Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Machine Learning-Based Contamination Detection in Water Distribution System

A Thesis Submitted to Addis Ababa University School of Graduate Studies in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Engineering

By

Akalewold Fikre

Advisor

Getachew Alemu (Ph.D.)

June, 2020

Addis Ababa, Ethiopia

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Machine Learning-Based Contamination Detection in Water Distribution System

Approved by Board of Examiners

Name	**Signature**	**Date**
_____	_____	_____
(Dean, School of Graduate committee)		
_____	_____	_____
Advisor		
_____	_____	_____
Internal Examiner		
_____	_____	_____
External Examiner		

Declaration

I, the undersigned declare that this thesis is my original work and has not been presented for a degree in this or any other university, and all sources of materials used for the thesis have been fully acknowledged.

Name	**Signiture**
Akalewold Fikre	_____

Place: Addis Ababa Institute of Technology, AAiT
Addis Ababa University, AAU
Addis Ababa,
Ethiopia.

Submitted in: June, 2022

This thesis has been submitted with my approval as a university advisor

Name	**Signiture**
Getachew Alemu (Ph.D.)	_____

Acknowledgment

"It is good that a man should both hope and quietly wait for the salvation of the LORD".
Lamentations 3:26

Above all, Glory to Almighty God and the savior, Jesus Christ, for giving me all the courage and strength in the course of the study and for the completion of this research. I know any good that has or will come from the above.

I am grateful to all those who participated in this study, specifically to the managers of different departments of AAWSA, and expertise as well as family members that are willing to participate in the research. Obviously, without their openness to share their knowledge and experience, this study would not have been possible. I also want to thank the institution, Addis Ababa water, and sewerage Authorities, for their kind support in accessing the laboratory, manuals, and different books and documents. I am also very pleased to thank the whole Staff, at Addis Ababa water and sewerage authorities, for their devotion to providing the help I needed throughout the research process.

My sincere thank also goes to Getachew Alemu (Ph.D.), my supervisor, for all his inspiring comments, friendly advice, suggestions, and scholarly critics from the beginning to the end has made it possible the completion of the thesis writing. I would also thank Menor Tekeba for his time and comments that helped to improve the thesis.

I am also indebted to thank my family for their prayers, advice, and support. My special thanks go to big Sister Amelework Fikre, Solomon, Leykun Gashawbeza, Firehiwot G/Silase, Mustafa Mehedi, Tekel Tufa, Bekele Desta, Biruk Teshome, Sultan Hasen, and Hana Biruk for all kinds of help and advice.

Last, but not least, it is my wish to thank all my Sunday BS team members, for their all-rounded help and prayers

Abstract

Water is a necessary component of all human activities. According to the United Nations World Water Assessment Program, every day, 2 million tons of sewage, manufacturing, and agricultural waste are discharged into the world's water. Due to population demands and dwindling clean water supplies as well as available water pollution management mechanisms, there is an urgent need to use computational methods to intelligently manage available water. To ensure the protection of drinking water, accurate detection of natural or deliberate pollution events in water delivery pipes is essential. Companies that have water must ensure that it is safe to drink. To resolve the global issue of rising water contamination, the design of water contamination detector models has monitored the security of water in pipelines when concentrations of water quality variables in the pipes surpass their maximum threshold is presented in this paper. This paper proposes artificial neural networks, specifically Convolutional Neural Networks, for automated water impurity, detection to refine the model must a picture of turbid water in the pipe is used to detect events. The algorithm of deep learning achieved 96.3 percent accuracy after extensive training with a dataset of 4220 images reflecting various levels of contamination. Besides that, the machine learning algorithm uses an efficient study of water turbidity and transparency levels to estimate the level of pollution in a specific sample of water. As the established model is combined with the current framework, it will provide a cost-effective way for the water company to obtain an estimate of water quality, alerting local and national governments to take action, and potentially saving millions of people throughout the world.

Table of Contents

List of Figures

List of Tables

Acronyms

AAWSA – Addis Ababa Water and Sewerage Authorities

BCM - Billion Cubic Meters

CNN - Convolutional Neural Network

NTU – Nephelometric Turbidity Unit

SVM - Support Vector Machine

UNDP - United Nations Development Program

UNEP - United Nations Environmental Protection

UNESCO – United Nations Educational Science and Cultural Organization

U.S.EPA – US Environmental Protection Agency

US – United States

WDS – Water Distribution System

WHO – World Health Organization

WWF - World Wildlife Fund

Chapter One

1.1 Background of the Study

The water resource covering 70% of its surface is a source of supply for all kinds of life on the planet (Jeremiah Castelo, 2021). When you consider how much more water is still in the ocean and how much of the earth's atmosphere is filled with it, water scarcity will not be a big problem for most people. However, the World Wildlife Fund for Nature estimates that over 1 billion people lack healthy and clean water. On the other hand, Africa, especially Sub-Saharan Africa and South Asia, is home to the majority of the world's people—more than 2.6 billion people—who don't even have access to improved sanitation (WHO/UNESCO, 2010). According to other reports, over 2.7 billion people are affected by water shortages at least once a year (Jeremiah Castelo, 2021).

It is only 3% of the world's water is considered freshwater, even though it occupies 70% of the planet's surface. Furthermore, humans are unable to enter about 2.6 percent of this freshwater. They are, either trapped in polar glaciers or ice caps, frozen in the atmosphere or soil, heavily polluted, or too far under the surface of the earth to be removed. It leaves about 0.4 percent of the water in the world that is available and drinkable to share among the planet's 7 billion people (World Atlas, 2018). Even so, most of this 0.4 percent is difficult to reach. The majority is located within underground aquifers that can be reached by digging wells. The remainder is found in rivers and streams, which we call surface water. With such a small proportion of clean water on the planet's surface, most of the population of the world is suffering (Jeremiah Castelo, 2021).

Since surface water is easier to reach, it has become the most popular method for humans to obtain healthy drinking water. Every day, nearly 321 billion gallons of surface water and 77 billion gallons of groundwater globally. Surface water includes lakes, rivers, streams, and reservoirs located on the earth's surface. Surface water accounts for 80% of global daily water use, and the bulk of that water is for irrigation and public supply. Oceans are the main source of surface water, accounting for the world's 97 percent, but it is unusable for humans due to their high salinity. (Jeremiah Castelo, 2021).

Drinking pure water is an issue for water supply companies around the world, and it is currently a well-known problem due to numerous vulnerable threats. Changes in water chemistry can occur before earthquakes, because of terrorist acts, or other man-made pollutants. The changes in water chemistry caused by an earthquake can be an excellent indicator of earthquake predictions. However, it can also be dangerous to human health because some changes are toxic. When a system event occurs, such as large-scale water contamination because of an accident or a malicious attack, it can affect society and the economy. One such instance is the 1993 contamination event in Milwaukee, which affected 403,000 people, resulting in thousands of hospitalizations and a handful of fatalities, at $96.2 million in medical and productivity costs (P. Corso, M. Kramer, K. Blair, D. Addiss, J. Davis, A. Haddix, 2003). Another incident occurred in 2014 in West Virginia (USA) when the drinking water distribution system was contaminated accidentally with crude MCHM, an industrial chemical (W. J. Cooper, 2014). It affected 300,000 people.

Because of all the threats to public health, water system pollution detection is critical. As a result, it is serious that water utilities assess the hydraulic and quality features of the water distribution system to ensure that all customers have water that is safe to consume. Nowadays, contamination-warning systems are used by water utilities to control the quality of drinking water. They monitor water quality and environmental data at multiple measuring stations using sensors. Despite these measurements, the system also requires a detection system that can accurately alert users to changes in water quality based on the data collected. Machine learning may produce good results for detecting contaminants in water quality, and this behavior motivated the development of a model. The number of erroneous predictions can be reduced dramatically using machine learning methods. The use of artificial neural networks for modeling water quality to compute estimation errors was recently examined, and the contamination event probability was calculated and compared to a given threshold to detect an event using a sequential Bayesian rule (L. Perelman, J. Arad, M. Housh, A. Ostfeld, 2012). Neural networks are a type of machine learning model that seeks to replicate characteristics of the human brain and its biological neural networks. They are the foundation of most deep learning approaches, a subset of machine learning that entails running data through numerous consecutive layers to conduct classification or pattern analysis. Computer vision, bioinformatics, and natural language

processing are just a few domains where deep learning has aided growth. Convolutional neural networks (CNNs) are a type of neural network, that is, a subset of deep learning. CNN's are large networks of nodes called "neurons" that make connections as they learn from data. Because CNN requires individuals to define specific features for a model to assess, they undergo supervised learning. As a result, the model's creator constructs the model's structure, spawning multiple layers that execute various functions and contribute to the model's performance in different ways. Researchers and data scientists have discovered CNNs' aptitude for numerous jobs, most notably image classification, in recent years, and they have increased in popularity as a result. The suggested system is an image-based water quality evaluation system. There are in-depth water quality measurements available, such as the index of water quality (WQI). The suggested model aims to offer a rapid estimate of water quality that does not require comprehensive chemical testing and can be performed anywhere and at any time.

Previous studies have found a relationship between turbidity levels and gastrointestinal disease, such as Schwartz and Levin in 1999 [Schwartz, J., Levin, R.]. As a result, the study believes that assessing the water's turbidity images is valid to determine the risk of drinking that water. Turbidity is an important indicator that may provide valuable information quickly, cheaply, and regularly. Turbidity may not always imply a direct threat to public health. However, it can indicate the presence of pathogenic microbes and serve as an early warning system for dangerous events occurring throughout the water supply system, from watershed use. Turbidity can be measured easily, accurately, and quickly and is often used for operational monitoring of control measures contained in water safety plans (WSPs), the WHO's recommended strategy for maintaining drinking-water quality (WHO, 2017). Instruments can only detect turbidities below 4 Nephelometric Turbidity Unit (NTU), but the murky, red-brown, or black suspension can be detected at 4 (NTU) and above, lowering the acceptability of drinking water. Turbidity can be used in operational settings. Water quality in distribution systems is monitored as an indicator of network integrity and proper operation and maintenance. The rapid changes in turbidity can indicate significant pollution incidents in surface water and groundwater catchments. However, some particles with turbidity values ranging from 5 to 13 Nephelometric Turbidity Units (NTU) are unseen by the naked human eye. This study aimed to develop a comprehensive machine learning model that only requires an image of water to assess whether or not contaminants are

present, making water quality analysis and determining the safety of drinking water much simple.

1.2 Statement of the Problem

Water distribution networks (WDN) are vital infrastructures that can be contaminated chemically, biologically, or radioactively, either intentionally or accidentally. Water utilities often take manual samples and perform chemical analyses regularly to monitor the water quality in the system. On the other hand, manual sampling can take days to discover, usually after a chemical analysis or consumer complaints. In recent years, aiming to monitor water quality, multi-parameter sensors in WDNs have been developed. Water quality sensors that are connected to the internet can be used to improve real-time monitoring of water quality. Because specialized water quality sensors are often linked with high prices, only a few sensors are put at strategic sites in the network in practice. This fact has prompted an extensive investigation into water quality sensor installation in recent years. Low-cost sensors measuring water quality parameters, such as oxidation-reduction potential, chlorine concentration, total organic carbon, turbidity, conductivity, and pH, provide an alternative to special sensors for pollution detection. In the past, various approaches were proposed to address pollution detection issues, including single or multiple-type measurements that are analyzed separately or in combination from one or more locations in the network, using model-based or model-free approaches. People who live in areas where clean water is scarce might benefit from technology that can identify the level of contamination in potential drinking water. This thesis, a deep learning approach to water quality analysis that utilizes the power of modern hardware and machine learning techniques, was created to provide this technology and quickly determine the level of contamination of water based on an image. Turbidity is a calculation of a liquid's relative visibility. Turbidity measures the concentration of light reflected by materials in the water; it increases when the materials found in the sample water rise. Clay, silt, organic and very tiny inorganic matter, algae, plankton, ingested colored organic compounds, and other microscopic species all contribute to the turbidity of water. As a result, this study aimed to develop a model-based approach for pollution detection throughout the water distribution system using the image of the turbid water that is closely correlated with the physicochemical properties of other water quality parameters.

1.3 Objective of the Research

This study's general objective was to detect contamination of water in water distribution systems.

1.3.1 Specific Objectives

- Describe the physicochemical properties of water in the water distribution system that is used for drinking and other purposes.
- Explore changes based on the turbidity of water quality measurement in physicochemical properties of water for contaminated and non-contaminated water
- Collect the data of water for both turbid and non-turbid water quality parameters and, based on the data, assess the image data difference.
- Developing the detection model for selected water quality parameters
- Observing the special characteristics obtained from the model and analyzing the output
- Offering a contamination detection model for the water distribution system to reduce the impact of water pollution

1.4 Scope of the Research

The research targets the detection of contaminated water in a water distribution system in Addis Ababa, Ethiopia, giving special attention to Legedadi Dam and Entoto groundwater sources. This site was chosen because most parts of the city get the water from the Legedadi Dam, and the network it covers is larger compared to other water sources, which makes it an appropriate place for the research. It also helped to collect more water samples that covered most areas of the city. The focus of this research was to detect pollutants in water distribution systems.

1.5 Limitations of the Study

This study has provided essential features regarding contamination detection in water distribution systems. However, there were some limitations recognized in this study.

First, the cameras used during the data collection period were Nikon d5200, Nikon d3200, and mobile telephone cameras, which affected the quality of the turbid water image. If the image is clear, the model identifies it correctly. It is because the image captured by the camera shows very tiny particles found in the water. However, the camera used for this study has a quality limit. For some images, similarities have been observed, especially in the case of low concentrations of

contaminated water images. Therefore, considering camera quality as a part of the problem, this research suggests future studies consider the image quality at a low concentration of pollutants.

Secondly, the turbid water image from five NTU to 15 NTU is very challenging to identify its type of contamination. In addition to that, the study did not include the turbid water image, especially from five NTU to eight NTU. Therefore, to keep the balance and credibility of the study, this research suggests future studies consider including the water image from five to eight NTU.

The final limitation of this study was lacking observing all behavioral changes in the chemical properties of the water during chemical pollution. It means that the study did not include all changes in the chemical properties of the water. Some chemical pollutants may not affect the turbidity of the water. Hence, future studies should consider all changes in the chemical properties of the water and the turbidity of the water in the case of different chemical contamination.

Chapter Two

Literature Review

This chapter focuses on presenting the research's theoretical and experimental roots, as well as incorporating them with the literature on water pollution in water delivery systems. Because of limited water resources, stringent budget requirements, a growing population, aging infrastructure, increasingly stringent regulations, and increased attention to safeguarding water supplies from accidental or deliberate contamination, drinking water utilities are facing new challenges in their real-time operations. Existing technologies, whether laboratory-based or online monitoring systems, are too slow to develop operational responses and do not provide a degree of public health safety in real time, necessitating the development of better water monitoring systems. Because of the potentially serious implications for human health, rapid identification (and response) to instances of contamination is important. Water is such an integral part of life that developing a new approach for assessing water quality and forecasting potential water quality patterns is critical.

The water delivery system has a wide body of work on simulation and model-based pollution detection. The Environmental Technology Confirmation program was developed by the US government in 1995 to investigate how changes in water quality parameters can be detected by real-time sensors. Academics, government departments, and private companies are all doing extensive research in close collaboration. On commercially produced hardware and software, these projects range from theoretical to applied science. Klise and McKenna (2006) and McKenna et al. (2008) planned three water quality detection methods based on a comparison of calculated and observed values and receiver operating characteristic (ROC) curves to test each technique. The US Environmental Protection Agency (EPA) has tested over 30 pollutants (including pesticides, insecticides, metals, and bacteria) that may be used in deliberate acts of water pollution (U.S.EPA, 2005a; U.S. EPA, 2005b). Hall et al. (2007) published the results of experiments that looked at how commercial water quality sensors of various designs and technologies responded to biological and chemical loads. The most responsive metrics to the common pollutants were identified to have been free chlorine, total organic carbon (TOC), oxidation-reduction potential (ORP), conductivity, and chloride, according to these studies.

Byer and Carlson (2005) were the first to develop and evaluate an online drinking water delivery monitoring system. They mixed four credible hazardous drinking water pollutants (sodium fluoroacetate, sodium arsenate, sodium cyanide, and aldicarb) with water from the tap and tested the detection accuracy in a delivery system at various concentrations. Ph., chlorine residual, total organic carbon, and turbidity values were measured before and after the introduction of these pollutants using benchtop analysis and online monitoring equipment. All four pollutants may be observed at low concentrations, according to the findings. Three of the four chemicals were found in concentrations that were not harmful to human health.

Yang et al. (2009) used a pilot-scale pipe system to monitor 11 pollutants at various concentrations. Background noise was minimized and contaminant signals were improved using the proposed adaptive transformation of sensory measurements. Pollutant classification and detection based on chlorine kinetics were made possible as a result of this.

Event detection systems were proposed by Arad et al. (2013), Perelman et al. (2012), and Murray et al. (2010). These experiments also show a correlation between water quality indexes. Some machine learning methods predict the expected calculated value of the next time phase by learning the behavior of each parameter time sequence. The models can then classify measurement outliers and define deviations from expected behavior. Most of the predictor variables are combined to determine the likelihood of the risk occurring.

The system developed by Guepie et al. (2012) was based on residual chlorine decay. They hypothesized that a contaminant in the WDS would absorb a large portion of the measured chlorine and that this single parameter would provide enough information. The methods used in the previous studies were supervised classification methods. According to a given training data set, the classifier learns to differentiate between regular operation and event time measurements using these methods. When there are no real-time measurements of contamination events, the models must be trained and tested using simulated contamination events. The models add some random disturbances to the measured data to reflect the contaminant effect to preserve generality and then in the absence of sufficient information.

Depending upon pollutant reaction kinetics and uncertainty, Uber et al. (2007) presented recommendations for event simulation. The foregoing studies applied simultaneous processes of unvaried analysis to water quality time series to detect pollution events. Following the concurrent methods, the results were combined to determine the likelihood of an occurrence occurring. A multivariate analysis varies from previous research in that it requires the simultaneous observation and analysis of several parameters.

Jonathan Arad, Mashor Housh, Lina Perelman, and Avi Ostfeld (2013) suggest a complex threshold scheme for water delivery system contamination event detection. Off-line and web phases are used to establish the technique. The algorithm of genetic (GA) is used to tune five decision variables throughout the off-line stage: positively and negatively filters, positively and negatively dynamic thresholds, and window size. Using the five design parameters, a recursive Bayes' rule is used to detect actual online events during the online point. Both visual and statistical signs of contamination incidents can be obtained using this technique. These conclusions are focused on a single or several criteria that are regularly calculated. Contamination incidents were simulated, but with different intensity severity effect characteristics, since the exact behavior of pollutants in the actual water delivery system is unknown.

The architecture and implementation of a coupled SVM-evolutionary optimization methodology for water quality event detection in water delivery systems is proposed by (Nurit Oliker and Avi Ostfeld, 2013). Since the suggested model is based on simultaneous conjunction analysis of the entire water quality time series data set, it differs from parallel one-dimensional data analysis. The multivariate analysis looks at the connections between water quality indexes and looks for shifts in their patterns. All measurements were taken in a typical working setting. The model is made up of time series data for six water quality indexes: total chlorine, electrical conductivity, total organic carbon, temperature, pH, and turbidity, as well as simulated events. The model's ability to identify events that were only partially reflected in the data is noteworthy (i.e., affecting only some of the measured parameters). In comparison to the previous model, the proposed model has a higher accuracy and detection ability. A minimum volume ellipsoid is developed and validated for the detection of outliers' measurements. The ellipsoid could take the place of the current model's SVM point.

Nurit Oliker and Avi Ostfeld (2014) propose a two-step classification model for outlier identification and event classification based on a minimal volume ellipsoid classifier and a corresponding sequence analysis. The ellipsoid construction only uses normal vectors (measured in normal operation time) in the known data set and does not include abnormal (events representing vectors) examples. All measurements were taken in a typical working setting. Total chlorine, electrical conductivity, pH, temperature, Total organic carbon, and turbidity are all data that are similar to SVM-evolutionary data. The proposed model includes a full decision-making method for detecting contamination events. The model also uses applied multivariate analysis of the water quality metrics data, which differs from most previous studies' parallel analysis of the parameters. The model produces strong results in terms of accuracy and identification, but it should be noted that it is susceptible to any sudden changes.

Advanced strategies for improving source separation in a composite sensor signal and thus improving sensor detection capability are discussed in a variety of publications (e.g., Choi et al., 2005; Feintuch, 1990; Lin et al., 2007). Using advanced sensor data processing techniques or evolving hardware, such as new generations of optical sensors for bacterial and chemical contaminations, improves detection accuracy and contaminant selectivity in traditional water quality sensors (e.g., U.S. EPA, 2005; NRC, 1995). Data analysis and other related statistical techniques such as multi-variant cluster analysis (Klise and McKenna, 2006), adaptive data treatment (Yang et al., 2007), and control chart-based robust statistics are used in the first approach (Kroll and King, 2006). In pilot-scale pipe flow experiments, Szabo et al. (2007) recorded observable sensor responses to the presence of multiple pollutants. The minimum concentration threshold of pollutants above which their presence can be registered in water quality sensors was stated by Haught et al. (2005) based on bench-scale research.

Yang et al. (2007) introduced an iterative detection system for actual occurrence adaptive detection, recognition, and alert (READiw). In a field application, water quality measurements are sent to a processing unit, which analyzes the sensor signals in real-time to distinguish contaminants from natural history. A statistical comparison of new measurements to an adaptively back-tracked backdrop in a moving time window governs real-time anomaly detection.

The researchers have continued to work on developing pollutant detection systems that incorporate multiple sensor responses. Kroll (2006) defined the Hach HST method, which uses multiple sensors to detect events and identify contaminants. While replies from numerous sensors are used in Kroll's system, their internal relationship is not explored. Arad et al. (2013) provided an overall approach for detecting quality defects in water delivery systems using multivariate water quality time series that combines a data-driven estimation model with sequential probability updating.

Centered on the enhancement and development of a Dempster–Shafer (D-S) proof principle, Dibo Hou, Huimei He, Pingjie Huang, Guangxin Zhang, and Hugo Loaiciga (2013) propose a water-contamination event detector. The advanced D-S theory for the monitoring of water-contamination events, according to their research, is based on time series of residuals of water-quality variable predictions, including the use of weighted-averaging and time-dimension details to address conflicts or ambiguities that occur when attempting to detect water-contamination events. They tested potassium ferricyanide (K3 [Fe (CN) 6]) and ferric ammonium sulfate (NH4Fe (SO4) 2), two toxic compound pollutants. Virtual and real-time controlled water-quality data were used to evaluate the proposed process. The findings show that using actual water quality monitoring; the suggested extended D-S fusion system can sense water-contamination events with simulated strengths greater than 1.2. However, applying the future extended D-S fusion approach to drinking water infrastructure in combination with actual early warning systems would necessitate (1) the development of an accurate library of contaminant-sensitive variables and (2) the integration of regular operational systems to minimize false alarms.

Eliades et al. (2014) looked into the issue of water quality by using a model-based method for detecting pollution events in water treatment systems. The suggested system takes into account well-known chlorine input signals and produces bounds of the predicted chlorine concentration at different sensing locations at each time stage by running several Monte-Carlo simulations in parallel with the real system. The sensor measurements are then compared to the approximate limits, and if these bounds are violated in compliance with a guideline, event logic raises an event warning flag. The result demonstrates that the system will adjust the detection bounds as the chlorine concentration input varies. Chlorination is the method of adding chlorine to drinkable water to disinfect it and destroy germs. It fixes one of many issues here.

Contamination, on the other hand, occurs for a variety of causes and has varying effects on various parameters. As a result, detecting pollution based on chlorine concentration does not provide a complete control mechanism.

From Kuhnert et al. (2014), an online security monitoring toolkit for a water delivery method based on sensor measurements of water quality should be created. They use an easy-to-parameterize event detection algorithm based on principal component analysis to implement the proposed process. To calculate the threshold, the algorithm only needs two parameters: the number of principal components and the sigma environment. They put the algorithm through its paces for both clean and contaminated water. They also measure the water's conductivity, or the concentration of ions in the water, here. Dissolved salts and inorganic materials such as alkalis, chlorides, sulfides, and carbonate compounds have these conductive ions. Consequently, treating one parameter of water quality affects the other, necessitating careful consideration.

(Mohammed et al., n.d.) Introduce Adaptive neuro-fuzzy inference system (ANFIS) models for determining the protection of water in pipe networks when concentrations of water quality variables in the pipes surpass their maximum thresholds. During the testing stage, the proposed model will correctly detect between 92 and 96 percent of the water in the pipe network's safety state, with a 1 percent false alarm rate. This work prompted me to redesign the algorithm for certain water quality parameters that are highly correlated with one another and have an impact on the algorithm's efficiency, i.e. some algorithms perform poorly on highly correlated parameters. So it's intriguing to find out which parameters are closely correlated and then remodel the algorithm.

Pollutant Detection Using Multiple Conventional Water Quality Sensors in an Early Warning System for detecting the safety status of source water (Che & Liu, 2014). This approach proposes a method for detecting pollutant events in a short period using data series from multivariate parameters. Conductivity, UV-254, Ph., temperature, oxidation-reduction potential, turbidity, nitrate-nitrogen, and phosphate are the eight parameters used in their proposed method, and the most widely used herbicide, glyphosate, is chosen as the test contaminant. Numerous sensors of water quality are presumed to react to a contaminant simultaneously in the proposed process. Their findings indicate that with the use of feedback from online sensors of water

quality, the proposed method could detect glyphosate contamination 1 minute after the contaminant was introduced at a concentration of 2 mg/l. The intensity of sensor correlative reactions was also linked to pollutant concentration, and the patterns of sensor correlative responses were contaminant-based. Since the orders of magnitude of the sensors' responses were linked to contaminant concentrations, all associated sensors do not react at the same stage.

Christos C. Anastasiou, Marios M. Polycarpou, Christos G. Panayiotou, and Theofanis P. Lambrou (2014) suggest the development and operation of such a low-cost device that can be used in consumers' homes to continuously track qualitative water parameters and fuse multi-parametric sensor responses to determine the risk of water consumption. They put their method to the test by injecting different concentrations of two critical pollutants (Escherichia coli bacteria and arsenic) at distinct time intervals and evaluating the efficiency of the event detection algorithms in real-time. When microbiological pollutants (E.coli) were injected into chlorinated potable water, the TU and EC sensors performed well. Almost all sensors in the case of chemically (arsenic) contaminated drinking water did not react at low levels of arsenic contamination. However, pH and ORP sensors responded well at concentrations greater than 25 g/L, and all sensors responded well at higher concentrations (greater than 500 g/L).

Mashor Housh and Avi Ostfeld (2015) propose an integrated logit model for pollution event detection in water delivery systems, to improve an early event detection model called the dynamic threshold method (DTM) by optimally incorporating detection from all different water quality parameters into the event detection structure. Total chlorine (mg/L), electrical conductivity (EC) (mS/cm), pH (_), temperature (_C), total organic carbon (TOC) (ppb), and turbidity are the six water quality parameters they use to evaluate their device (NTU). The developed approach is linked to data collected from a single monitoring station site, and it does not take into consideration data from multiple sensors. However, although the model detects events with greater likelihood, the system's ability to incorporate multiple sensor data into event detection is constrained.

Shuming Liu, Han Che, Kate Smith, and Tian Chang (2015) created a real-time contaminant classification system based on quantitative analysis of data from a variety of traditional water quality sensors. This suggested approach is tested with data from laboratory pollutant dosing

experiments. Turbidity, temperature, oxidation-reduction potential (ORP), pH, conductivity, phosphate, UV-254, and nitrate can all be measured simultaneously and continuously. They used heavy metals (nickel nitrate, cadmium nitrate), herbicides (glyphosate, atrazine), and inorganic salts to assess the three classes with the most common six toxic substances: herbicides (atrazine, glyphosate), heavy metals (cadmium nitrate, nickel nitrate), and inorganic salts (sodium fluoride, sodium nitrate). Additionally, clustering is used to describe the type of contaminant. And they discovered that sensor response varied from contaminant to contaminant in their experiment. In terms of categorizing the kind of pollutant in real-time, the proposed approach has great performance and robustness. Whenever a glyphosate solution of 2.8 mg/l (4 times the national limit) is detected, for example, the suggested method would accurately classify a pollutant within 4 minutes after detection, with such a true positive rate of 0.88.

Mohammadpour et al. (2015) used three separate algorithms, SVM, and two artificial neural network approaches to investigate the issue of water quality. R2, RMSE, and MAE are used to compare results. The SVM algorithm is competitive with neural networks in terms of results.

Pingjie Huang, Yu Jin, Dibo Hou, Jie Yu, Dezhan Tu, Yitong Cao, and Guangxin Zhang (2017) propose a multi-classification probability output-based pollution classification system (MCPO). This method takes a binary SVM and transforms it into a multi-classification SVM with a likelihood of production. Throughout the process of developing a pattern library, the proposed approach decreases the effect of pollutant concentrations. The proposed method will determine if a sample feature is readily visible. A proposed model will determine if a sample feature is readily visible. The samples with readily evident classification characteristics can decide their category; the samples with non-apparent classification characteristics can avoid making a single classification decision. As a consequence, samples with a broad maximum likelihood and standard deviation have a distinct classification function, resulting in a highly credible single-classification result. The classification function is not as evident in the early stages of contaminant injection in samples with limited maximum likelihood and standard deviation. However, for certain pollutants, high-concentration samples will accurately represent the properties of material while having little impact on the construction of a classification decision boundary.

Yue, Wong, Zhu, and Zhang (2017) designed a lightweight anomaly-finding methodology for water quality parameters that utilizes dual time-moving windows to distinguish anomalous data from historical trends in real time. The algorithm uses statistical models, specifically the autoregressive linear combination model. They put the algorithm to the test using 3-month PH data from a real water quality monitoring station in a river system. Experiments show that their algorithms have a lower rate of false positives and do better than the AD and ADAM algorithms at detecting anomalies.

Xie, Gao, and Kang (n.d.) have created a model for predicting water quality. Artificial Neural Network with Nonlinear Autoregressive showed the best results. In 2009, Xiang & Jiang used a combination of least squares support vector machine (LS-SVM) and particle swarm optimization methods to predict water quality and overcome the limitations of traditional back-propagation algorithms, such as being moderately difficult to reach and easy to achieve the extreme minimum value. They also realized that the model is very good at determining the quality of the water of the Liuxi River Xiang and Jiang via simulation testing (2009). Recently, the Recurrent Neural Network has been used in large-scale vision and speech problems as a dynamical device whose next state and output are dependent on the current network state and input (Gregor, Danihelka, Graves, Rezende, & Wierstra, 2015). RNNs and LSTMs excel at extracting patterns from input feature space, even when the data spans long sequences. They can nearly seamlessly system problems with numerous input variables, which is particularly useful in time series forecasting, where traditional linear models can be difficult to adapt to multivariate or multiple input problems (Che, Purushotham, Cho, Sontag, & Liu, 2018).

Revathi, S. Kavi, and G. Shenbagalakshmi (2018) suggest using a wireless sensor network to develop and implement an actual system of water quality control for drinking water. The pollution finding algorithm and fuzzy rules assist in defining and classifying water based on pollution in the pipeline. The smart sensors' low-frequency communication connection is a pipe-to-pipe data transfer that is also long distance. Signal attenuation is a significant factor in signal propagation through a drenched medium at high frequencies. Therefore, the key challenge is wireless communication between the sensors and signal transmission from the sensors. The proposed system is low in cost, lightweight, and consumes little power.

Jeniffer Jepkoech and Consolata Gakii (2019) created a classification model based on a decision tree to analyze water quality and determine whether it is safe or not. Something like a decision tree as just a data mining technique to predict clean water based on water quality parameters will help the laboratory technologist's job by predicting which water samples can go on to the next stage of analysis. Conductivity, pH level, and water alkalinity analysis can all help to determine water quality. The model was developed using five decision tree classifiers: Random forest, LMT, J48, Hoeffding tree, and Decision Stump, and the precision was compared. Decision Stump had the lowest accuracy of 83 percent, while the J48 decision tree had the highest accuracy of 94 percent.

Florin Leon, Doina Logofătu, and Fitore Muharemi (2019) suggest finding the best performing model for water quality data for event detection on time series data, to find the best model for anomaly detection on water quality systems. To see if machine learning models are more accurate than logistic regression, they use the F-score metric to evaluate the output of each algorithm. For just a long period, time series have been studied using statistical algorithms. Machine learning Algorithms are increasingly common for their ability to perform well on time series data sets. The results indicate that the whole algorithms are vulnerable, with ANN, logistic regressions, and SVM being the least vulnerable, while LSTM, RNN, and DNN being the most vulnerable.

Automated Aqua Sight water pollution detection, proposed by Elliott Ruebush and Ankit Gupta (2019), uses a picture to assess the level of contamination in water. Believe that determining the turbidity of a water picture is a viable method for determining the risk of consuming that water. They have used a convolutional neural network to do this, which involves an image of water to decide whether or not pollutants are present. They checked the algorithm and found it to be accurate 96.2 percent of the time. However, the dataset they used to arrive at this conclusion contains only 105 water images, and the data is not balanced, i.e. they only focus on the physical properties of the water and they are not considering other properties. They concentrate on the physical properties of the water, which implies that the contaminants applied to the water to alter the turbidity in order were: sand, sand and salt, sand, salt, and black pepper, sand, salt, black pepper, and oil-paint. The chemicals, however, affect the turbidity of the water. As a

consequence, during the data collection process, the chemical properties of the water are ignored. Finally, all of the differences have an impact on the model's preparation and outcome.

Chapter Three

Research Methodology

3.1 Introduction

This chapter focuses on the methodology that the study followed in the research. It begins with a brief description of the research area, an account of the research participants, and the techniques used to recruit follows. The chapter also makes a brief description of the research methods used to gather the necessary data and achieve the research objective. The chapter also briefly describes the method employed for analyzing the collected data. In the end, the ethical issues are presented.

3.2 Research Site

Ethiopia is one of Africa's largest countries, with more than a population of 110 million people and a capital city of about 6 million people, Addis Ababa (UN, 2017). The capital city, Addis Ababa, is situated on the banks of the Akaki River. Entoto Mountain Range forms the town's northern border. Mt. Furi, to the southwest of a town, and Mt. Yerer, to the southeast, are high massive volcanic centers rising to elevations of 2,839 and 3,100 meters, respectively. The east and west drainages that divide the Akaki River are formed by such two peaks.

The location of town is in the highlands, at a height of 2,200 to 2,500 meters above sea level. The city's river network can be divided into two catchments: the Great Akaki catchment (900 km2) and the Little Akaki catchment (540 km2), all of which drain into the lake Aba Samuel (Aschale et al., 2017). The River of Kebena and the Great Akaki River are two main river branches that originate from the Great Akaki catchment, each with its network of smaller tributaries. The River of Kebena flows through the city's densely populated areas before joining the Great Akaki River, creating a sub-basin of the Great Akaki catchment. Apart from the Kebena River, the Kebena catchment has only one major river, the Bantyketu.

As the Addis Ababa Observatory, the long-term average annual rainfall is 1254mm. The maximum temperature fluctuates between 20° C in the wet season and 25° C in the dry season, while the minimum temperature fluctuates between 7° C and 12° C all year. In the county, the speed of wind is usually moderate, with average values ranging from 0.5 to 0.9m/s. In November

and December, the average sunshine hours per day reach 9.5 hours, while in July and August, this number drops to 3 hours or less. Monthly pan evaporation records collected from previous studies show that the average monthly pan evaporation at Addis Ababa Observatory during the dry season (November) is around 180 mm and this value decreases to around 75 mm during the wet season (July).

The city is rapidly expanding because of urbanization, and it is undergoing massive infrastructure growth. There are over 2,000 industries in Addis Ababa, including potable water, cement, clothes, beer and alcohol, tobacco, leather, tannery, rubber, and food factories. The city serves as the country's industrial, cultural, governmental, commercial, and modern nerve center (Aschale, 2016). With its many foreign organizations and agencies, the city is also one of Africa's hubs. Addis Abeba is the center for the United Nations Economic Commission for Africa (ECA) and over a hundred embassies. It is said to be Africa's diplomatic capital and a symbol of the continent's current humanitarian development.

Addis Ababa was chosen partly because of the natural springs on the site when it was created. As the city expanded, new water sources were required. Therefore, in 1938, a plant was built to treat water from the nearby rivers. The city began to develop, and the first dam (Gefersa I) was installed in 1944 to meet the increased demand. Many springs degraded in quality during the 1950s and were withdrawn from service. The Gefersa treatment plant, which has a capacity of 30,000 m3/day, was constructed in the 1960s and another dam (Gefersa III) was built in 1966 to increase capacity and serve as a sediment trap. Two transmission pipelines, each 400mm in diameter, ran from the Gefersa treatment plant to Addis Ababa's distribution reservoirs.

The Legadadi Dam, treatment plant, and transmission pipeline were completed in 1970. The water distribution system for this lower dam required a pumping system. The Legadadi treatment plant's capacity was increased from 50,000 to 150,000 m3/day in the 1980s, and a second transmission pipeline was constructed. Due to Addis Ababa's population, the engineers knew that these sources would only be adequate until 1992. They looked at various locations for the next great dam, which would supply water until 2020. This project was delayed (and is still being postponed) due to numerous political and economic instability and two emergency projects, the building of the dire dam and the Akaki area, had to be implemented instead.

The Legadadi treatment plant was expanded in tandem with the building of the Dire dam, reaching its current capacity of 165,000 m3/day (Sime et al., 1998). Its capacity has increased to 195,000 m3/day. The Akaki well field has a 43,000 m3/day range. The American Association of Women's Sports Administrators (AAWSA) published a report in 2011.

The city's factories are situated along the River of Little Akaki and its tributaries, which influence the river. While the Great Akaki's catchment area is sparsely populated, the upstream areas are highly populated with residential and commercial areas. Besides, it passes through agricultural fields before joining the Kebena River. Before leaving the city and entering another agricultural area, the river flows through a denser city area with industries present. Aba Samuel is a lake about 53 kilometers from Addis Ababa's city center. It was designed in the late 1930s as part of a dam project to produce electricity. Now, it is highly contaminated by industrial and urban waste wastewater Little Akaki or the Great Akaki (Yohannes & Elias, 2017).

Residential, leisure, urban agriculture, manufacturing, commercial, open-market, and green areas are all part of the city's land use. Many riverbank areas and surrounding areas in the city have settlers living there without permission. In Addis Ababa, illegal settlement is one of the problems and the river is used as a garbage dump and toilet, and a washing machine. Domestic waste is discharged into rivers without being treated, and pipelines from different sources have outputs that discharge directly into rivers (Yohannes & Elias, 2017).

3.3 Data collection tools

This research used several techniques to gather the data needed to fulfill the study's target. The goal of combining several research instruments to investigate the same phenomenon made it crucial to deepen our understanding of how to regulate water quality in water delivery systems. Furthermore, combining multiple research tools allowed researchers to triangulate their findings and compensate for the shortcomings of one approach with the strength of another. As a result, in this study, data collection instruments such as laboratory tests, secondary data sources and the researcher's field notes were used to gather the information needed to achieve the goal of the research.

3.3.1 Secondary Data Sources

The study also collected secondary data from a variety of sources, including books, journals, research papers, and other related unpublished documents, such as information from the Addis Ababa Water and Sewage Authority (AAWSA). It focused on the scope of the water pollution issue and assessed whether any intervention initiatives were implemented to address it. The opportunity exists via evaluating historical information regarding water quality and pollution events that occur during various periods and their cumulative effect on human health.

3.4 Method of Data Analysis

The first step in data processing is to obtain water samples from various locations. The most important aspect of data processing was making a list of coding categories. This was achieved by searching for trends and continuity in the collected data and recognizing the themes. Furthermore, coding categories were generated using keywords and phrases that reflect the topics and themes. Creating a list of themes and subthemes made processing the collected data much easier, which was essential for the data analysis. The meanings of the code were listed and then grouped into common categories or themes. The significant themes of the phenomena were then developed using the clustered themes and meanings.

3.4.1 Collection of water samples

From various locations, a total of 216 water samples were collected (Figure 3.1-3.3). The samples are collected throughout different phases of a treatment plant, including raw water, during sedimentation, and finally from the reservoir, city reservoir, and household. Parts of the city where water pipes were exposed to domestic waste were chosen as sampling sites. Another criterion used in the selection process was proximity to contamination sources like industrial wastewater and hospital effluent outlets. Since the water pipe could not be accessed anywhere, locations had to be selected where it would be convenient to take a sample if it broke down due to construction or some other reason.

The sampling took place over eleven months in 2020, from January 8 to December 12. The majority of samples were taken in the morning. The collection protocol followed the Addis Ababa Water and Sewage Authority's procedures (AAWSA). In summary, the 600 ml sampling bottle was used to extract water from the sample points in both clean and contaminated

situations. Water samples were directly stored in a cooler after being sampled and kept refrigerated in the laboratory to avoid any reactions or microbiological activity.

Physicochemical parameters were studied in the samples (Turbidity, pH, EC, TDS, Total Alkalinity, Calcium Hardness, Total Hardness, Magnesium Hardness, Ammonia, Nitrate, Nitrite, Phosphate, Fluoride, Iron, Manganese, Silica, Chloride, and Bicarbonate Alkalinity). Following the sampling, all measurements were carried out in line with WHO requirements (WHO). All sample processing is carried out at the Addis Ababa Water and Sewage Authority's regular laboratory (AAWSA). Appendix C and D contain the results of all samples taken from a real network of water delivery systems for both safe and polluted water.

Figure 3. 1 The raw water at the time of sampling

Figure 3. 2 Water Reservoir at the time of sampling.

Figure 3. 3 Customer Tap water at the time of sampling.

However, laboratory findings indicate that most physical and bacteriological parameters have changed in samples taken from various points along with the water distribution network. Chemical parameters change just slightly. It is not to say there is no chemical pollution. This analysis is focused on physicochemical parameters and excludes bacteriological findings. As a result, it is risky to add chemicals to the water distribution line, to better understand chemical contamination, so it is important to consider changing the physical and chemical parameters by simulating the drinking water line and incorporating different chemicals that are harmful to human health.

3.4.2 Preliminary-scale contaminant injection and monitoring system

The preliminary-scale device used in this study is a water delivery system simulator to learn how physicochemical parameters change when harmful chemicals are injected into the system. Figure 3.4 illustrates the process flow of the Preliminary-scale method, which was used for baseline establishment and multi-pass contaminant checks. The water tank is about 60 cm high with 40 cm in diameter, with a total capacity of 20 L. 20L of water supply flows via the multi-direction to the end line in this mode. A screwed valve is located between the source and the end line and used to inject a chemical into the device. Finally, a 600 ml sampling bottle was used to extract the sample for all types of contamination, and it was weighed in the laboratory in the same way as the previous sample and the adjustment of parameters for the proposed method was studied.

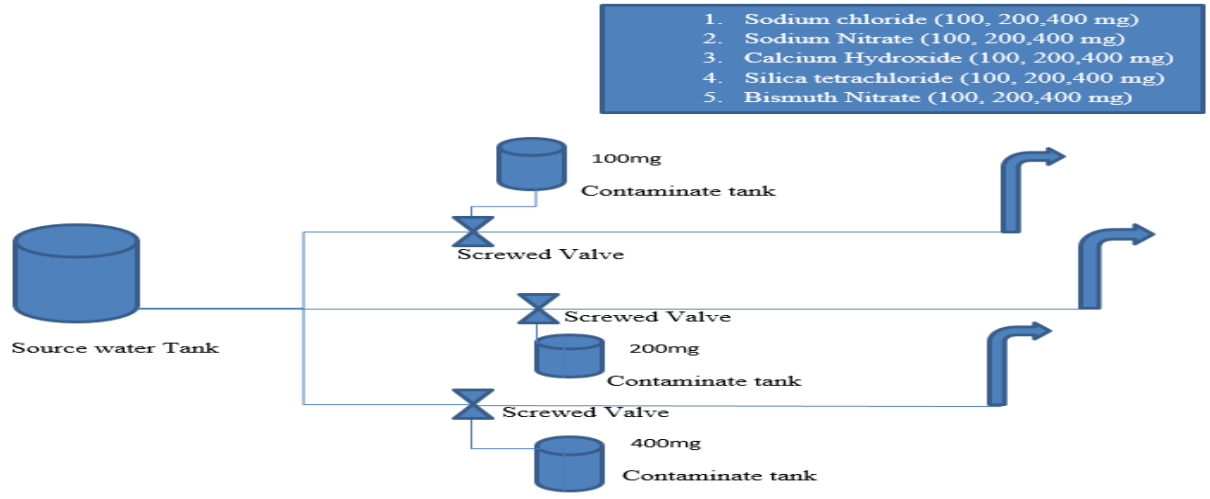

Figure 3. 4 A process flow schematic of the Preliminary-scale system

The system can measure the following eight parameters: pH, turbidity, TDS, Total Alkalinity, Calcium Hardness, iron, manganese, Chloride, nitrate-nitrogen, and Silica.

3.4.3 Developing the dataset

We looked at the overall method of collecting data in the previous portion. In addition, lab tests were conducted on samples from different distribution lines, including clean and polluted water samples, to determine their physicochemical properties in detail. The outcome, on the other hand, is not the target in and of itself, but it does point in the right direction. Designers should consider changes throughout the physicochemical properties of water, concentrating on the turbidity of the water that is influenced by the chemical and biological particles of the contaminant, based on the findings of the previous parts. Turbidity is a term that refers to the cloudiness of water caused by suspended particles like clay and silts, chemical precipitates like manganese and iron, and organic particles like plant debris and species. As a result, a dataset for the proposed method is generated based on the properties of the turbidity of the water. As the turbidity of the water rises, it decreases the purity of the water to transmit light by scattering and absorbing the light. A total of 216 collected water samples image were used in the proposed system's dataset. But it doesn’t mean the whole dataset is only 216 water images. It was divided into three parts. The first 121 images contained both safe and contaminated water in various forms. The second 69 images were all of the clean water, and they added a variety of light levels to the dataset. A total of 26 images were added to the dataset from the third round of data collection, all of which were of polluted water with a broader variety of light levels. The

proposed approach considers different effects of light on the collected water to create the image dataset based on the turbidity of the water.

From those methods:

- The image is taken with 5 colored lights: light Blue, Blue, Green, Yellow, and Red.
- Image was taken with light-Off
- Image taken with varying levels of brightness Images
- taken at different positions

(a) (b)

Figure 3. 5 (a) is clean water with normal light and (b) is clean water with blue light.

(a) (b)

Figure 3. 6: (a) is contaminated water with 100mg of sodium nitrate and (b) is contaminated water with 400mg of sodium nitrate.

(a) (b)

Figure 3. 7: (a) is contaminated water with clay at low concentration and (b) is contaminated water with clay at high concentration.

(a) (b)

Figure 3. 8: (a) is contaminated water with sand at low concentration and (b) is contaminated water with sand at high concentration.

(a) (b)

Figure 3. 9: (a) is contaminated water with silt at low concentration and (b) is contaminated water with silt at high concentration.

Figure 3. 10: Contaminated water with 34mg/l of Iron.

3.5 Model Design

Today's water supply companies and people in the city need a more advanced and easy-to-use and access purified water management system that will benefit from technology that allows them to assess how polluted potential drinking water is. This research proposes a deep learning approach to water quality analysis that utilizes the power of modern hardware and machine learning techniques to provide this technology and quickly determine the degree of pollution of water based on a picture.

3.5.1 Architectural design of the Proposed System

Machine Learning (ML) algorithms are well-known for learning the underlying relationships in data and making decisions without the need for explicit instructions. Convolutional neural networks (CNNs) are a form of neural network that is used in deep learning. CNNs are large networks of nodes called “neurons” that create connections as they learn from data. CNNs are one of the most powerful learning algorithms for comprehending image information, with excellent results in image segmentation, classification, identification, and retrieval tasks (Ciresan et al. 2012; Liu et al. 2019).

The ability of CNN to manipulate spatial or temporal similarity in data is one of its most appealing features. CNN is divided into several learning levels, each of which consists of a mixture of convolutional layers, nonlinear processing units, and subsampling layers (Jarrett et al. 2009). CNN is a feed-forward multilayered hierarchical network in which each layer performs multiple transformations using a bank of convolutional kernels (LeCun et al. 2010). The convolution operation aids in the extraction of useful features from data points that are globally

correlated. The non-linear processing unit (activation function) receives the output of the convolutional kernels, which not only aids in learning abstractions but also embeds non-linearity in the feature space. This non-linearity results in various activation patterns of different reactions, making it easier to understand semantic differences in images. Subsampling is commonly applied to the output of the non-linear activation function, which aids in the summarization of the results while also making the input invariant to geometrical distortions (Scherer et al. 2010; LeCun et al. 2010). The ability of CNN to automatically extract features eliminates the need for a separate feature extractor (Najafabadi et al. 2015). As a result, CNN can learn a decent internal representation from raw pixels without undergoing extensive processing. Learning with Hierarchical, multi-tasking, automatic feature extraction, and weight sharing are among CNN's interesting features (Guo et al. 2016; Liu et al. 2017; Abbas et al. 2019).

During preparation, CNN learns by controlling the shift in weights according to the goal using a back-propagation algorithm. The optimization of an objective function with a back-propagation algorithm is close to the human brain's response-based learning. Deep CNN's multilayered, hierarchical structure allows it to retrieve high-, mid-, and low-level features. Mid and low-level features are joined to make high-level features (more abstract features). CNN's hierarchical feature extraction ability mimics the Neocortex throughout the human brain's deep and layered processes of learning, which dynamically learns features from raw data (Bengio 2009). CNN's fame stems largely from its ability to collect hierarchical features.

3.5.1.1 Basic CNN components

CNN can learn representations from grid-like data, and it has recently demonstrated significant performance improvements in a variety of machine learning applications. Figure 2 depicts a typical block diagram of an ML system. A typical CNN architecture consists of pooling and convolution layers alternated with one or more completely connected layers at the end. A fully connected layer may be replaced with a global average pooling layer in some cases. Various regulatory units like dropout and batch normalization are also implemented to improve CNN performance, in addition to various mapping functions (Bouvrie 2006). The arrangement of CNN components is critical during the implementation of the new design architecture to achieve improved performance.

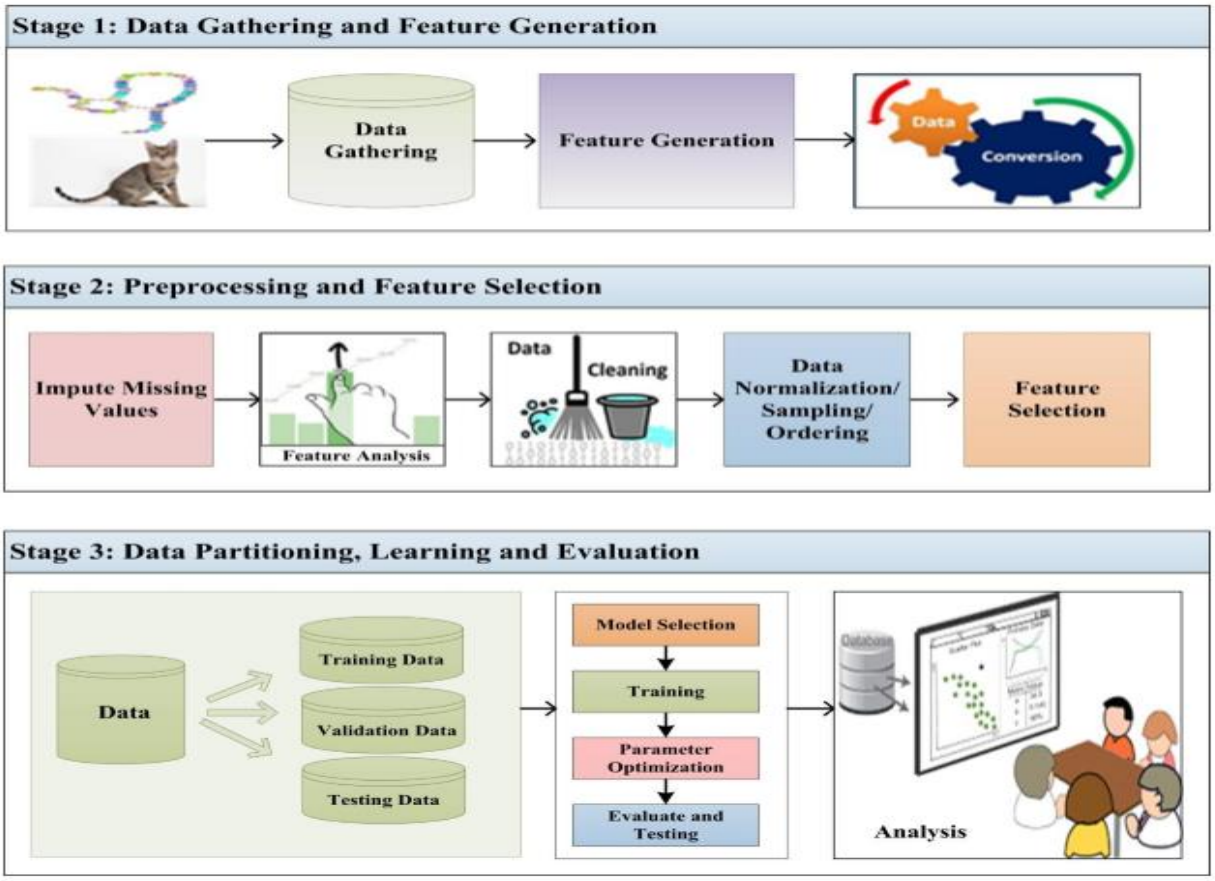

Figure 3. 11 Basic layout of a typical ML system has several stages. (Source: A Khan et al. (2020)).

While they could be used with three-dimensional data and one-dimensional, CNN is a special form of neural network model designed for interacting with two-dimensional image data.

3.5.1.2 Convolutional layer

The convolutional layer gives its network its name and serves as the heart of CNN. This layer executes a process called "convolution." Every neuron performs as a kernel throughout the convolution layers, which would be made up of a collection of convolution kernels. The operation of convolution becomes a correlated operation if the kernel is symmetric. The convolutional kernel divides the image into small slices, known as receptive fields, as shown in fig. Extracting feature motifs is easier when an image is divided into small blocks. Like an older neural network, a convolution involves a linear operation multiplying a set of inputs with the weights. When multiplication is performed with an array of input data and the weights of a two-

dimensional array, called a kernel or a filter, the method was developed for two-dimensional input (Khan et al., 2020).

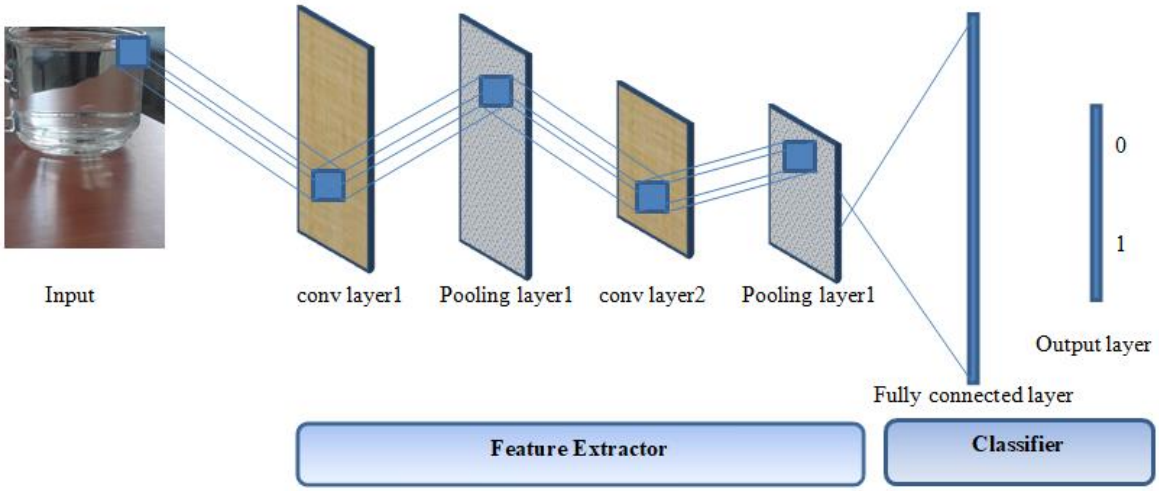

Figure 3. 12 Architecture of the Convolutional Neural Network.

Convolution operations can be expressed as follows:

$$F_l^k(p, q) = \sum_c \sum_{x,y} i_c(x, y).e_l^k(u, v) \qquad (4.1)$$

Where, $i_c(x, y)$: an element of the input image tensor

I_c,: is element-wise multiplied by $e_l^k(u, v)$

$e_l^k(u, v)$: index of the k^{th} convolutional kernel k_l of the l^{th} layer.

Whereas the output feature-map of the k^{th} convolutional operation can be expressed as $F_l^k = [f_l^k(1, 1)\ldots f_l^k(p, q)\ldots f_l^k(P, Q)]$.

The input data is greater than the filter, and the dot product is used to multiply a filter-sized patch of the input with the filter. A dot product is the element-wise multiplication of the input and filter's filter-sized patch, which is then averaged, always yielding a single value. The operation is often referred to as the "scalar product" because it produces a single value.

Consider a 5 x 5 image with pixel values of 0, 1 and a 3 x 3 filter matrix, as shown below.

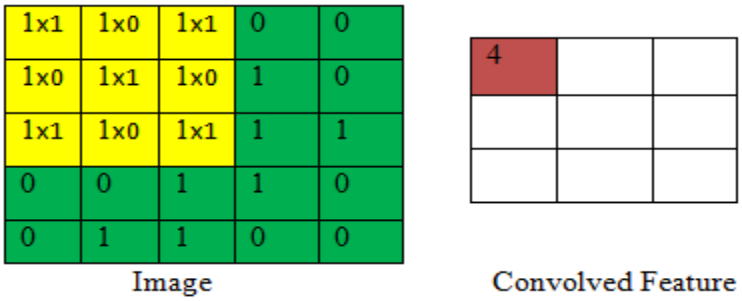

Figure 3. 13 3x3 Output matrix

Various kinds of features inside of an image can be extracted by sliding kernels with the same set of weights on the image, making CNN parameters more powerful than fully connected networks. Convolution operations can be classified based on the number and size of filters, the type of padding, and the direction of convolution.

- **Strides**

The number of pixels shifted over the input matrix is referred to as the stride. When the stride is set to 1, the filters are moved one pixel at a time. We switch the filters two pixels at a time when the stride is two, and so on. Convolution will function with a stride of 1, as seen in the diagram.

- **Padding**

Filters do not always perfectly fit the input image. There are two possibilities:

- To make the image fit, pad it with zero (zero padding).
- Remove the section of the image where the filter didn't fit. This is referred to as valid padding, and it preserves only the valid portions of the image.

3.5.1.3 Pooling layer

Feature motifs, which appear in the image as a result of the convolution process, may appear in a variety of places (see Fig. 1). Once features have also been retrieved, their actual position is less important as long as their relative position to others is preserved. Down-sampling, also known as pooling, is an intriguing local process. It compiles similar data in the general neighborhood of the receptive field and outputs the dominant response for that area.

$$Z_i^k = g_p\,(F_l^k) \qquad (4.2)$$

Equation (4.2) depicts the pooling process, where Z_i^k denotes the pooled feature-map of the lth layer for the kth input feature-map F_l^k, and $g_p(.)$ denotes the pooling operation form. The pooling operation aids in the extraction of a set of features that are insensitive to translational shifts and minor distortions. Eliminating the feature-map size to an invariant feature set not only governs the network's complexity, but it also aids generalization by reducing over-fitting.

Spatial Pooling can be of different types:

- Max pooling
- Average Pooling
- Sum Pooling

Max pooling takes the largest element from the rectified feature map. Taking the largest element could also take the average pool. The sum of all elements in the feature map is called sum pooling.

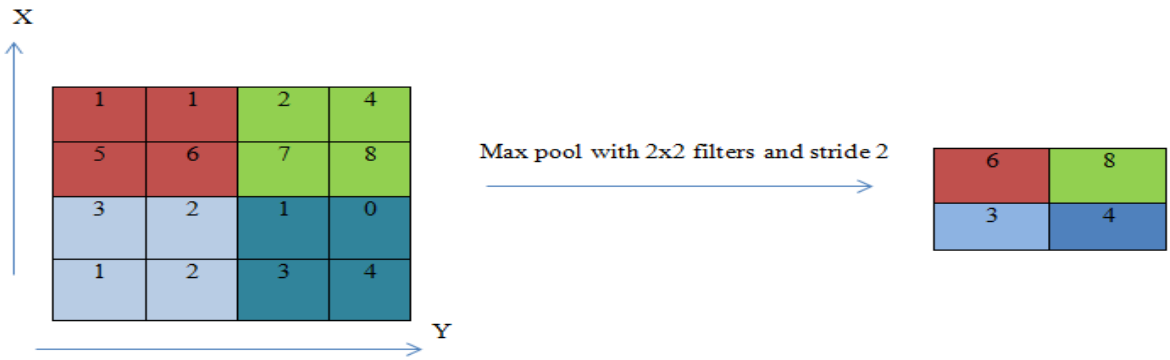

Figure 3. 14 Max Pooling.

3.5.1.4 Activation function

The decision-making function is an activation function that aids in the learning of complex patterns. The use of the right activation function would speed up the learning process. Equation (4.3) defines the activation function for a convolved feature map.

$$T_i^k = g_a(F_l^k) \qquad (4.3)$$

Where F_l^k: is a convolution output that is allocated to the activation function

$g_a(.)$: introduces non-linearity and returns a transformed output

T_i^k: for the l^{th} layer.

To inculcate non-linear combinations of features, various activation functions such as sigmoid, tanh, maxout, SWISH, ReLU, and variants of ReLU such as leaky ReLU, ELU, and PReLU are used. ReLU and its variants, on the other hand, are preferred because they help overcome the vanishing gradient problem. For a non-linear operation, ReLU stands for Rectified Linear Unit. The output is $f(x) = max(0, x)$. Why ReLU is important: ReLU's purpose is to introduce non-linearity in our ConvNet. Since the real-world data would want the ConvNet to learn would be non-negative linear values.

3.5.1.5 Batch normalization

The issue of internal covariance change within feature maps is addressed using batch normalization. The internal covariance shift is a change in the value distribution of hidden units that delays convergence (by pushing the learning rate to a small value) and necessitates cautious parameter initialization. In equation (4.4), batch normalization for a transformed feature-map F_l^k is shown.

$$\mathbf{N}_l^k = \frac{\mathbf{F}_l^k - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{4.4}$$

Where, N_l^k: represents a normalized feature-map

F_l^k is the input feature-map,

μ_B and σ_B^2 depict the mean and variance of a feature map for a mini batch respectively.

To avoid division by zero, ε is added for numerical stability. By setting the distribution of feature-map values to zero mean and unit variance, batch normalization unifies the distribution. Furthermore, it smoothest the gradient flow and serves as a controlling factor, thus assisting in the network's generalization.

3.5.1.6 Dropout

Dropout adds regularization to the network, which increases generalization by skipping certain divisions or connections at a certain probability at random. Multiple connections learning a non-linear relation in NNs are often co-adapted, resulting in over-fitting. This random dropping of certain connections or units results in several thinned network architectures, from which one

representative network is chosen with small weights. After that, the chosen architecture is used to estimate all of the suggested networks.

3.5.1.7 Fully connected layer

The fully connected layer is typically used for classification at the network's end. It is a global operation, unlike pooling and convolution. It takes data from the feature extraction stages and analyzes the output of all the layers before it. As a result, it creates a non-linear combination of selected features that are used to classify data (Khan et al., 2020).

3.5.2 Experimental Setup

A collection of tools and libraries is required to implement the framework.

- **Python**
- **TensorFlow**: Machine learning models can be implemented, trained, and deployed using this open-source platform.
- **Keras**: An open-source library that can be used to implement neural network architectures on both CPUs and GPUs.
- **Pandas**: Data analysis and modification library.
- **Matplotlib**: In Python, this tool is used to generate visualization plots such as charts, graphs, and more.
- **Numpy**: Array data structures can be used to perform a variety of mathematical computations and operations.

After installing all of the necessary software and libraries, the data collected in the previous section must be classified into chemical, clay, clean, sand, and silt and vegetable categories. Image processing techniques are used to detect contamination in water distribution systems in four phases: image acquisition, pre-processing (noise removal and image enhancement), image segmentation, and image analysis. Each of the foregoing processes has been improved to produce a high-quality image. Figure 3. 15 depict a typical image processing block diagram.

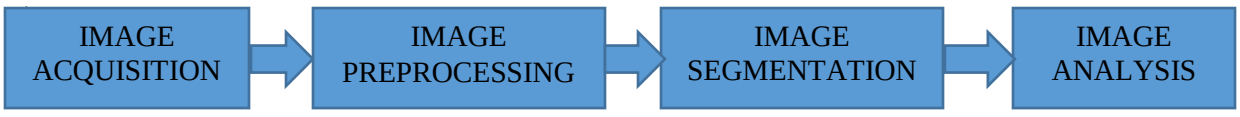

Figure 3. 16 Image processing and analysis.

Image Acquisition: This step is crucial and is regarded as the foundation of a proper image processing system. Three processes are involved in the acquisition: the region of interest object reflection, an optical system for focusing the energy, and sensors for detecting the amount of energy. After that, the image is saved on a storage medium. The dataset, which includes 2,000 images of water, was generated by the researcher. Of the 2,000 water images, 71% of the images are collected from sampled water, and another 29% of the images are collected from different sources. It comprises 1,600 training image datasets and 400 testing image datasets, all of which are colored images of various scale measurements grouped into five groups. However, the entire set of categorized images is not numbered, making it difficult to proceed to the next level. Labeling usually takes a collection of unlabeled data and adds meaningful, descriptive tags to each piece of that unlabeled data. As a consequence, any image in the category must be labeled.

Image Preprocessing- After sample collection, preprocessing ensures that the sample is clean and devoid of unnecessary deformities, as well as improves the sample's critical parameters for subsequent processing. Enhancement of the image is the process of increasing the visual appeal of an image by altering the pixel brightness. It consists of approaches for transforming images into a more understandable format for both humans and machines. Image enhancement focuses on and improves the image's certain features or attributes rather than increasing the image's intrinsic details. Based on the fact that once all images are labeled, the next part that is important for the implementation is selecting a standardized image size. The images will have to be reshaped before modeling so that all images have the same shape. This is often a small square image. If the source data has a variety of aspect ratios, some portraits, some landscapes, and the target object is typically in the center, a square crop from the middle may be a good compromise. There are many ways to accomplish this, but the most popular is to use a simple resize operation to stretch and distort an image's aspect ratio and push it into a new shape. When loading all of the images, I could examine the distribution of picture widths and heights, after which I created a new image size that better represents what it is most likely to see in person. Smaller inputs indicate a faster-to-train model, and this is generally the driving force in image size selection. In this case, the method is used and set to a fixed size of 256x256 pixels, which can be found in Appendix A code 1. The next step in the preprocessing phase is data cleaning and data augmentation.

Data cleaning is the process of creating a homogeneous input data set to prevent the model from being biased against some misinterpretations of images during training and therefore always predicting certain categories.

Image Data Augmentation: Image Data Augmentation is a technique for artificially increasing the size of a training dataset by creating altered versions of the images in the dataset. Deep learning neural network models that are trained on more data become more skilled, and augmentation techniques may generate variations of the images that enhance the fit models' ability to generalize what they've learned to new images. Data augmentation can also be used as a regularization strategy by introducing noise to the training data and allowing the model to learn the same features regardless of where they appear in the input. Small shifts and horizontal flips to the input images of clean and contaminated water may be useful for this issue. The ImageDataGenerator used for the training dataset will take these augmentations as arguments. Since we want to assess the model's output on the unaltered images, the augmentations should not be used for the test dataset. In this case, images in the training dataset will be augmented with small (10%) random horizontal and vertical shifts, rotated with (30%), zoomed with (20%), and random horizontal flips that create a mirror image of a photo, which can be found in Appendix A code 2. Images in both the training and test steps will have their pixel values scaled in the same way.

Image Segmentation Process: - The process of breaking an image into smaller portions, or segments, is known as image segmentation (Figure 3. 17). The approach is used to isolate an item or region of interest (ROI) from the visual background, allowing for easy assessment of the most essential and meaningful region. In addition to that, defining image properties is important.

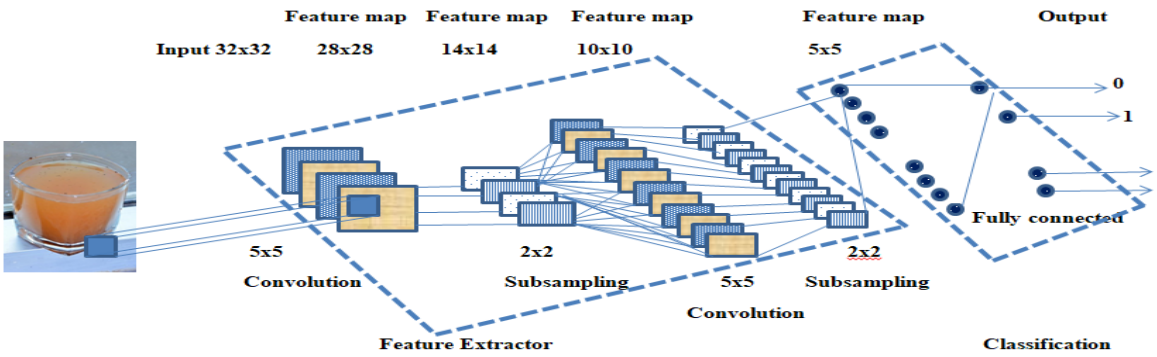

Figure 3. 18 Segmentation Process

Since data augmentation increases the training data, the image size has been redefined, which can be found in Appendix A code 3. This has an impact on the model's training time and requires more hardware resources. There are also different phases in the segmentation process.

Data preparation is an iterator that must load images for a single dataset in a sequential manner. It also helps to customize more information on how images are loaded. The "target size" claim is noteworthy because it allows you to resize all images to a particular size, which is often needed when modeling. Furthermore, the "class mode" argument allows you to determine the type of classification task, which can be either "binary" or "categorical," which is a multi-class classification. This is a multi-class classification analysis. When practicing, the default "batch size" is 32, which means 32 randomly selected images from around the dataset's classes will be returned in each batch, which can be found in Appendix A code 4. It is also possible to save augmented images to a file at the same time. Keras performs the data processing and augmentation in real-time. This is good for recall, but it might need the same images used during training. For example, you might want to use them later in a different software package, or you might just want to generate them once and use them on several deep learning models or configurations.

Create the model: - The model is made up of four convolution blocks, each with a max pool layer. A relu activation feature is used to enable a completely connected layer with 512 units on top. This model hasn't been fine-tuned for high precision.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 73, 73, 64)        1792
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 34, 34, 64)        36928
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 17, 17, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 15, 15, 64)        36928
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 7, 7, 64)          0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 5, 5, 64)          36928
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 2, 2, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 256)               0
_________________________________________________________________
dense (Dense)                (None, 512)               131584
_________________________________________________________________
dense_1 (Dense)              (None, 5)                 2565
=================================================================
Total params: 246,725
Trainable params: 246,725
Non-trainable params: 0
_________________________________________________________________
```

Image Analysis: - The method of gathering useful information from a digital image is known as image analysis. Scene analysis, picture description, image interpretation, pattern recognition, and computer/machine vision are examples of automatic or semi-automatic methods for extracting information.

3.6 Ethical Consideration

Ethical questions arise often in all research activities involving a wide range of individuals. Nonetheless, because of the sensitivity and ethical complexities of researching certain research topics, researchers are more careful. To ensure voluntary informed consent and protect the confidentiality and privacy of all companies and persons concerned, sampling and data collection must be conducted with caution. This was primarily done to safeguard the researcher's emotional, physical, and mental well-being. As a result, ethical values such as obtaining consent, maintaining privacy, and confidentiality were heavily considered during the research process.

3.6.1 Informed Consent

Securing the consent of participants and organizations with a stake in and responsibility for the potential research participants with whom the researcher aims to perform the study is one of the fundamentals of research activities, especially fieldwork. Informed consent means that a study participant has decided to participate in research after learning about and knowing the researcher, the research's goals, methodology and procedures, subjects, data use, and the ability to withdraw from the study at any time. However, obtaining consent entails not only providing facts but also

doing so in terms that participants understand. Potential research participants were told about the overall aspects of the study, including the study's purpose, methodology, and procedures, as well as their right to withdraw from the study at any time if they felt uncomfortable.

3.6.2 Respecting for Person

In this research, different participants are involved, including Addis Ababa Water and Sewage Authority, (AAWSA). This means, that individuals (the primary research participants) were contacted and then asked to provide personal experiences and views on water quality as well as the Authority data used for the research. Accordingly, privacy, avoiding undue interference in personal affairs, was taken as a vital ethical issue that should be considered in this research activity. This included recognizing the rights of research participants to control all personal information they provide. The standards focused on protecting participants from any harm during the research, ensuring that consent was obtained, and then the risks posed were modest. The dignity of all research participants and the privacy of the institution were respected. Thus, adherence to this principle ensures that people are not used simply as a means to achieve research objectives. Therefore, in this study, the research participants were informed not to talk and expose matters that they have made private to avoid negative feelings for themselves and others as well. Furthermore, as part of the research responsibility, the research confines itself to social and ethical principles so that it will not infringe on the private matters of the research participants and the Authority.

Chapter Four

Result and Discussion

The outcomes of a proposed technique are discussed in this chapter. First, we look at the overall model accuracy of the proposed architecture's training and validation results. Second, we show how different architectures were optimized to analyze the model that was used to find the best accuracy during the validation data identification phase.

4.1. System Models Analysis

The error rates are significant indicators in assessing the model's success in the proposed architecture. The accuracy indicates the likelihood that the picture will fit the target mark correctly. The accuracy level of the model varies throughout the training cycle due to the dataset used for training and validating the model. In 100 epochs, the model was conditioned. The batch size is 32 files. The overall accuracy of the training was 95%.

```
Epoch 95/100
57/57 [==============================] - 19s 337ms/step - loss: 0.1260 - accuracy: 0.9500 - val_loss: 0.6005 - val_accuracy:
0.8683
Epoch 96/100
57/57 [==============================] - 19s 337ms/step - loss: 0.1073 - accuracy: 0.9588 - val_loss: 0.7607 - val_accuracy:
0.8585
Epoch 97/100
57/57 [==============================] - 19s 338ms/step - loss: 0.1272 - accuracy: 0.9560 - val_loss: 0.9151 - val_accuracy:
0.8098
Epoch 98/100
57/57 [==============================] - 19s 338ms/step - loss: 0.1452 - accuracy: 0.9489 - val_loss: 1.0823 - val_accuracy:
0.8146
Epoch 99/100
57/57 [==============================] - 19s 337ms/step - loss: 0.1366 - accuracy: 0.9511 - val_loss: 0.9169 - val_accuracy:
0.8293
Epoch 100/100
57/57 [==============================] - 19s 339ms/step - loss: 0.1316 - accuracy: 0.9500 - val_loss: 0.8229 - val_accuracy:
0.8049
```

Figure 4. 1 Accuracy, loss, validation accuracy, and validation loss for the proposed model.

The model which is created in the previous chapter is made up of 4 convolutional layers, 4 max-pooling layers, and 1 dense layer is used. The accuracy of validation is 80.49 percent. This means we can correctly classify 80.49 percent of the images in the validation collection that the model missed.

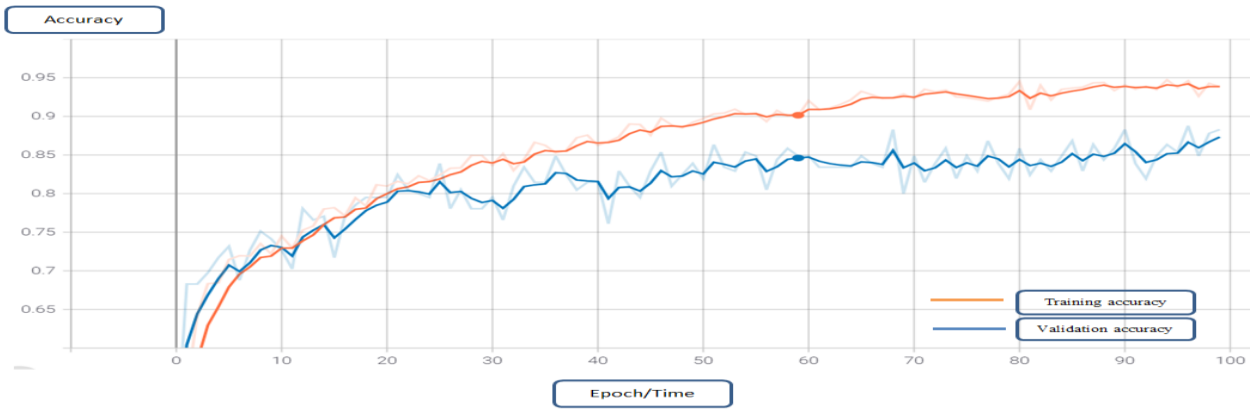

Figure 4. 2 Training and validation accuracy for the proposed model

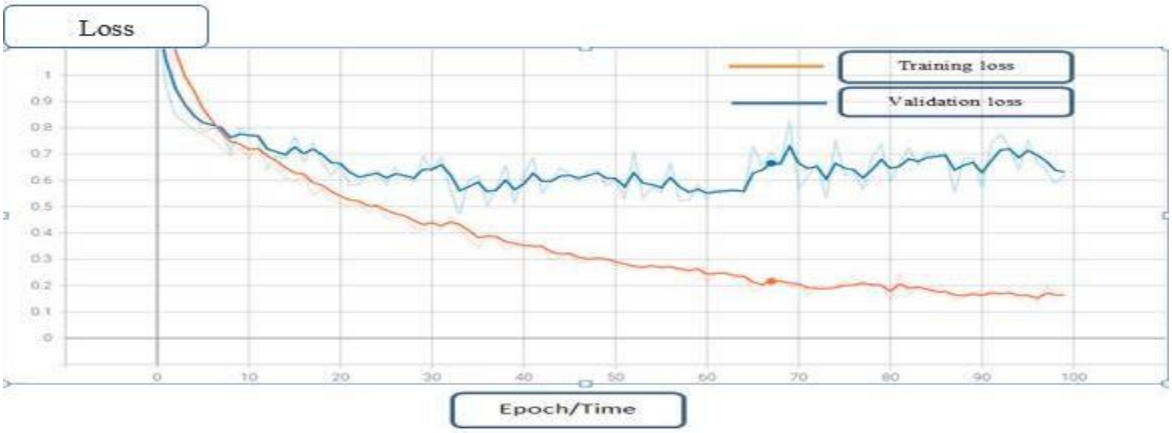

Figure 4. 3 Training and validation loss for the proposed model

From Figures 5.4 and 5.5, we can see that the training accuracy is above the validation accuracy and training loss is way below the validation loss, especially after the 20th epoch. When there are few training datasets, the model can learn from sounds or unwanted information in the training dataset, which can harm the model's output on the new dataset. This phenomenon is known as over-fitting. It means that the model will have a difficult time generalizing on the new dataset

There are many methods for combating over-fitting during the training period. Using data augmentation and adding Dropout to the model are the two main methods.

Data augmentation is a technique to produce further training data from an existing dataset by augmenting it with random transformations that result in believable-looking photos. This allows the model to be exposed to more facets of the data and generalize more effectively. The updated

model was trained with 4220 images, and its testing dataset is 1055 images, based on the principle of data augmentation.

Dropout, the kind of regularization, is another technique for reducing overfitting in the network. When Dropout is applied to a layer, it randomly removes several output units from the layer during the training phase (by setting the activation to zero). Dropout accepts fractional numbers in the manner of 0.1, 0.2, 0.4, and so on as data. This means that 10%, 20%, or 40% of its output nodes from the applied layer would be randomly removed. After applying the dense layer, the updated model randomly drops out twice 20 percent of the output units.

A little modification on the model which is created in the previous section is used to train the model with the new dataset, i.e., 4 convolutional layers, 4 max pooling, and 2 dense layers are used to achieve the accuracy level, followed by 2 dropouts to prevent over-fitting. The validation accuracy is 90.6 percent, and with a good balance of training and validation metrics, the accuracy increases.

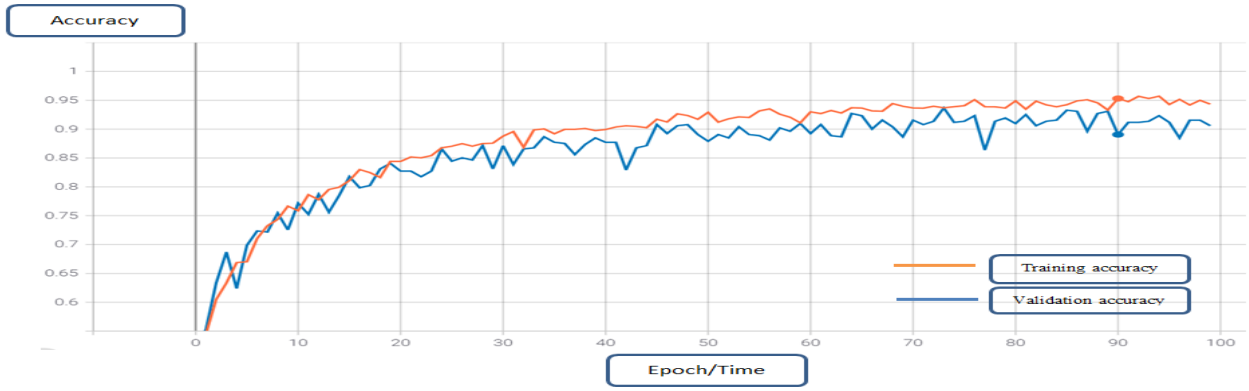

Figure 4. 4 Training and validation accuracy for the modified proposed model

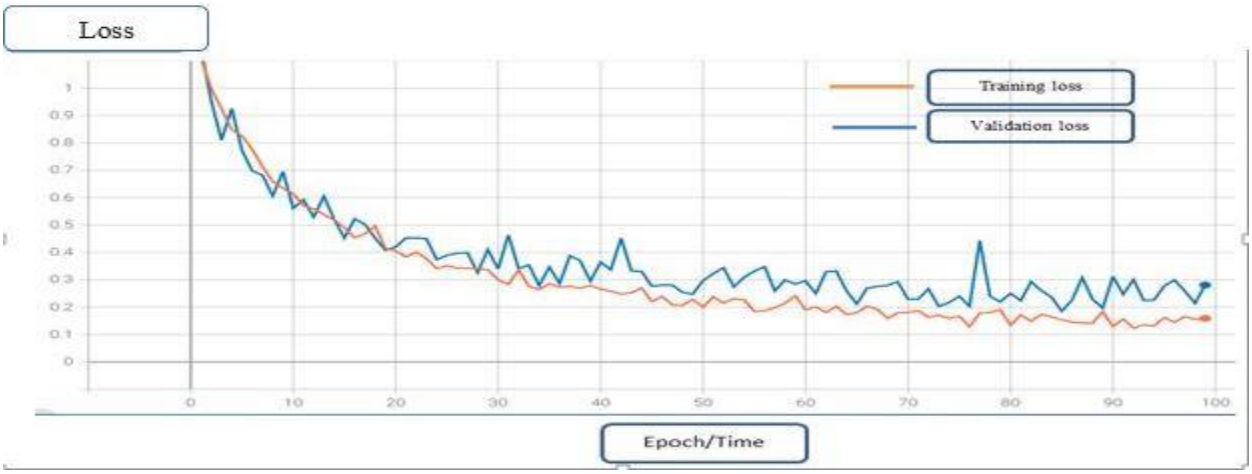

Figure 4. 5 Training and validation loss for the modified proposed model

The model improvement phase will go on until the training and validation precision, as well as training and validation loss, is almost perfectly aligned.

4.2. Optimization

We discussed the model accuracy and how to improve it using augmenting the data and dropout in the previous part. In this section, we will keep optimizing the model until it achieves the best accuracy on training and validation, as well as relatively low loss model architecture. Layers and nodes per layer, as well as 0, 1, or 2 dense layers, are the simplest things to change in the model. As a result, we can create 27 different models based on layers, nodes per layer, and dense layers.

The same dataset is used to train and test all models. Table 4.1 shows the accuracy, loss, validation accuracy, and validation loss of the produced models. Models contain some randomness. There will be no two rounds of optimizations that are the same. They can resemble each other but not be identical. Random weights are also used to start models. This can have a major impact on models, particularly when there are fewer epochs or a limited training set.

Table 4. 1 Model's accuracy, loss, and validation accuracy and validation loss

Models		Metrics			
		Accuracy	Loss	Validation Accuracy	Validation Loss
1	1-Conv-32-nodes-0-dense	0.2057	nan	0.2054	nan
2	1-Conv-32-nodes-1-dense	0.2057	nan	0.2054	nan
3	1-Conv-32-nodes-2-dense	0.9246	0.2206	0.9098	0.3099
4	1-Conv-64-nodes-0-dense	0.2057	nan	0.2054	nan
5	1-Conv-64-nodes-1-dense	0.9552	0.1282	0.9175	0.3176
6	1-Conv-64-nodes-2-dense	0.9392	0.1587	0.9136	0.2594
7	1-Conv-128-nodes-0-dense	0.2057	nan	0.2054	nan
8	1-Conv-128-nodes-1-dense	0.9392	0.1718	0.9213	0.2492
9	1-Conv-128-nodes-2-dense	0.9229	0.2133	0.8983	0.2651
10	2-Conv-32-nodes-0-dense	0.2057	nan	0.2054	nan
11	2-Conv-32-nodes-1-dense	0.9409	0.1588	0.9175	0.2249
12	2-Conv-32-nodes-2-dense	0.9399	0.1774	0.8829	0.3168
13	2-Conv-64-nodes-0-dense	0.2057	nan	0.2054	nan
14	2-Conv-64-nodes-1-dense	0.9656	0.1158	0.9487	0.1719
15	2-Conv-64-nodes-2-dense	0.9337	0.1986	0.8925	0.314
16	2-Conv-128-nodes-0-dense	0.2057	nan	0.2054	nan
17	2-Conv-128-nodes-1-dense	0.9614	0.1122	0.9367	0.185
18	2-Conv-128-nodes-2-dense	0.942	0.1736	0.9098	0.2338
19	3-Conv-32-nodes-0-dense	0.2057	nan	0.2054	nan
20	3-Conv-32-nodes-1-dense	0.9243	0.1962	0.906	0.2487
21	3-Conv-32-nodes-2-dense	0.9205	0.2257	0.8983	0.2875
22	3-Conv-64-nodes-0-dense	0.2057	nan	0.2054	nan
23	3-Conv-64-nodes-1-dense	0.9473	0.1501	0.9175	0.231
24	3-Conv-64-nodes-2-dense	0.9291	0.1919	0.9155	0.3459
25	3-Conv-128-nodes-0-dense	0.2057	nan	0.2054	nan
26	3-Conv-128-nodes-1-dense	0.9464	0.1504	0.9328	0.2184
27	3-Conv-128-nodes-2-dense	0.9119	0.2605	0.9002	0.3076

It may be tempting to choose the model with the highest validation accuracy, but it is preferable to choose the best (lowest) validation loss model. When it comes to models, there is some randomness, as previously said, but you should note patterns. For example, I noticed that models with 1 dense layer performed better overall, while models with 0 dense layers performed the worst. Thus, depending on such a validation accuracy of greater than 91 percent, this was chosen because the previous component had already achieved 90 percent validation accuracy. Now let

us start by looking at some of the better ones by zooming in on the validation accuracy graph. The following are the top 9:

Table 4. 2 Shows the top 9 models in terms of validation accuracy and validation loss.

Models	Validation Accuracy	Validation Loss
2-Conv-64-nodes-1-dense	0.9487	0.1719
2-Conv-128-nodes-1-dense	0.9367	0.185
3-Conv-128-nodes-1-dense	0.9328	0.2184
1-Conv-128-nodes-1-dense	0.9213	0.2492
1-Conv-64-nodes-1-dense	0.9175	0.3176
2-Conv-32-nodes-1-dense	0.9175	0.2249
3-Conv-64-nodes-1-dense	0.9175	0.231
3-Conv-64-nodes-2-dense	0.9155	0.3459
1-Conv-64-nodes-2-dense	0.9136	0.2594

When only training accuracy and training loss are considered, the model's output is 96.14 percent and 0.112 percent, respectively. However, when the validation accuracy and validation loss of this model are considered, it ranks second. From this one, I assume the model has been comfortable with 1 dense and 2 convolutional layers, as each version of those 2 options has proven to be better than the others. The top three models in terms of validation accuracy and validation loss are as follows:

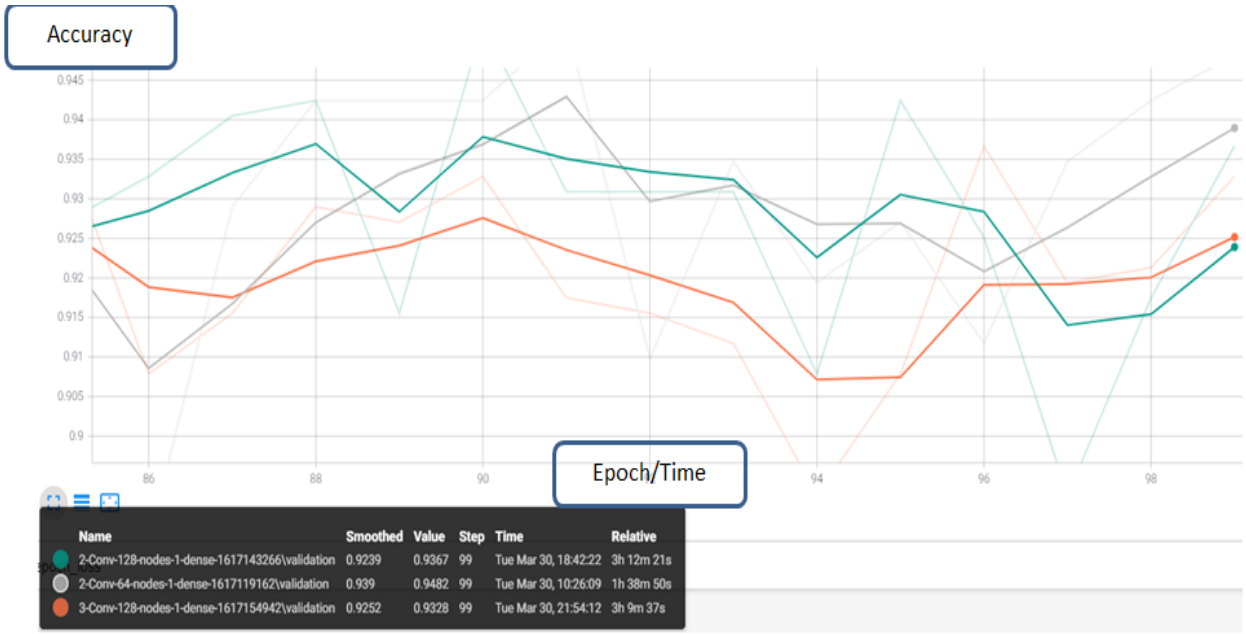

Figure 4. 6 Validation accuracy for the top 3 models

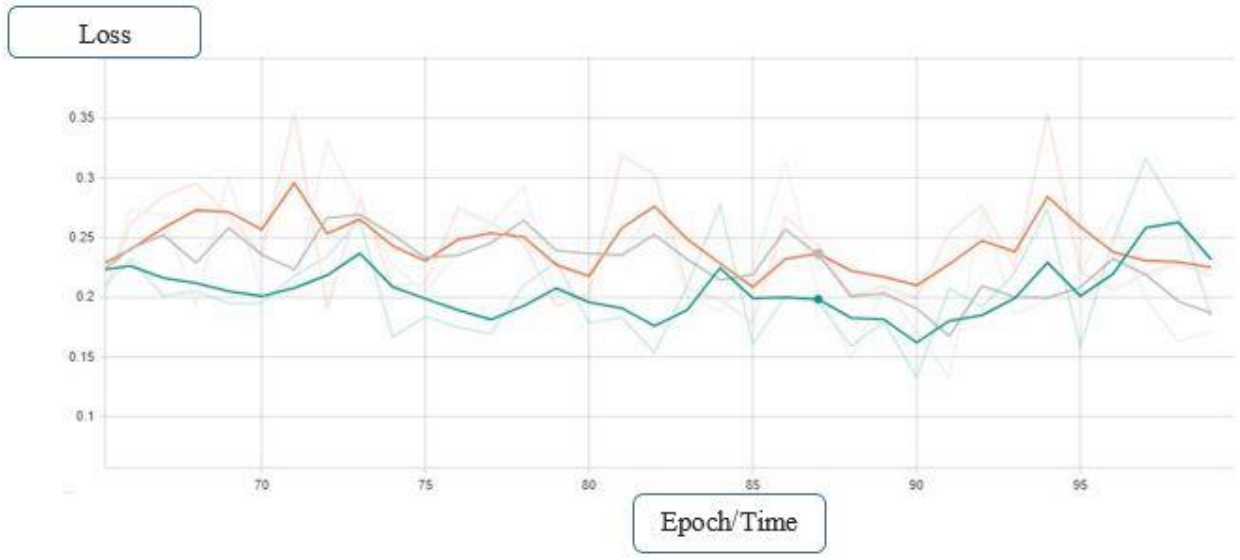

Figure 4. 7 Validation losses for the top 3 models

Finally, the model achieved an almost match with well-balanced training and validation metrics, with a validation accuracy of 94.82 percent and the highest (lowest) validation loss of 0.1719. The result was achieved using two convolutional layers, each with 64 nodes, 2 max-pooling layers, and one dense layer followed by a single dropout.

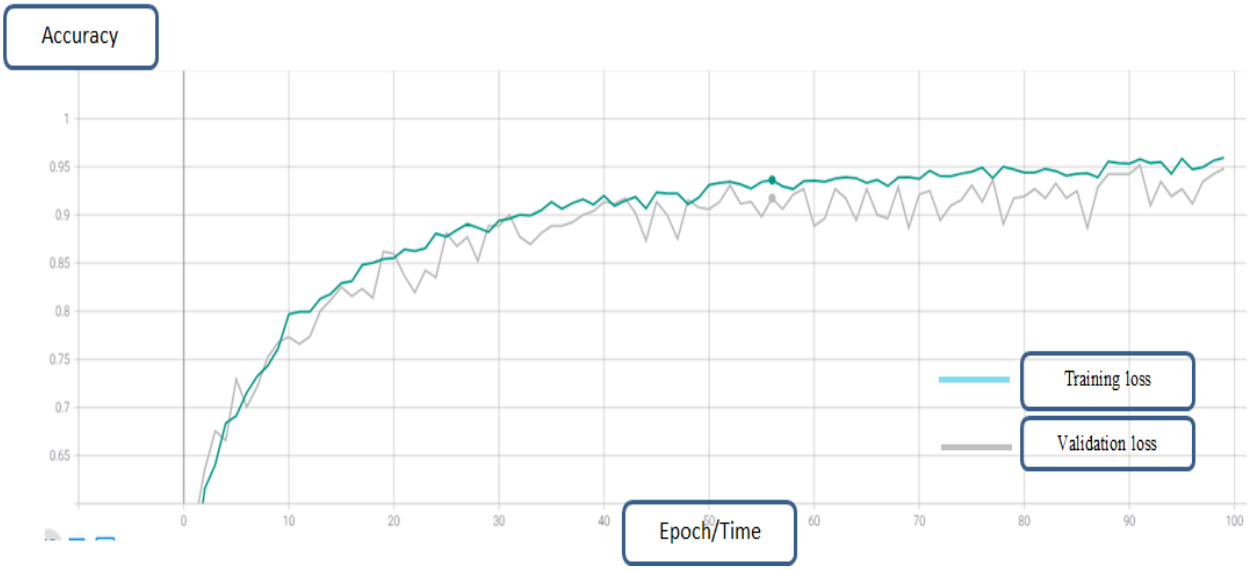

Figure 4. 8 Training and validation accuracy for the best models

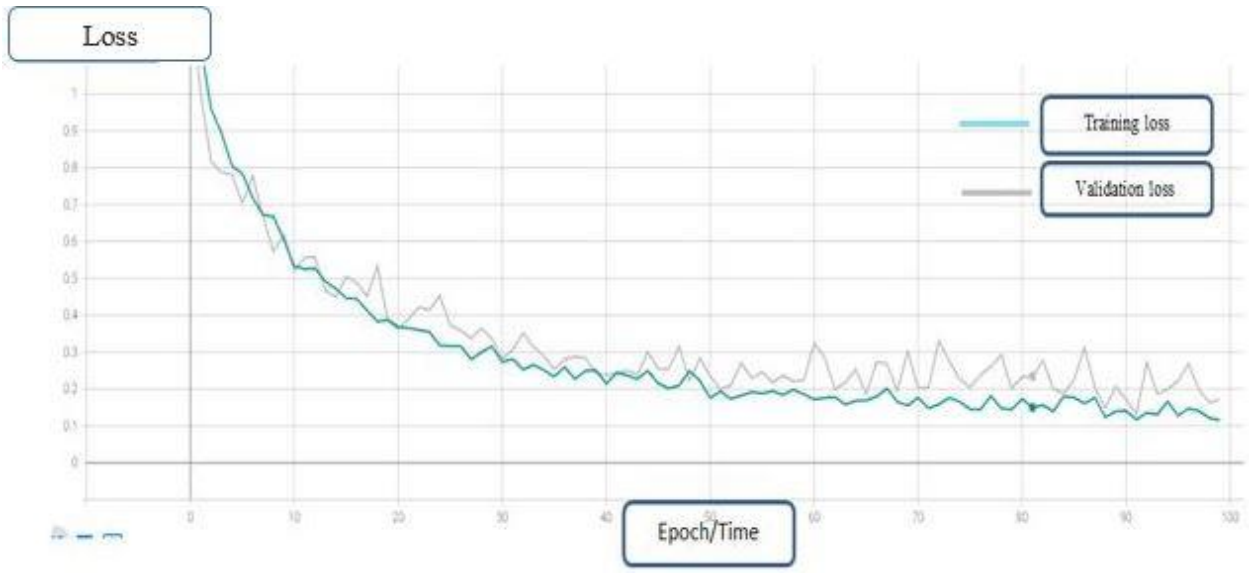

Figure 4. 9 Training and validation loss for the best models

If there are an unequal number of observations in each class, or if the dataset has more than two classes, classification accuracy alone can be misleading. The fundamental issue with classification accuracy is that it masks the information needed to properly evaluate your classification model's performance. There are two circumstances in which this problem is most likely to occur:

- If there are even more than two classifications in a dataset, with three or more classes, the classification accuracy may reach 80%, although it's unclear whether this is because all classes are predicted equally well or because one or two classes are ignored by the model.

- When the number of classes in the dataset is not even, the model may get a 90% or higher accuracy rate, but this isn't a good score if 90 of every 100 records belong to one class, and the model does this by always predicting the most prevalent class value.

In general, classification accuracy can mask the information essential to diagnose your model's performance. As a result, generating a confusion matrix can help you figure out what the classification model is getting right and where it is going wrong. Below is a multiclass confusion matrix showing TP, TN, FP, and FN.

- The total number of test images of any class would be the sum of the corresponding rows (i.e. the T_P+F_N for that class)
- The total number of F_N's for a class is the sum of values in the corresponding Row (excluding the T_P)
- The total number of F_P's for a class is the sum of values in the corresponding Column (excluding the T_P)
- The total number of T_N's for a certain class will be the sum of all columns and rows excluding that class column and row

Table 4. 3 *A multiclass Confusion Matrix of the classes*

	Predicted					
	Class	Chemical	Clay	Clean	Sand	Silt
Actual	Chemical	208	0	0	0	2
	Clay	0	206	0	5	1
	Clean	0	3	203	3	1
	Sand	1	1	5	201	4
	Silt	2	7	0	4	198

There are four machine learning-specific statistical measures. These tests were Accuracy, Precision, recall, and F1 score. The equations for calculating each of the statistical measures are shown below.

$$\text{F1-Score of Class } C_i \text{ F1 } (C_1) = \frac{= 2 * \text{TPR(Ci)} * \text{PPV(Ci)}}{(\text{TPR(Ci)} + \text{PPV(Ci)}} \quad (5.1)$$

$$\text{Recall of Class } C_i \text{ TPR } (C_i) = \frac{\text{TP(Ci)}}{(\text{TP(Ci)} + \text{FN(Ci)})} \quad (5.2)$$

$$\text{Precision of Class } C_i \text{ PPV } (C_i) = \frac{\text{TP(Ci)}}{(\text{TP(Ci)} + \text{FP(Ci)})} \quad (5.3)$$

$$\text{Accuracy } (A_{reduced}) = \frac{\sum \text{N i=1 TP(Ci)}}{\sum \text{N i=1} \sum \text{N j=1 Ci,j}} \quad (5.4)$$

$$\text{Accuracy} = \frac{208+206+203+201+198}{1055} = \mathbf{0.963}$$

Table 4. 4 *Precision, Recall and F1 scores of the classes*

	Precision	Recall	F1-score
Chemical	0.991	0.986	0.988
Clay	0.972	0.949	0.960
Clean	0.967	0.976	0.971
Sand	0.948	0.944	0.946
Silt	0.929	0.961	0.945

The F1-scores of each class can be combined to create a single measure for the entire model. There are a few methods

1. **Micro F1**:- It is calculated by taking into account the model's total FP, total TP, and total FN. It does not take each class into account separately; instead, the metrics are calculated as a whole.

   Total TP = (208 + 206 + 203 + 201 + 198) = 1016

   Total FP = (1+2) + (3+1+7) + (5) + (5+3+4) + (2+1+1+4) = 39

   Total FN = (2) + (5+1) + (3+3+1) + (1+1+5+4) + (2+7+4) = 39

Hence,

Precision = 1016/1016+39 = 0.963

Recall = 1016/1016+39 = 0.963

We can now use the regular F1-score algorithm to calculate the Micro F1-score using the precision and recall values listed above.

$$\text{Micro F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * 0.963 * 0.963}{0.963 + 0.963}$$

Micro F1 = **0.963**

As you can see, when we calculate the metrics globally, all of the variables are equal. That means Precision, Recall, Micro F1, and Accuracy all are equal. The accuracy calculated above is also the same.

Precision = Recall = Micro F1 = Accuracy

2. **Macro** F1:- It calculates metrics for each class separately and takes an unweighted average of the results. As we have seen from table 4.4 Precision, Recall, Micro F1 for each class,

   Class chemical F1-score = 0.988 Class clay F1-score = 0.960

   Class clean F1-score = 0.971 Class sand F1-score = 0.946

   Class silt F1-score = 0.945

Hence,

Macro F1 = (0.988+0.960+0.971+0.946+0.945)/5 = **0.963**

3. **Weighted F1**:- It uses a weighted mean of the metrics, unlike Macro F1. The total number of samples in each class determines the weights. Since we had 211 chemicals, 217 clay, 208 clean, 213 sand, and 206 silt samples.

   Weighted F1 = $\frac{(0.988*211)+ (0960*217) + (0.971*208) + (0.946*213) + (0.945*206)}{1055}$ = **0.963**

Figure 4. 10 Graphical user interfaces

After evaluating the model on the testing image dataset, it had an accuracy of 96.3 percent (1055/39 images) and a loss of 0.1719. The 39 pictures that were missing were categorized as follows. While all 210 images of chemically polluted water were marked as contaminated, two of them were incorrectly labeled as silt contamination. Six of the 212 clay-contaminated water images were misidentified. Of those six images, one image is classified as silt and five as sand pollution. Seven of the images that were missing were clean images that were misclassified as contaminated. From these seven missing images, three images were classified as clay, three images as sand, and one image was classified as silt pollution. Unlike the other contaminated classes, sand contaminated images are misclassified as clean water. Off all 212 sand contaminated water images, eleven images were classified wrongly, i.e. five images identified as clean, one image identified as clay, one image identified as chemical, and 4 images classified as silt contamination. Finally, all silt and vegetation testing images were categorized as polluted, but thirteen images were incorrectly classified, including four images identified as sand, two images identified as chemical, and seven images identified as clay, similar to other contaminated water images.

4.3. Discussion

The issue that this project looked into was finding easy-to-use methods of classifying water as safe or polluted in a water delivery system that was open to both the water company influences. This project aims to introduce simple and fast additional monitoring mechanisms to an existing system, assisting a water company and people who do not have a basic way to determine whether or not the distributed drinking water is polluted. Contaminated water is a major health risk, and many people who are affected by it are unaware of the full extent of the risk. Furthermore, water quality analysis can be difficult to obtain, especially in developing countries where costly chemical testing kits or other advanced methods of pollution testing are unavailable. This project, on the other hand, proposes a solution based on machine learning that enables the water company to monitor the water delivery system while also providing people in such circumstances with a simple way to evaluate water in their communities, resulting in bilateral control mechanisms. As discussed in the results section, the model had relatively high confidence in its predictions. The high accuracy value of 96.3 percent reflects this belief. The model predicted more false positives than false negatives, with seven clean photos listed as

contaminated, and five of the contaminated images were incorrectly classified as clean. One possible explanation for this phenomenon is that the model predicts three of the images as clay, three of the images as sand, and the last as silt, based on these seven frames. Because of this, when we saw the first image, the container containing the water was so close to the clay color, and because of this, its prediction was incorrect. Because of water transference, when we saw the other two images, the background was reflected through the water, and the background color looked like clay. As a result, the model was predicted inaccurately. The image, which the model assumed was sand, was taken from a distance. When we modify the image to some extent, for example, by zooming the image, and the model predicts that image, it accurately classifies the image. During the augmentation of the image, it changes some characteristics of the original image. In addition to that, when a photo is taken far from the object, the other things in the area are also included. These factors affect the classification process. One of the two photos, which were misclassified as contaminated with sand, was taken on the front side, and the other was taken from the top side. When we saw both images, they were close to being contaminated with sand at a low concentration. It accurately estimates the same photos taken in the other direction. When we see the last image, the two glasses overlap each other. For this reason, the model's prediction is incorrect. But when we take the main image at different angles, like from a top view or side view, and when the model predicts with those images, it perfectly classifies the image.

When we look at the other four groups, we can see that they have all incorrectly labeled the picture in various sizes. However, only one class i.e. sand class can classify the contaminated water as pure water, even though they classify the picture incorrectly. Let’s look at each one in detail.

1. Chemical class:-Two of the 210 test images were incorrectly identified as silt contamination. When we look at these two photographs, we can see that one of them is a combination of chemicals, silt, and vegetation, but the chemical concentration is higher than the silt and vegetation, so it is classified as a chemical. The classification is misleading, however, since these two groups are very identical. The other is a top-down image that has been contaminated with a small chemical. When we take this image from a side angle and create a model to classify it, it is accurately classified, not only for these images but also for other images taken from different angles.

2. Clay class: - Six pictures out of 212 were incorrectly identified as sand and silt contamination. When we look at all these images, we can see that they are polluted with a very low concentration of clay, which the light reflected into the water is a yellow and light blue color, respectively, and that they are near other classes that are also contaminated with a small amount. As a result, it incorrectly classifies the picture. However, the model correctly classifies the image when we adjust the color of the light that is reflected on the water.
3. Sand class: - Out of 212 test images, eleven were incorrectly classified: five as clean, one as clay, one as chemical, and four as silt. All five photos, classified as pure water, have the same container containing the water, and they are polluted with a very low concentration of sand. The images were taken from a distance. For all images, the light reflected into the water is a light blue color, which is very close to the container of the water. If we take an image with other lights, the model will classify perfectly. On the other hand, when we look at the image as sand pollution, we can see that it is polluted with a very low concentration of sand, that the light reflected into the water is blue, and that it is close to other groups that are contaminated with a small amount. The other images that were taken as chemical pollution are also contaminated with a small amount, similar to the previous one. However, when we take this picture from a different point of view and use a model to identify it, it is correctly classified. When we look at the other four pictures, we can see that two of them are contaminated with a low concentration and the other two are highly polluted and fall into the silt class. As a consequence, the model makes wrong predictions.
4. Silt class ꞉ As compared to other classes, this one incorrectly classifies the image used for testing purposes the most. Thirteen out of 211 test images were incorrectly categorized, with seven classified as clay, four as sand, and two as chemical. When we examine the seven photos taken as clay contamination, we can see that four images out of seven are contaminated with a very low concentration of silt and vegetation, with turbidity values of 8.45NTU and the rest three images are highly contaminated but when we look the images there is a similarity with clay. The other two images that are taken as chemical pollution is a top view image, close to the chemical contamination as seen above. However, when we take this picture from the side and use a model to identify it, it is

perfectly classified. The remaining four images are thought to be contaminated with sand. Some images are heavily polluted, while others are lightly polluted, and the remaining two images are taken from above.

Chapter Five

Conclusion and Recommendation

5.1. Conclusion

The objective of this research is to develop a model for detecting pollution in the water delivery system using an image of turbidity, a water quality parameter that is closely related to the Physico-chemical properties of water quality parameters. The suggested model, in particular, is a CNN that introduced an outlier detector in a drinking water delivery system and compared its output to sensor-based water quality models. A total of 2000 images were obtained from both clean and contaminated water to implement the model. In this study, polluted water was divided into four groups: chemicals, clay, sand and silt, and vegetation. The images obtained are insufficient to train a model to achieve high detection accuracy. Since the model accuracy improves as the amount of training data grows. Using data augmentation, the number of images obtained in this study was artificially increased to 5275. Of 5275 images, 4220 images are used for training, and 1055 images are used to test the model. The developed model not only detects whether it is clean or contaminated water, but it also identifies the source of contamination, i.e. chemicals, clay, sand and silt, and vegetation. This helps for treatment purposes. As a result, the established CNN model was effective in a variety of ways:

- The model achieved an accuracy of 96.2 percent and a loss of 0.1719, analyzing each sample in only 30s, much more efficient than other models like sensor-based contamination detector.
- The model detects contamination at a very low concentration of pollutants, i.e. it detects water pollution at the turbidity level of 6.45NTU and above.
- The model identifies the type of contamination that is used for treatment.
- The model provided easily understandable results, meaning that it uses a graphical user interface to enable users to interpret the results, and the prediction values allowed us to evaluate the model's confidence in its decision, while also providing a variety of contamination types.

In conclusion, this study aids in the development of an effective system for water quality management that is both efficient and simple for those water supply organizations that would assess the protection of their water supplies.

5.2. Recommendation

Expand the research in the future to provide analysis of microscopic water images to monitor for the existence of harmful bacteria and other types of a pollutant not covered in this study. I will keep refining the model and developing the mobile application, which is incorporated with the system and can be used by individuals to monitor water quality from anywhere, eventually leading to a mature system that allows for successful water quality inspection without the use of complicated chemical testing procedures.

References

AAWSA. (2011). A brief summary, accessible as hardcopy in AAWSA Library. Unpublished document.

Abbas Q, Ibrahim MEA, Jaffar MA. (2019). A comprehensive review of recent advances on deep vision systems. Artif Intell Rev 52:39–76. doi: 10.1007/s10462-018-9633-3

Ademe, A. S. (2014). Source and Determinants of Water Pollution in Ethiopia: Distributed Lag Modeling Approach. *Intellectual Property Rights: Open Access*, *2*(2). https://doi.ortests.4172/2375-4516.1000110

APHA. Standard methods for the examination of water and wastewater, 20th ed., American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC, USA. 1998-99

Ara S, MA Khan, MY Zargar, et al. Physico-chemical characteristics of Dal Lake water. Aqua Environ Toxicol. 2003; 12:129-134.

Arad, J., Housh, M., Perelman, L., Ostfeld, A. (2013). A dynamic thresholds scheme for contaminant event detection in water distribution systems. Water Res. 47 (5), 1899e1908. http://dx.doi.org/10.1016/j.watres.2013.01.017.Ben-Hur, A., Weston, J., 2010. A user's guide to support vector machines. In: Data Mining Techniques for the Life Sciences, Methods in Molecular Biology, vol. 609. Humana Press, pp. 223e239 a part of Springer Science þ Business Media, LLC 2010, http://dx.doi.org/10.1007/978-1-60327-241-4_13

Aschale, M., Sileshi, Y., Kelly-Quinn, M. & Hailu, D. (2016). Evaluation of potentially toxic element pollution in the benthic sediments of the water bodies of the city of Addis Ababa, Ethiopia. Journal of Environmental Chemical Engineering, 4(4, Part A), pp 4173–4183. Available: http://www.sciencedirect.com/science/article/pii/S2213343716303207.

Aschale, M., Sileshi, Y., Kelly-Quinn, M. & Hailu, D. (2017). Pollution Assessment of Toxic and Potentially Toxic Elements in Agricultural Soils of the City Addis Ababa, Ethiopia. Bulletin of Environmental Contamination and Toxicology, 98(2), pp 234–243. Available: https://doi.org/10.1007/s00128-016-1975-4.

Bengio Y. (2009 Learning Deep Architectures for AI. Found Trends® Mach Learn 2:1–127. doi:10.1561/2200000006.

Bouvrie J. (2006). Introduction Notes on Convolutional Neural Networks. doi: http://dx.doi.org/10.1016/ j.protcy.2014.09.007

Byer, D., & Carlson, K. H. (2005). Real-time detection of intentional chemical contamination in the distribution system. Journal-American Water Works Association, 97(7).

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Scientific Reports, 8(1), 6085.

Che, H., & Liu, S. (2014). Contaminant detection using multiple conventional water quality sensors in an early warning system. *Procedia Engineering*, *89*, 479–487. https://doi.org/10.1016/j.proeng.2014.11.239

Choi, S., Cichocki, A., Park, H.-M., Lee, S.-Y. (2005). Blind source separation and independent component analysis: a review. Neural Info Proc – Let and Rev. 6(1), 1-57.

Ciresan D, Giusti A, Gambardella LM, Schmidhuber J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems. pp 2843–2851

Eliades, D. G., Lambrou, T. P., Panayiotou, C. G., & Polycarpou, M. M. (2014). Contamination event detection in water distribution systems using a model-based approach. *Procedia Engineering*, *89*, 1089–1096. https://doi.org/10.1016/j.proeng.2014.11.229

Feintuch, P.L. (1990). Active Adaptive Noise Canceller without Training Mode. U.S. Patent, No. 5117401. U.S. Patent Office

Gakii, C., & Jepkoech, J. (2019). A Classification Model for Water Quality Analysis Using Decision Tree. *Journal of Chemical Information and Modeling*, *7*(3), 1–8.

Gangwar RK, Khare P Singh, J Singh, et al. (2012). Assessment of physicochemical properties of water. River Ramganga at Bareilly, UP, Journal of Chemical and Pharmaceutical Research, 2012; 4(9):4231-34. https://pubs.usgs.gov/wri/wri004139/pdf/wrir00-4139.pdf

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623

Guepie, B.K., Fillatre, L., Nikiforov, I. (2012). Sequential monitoring of water distribution network. In: Paper Presented at the IFAC Proceedings Volumes (IFAC-papers Online), vol. 16, pp. 392e397. PART 1, http://dx.doi.org/10.3182/20120711-3-BE-2027.00114.

Guo Y, Liu Y, Oerlemans A, et al. (2016). Deep learning for visual understanding: A review. Neurocomputing 187:27–48. doi: 10.1016/ j.neucom.2015.09.116

Gupta, A., & Ruebush, E. (2019). *AquaSight: Automatic Water Impurity Detection Utilizing Convolutional Neural Networks*. http://arxiv.org/abs/1907.07573

Hall, J., Zaffiro, A.D., Marx, R.B., Kefauver, P.C., Krishman, E.R.,Haught, R.C., Herrmann, J.G. (2007). On-line water quality parameters as indicators of distribution system contamination. Journal of the American Water Works Association 99 (1), 66e77.

Haught, R., Hall, J., Rahman, M., Richardson-Coy, R., Piao, H. (2005). The Bench-Scale Minimum Threshold Experiment. WQTC, Toronto, Canada.

Hou, D., He, H., Huang, P., Zhang, G., & Loaiciga, H. (2013). Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster-Shafer method. *Measurement Science and Technology*, *24*(5). https://doi.org/10.1088/0957-0233/24/5/055801

Housh, M., & Ostfeld, A. (2015). An integrated logit model for contamination event detection in water distribution systems. *Water Research*, *75*, 210–223. https://doi.org/10.1016/j.watres.2015.02.016

Huang, P., Jin, Y., Hou, D., Yu, J., Tu, D., Cao, Y., & Zhang, G. (2017). Online classification of contaminants based on multi-classification support vector machine using conventional

water quality sensors. *Sensors (Switzerland)*, *17*(3). https://doi.org/10.3390/s17030581

Kang, G. K., Gao, J. Z., & Xie, G. (n.d). Data-driven water quality analysis and prediction: A survey.

Kavi Priya, S., Shenbagalakshmi, G., & Revathi, T. (2017). Design of smart sensors for real time drinking water quality monitoring and contamination detection in water distributed mains. *International Journal of Engineering & Technology*, *7*(1.1), 47. https://doi.org/10.14419/ijet.v7i1.1.8921

Khan MA. Recent bio-limnological pollution trends in Kashmir Himalayan Dal Lake ecosystem. Development of red bloom. Ecol Energy. 1996; 22:41-77.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6

Klise, K.A., McKenna, S.A. (2006). Water quality change detection: multivariate algorithms. In: Saito, T.T., Lehrfeld, D. (Eds.), Proc. SPIE, 6203, pp. J1–J9.

Klise, K.A., McKenna, S.A. (2006). Multivariate application for detecting anomalous water quality. In: Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium. WDSA, Cincinnati, Ohio, USA, pp. 1e11. https://doi.org/10.1061/40941 (247) 130

Kroll, D., King, K. (2006). Real world operational testing and development of an on-line water security monitoring system. In: Eighth Ann WDSA Symp. Cincinnati, USA.

Kroll, D. (2006). Securing our water supply: protecting a vulnerable resource. Tulsa: Pennwell.

Kühnert, C., Bernard, T., Montalvo Arango, I., & Nitsche, R. (2014). Water quality supervision of distribution networks based on machine learning algorithms and operator feedback. *Procedia Engineering*, *89*, 189–196. https://doi.org/10.1016/j.proeng.2014.11.176

L. Perelman, J. Arad, M. Housh, A. Ostfeld, Event detection in water distribution systems from multivariate water quality time series, Environmental Science & Technology 46 (2012) 8212–8219

LeCun Y, Kavukcuoglu K, Farabet CC, others. (2010). Convolutional networks and applications in vision. In: ISCAS. IEEE, pp 253–256

Lin, Q.-H., Zheng, Y.-R., Yin, F.-L., Liang, H., Calhoun, V.D. (2007). A fast algorithm for one-unit ICA-R. Inf. Sci. 177, 1265–1275.

Liu, S., Che, H., Smith, K., & Chang, T. (2015). A real time method of contaminant classification using conventional water quality sensors. *Journal of Environmental Management*, *154*, 13–21. https://doi.org/10.1016/j.jenvman.2015.02.023

Liu W, Wang Z, Liu X, et al. (2016). A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26. doi: 10.1016/j.neucom.2016.12.038

Liu X, Deng Z, Yang Y. (2019). Recent progress in semantic image segmentation. Artif Intell

Rev 52:1089–1106. doi: 10.1007/s10462-018-9641-3

McKenna, S.A., Wilson, M., Klise, K.A. (2008). Detecting changes in water quality data. Journal of the American Water Works Association 100 (1), 74e85.

Mohammed, H., Hameed, I. A., & Seidu, R. (n.d.). *Machine Learning – Based Detection of Water Contamination in Water Distribution Systems*. 1664–1671.

Mohammadpour, R., Shaharuddin, S., Chang, C. K., Zakaria, N. A., Ab Ghani, A., & Chan, N. W. (2015). Prediction of water quality index in constructed wetlands using support vector machine. Environmental Science and Pollution Research, 22(8), 6208–6219.

Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294–307. https://doi.org/10.1080/24751839.2019.1565653

Murray, R., Haxton, T., McKenna, S.A., Hart, D.B., Klise, K.A.,Koch, M., Vugrin, E.D., Martin, S., Wilson, M., Cruze, V.A.,Cutler, L. (2010). Water quality event detection system for drinking water contamination waring system: Development Testing and Application Of CANARY. EPA/600/R-10/036. U.S. Environmental Protections Agency, Office of Research and Development, National Homeland Security Research Center, Cincinnati, Ohio,USA http://cfpub.epa.gov/si/si_public_record_report.cfm?address¼nhsrc/&dirEntryId¼221394.

Najafabadi MM, Villanustre F, Khoshgoftaar TM, et al. (2015). Deep learning applications and challenges in big data analytics. J Big Data 2:1–21. doi: 10.1186/s40537-014-0007-7.

National Research Council. (1995). Expanding the Vision of Sensor Materials. Academies Press, Washington, D.C., 134 p.

Oliker, N., & Ostfeld, A. (2014a). A coupled classification - Evolutionary optimization model for contamination event detection in water distribution systems. *Water Research*, *51*, 234–245. https://doi.org/10.1016/j.watres.2013.10.060

Oliker, N., & Ostfeld, A. (2014b). Minimum volume ellipsoid classification model for contamination event detection in water distribution systems. *Environmental Modelling and Software*, *57*, 1–12. https://doi.org/10.1016/j.envsoft.2014.03.011

Olson E. (2004). Grading Drinking water in U.S cities what's on Tap? National resource defense Counsel, New York City, and Washington, D.C., Los Angeles, and San Francisco.

P. Corso, M. Kramer, K. Blair, D. Addiss, J. Davis, A. Haddix, Cost of illness in the 1993 waterborne cryptosporidium outbreak, Milwaukee, Wisconsin, Emerging Infectious Diseases 9 (2003) 426–431.

Perelman, L., Arad, J., Housh, M., Ostfeld, A. (2012). Event detection in water distribution systems from multivariate water quality time series. . Environ. Sci. Technol. 46 (15), 8212e8219. http:/dx.doi.org/10.1021/es3014024.

Scherer D, Müller A, Behnke S. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. In: Artificial Neural Networks--ICANN 2010. Springer,

pp 92–101

Schwartz, J., Levin, R. (1999). Drinking water turbidity and health. Epidemiology, 86-90.

Sime, I. (1998). Addis Ababa water supply stage III-A project, Ethiopian Association of Civil Engineers Bulletin, vol. 1, no. 1, Ethiopian Association of Civil Engineers, Addis Ababa.

Skadsen, J., Janke, R., Grayman, W., Samuels, W., Tenbroek, M., Steglitz, B., & Bahl, S. (2008). Distribution system on-line monitoring for detecting contamination and water quality changes. *Journal / American Water Works Association*, *100*(7), 81–94. https://doi.org/10.1002/j.1551-8833.2008.tb09678.x

Szabo, J.G., Hall, J.S., Meiners, G. (2007). Water quality sensor responses to contamination in a single pass water distribution system simulator. EPA/600/R-07/001, p. 35.

The Water Project (Water in Crisis - Spotlight Ethiopia). (n.d.). Retrieved, 2015, from http://thewaterproject.org/water-in-crisis-ethiopia

Uber, J.G., Murray, R., Magnuson, M., Umberg, K. (2007). Evaluating real-time event detection algorithms using synthetic data. In: Paper Presented at the Restoring Our Natural Habitat-Proceedings of the 2007 World Environmental and Water Resources Congress. http://ascelibrary.org/doi/abs/10.1061/40927%28243%29499.

UN (United Nations). (2010). United Nations Millennium Development Goals. Accessed at https:// www.un.org/millenniumgoals/environ.shtml .

UN. (2017). World Population Prospects - Population Division - United Nations. Available https://population.un.org/wpp/Download/Standard/Population/

UN. (2018). World Urbanization Prospects - Population Division - United Nations. Available https://population.un.org/wup/

UNDP (United Nations Development Program). (2006). Human development report 2006. New York: UNDP.

U.S. Environmental Protection Agency. (2005). Technologies and Techniques for Early Warning Systems to Monitor and Evaluate Drinking Water Quality: State-of the- Art Review. ORD, NHSRC, Washington, DC.

U.S. EPA. (2005a). Water Sentinel: Online Water Quality Monitoring as an Indicator of Drinking Water Contamination. Available at: http://www.epa.gov/watersecurity/pubs/watersentinel_wq_monitoring.pdf

U.S. EPA. (2005b). Water Sentinel: System Architecture. Available at: http://www.epa.gov/watersecurity/pubs/watersentinel_system_architecture.pdf

W.G. Wright Jr, and A.J. Thomas III. (2000). the Federal/Arkansas Water Pollution Control Programs: Past, Present, and Future. UALR L. Rev., 23: 541 (2000).

W. J. Cooper, (2014). Responding to crisis: The West Virginia chemical spill, Environmental Science & Technology 48 (2014) 3095–3095.

Worldwildlife. (2021) Water Scarcity | Threats | WWF. (n.d.). Retrieved May 29, 2021, from

https://www.worldwildlife.org/threats/water-scarcity

Jeremiah Castelo. (2021) *What is the Percentage of Drinkable Water on Earth? | World Water Reserve. (n.d.).* Retrieved May 29, 2021, from https://worldwaterreserve.com/water-crisis/percentage-of-drinkable-water-on-earth/

WHO/UNESCO. (2010). Progress on Sanitation and Drinking water: 2010 Update. Geneva, WHO press JMP, Update, 2010.

WHO, W. H. O. (2006). Guidelines for the safe use of wastewater, excreta and greywater - Volume 1. (WHO), http://www.who.int/water_sanitation_health/publications/gsuweg1/en/

World Health Organization. (2011). Nitrate and Nitrite in Drinking-Water. https://www.who.int/water_sanitation_health/dwq/chemicals/nitratenitrite2ndadd.pdf.

Xiang, Y., & Jiang, L. (2009). Water quality prediction using ls-svm and particle swarm optimization. Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on (pp. 900–904). IEEE.

Yang, Y.J., Haught, R.C., Hall, J., Szabo, J., Clark, R.M., Meiners, G. (2007). Adaptive water sensor signal processing: experimental results and implications for online contaminant warning systems. In: Wat Env and Wat Res Congress, Tampa, Florida

Yang, J.Y., Haught, C.R., Goodrich, A.J. (2009). Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: techniques and experimental results. Journal of Environmental Management 90 (8), 2494e2506.

Yohannes, H. & Elias, E. (2017). Contamination of Rivers and Water Reservoirs in and Around Addis Ababa City and Actions to Combat It. Environment Pollution and Climate Change, 01.

Zhang, J., Zhu, X., Yue, Y., & Wong, P. W. (2017). A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. Seventh international conference on innovative computing technology (INTECH) (pp. 36–41). IEEE.

Appendix

Appendix: A Sample code

```
In [9]: from PIL import Image
        import os
        src_path = r'C:\Users\Administrator\Desktop\New folder\infected/'
        def resize_multiple_images(src_path, dst_path):
            # Here src_path is the location where images are saved.
            for filename in os.listdir(src_path):
                try:
                    img=Image.open(src_path+filename)
                    new_img = img.resize((256,256))
                    if not os.path.exists(dst_path):
                        os.makedirs(dst_path)
                    new_img.save(dst_path+filename)
                    print('Resized and saved {} successfully.'.format(filename))
                except:
                    continue
        src_path = r'C:\Users\Administrator\Desktop\Measured and\Clay/'
        dst_path = r'C:\Users\Administrator\Desktop\New folder (2)\clay/'
        resize_multiple_images(src_path, dst_path)

        Resized and saved DSC00008.JPG successfully.
        Resized and saved DSC00009.JPG successfully.
        Resized and saved DSC00010.JPG successfully.
```

Code 1

```
In [3]: training_datagen = ImageDataGenerator(rescale=1./255,
                                              featurewise_center=False,  # set input mean to 0 over the dataset
                                              samplewise_center=False,  # set each sample mean to 0
                                              featurewise_std_normalization=False,  # divide inputs by std of t
                                              samplewise_std_normalization=False,  # divide each input by its s
                                              zca_whitening=False,  # apply ZCA whitening
                                              rotation_range = 30,  # randomly rotate images in the range (degr
                                              zoom_range = 0.2, # Randomly zoom image
                                              width_shift_range=0.1,  # randomly shift images horizontally (fra
                                              height_shift_range=0.1,  # randomly shift images vertically (frac
                                              horizontal_flip = True,  # randomly flip images
                                              vertical_flip=False)  # randomly flip images
```

Code 2

```
In [5]: # Model configuration
        batch_size = 32
        img_width, img_height, img_num_channels = 75, 75, 3
        no_classes = 5
        no_epochs = 10
        optimizer = Adam()
        verbosity = 1
```

Code 3

```
In [8]: test_set_folder     = r'C:\Users\Administrator\Desktop\Research_Four\test/'
        validation_gen = validation_datagen.flow_from_directory(test_set_folder,
                                                                target_size=(75, 75),
                                                                save_to_dir = r'C:\Users\Administrator\Desktop\Researc
                                                                save_format='jpg',
                                                                batch_size=batch_size,
                                                                class_mode = "categorical")
```

Code 4

Appendix: B Data Analysis Results of clean water

Table 1 Results from conducted analyses of collected clean water samples.

Parameters	Turbidity	PH	TDS	EC	Total Alkalinity	Total Hardness	N (Ammonia)	N (Nitrite)	N (Nitrate)	SO4 (Sulfate)	PO4 (Phosphate)	F- (Fluoride)	Fe (Iron)	Mn (Manganese)	SiO2 (Silica)	Cl (Chloride)	Bicarbonate Alkalinity
Units																	
Date	NTU		mg/l	µS	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l
Jan 8, 2020	0.56	7.75	74	134.1	44	52	0.162	0.0055	0.48	0.1	0.086	0.27	0.122	0.018	11.6	6.5	44
Jan 12, 2020	1.2	7.61	72	127	47.4	49.6	0.133	0.0022	0.32	1.8	0.053	0.13	0.064	0.058	7.4	5.8	47.4
Jan 15, 2020	0.96	7.73	70	132.4	45	49.9	0.855	ND	0.2	3.4	ND	0.051	0.077	0.017	ND	5	45
Jan 19, 2020	0.98	7.58	76	137.6	51	48	0.025	0.0014	0.15	ND	0.294	ND	2.543	0.42	1.4	7	51
Jan 21, 2020	1.1	7.71	78	141.4	50.8	92.8	0.755	0.0043	0.42	1.5	1.363	ND	2.077	0.012	ND	3.6	50.8
Jan 24, 2020	0.78	7.15	74	135.5	63.2	58.6	0.336	ND	0.65	2.3	0.889	0.086	2.088	0.028	2.7	5.1	63.2
Jan 26, 2020	1.23	7.54	72	131.4	48	45.2	ND	ND	0.8	ND	0.129	ND	1.993	0.017	5.6	5.4	48
Jan 28, 2020	0.75	6.76	71	129.4	52.2	54.4	0.325	0.0062	0.45	3.3	0.136	0.14	0.137	0.026	15.6	6.3	52.2
Jan 30, 2020	0.88	7.17	68	134.9	50.8	92.8	0.544	0.0043	0.42	1.5	1.363	ND	2.077	0.012	8.3	3.6	50.8
Jan 31, 2020	0.89	6.98	65	128.6	50.4	58.2	0.753	0.0026	1.1	0.9	0.103	0.18	1.662	0.349	12.3	4.1	50.4
Feb 2, 2020	0.91	7.05	64	110.6	45.2	56.4	0.887	0.0043	1.2	0.7	0.133	0.14	1.582	0.488	14.7	5.1	45.2
Feb 4, 2020	0.95	7.37	63	117.7	48.8	62.6	0.447	0.0037	0.89	0.9	0.095	0.19	0.464	0.228	9.3	6.4	48.8
Feb 7, 2020	0.72	7.57	66	118.3	44.6	54.8	0.262	0.063	0.75	0.6	0.077	0.14	0.263	0.118	10.6	6.2	44.6
Feb 10, 2020	0.45	7.65	73	138.4	45.4	51.2	0.146	0.0123	0.52	1.2	0.064	0.18	0.164	0.085	8.6	5.9	45.4
Feb 12, 2020	1.1	7.73	71	135.4	46	48	0.035	0.0146	0.28	2.5	0.188	0.074	1.273	0.116	2.3	6.1	46
Feb 15, 2020	0.99	7.76	78	144.5	62.2	108	0.034	0.0223	0.13	0.8	0.394	0.033	2.245	0.34	0.9	6.7	62.2
Feb 18, 2020	0.89	7.68	64	118.6	54.8	88.6	0.63	0.0034	0.24	1.8	1.133	0.041	2.144	0.025	1.9	3.5	54.8
Feb 21, 2020	0.85	7.21	65	138.5	61.2	61.6	0.146	0.0017	0.57	1.9	0.995	0.063	2.011	0.032	3.7	4.2	61.2
Feb 24, 2020	0.79	7.54	62	131.4	49	45.2	ND	ND	0.8	ND	0.129	ND	1.993	0.017	5.6	5.4	49

Mar 1, 2020	0.77	7.51	75	142	47.2	76.4	0.074	ND	3.7	0.2	0.159	ND	1.484	0.434	13	2.3	47.2
Mar 4, 2020	0.88	7.26	71	149.5	63	65	0.876	0.009	0.11	0.3	0.071	0.32	1.771	0.644	13.2	6.3	63
Mar 10, 2020	0.91	7.08	73	136.4	53.2	60.2	0.785	0.001	1.4	0.7	0.098	0.29	1.887	0.457	14.1	5.3	53.2
Mar 13, 2020	1.1	6.95	71	135	43.4	54.8	0.336	0.006	1.9	0.5	0.121	0.01	1.982	0.557	15.3	4.5	43.4
Mar 16, 2020	1.2	7.57	70	133.2	48.4	64.8	0.385	0.0035	0.36	1.8	0.198	0.15	0.228	0.023	8.3	5.7	48.4
Mar 23, 2020	0.99	7.75	74	134.4	42	54	0.162	0.0055	0.48	0.1	0.086	0.27	0.122	0.018	11.6	6.5	42
Mar 26, 2020	0.95	7.61	72	137	47.4	50.6	0.133	0.0022	0.32	1.8	0.053	0.13	0.064	0.058	7.4	5.4	47.4
Mar 29, 2020	0.93	7.73	70	134.3	47	49.9	0.23	ND	0.2	3.4	ND	0.051	0.077	0.017	ND	4.8	47
Apr 2, 2020	0.92	7.58	76	137.6	51	48	0.025	0.0114	0.15	ND	0.294	ND	2.543	0.42	1.4	7	51
Apr 5, 2020	0.56	7.71	78	131.4	52.8	92.8	0.662	0.0043	0.42	1.5	1.363	ND	2.077	0.012	ND	4.2	52.8
Apr 11, 2020	0.75	7.15	74	135.5	64	58.6	0.26	ND	0.65	2.3	0.889	0.086	2.088	0.028	2.7	4.1	64
Apr 14, 2020	0.85	7.54	72	131.4	53	45.2	ND	ND	0.8	ND	0.129	ND	1.993	0.017	5.6	5.3	53
Apr 17, 2020	0.86	6.76	71	129.4	57	54.4	0.325	0.0145	0.45	3.3	0.136	0.14	0.137	0.026	15.6	7.1	57
Apr 23, 2020	0.89	7.17	68	134.9	49	92.8	0.84	0.0054	0.42	1.5	1.363	ND	2.077	0.012	8.3	3.6	49
Apr 26, 2020	0.91	6.98	71	138.6	52	58.2	0.532	0.0233	1.1	0.9	0.103	0.18	1.662	0.349	12.3	4.1	52
Apr 29, 2020	0.99	7.05	74	130.6	45	56.4	0.322	0.0078	1.2	0.7	0.133	0.14	1.582	0.488	14.7	5.3	45
May 3, 2020	1.1	7.37	73	137.7	47	62.6	0.447	0.0056	0.89	0.9	0.095	0.19	0.464	0.228	9.3	6.4	47
May 6, 2020	1.02	7.57	76	138.3	45.4	54.8	0.346	0.0025	0.75	0.6	0.077	0.14	0.263	0.118	10.6	5.2	45.4
May 12, 2020	0.95	7.65	73	135.4	46.2	51.2	0.453	0.0089	0.52	1.2	0.064	0.18	0.164	0.085	8.6	6.3	46.2
May 15, 2020	0.88	7.73	71	133.8	48	48	0.335	0.0045	0.28	2.5	0.188	0.074	1.273	0.116	2.3	5.8	48
May 21, 2020	0.82	7.76	68	124.5	59	108	0.078	0.0156	0.13	0.8	0.394	0.033	2.245	0.34	0.9	5.6	59
May 24, 2020	0.78	7.68	74	138.5	56.8	88.6	0.583	0.0085	0.24	1.8	1.133	0.041	2.144	0.025	1.9	4.3	56.8
May 27, 2020	0.92	7.21	65	127.3	60.4	61.6	0.143	0.0074	0.57	1.9	0.995	0.063	2.011	0.032	3.7	4.4	60.4
May 30, 2020	0.65	7.54	72	131.4	49	45.2	0.089	0.0045	0.8	ND	0.129	ND	1.993	0.017	5.6	5.2	49
Jun 3, 2020	0.67	7.51	78	141.2	47.2	76.4	0.074	0.0053	3.7	0.2	0.159	ND	1.484	0.434	13	3.1	47.2
Jun 8, 2020	0.89	7.26	71	143.5	63	65	0.765	ND	0.11	0.3	0.071	0.32	1.771	0.644	13.2	5.9	63
Jun 9, 2020	0.92	7.08	73	136.4	53.2	60.2	0.326	0.001	1.4	0.7	0.098	0.29	1.887	0.457	14.1	4.5	53.2
Jun 11, 2020	0.78	6.95	71	135.6	43.4	54.8	0.852	0.006	1.9	0.5	0.121	0.01	1.982	0.557	15.3	4.6	43.4
Jun 13, 2020	0.99	7.57	70	133.2	48.4	64.8	0.432	ND	0.36	1.8	0.198	0.15	0.228	0.023	8.3	6.8	48.4
Jun 16, 2020	1.1	7.75	74	134.1	42	52	0.162	0.0055	0.48	0.1	0.086	0.27	0.122	0.018	11.6	6.5	42

Appendix: C Data Analysis Results of contaminated water

Table 2 Results from conducted analyses of collected contaminated water samples.

Parameters	Turbidity	PH	TDS	Total Alkalinity	Ca Hardness	N (Nitrite)	N (Nitrate)	Fe (Iron)	Mn (Manganese)	SiO2 (Silica)	Cl (Chloride)
Units											
Date	NTU		mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l
Jan 8, 2020	33.7	7.75	524	63	40.6	0.006	1.1	1.982	0.644	11.6	6.7
Jan 12, 2020	62.2	7.61	574	53.2	35	ND	1.2	0.228	0.488	7.4	3.5
Jan 15, 2020	73.12	7.73	578	43.4	42.4	0.0055	0.89	0.122	0.457	ND	4.2
Jan 19, 2020	77.1	6.02	547	48.4	33.2	0.0022	57.12	0.064	0.434	1.4	5.4
Jan 21, 2020	57.41	7.71	562	42	35.8	ND	0.52	0.077	0.349	ND	2.3
Jan 24, 2020	64.52	7.15	521	47.4	32.2	0.0231	0.28	2.543	0.34	2.7	6.3
Jan 26, 2020	62.62	4.23	423	47	67.5	0.0056	0.13	22.1	3.2	100	300
Jan 28, 2020	59	6.76	200	51	50.2	ND	0.24	24.2	0.118	15.6	4.5
Jan 30, 2020	41.4	7.17	159	52.8	42.4	ND	0.57	1.993	0.116	8.3	5.7
Jan 31, 2020	71.3	6.98	320	64	39.2	0.0062	0.8	0.137	0.085	12.3	6.5
Feb 2, 2020	65.6	7.05	331	53	47	0.0045	3.7	2.077	0.032	14.7	5.4
Feb 4, 2020	87.9	7.37	323	48.8	44	0.0156	0.11	1.662	0.026	9.3	4.8
Feb 7, 2020	88.7	5.91	351	44.6	42.2	0.0085	62.3	16.5	0.025	10.6	7
Feb 10, 2020	83.3	7.65	372	45.4	34.2	0.0074	1.9	0.464	0.017	8.6	4.2
Feb 12, 2020	64.4	7.73	311	46	42.6	0.0045	0.36	0.122	0.012	2.3	4.1
Feb 15, 2020	77.6	7.76	301	62.2	33.2	0.0053	0.13	1.993	0.644	14.1	5.3
Feb 18, 2020	72.2	7.68	306	54.8	35.8	ND	0.24	17.64	0.488	15.3	7.1
Feb 21, 2020	66.8	7.21	323	61.2	32.2	0.001	0.57	1.771	0.434	8.3	3.6
Feb 24, 2020	87.8	7.54	410	49	60.4	0.006	0.8	1.887	9.1	11.6	4.1
Mar 1, 2020	99.12	7.51	387	47.2	50.2	ND	3.7	1.982	0.34	53.7	5.3
Mar 4, 2020	102.3	7.26	315	44	71.3	0.0034	0.11	13.56	0.228	ND	6.4

Mar 10, 2020	120.12	7.08	425	47.4	39.2	0.0017	1.4	0.122	0.118	1.4	5.2
Mar 13, 2020	92.1	6.95	362	45	47	ND	1.9	0.064	0.116	ND	6.3
Mar 16, 2020	78.4	7.57	303	42.2	44	ND	0.36	0.077	0.085	2.7	5.8
Mar 23, 2020	95.6	7.75	299	51	42.2	0.009	0.48	2.543	0.032	5.6	5.6
Mar 26, 2020	96.4	7.61	342	52.8	34.2	0.001	0.32	2.077	0.026	15.6	4.3
Mar 29, 2020	89.7	7.73	344	64	42.6	0.006	1.1	19.9	17.2	8.3	287
Apr 2, 2020	63.8	5.99	356	53	24	0.0035	69.2	1.993	0.017	12.3	5.2
Apr 5, 2020	77.1	7.71	361	57	37.4	0.0055	0.89	0.137	0.017	14.7	3.1
Apr 11, 2020	56.5	7.15	320	49	34.2	0.0022	0.48	2.077	0.012	9.3	5.9
Apr 14, 2020	65.3	7.54	350	52	38.4	0.0156	0.32	1.993	0.018	0.9	4.5
Apr 17, 2020	52.5	6.76	301	45	42.2	ND	0.2	0.137	0.488	1.9	4.6
Apr 23, 2020	56.1	5.3	298	47	40.6	0.0563	0.15	13.2	0.42	76.4	6.8
Apr 26, 2020	57.9	6.98	305	45.4	35	0.0085	0.42	1.662	0.349	5.6	4.8
Apr 29, 2020	58.7	7.05	302	46.2	66	0.0055	0.65	1.582	0.34	13	6.4
May 3, 2020	66.75	7.37	312	47	22	0.0022	0.8	0.464	0.228	13.2	6.5
May 6, 2020	56.6	6.05	300	45.4	36.4	ND	65.4	0.263	10.2	14.1	5.8
May 12, 2020	45.4	7.65	320	46.2	34.2	0.0014	0.42	0.164	0.116	15.3	5
May 15, 2020	41.2	7.73	273	48	34	0.0043	1.1	14.4	0.085	8.3	7
May 21, 2020	42.3	7.76	363	59	42.2	ND	1.2	2.245	0.058	11.6	3.6
May 24, 2020	36.1	7.68	356	56.8	40.6	0.0085	0.89	2.144	0.028	7.4	5.1
May 27, 2020	39.9	7.21	307	48	39.2	0.0074	0.57	2.011	0.026	ND	5.4
May 30, 2020	46.21	7.54	193	59	38.4	0.0045	0.8	1.993	0.017	1.4	6.3
Jun 3, 2020	78.7	7.51	255	56.8	42.2	0.0053	3.7	1.484	0.017	ND	3.6
Jun 8, 2020	98.9	7.26	314	60.4	87.5	ND	0.11	1.771	0.012	2.7	305.1
Jun 9, 2020	88.4	7.08	361	49	42.2	0.001	1.4	1.887	0.012	5.6	5.1
Jun 11, 2020	83.5	6.95	323	47.2	34.2	0.0055	1.9	0.122	1.5	15.6	6.4
Jun 13, 2020	95.1	5.6	311	63	42.6	0.0022	74.2	0.064	0.025	8.3	6.2
Jun 16, 2020	235.2	7.75	547	53.2	22	ND	0.48	0.077	0.032	65	5.9
Jun 19, 2020	195.3	7.61	465	43.4	36.4	0.0014	0.32	2.543	0.017	14.7	6.1
Jun 22, 2020	311.1	7.73	700	48.4	34.2	0.0043	0.2	2.077	0.434	9.3	5