-1123. https://doi.org/10.1126/science.167.3921.1121

SINGH, B., SIHAG, P. & SINGH, K. 2017a. Modelling of impact of water quality on infiltration rate of soil by random forest regression. *Model. Earth. Syst. Environ.,* 3(3): 999-1004. https://doi.org/10.1007/s40808-017-0347-3

SINGH, K.P., BASANT, A., MALIK, A. & JAIN, G. 2009. Artificial neural network modeling of the river water quality — a case study. *Ecol. Modell.,* 220(6): 888-895. https://doi.org/10.1016/j.ecolmodel.2009.01.004

SINGH, R., SINGH, E. & NALWA, H.S. 2017b. Inkjet printed nanomaterial based flexible radio frequency identification (RFID) tag sensors for the internet of

nano things. *RSC Adv.,* 7(77): 48597-48630. https://doi.org/10.1039/C7RA07191D

SISHI, M.N. & TELUKDARIE, A. 2017. Implementation of industry 4.0 technologies in the mining industry: A case study. In: Shi, J., Kujala, J., Lam, J.S.L. & Lim, T.J. (eds.). *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).* Singapore: IEEE: 201-205

SKANSI, S. 2018. *Introduction to Deep Learning: from logical calculus to artificial intelligence.* Cham: Springer. https://doi.org/10.1007/978-3-319-73004-2

SONG, H., SUI, S., HAN, Q., ZHANG, H. & YANG, Z. 2020. Autoregressive integrated moving average model–based secure data aggregation for wireless sensor networks. *Int. J. Distrib. Sens. Netw.,* 16(3): 1550147720912958. https://doi.org/10.1177/1550147720912958

SONG, J.-L., GAO, H.-W. & SONG, Y.-J. 2010. Research on transceiver system of WSN based on V-MIMO underground coal mines. In: Abbott, D. (ed.). *2010 International Conference on Communications and Mobile Computing.* Shenzhen: IEEE: 374-378. https://doi.org/10.1109/CMC.2010.287

SRI CONSULTING BUSINESS INTELLIGENCE. 2008. The Internet of things (background). In: Abdi, F., Adams, W., Alarcon, E. & Ansari, M. (eds.). *TechConnect World Innovation Conference and Expo.* United States of America: National Intelligence Council:

STOCK, T. & SELIGER, G. 2016. Opportunities of sustainable manufacturing in industry 4.0. *Procedia CIRP,* 20: 536-541. https://doi.org/10.1016/j.procir.2016.01.129

SUDHEER, C. & MATHUR, S. 2010. Modeling uncertainty analysis in flow and solute transport model using adaptive neuro fuzzy inference system and particle swarm optimization. *KSCE J. Civil Eng.,* 14(6): 941-951. https://doi.org/10.1007/s12205-010-0865-2

SUN, B., AHMED, F., SUN, F., QIAN, Q. & XIAO, Y. 2016. Water quality monitoring using STORM 3 Data Loggers and a wireless sensor network. *Int. J. Sens. Netw.,* 20(1): 26-36. https://doi.org/10.1504/ijsnet.2016.074270

SUN, E., ZHANG, X. & LI, Z. 2012. The internet of things (IOT) and cloud computing (CC) based tailings dam monitoring and pre-alarm system in mines. *Saf. Sci.,* 50(4): 811-815. https://doi.org/10.1016/j.ssci.2011.08.028

SUNNY, A.I., TIAN, G.Y., ZHANG, J. & PAL, M. 2016. Low frequency (LF) RFID sensors and selective transient feature extraction for corrosion characterisation. *Sensor. Actuat. A-Phys.,* 241: 34-43. https://doi.org/10.1016/j.sna.2016.02.010

SUTTON, S.G., HOLT, M. & ARNOLD, V. 2016. “The reports of my death are greatly exaggerated”—Artificial intelligence research in accounting. *Int. J. Account. Inf. Syst.,* 22: 60-73. https://doi.org/10.1016/j.accinf.2016.07.005

SVOZIL, D., KVASNICKA, V. & POSPICHAL, J. 1997. Introduction to multi-layer feed-forward neural networks. *Chemom. Intell. Lab. Syst.,* 39(1): 43-62. https://doi.org/10.1016/S0169-7439(97)00061-0

SWAMYNATHAN, M. 2017. *Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python.* New York: Springer. https://doi.org/10.1007/978-1-4842-2866-1

SWEENEY, P.J. 2010. *RFID for Dummies.* Jersey City: Wiley.

TABAKIN, F. 2017. Model Dynamics for Quantum Computing. *Ann. Phys.,* 383: 33-78. https://doi.org/10.1016/j.aop.2017.04.013

TAHIR, A., BOLING, J., HAGHBAYAN, M.H., TOIVONEN, H.T. & PLOSILA, J. 2019. Swarms of Unmanned Aerial Vehicles – A Survey. *J. Ind. Inf. Integr.,* 16: Article 100106. https://doi.org/10.1016/j.jii.2019.100106

TEKINER, F. & KEANE, J.A. 2013. Big data framework. In: Abbott, D. (ed.). *2013 IEEE International Conference on Systems, Man, and Cybernetics.* Manchester: IEEE: 1494-1499. https://doi.org/10.1109/SMC.2013.258

TENG, M. 2010. Anomaly detection on time series. In: Wang, Y. & Luo, Y. (eds.). *International Conference on Progress in Informatics and Computing*. IEEE: 603-608

TORRENCE, C. & COMPO, G.P. 1998. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.,* 79(1): 61-78. https://doi.org/10.1175/1520-0477(1998)079

TUNA, G., ARKOC, O. & GULEZ, K. 2013. Continuous Monitoring of Water Quality Using Portable and Low-Cost Approaches. *Int. J. Distrib. Sens. Netw.,* 2013: Article ID 249598. https://doi.org/10.1155/2013/249598

TZOUNIS, A., KATSOULAS, N., BARTZANAS, T. & KITTAS, C. 2017. Internet of Things in agriculture, recent advances and future challenges. *Biosyst. Eng.,* 164: 31-48. https://doi.org/10.1016/j.biosystemseng.2017.09.007

UR REHMAN, M.H., YAQOOB, I., SALAH, K., IMRAN, M., JAYARAMAN, P.P. & PERERA, C. 2019. The role of big data analytics in industrial Internet of Things. *Future Gener. Comp. Syst.,* 99: 247-259. https://doi.org/10.1016/j.future.2019.04.020

USHER, B., STRAND, R., STRACHOTTA, C. & JACKSON, J. 2010. Linking fundamental geochemistry and empirical observations for water quality predictions using Goldsim. In: Wolkersdorfer, C. & Freund, A. (eds.). *Mine Water and Innovative Thinking—International Mine Water Association Congress.* Sydney, Canada: 313-317

VADAPALLI, V.R.K., SAKALA, E., DUBE, G. & COETZEE, H. 2020. Mine Water Treatment and the Use of Artificial Intelligence in Acid Mine Drainage Prediction. In: Fosso-Kankeu, E., Wolkersdorfer, C. & Burgess, J. (eds.). *Recovery of Byproducts from Acid Mine Drainage Treatment*. Scrivener, 23-48. https://doi.org/10.1002/9781119620204.ch2

VAN ROSSUM, G. 1991. *Python*. Available from: www.python.org. Accessed: 2019-03-27.

VEER, K. & AGARWAL, R. 2015. Wavelet and short-time Fourier transform comparison-based analysis of myoelectric signals. *J. Appl. Stat.,* 42(7): 1591-1601. https://doi.org/10.1080/02664763.2014.1001728

VERBURG, R. 2011. Mitigating Acid Rock Drainage. In: Darling, P. (ed.). *SME Mining Engineering Handbook*. 3 ed. ed., Vol. 1. Littleton: Society for Mining Metallurgy and Exploration, 1721-1732.

VOULODIMOS, A.S., PATRIKAKIS, C.Z., SIDERIDIS, A.B., NTAFIS, V.A. & XYLOURI, E.M. 2010. A complete farm management system based on animal identification using RFID technology. *Comput. Electron. Agr.,* 70(2): 380-388. https://doi.org/10.1016/j.compag.2009.07.009

WACKERLY, D., MENDENHALL, W. & SCHEAFFER, R.L. 2014. *Mathematical Statistics with Applications.* 7th ed. Belmont: Thomson.

WAGNER, H.N.R., KÖKE, H., DÄHNE, S., NIEMANN, S., HÜHNE, C. & KHAKIMOVA, R. 2019. Decision tree-based machine learning to optimize the laminate stacking of composite cylinders for maximum buckling load and minimum imperfection sensitivity. *Compos. Struct.,* 220: 45-63. https://doi.org/10.1016/j.compstruct.2019.02.103

WALTERS, L.P. 2011. *Applications of Swarm Intelligence.* New York: Nova Science. Engineering Tools, Techniques and Tables.

WANG, P. & OLIPHANT, T. 2012. *Anaconda*. Available from: www.anaconda.com. Accessed: 2020-01-15.

WANG, Y.-M., WANG, Y.-S. & YANG, Y.-F. 2010. Understanding the determinants of RFID adoption in the manufacturing industry. *Technol. Forecast Soc.,* 77(5): 803-815. https://doi.org/10.1016/j.techfore.2010.03.006

WOLFGANG, E. 2011. *Introduction to artificial intelligence.* Cham: Springer. Undergraduate Topics in Computer Science. https://doi.org/10.1007/978-3-319-58487-4

WOLKERSDORFER, C. 2008. *Water Management at Abandoned Flooded Underground Mines – Fundamentals, Tracer Tests, Modelling, Water Treatment.* 1st ed. Heidelberg: Springer. https://doi.org/10.1007/978-3-540-77331-3

WOLKERSDORFER, C. 2013. Management von Grubenwasser 3.0 – Blick in die Zukunft. *Wiss. Mitt. Inst. Geol.*(44): 105-113

WOLKERSDORFER, C., MUGOVA, E., DAGA, V.S., CHARVET, P. & VITULE, J.R.S. 2022. Effects of Mining on Surface Water – Case Studies. In: Irvine, K., Chapman, D. & Warner, S. (eds.). *The Encyclopedia of Inland Waters*. 2nd ed., Vol. 4. Oxford: Elsevier, 210-224. https://doi.org/10.1016/B978-0-12-819166-8.00036-0

WONG, B., MILNER, T.E., ANVARI, B., SVIRIDOV, A., OMEL'CHENKO, A., BAGRATASHVILI, V., SOBOL, E. & NELSON, J. 1998. Measurement of radiometric surface temperature and integrated backscattered light intensity during feedback-controlled laser-assisted cartilage reshaping. *Laser Med. Sci.,* 13(1): 66-72. https://doi.org/10.1007/BF00592961

WORTMANN, F. & FLÜCHTER, K. 2015. Internet of things. *Bus. Info. Syst. Eng.,* 57(3): 221-224. https://doi.org/10.1007/s12599-015-0383-3

XIANG, W., ZHENG, K. & SHEN, X.S. 2017. *5G mobile communications.* Cham: Springer. https://doi.org/10.1007/978-3-319-34208-5

XIANMING, Q., ZHI NING, C. & AILIAN, C. 2007, 9-15 June 2007. Multi-loop antenna for high frequency RFID smart shelf application. In: Burkholder, R.J., Peterson, A. & Hong, W. (eds.). *2007 IEEE Antennas and Propagation Society International Symposium.* IEEE: 5467-5470. https://doi.org/10.1109/APS.2007.4396785

XIE, M., GOH, T. & TANG, X. 2000. Data transformation for geometrically distributed quality characteristics. *Qual. Reliab. Eng. Int.,* 16(1): 9-15. https://doi.org/10/ctw9qp

YANG, M., KIANG, M. & SHANG, W. 2015. Filtering big data from social media–Building an early warning system for adverse drug reactions. *J. Biomed. Inform.,* 54: 230-240. https://doi.org/10.1016/j.jbi.2015.01.011

YIN, Y., STECKE, K.E. & LI, D. 2018. The evolution of production systems from Industry 2.0 through Industry 4.0. *Int. J. Prod. Res.,* 56(1-2): 848-861. https://doi.org/10.1080/00207543.2017.1403664

YOUNGER, P. 1997. The longevity of mine water pollution – a basis for decision making. *Sci. Total Environ.,* 194-195: 457-466. https://doi.org/10.1016/S0048-9697(96)05383-1

YOUNGER, P.L., BANWART, S.A. & HEDIN, R.S. 2002. *Mine water: hydrology, pollution, remediation.* Dordrecht: Kluwer Academic/Springer.

YU, Q., JIBIN, L. & JIANG, L. 2016. An improved ARIMA-based traffic anomaly detection algorithm for wireless sensor networks. *Int. J. Distrib. Sens. Netw.,* 12(1): 9653230. https://doi.org/10.1155/2016/9653230

YUAN, H., XIONG, F. & HUAI, X. 2003. A method for estimating the number of hidden neurons in feed-forward neural networks based on information entropy. *Comput. Electron. Agr.,* 40(1-3): 57-64. https://doi.org/10.1016/S0168-1699(03)00011-5

ZADEH, L.A. 1965. Fuzzy sets. *Inform. Control,* 8(3): 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

ZHANG, F., KEIVANLOO, I. & ZOU, Y. 2017. Data Transformation in Cross-project Defect Prediction. *Empir. Software Eng.,* 22(6): 3186-3218. https://doi.org/10.1007/s10664-017-9516-2

ZHANG, G., EDDY PATUWO, B. & Y. HU, M. 1998. Forecasting with artificial neural networks: The state of the art. *Int. J. Forecast.,* 14(1): 35-62. https://doi.org/10.1016/S0169-2070(97)00044-7

ZHANG, G.P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomput.,* 50: 159-175. https://doi.org/10.1016/S0925-2312(01)00702-0

ZHANG, Y. & HAGHANI, A. 2015. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.,* 58: 308-324. https://doi.org/10.1016/j.trc.2015.02.019

ZHANG, Z., MAYER, G., DAUVILLIERS, Y., PLAZZI, G., PIZZA, F. & FRONCZEK, R. 2018. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning. *Sci. Rep.,* 8(1): 10628. https://doi.org/10.1038/s41598-018-28840-w

ZHU, X., MUKHOPADHYAY, S.K. & KURATA, H. 2012. A review of RFID technology and its managerial applications in different industries. *J. Eng. Technol. Manage.,* 29(1): 152-167. https://doi.org/10.1016/j.jengtecman.2011.09.011

ZHUIYKOV, S. 2012. Solid-state sensors monitoring parameters of water quality for the next generation of wireless sensor networks. *Sens. Actuators B Chem.,* 161(1): 1-20. https://doi.org/10.1016/j.snb.2011.10.078

ZIKOPOULOS, P. & EATON, C. 2011. *Understanding big data: Analytics for enterprise class hadoop and streaming data.* New York: McGraw-Hill.

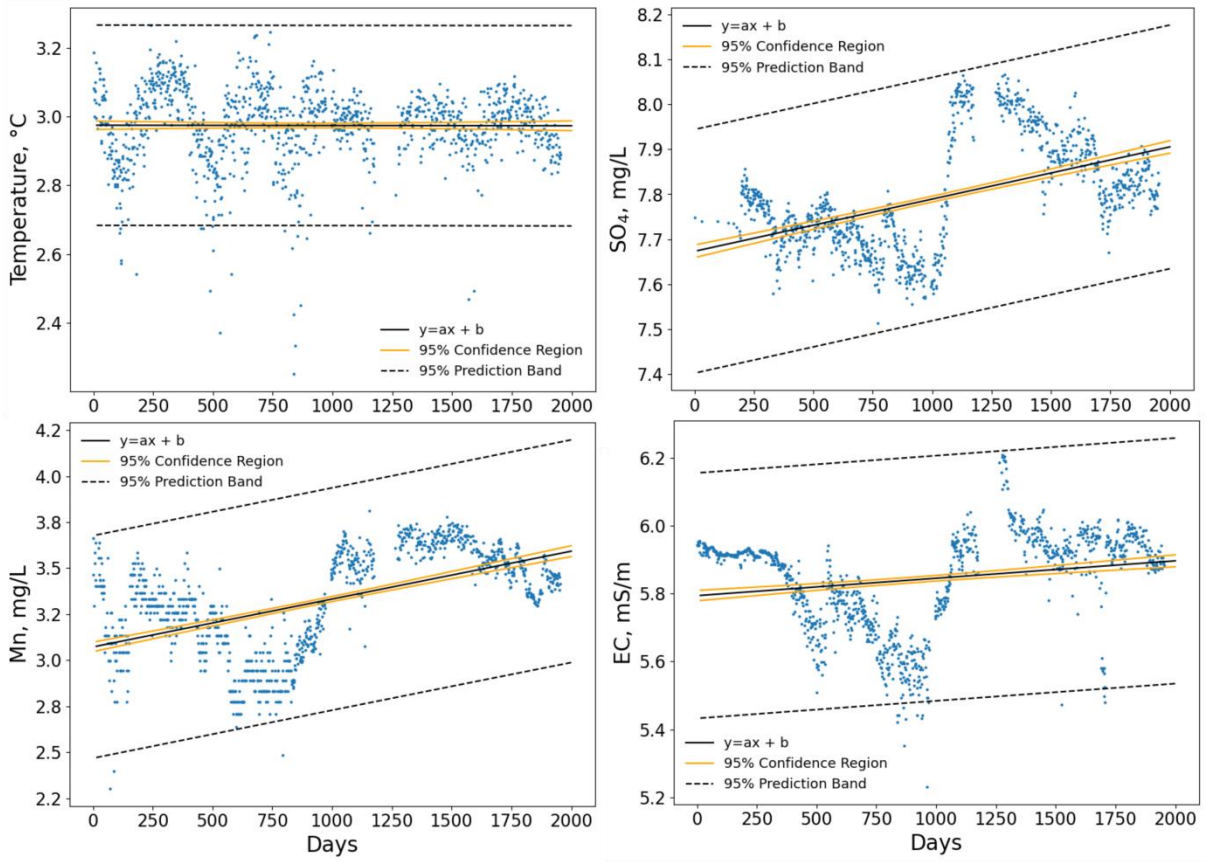
ZUEHLKE, D. 2010. Smart Factory — Towards a factory-of-things. *Annu. Rev. Control,* 34(1): 129-138. https://doi.org/10.1016/j.arcontrol.2010.02.008

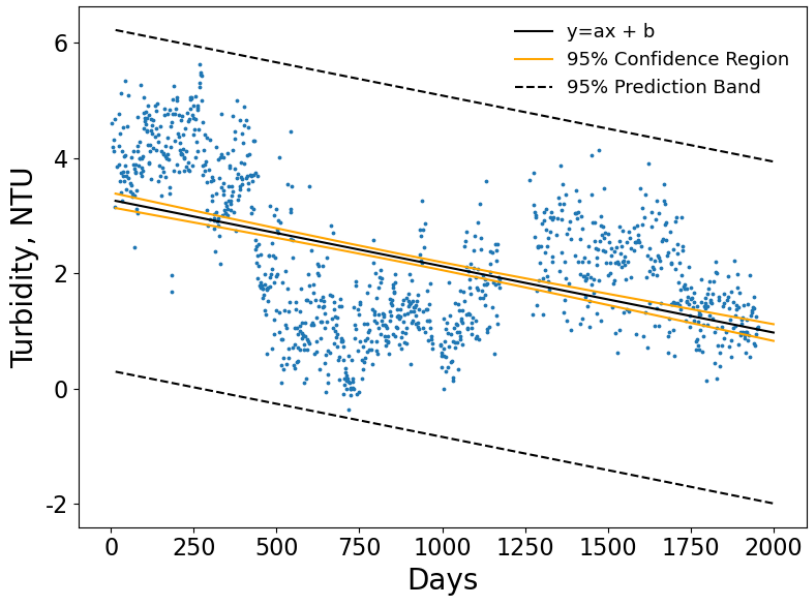
ZULKIFLI, S.N., RAHIM, H.A. & LAU, W.-J. 2018. Detection of contaminants in water supply: A review on state-of-the-art monitoring technologies and their applications. *Sens. Actuators B Chem.,* 255: 2657-2689. https://doi.org/doi.org/10.1016/j.snb.2017.09.078

ZUO, R. 2017. Machine learning of mineralization-related geochemical anomalies: A review of potential methods. *Nat. Resour. Res.,* 26(4): 457-464. https://doi.org/10.1007/s11053-017-9345-4

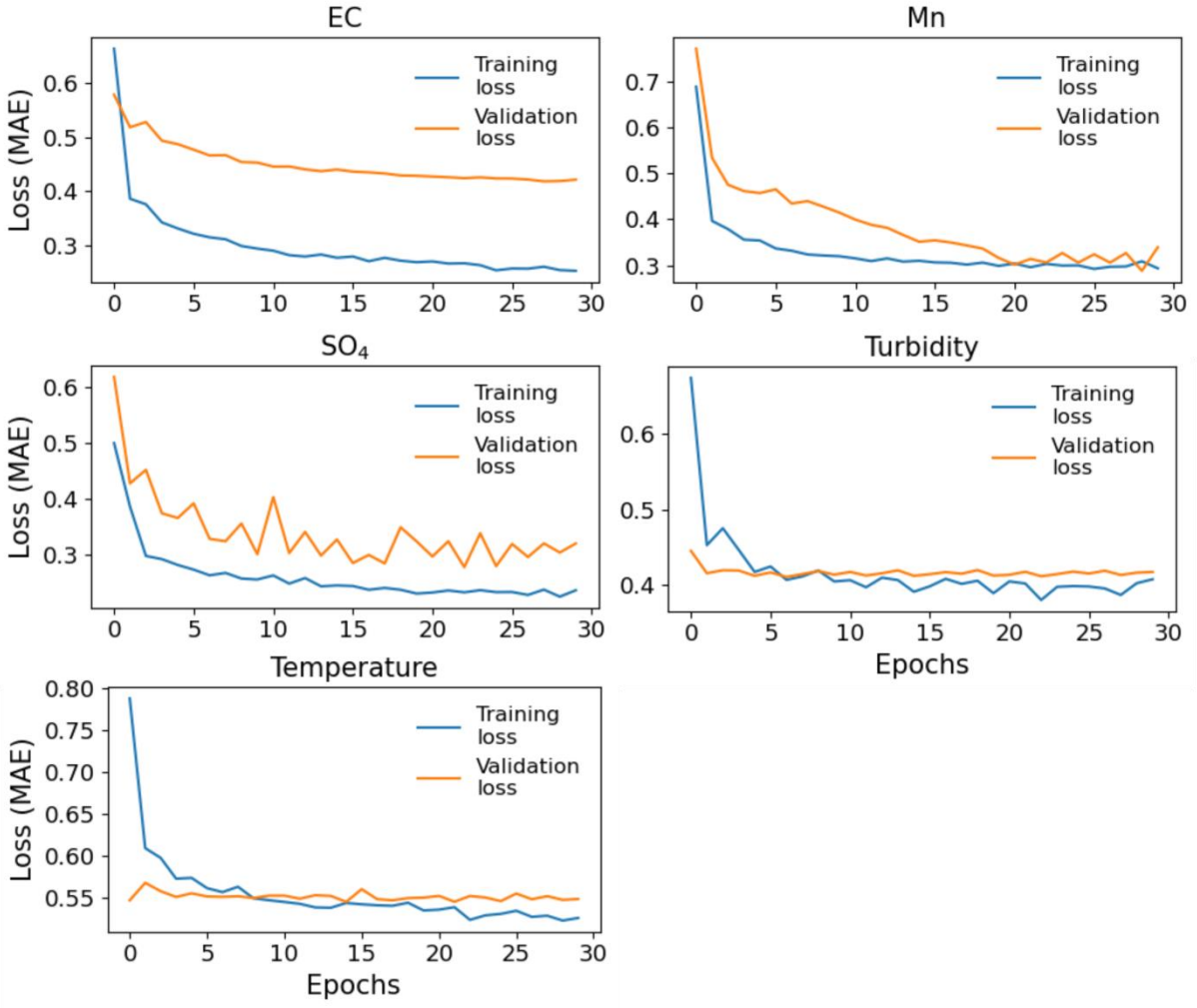
Annexure A Gradient boosting regressors to detect anomalies

Annexure B Robust linear regression with upper and lower bands of 95% confidence and prediction levels for anomaly detection

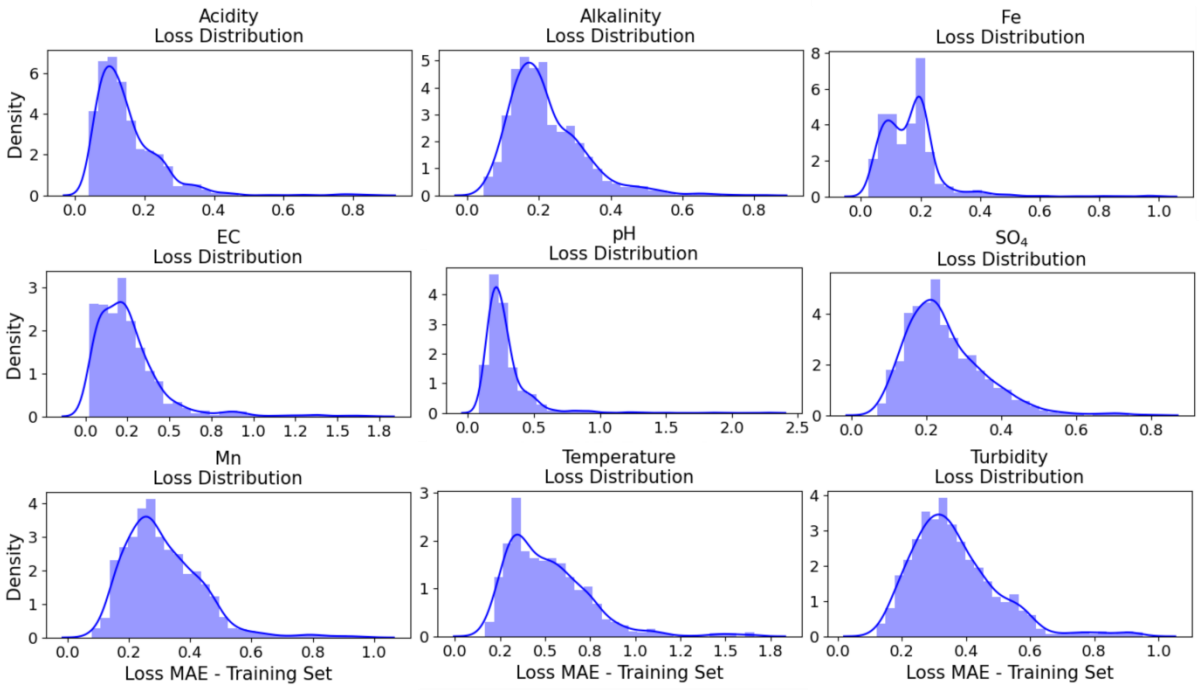


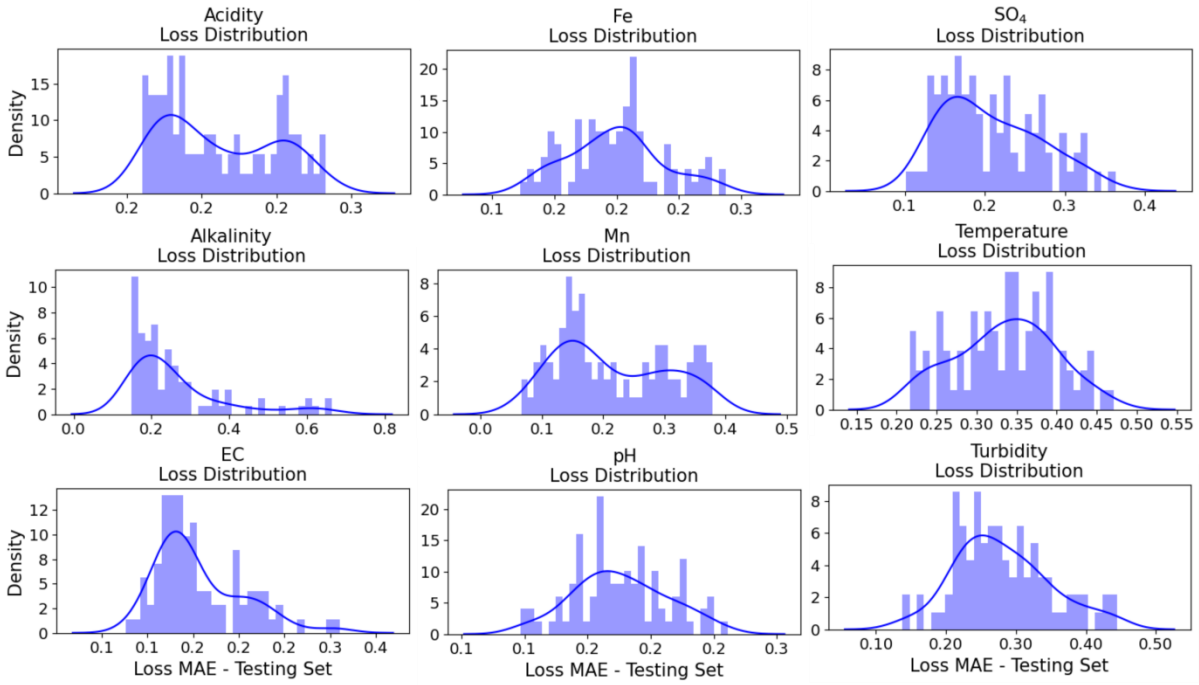


Annexure C LSTM Autoencoder Model performances for anomaly detection

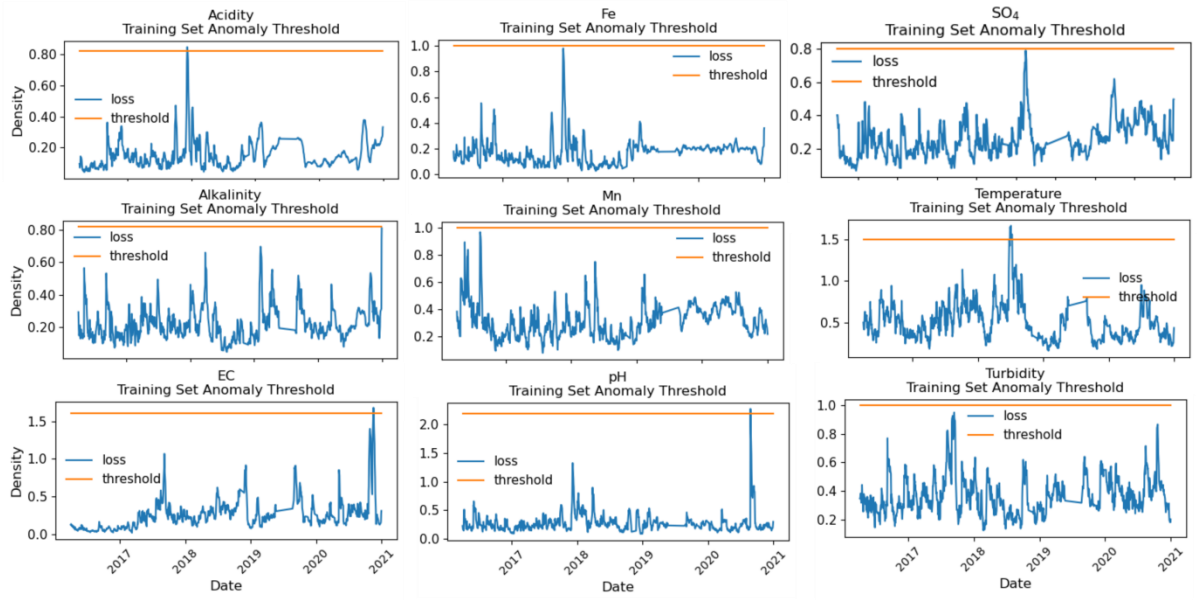


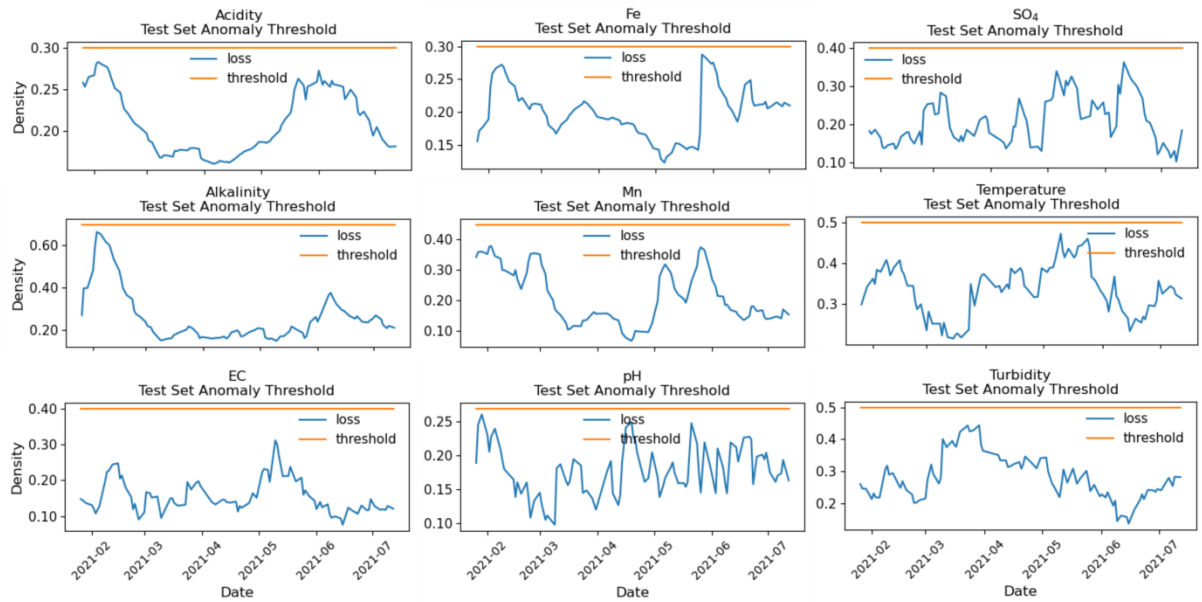
Annexure D LSTM Autoencoder Loss distribution (training and testing sets)



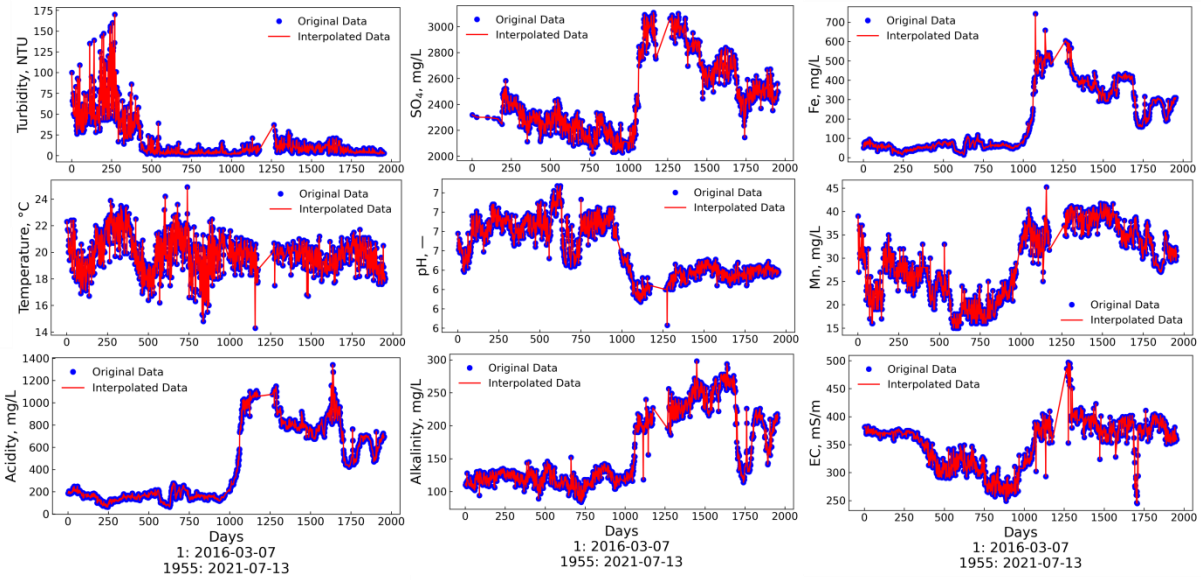


Annexure E LSTM Autoencoder Anomaly threshold (training and testing sets)





Annexure F Interpolated missing values using B-Splines



Annexure G Forecasting error analysis using the multivariate LSTM results

Error analysis for the Alkalinity and pH parameters

Measured Alkalinity, mg/L $CaCO_3$	LSTM Forecast 1	Error %	LSTM Forecast 2	Error %	LSTM Forecast 3	Error %	Measured pH, —	LSTM Forecast 1	Error %	LSTM Forecast 2	Error %	LSTM Forecast 3	Error %
212	212	0.1	207	2.4	202	4.6	6.38	6.38	0.0	6.35	0.4	6.33	0.8
214	210	2.0	206	3.9	201	6.3	6.39	6.37	0.3	6.35	0.7	6.32	1.1
216	205	5.2	202	6.3	198	8.6	6.32	6.34	0.4	6.33	0.2	6.31	0.2
214	206	3.7	203	5.1	201	6.2	6.32	6.35	0.5	6.33	0.2	6.32	0.0
226	203	10.1	201	11.2	198	12.2	6.40	6.34	1.0	6.32	1.2	6.31	1.4
204	206	1.0	202	0.8	200	2.0	6.37	6.35	0.3	6.33	0.6	6.32	0.8
196	207	5.5	202	3.3	200	2.3	6.37	6.35	0.3	6.33	0.6	6.32	0.8
216	206	4.4	202	6.4	200	7.4	6.21	6.35	2.3	6.33	1.9	6.32	1.8
214	204	4.4	202	5.8	200	6.6	6.36	6.34	0.3	6.33	0.5	6.32	0.7
204	202	1.2	200	2.0	197	3.2	6.36	6.33	0.5	6.32	0.6	6.31	0.8
200	202	0.9	200	0.1	200	0.2	6.32	6.33	0.1	6.32	0.0	6.32	0.0
202	202	0.1	200	0.9	201	0.6	6.26	6.33	1.1	6.32	1.0	6.32	1.0
214	201	6.2	199	7.0	200	6.4	6.37	6.32	0.7	6.31	0.9	6.32	0.8
208	201	3.3	198	4.8	203	2.6	6.34	6.33	0.2	6.31	0.5	6.33	0.1
212	202	4.7	199	6.3	203	4.1	6.26	6.33	1.1	6.31	0.8	6.34	1.2
204	203	0.7	199	2.4	205	0.3	6.39	6.33	0.9	6.31	1.2	6.34	0.7
200	207	3.6	204	1.8	210	5.2	6.41	6.36	0.8	6.34	1.1	6.37	0.6
202	208	2.9	204	1.1	211	4.4	6.34	6.36	0.3	6.34	0.0	6.37	0.5
200	207	3.7	204	1.9	210	5.0	6.32	6.36	0.6	6.34	0.3	6.37	0.8
182	209	14.7	204	12.2	211	15.9	6.31	6.36	0.8	6.34	0.5	6.37	1.0
178	210	17.7	205	14.9	211	18.4	6.31	6.37	0.9	6.34	0.5	6.37	1.0
182	210	15.4	205	12.4	210	15.4	6.37	6.37	0.0	6.34	0.4	6.37	0.0
190	209	10.2	204	7.6	210	10.5	6.34	6.37	0.4	6.34	0.0	6.37	0.5

Error analysis for the SO_4 parameter

Measured SO_4, mg/L	LSTM Forecast 1	Error %	LSTM Forecast 2	Error %	LSTM Forecast 3	Error %
2500	2459	1.7	2433	2.7	2409	3.7
2372	2448	3.2	2427	2.3	2400	1.2
2504	2423	3.2	2410	3.8	2385	4.8
2532	2429	4.1	2414	4.7	2401	5.2
2576	2414	6.3	2402	6.8	2390	7.2
2520	2429	3.6	2410	4.4	2397	4.9
2348	2433	3.6	2410	2.7	2400	2.2
2476	2431	1.8	2409	2.7	2398	3.2
2444	2421	1.0	2406	1.6	2397	1.9
2460	2406	2.2	2397	2.6	2385	3.1
2400	2407	0.3	2398	0.1	2395	0.2
2460	2407	2.2	2398	2.5	2401	2.4
2476	2401	3.0	2393	3.4	2399	3.1
2524	2403	4.8	2388	5.4	2412	4.5
2404	2408	0.2	2390	0.6	2414	0.4
2388	2411	1.0	2393	0.2	2422	1.4
2420	2435	0.6	2416	0.2	2451	1.3
2468	2438	1.2	2420	2.0	2453	0.6
2500	2436	2.6	2417	3.3	2449	2.0
2364	2443	3.3	2419	2.3	2455	3.8
2416	2447	1.3	2421	0.2	2453	1.5
2256	2449	8.6	2421	7.3	2449	8.6
2276	2446	7.5	2421	6.4	2449	7.6