

**A TIME SERIES MODEL FOR FORECASTING LAKE EXPANSION: CASE STUDY  
OF LAKE BARINGO**

**BENEDICT ODHIAMBO**

**21/08516**

**A RESEARCH DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF  
THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE DEGREE  
IN DATA ANALYTICS IN THE SCHOOL OF TECHNOLOGY AT KCA  
UNIVERSITY**

**2023**

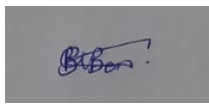
## DECLARATION

I declare that this dissertation is my original work and has not been previously published or submitted elsewhere for the award of a degree. I also declare that this contains no material written or published by other people except where due reference is made and the author duly acknowledged.

**Student Name: Benedict Odhiambo**

**Reg No 21/08516**

**Sign:**



**Date:**

**26/10/2023**

I do hereby confirm I have examined the master's Dissertation of Benedict Odhiambo and have approved it for examination.



Recoverable Signature

X



Date:

Dr. Lucy Waruguru

mburul@kcau.ac.ke

Signed by: Dr. Lucy Waruguru

**Dr. Lucy Waruguru**

## ABSTRACT

A flood is a natural disaster that refers to the temporal overflow of water on top of land that was previously not inhabited by water. It can be caused by too much precipitation or even outbursts of water reservoirs due to other reasons. Severe Floods have occurred in the Rift Valley lakes since 2011 due to lake expansion. Floods in the Lake Baringo area have occurred due to overflows of the lake and is a dangerous disaster leading to many pros rather than cons. It is due to the major problems experienced that the need for the use of Machine Learning, GIS, and Remote Sensing arose to help in monitoring, and creation of a forecast model to help create awareness of the area that is likely to be affected by floods in the future years. The research was guided by three objectives: Determining factors leading to the expansion of Lake Baringo, mapping spatial-temporal change in the Lake Baringo region to help compare the changes, comparing the time-series algorithms (LSTM and GRU) efficiency in the training of the dataset and lastly developing a time-series model for forecasting the area growth of Lake Baringo. Earlier researchers had used GIS and RS for the monitoring of similar cases but the element of prediction was not well looked into. Machine Learning methods have also been used to create prediction models but in the case of lake area expansion limited researchers had explored, hence the identified gaps arose. The research design used was longitudinal and it comprised two sets of data mainly satellite images and previously recorded data. Images were used for classification to map the changes over time and to visualize the lake's growth, the other form of dataset was used for analysis and creation of the model. GRU outperformed the LSTM algorithm as per metrics, it was found that Lake Baringo had expanded by 50% from the year 2011 mainly due to increased rainfall and reduced evaporation increasing the rate of sedimentation which led to the rising of the lake level. The study was limited by the available data and time used in the image analysis. The objectives were achieved and, in the future, better models could be developed for numerous lakes in Kenya and not only Lake Baringo.

**Keywords:** GIS, RS, LSTM, GRU, MSE and RMSE

## **ABBREVIATIONS AND ACRONYMS**

LSTM-Long Short-Term Memory

GRU-Gated Recurrent Unit

MLP- Multi-Layer Perceptron

CNN-Convolution Neural Network

LULC-Land Use Land Cover (Features covering the land and the purpose for which human beings are using the land.

RS-Remote Sensing-Gathering information from the earth without coming into physical contact with it (satellite images)

GIS-Geographic Information Systems

ANN-Artificial Neural Network

MAE-Mean Absolute Error

RMSE- Root Mean Squared Error

MSE- Mean Squared Error

OLI- Operational Land Imager

SVM-Support Vector Machine

GEE- Google Earth Engine

CART-Classification and Regression Trees

RNN-Relational Neural Network

MLA-Machine Learning Algorithms

## TABLE OF CONTENTS

DECLARATION .....	ii
ABSTRACT.....	iii
ABBREVIATIONS AND ACRONYMS .....	iv
CHAPTER ONE: INTRODUCTION.....	1
1.1 Background of the study .....	1
1.2 Problem Statement .....	2
1.3 Objectives .....	3
Main Objective.....	3
Specific Objectives .....	3
1.4 Research Questions .....	3
1.5 Motivation for Study.....	4
1.6 Justification of the Study .....	5
1.7 Study Area .....	6
.....	6
1.8 Scope of the Study .....	7
1.9 Significance of the Study .....	7
CHAPTER TWO: LITERATURE REVIEW .....	9
2.1 Introduction.....	9
2.2 Theoretical Review .....	10
2.2.1 Climate Change.....	11

2.2.3 Deforestation.....	13
2.2.4 Topography and Land Use Practices .....	15
2.3 Empirical Review.....	18
2.3.1 Remote Sensing and GIS Methods .....	18
2.3.2 Machine Learning Methods .....	33
2.3.3 LSTM and GRU.....	36
2.4 Knowledge Gaps.....	41
2.5 Conceptual Framework.....	44
2.5.1 Characteristics of Intervening Variables:.....	44
2.6 Operationalization of Variables .....	45
2.7 Summary .....	46
<b>CHAPTER THREE: RESEARCH METHODOLOGY .....</b>	<b>47</b>
3.1 Introduction.....	47
3.2 Current Methodological Approaches .....	47
3.3 Research Design.....	47
3.3.1 Dataset.....	52
3.3.2 Image Preprocessing and Data Preparation .....	54
3.3.3 Image Classification.....	55
3.3.4 Data Analysis .....	57
3.3.5 Modelling.....	57
<b>CHAPTER FOUR: DATA ANALYSIS FINDINGS AND DISCUSSIONS .....</b>	<b>60</b>

4.1 Introduction.....	60
4.2 Descriptive Statistics.....	60
4.3 Research Findings.....	61
4.3.1 Objective one Findings .....	61
4.3.2 Objective two Findings.....	67
4.3.3 Objective three Findings .....	74
4.4 Discussion of the Research Findings .....	76
CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS .....	81
5.1 Introduction.....	81
5.2 Conclusions.....	81
5.2.1 Objective One .....	81
5.2.2. Objective Two.....	82
5.2.3 Objective Three.....	82
5.3 Contributions of the study.....	83
5.4 Limitations of the Study.....	84
5.5 Recommendations for future research .....	85
REFERENCES .....	86
APPENDICES .....	92
Appendix 1: Budget and Resources .....	92
Appendix 2: Schedule .....	93

## LIST OF FIGURES

FIGURE 1: The Study Area Map .....	6
FIGURE 2: Image Classification Methods .....	29
FIGURE 3: LSTM Algorithm Structure .....	36
FIGURE 4: GRU Algorithm Structure .....	39
FIGURE 5: Conceptual Framework .....	44
FIGURE 6: Methodology Workflow .....	51
FIGURE 7: Knowledge Discovery in Databases (DBD, 2020).....	53
FIGURE 8: Summary of the Dataset Used .....	61
FIGURE 9: Spatial-temporal changes in LULC in the Lake Baringo and Neighbouring Area .....	63
FIGURE 10: LULC Change Summary .....	64
FIGURE 11: Visualization of Lake Baringo area change (2002 to 2023).....	65
FIGURE 12: Characteristics of lake Baringo Surface area, Rainfall, Lake Level and Temperature .....	66
FIGURE 13: LSTM Model Summary .....	68
FIGURE 14: GRU Model Summary.....	68
FIGURE 15: Baseline dataset against train and test Predicted Data by LSTM Model .....	70
FIGURE 16: Baseline Dataset Against Train and Test Predicted Data by GRU Model.....	71
FIGURE 17: Training and Validation Loss Curves for LSTM Model.....	72
FIGURE 18: Training and Validation Loss Curves for GRU Model .....	72
FIGURE 19: Time taken by LSTM to Predict Train and Test Data .....	75



FIGURE 20: Accuracy Measure of the LSTM Model.....	75
FIGURE 21: Time Taken by GRU to Predict Train and Test Data.....	75
FIGURE 22: Accuracy Measures of GRU Model .....	75
FIGURE 23: Actual Surface Area Vs GRU Model Predicted Surface Area.....	76
FIGURE 24: Schedule of Work.....	93

## **LIST OF TABLES**

TABLE 1: Landsat Satellite Images Characteristics .....	21
TABLE 2: Summary of Landsat Satellite Images Properties .....	22
TABLE 3: Summary of the Knowledge Gaps .....	42
TABLE 4: Operationalization of Variables .....	45
TABLE 5: Estimated budget of the Study .....	92

## CHAPTER ONE: INTRODUCTION

### 1.1 Background of the study

Lake Baringo in the Rift valley is one of the freshwater lakes in the region and serves many pastoralists with fresh water for their herds of cattle and for their day-to-day activities, it also serves as a source of food ranging from fishing products to a source of income for the fishermen. The locals also use the lake for Navigation to other parts bordering the lake. Due to climatic changes; since September 2010 the area and water level of Lake Baringo and seven other Rift valley lakes have been rising to unimaginable levels, this is according to a research by (Daniel Muia, 2021) , this has led to the displacement of people within the area and the submergence of infrastructure such as; roads, social amenities, grazing land, Farming land, Fishing land, and processing facilities due to flooding and even spread of diseases and fatalities.

The flooding in the region is a threat to human settlement and if not looked into it could result in adverse effects, it is better for the relevant authorities within the flood-prone areas within the lake region to understand their area and prepare mitigation measures or even relocation in the future when the flood may be predicted to occur. This can be done by analyzing the trend of various characteristics of the lake and even creating a prediction model that will help predict the time and area which is likely to be affected by the flood.

There are numerous machine learning time series algorithms ranging from Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Prophet, Exponential Smoothing State Space Model (ETS), Seasonal-Trend decomposition using LOESS(STL), Autoregressive Integrated Moving Average (ARIMA), XGBoost, LightGBM and CatBoost. Time series algorithms have been used in the past to predict instances of rainfall by Manoj (2020). The researcher compared Long Short-Term Memory (LSTM), Linear Regression, Gated Recurrent Unit (GRU), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and

Bidirectional Long Short-Term Memory (BiLSTM) based on the recorded parameters by the weather station(automatic) in the Bhutan region and found out that BiLSTM-GRU outperformed the other models. Predictive models are essential and important since they give us a rough idea of what the future might look like and hence preparations can be made early.

The purpose of this research was to monitor the trend and changes in the Lake Baringo characteristics over the years, compare different time-series algorithms and to use the most accurate to create a forecasting model that could be used to predict the Surface area size of the lake which would be affected by flooding due to the expansion in the future.

The scope of the study was not limited to the lake only but also to the surrounding of the lake to see how the expansion has affected the surrounding area. The study was guided by theories such as climate change, deforestation, Sedimentation and Topographical and Land Use practices. The study being time series the research design used was longitudinal research.

## **1.2 Problem Statement**

Lake Baringo serves many people in different ways such as; settlement, Fishing, farming, Education, and pastoralism (Omweni et al., 2021). Recently there has been negative effects on the day-to-day activities due to the swelling of the lake that has led to the displacement of many people and disruption of the activities in the area without prior warning. In research by (Daniel et al., 2021) floods have also led to human and animal conflict in cases where the water that bursts from the reservoir free or expands to the territory of wild animals such as crocodiles or even hippopotamuses. The disturbance has led to even loss of livelihood. At other times students have been left to travel by boat to their schools and others even drop out of school due to the struggle of reaching the school in search for education. (Nyakundi et al., 2023) Whenever floods occur there have been instances of the breakout of waterborne diseases such as; bilharzia, typhoid and Malaria or even injury to humans when they try to

swim or climb to high places to evade the hazards of Floods and even mental problems in other instances.

There are also knowledge gaps that could be addressed incase researchers would venture into the field of Data analytics and come up with ways to help reduce or mitigate the effects accrued whenever such cases occur. Some of the problems could be avoided if the relevant authorities could have prior planning for the occurrence of such events. This research can help in creating a forecasting model that could help in prior planning to reduce the adverse effects of flooding by determining the area where the lake growth is likely to affect the most.

### **1.3 Objectives**

#### **Main Objective**

The main objective of the study was;

- To propose a time-series model for forecasting the Lake Baringo area growth.

#### **Specific Objectives**

The specific objectives that guided the study were as follows;

1. To determine the factors leading to spatial-temporal (2002, 2009, 2016, and 2023) change in the Lake Baringo region.
2. To develop a time-series model (GRU/LSTM) with a better efficiency for forecasting the area growth of Lake Baringo.
3. To test and evaluate the developed model.

### **1.4 Research Questions**

The Research questions that were to be answered by the end of the study were as listed below;

1. What are the factors leading to spatial-temporal (2002, 2009, 2016, and 2023) change in the Lake Baringo region?

2. How can we develop a time-series model with a better efficiency for forecasting the area growth of Lake Baringo?
3. How do we test and evaluate the developed time series model?

### **1.5 Motivation for Study**

With the rising harsh economic times, there is a need to help people save on the cost they incur in relocating from one place to another whenever there is floods in the area where they call home. A lot of losses are also incurred when flooding occurs; the research could help minimize this due to proper planning and monitoring. There are students who have to strive to get an education in other schools that are affected by the floods and this could be minimized by enrolling in schools that are not easily affected by the expansion of the lake or even relocating and construction of schools in the area which is farther from the lake.

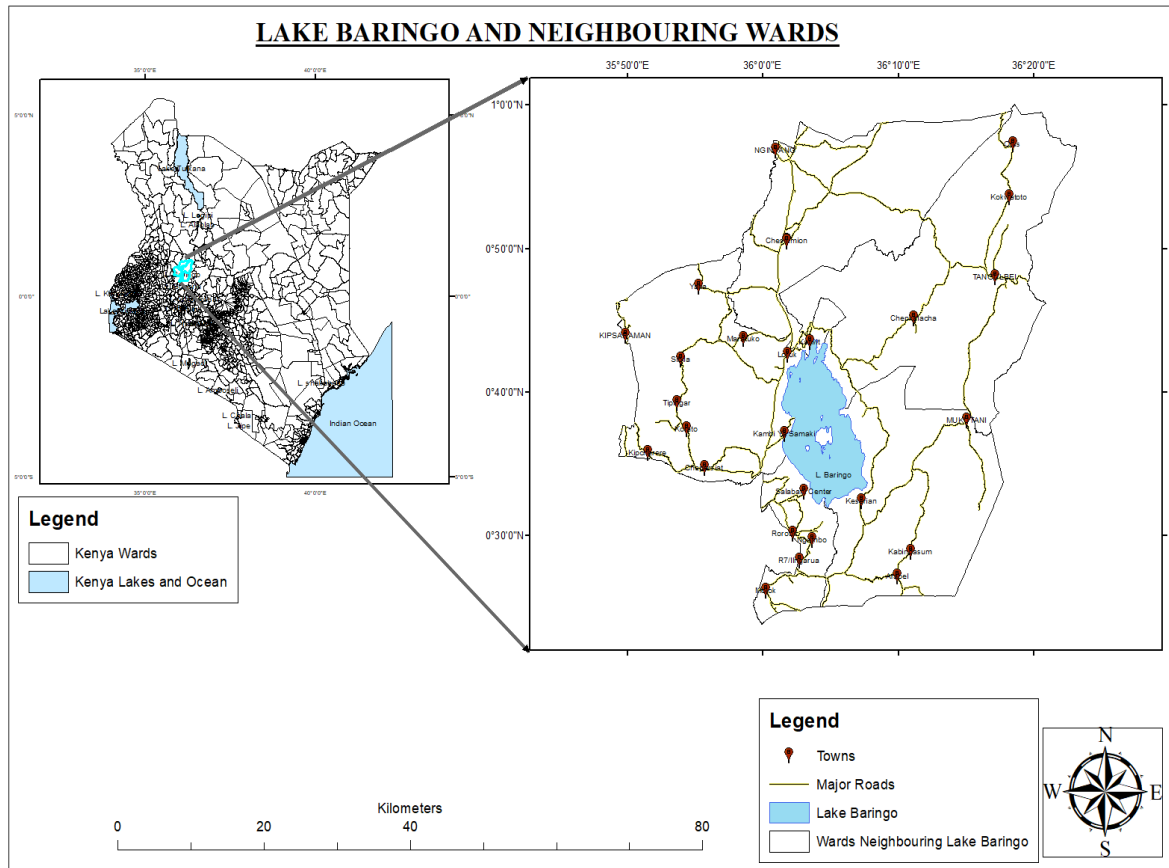
(Daniel, 2022) Reported that droughts and floods could cost the world economy \$5.6 trillion by the year 2050. The economic losses that are incurred due to flooding which include; Disruption of Businesses and destruction of infrastructure which help the traders to move from one place to another. These losses lead to loss of revenue to the country since taxes collected from such sales and movement tend to be less or no more. The cost of renovation due to the destructions also tend to be high when flooding recesses due to the massive destruction. These are also reasoning why there is need to pursue the research topic. There is also a quest to grow my knowledge while conducting this research and the curiosity to know how the two time-series algorithms behave on training of the dataset and to what extent will Lake Baringo grow in the future when the flood occurs.

## 1.6 Justification of the Study

Lake Baringo has been greatly expanding causing floods in the neighborhood area and there is need for a counter measure that can help create awareness of the probable areas to be affected in the future through a model that could show the increase. LSTM and GRU algorithms were chosen mainly because they are unique than other time series algorithms. Both the algorithms address the vanishing gradient problem in the traditional RNNs due to long-term dependencies in time series data, they also have capability of processing data sequentially that is learning and remembering the sequential patterns, they also have gating mechanisms that control the flow of information, making them adaptable to different time scales and patterns, they also have capability to learn different features and patterns directly from raw data which is particularly useful when the underlying patterns are complex or not well understood. Another justification for choosing LSTM and GRU is because the data used has irregular time interval and the algorithms can accommodate real-world scenarios where data points are not uniformly spaced. The two algorithms have also shown state-of-the-art performance in various time series prediction tasks, outperforming many traditional algorithms in research. Based on previous studies a lot of literature support the effectiveness of LSTMs and GRUs in diverse time series applications.

## 1.7 Study Area

**FIGURE 1: The Study Area Map**



According to report by (Baringo County | County Trak Kenya) Baringo County is situated in the Great Rift valley Region, which borders Turkana County and Samburu County to the north, Laikipia County to the East, Nakuru County to the south and Kericho County, Nandi County, Uasin Gishu County, Elgeiyo Marakwet County and West Pokot County to the west. The county covers a total area of 10976.4 Km<sup>2</sup>. According to (Adminusr, 2019) Kenyan population census the county has a population of 666,763 with a population density of 61 people per Km<sup>2</sup> and a yearly growth rate of 2.6%.the major economic activities of residents in the county include; Pastoralism, Agriculture, Fishing and Sand harvesting. Baringo county has six constituencies and 30 electoral wards.



The study focus was on Lake Baringo and the neighboring wards namely; Loiyamorok Ward, Tangelbei/Korossi Ward, Saimo/Soi Ward, Mukutani Ward and Lichamus Ward. The choice of the wards was based on the surrounding of the lake to see the behavior of the LULC.

### **1.8 Scope of the Study**

The focus of the study was on Lake Baringo and the neighboring area. Landsat Satellite images for the years (2002, 2009, 2016, and 2023) were subjected to classification analysis to produce the Land use Land Cover maps and their accuracies computed. The trend of the different classes of the region were then compared to identify the behavior. The lake extent from the classified satellite images were overlaid and the change per year well defined. Using the dataset, the behavior of; rainfall, lake depth, Temperature and lake area were analyzed and studied. Using the available dataset, the data was split to training and testing data and training performed the different algorithms: LSTM, and GRU. The accuracy of the algorithms was compared to get the one that performs the best.

From the performance of the best algorithm, a forecasting model was then created to help predict the total area of the Lake Baringo in future dates to help determine the extent of flooding.

### **1.9 Significance of the Study**

The research findings have the potential of providing useful insights to the Ministry of Water and Irrigation in Kenya, stakeholders in the water sector in Kenya, non-governmental institutions, domain experts, and the personnel in charge of policy formulation. Such insight includes knowing which measures to put in place by the government and stakeholders to curb the adverse effects of expansion of Lake Baringo. Having knowledge of the past state of the lake and the flooding extent per year could help to plan for the preparation measures and in the

stabilization of the economy of the area whenever floods occur. Stakeholders should also be able to put more emphasis on the importance of prediction studies of such disasters.

Data mining technology has not been widely explored in Kenya in the field of disaster management and therefore the findings of this study also have the potential of competitive advantage since businesses and nations can gain a competitive edge by implementing the latest IT practices derived from research findings, the study also has capability to provide evidence-based insights, aiding policymakers in creating effective regulations governing technology use and dissemination of data. Study findings establish best practices in IT, guiding professionals and organizations on importance of regular data collection and dissemination.

The data that is available yet not being fully utilized in institutions like the Ministry of Water and Irrigation will be put to better use in decision-making on matters of disaster. The findings would encourage more researchers to explore this area of knowledge and lead to new findings and innovations to make human life better. This study would also guide those who would like to research in areas similar to this as it would serve as a foundation or give an idea of how to further this research.

## CHAPTER TWO: LITERATURE REVIEW

### 2.1 Introduction

The Rift valley in Africa begins at the Red Sea in the Northern side to Mozambique in the Southern region. The East African Rift Valley on the other hand has eight lakes ranging from freshwater to saline water. Naivasha and Baringo are freshwater lakes in the region while Lake Turkana is semi-saline while Lakes Magadi, Elementaita, Nakuru, Bogoria, and Logpi are saline-alkaline. Lakes have importance, such as being home to water animals, and freshwater lakes support fishing and Agriculture (Daniel Muia, 2021).

Extreme floods caused by the expansion of the Rift Valley lakes is a serious concern that should be looked into. The rift valley lakes have been near empty in the past, but recently, the water has risen to higher levels. (Muita, 2021) This to some extent can be attributed to rainfall as reports indicate that lakes Nakuru, Bogoria, and Baringo have continued to rise to their highest levels in this decade. For instance, lake Baringo increased from 143.6km<sup>2</sup> in January 2010 to a high 219.8km<sup>2</sup> in December 2014 a whole increase of 53.1% in the area. (Daniel Muia, 2021)

Flooding is known to be a short disaster but fatal and destructive. Since 2012 lake Baringo has experienced swelling which has affected and destructed multiple facilities and day-to-day activities of the people within that area. Flooding has serious impacts on livestock, health, and livelihood. The floods have adverse effects such as Farmlands and grazing pastureland being submerged. Families and households' displacement. There is a need for early warning to be put in place to control such effects. (Lake Baringo Flood Resilience Project, 2023)

As regarded by (Moskolai et al., 2021) spatial-temporal analysis is the study of data collected across both space and time dimensions. It describes an event that took place in a particular place over time. With the development of powerful computational processors, Deep Learning techniques are widely applied today by data scientists for spatial-temporal data analysis, due to the development of powerful computational processors. The spatial-temporal prediction of snow cover, leaf area index, sea ice motion, vegetation, sea surface temperature, and others are among the specific applications. Most of the times indices are first extracted from images prior to their use. The spatial temporal prediction of snow cover, leaf area index, Arctic motion, vegetation, sea surface temperature and so forth are just some of the special applications. Most of the time, indices are extracted from photographs prior to their use. As a matter of fact, multichannel images are used to calculate the map indices. In addition, they point out a specific phenomenon that is changing the effect of an image and mitigate additional factors which change it. For example, healthy vegetation is represented by a light color in the normalized difference vegetation index (NDVI) image, while unhealthy vegetation has lower values.

This study aims at building a time-series forecasting model that would help to establish the extent of flooding of Lake Baringo in the coming years in Kenya. Chapter 2 of the study primarily focuses on the available literature in order to establish what areas in this domain have been covered thus far and how they relate to this study area. The chapter also revealed the gaps in knowledge that were picked from the review of literature, and how this study filled some of the gaps.

## **2.2 Theoretical Review**

The disaster of flooding in the neighborhood area of the Rift Valley lakes due to the expansion of the lakes recently has brought up discussions and researchers have been trying to

find out what could have caused the rising lake levels. Four main theories have been discussed to be the main cause of the disaster, including climate change, Sedimentation, Deforestation, and Topography and Land Use practices.

### **2.2.1 Climate Change**

(Tabari, 2020) explained the relation of climate change and flooding. The researcher explained that the hydrological cycle has high chance of intensifying with the severe changes of global warming, at its extreme would likely lead to an increase in the precipitation, increased precipitation would then lead to higher rainfall in other areas where there are rivers and the rivers would in turn feed the lakes with a lot of water hence the expansion, intensity and the risk of flooding. The changes are not often different from the theory that expects an increase in the water-holding capacity of the reservoir in hot conditions (when the sun is extreme the ice on the mountains and other areas would melt leading to an increase of water flow to the lake. The increase of water flow to the lakes leads to the increase in its surface area). As the sun is prolonged for a duration that is longer than the normal duration, the precipitation will also be prolonged and will be of high quantity than the usual amount. When the rain then starts it will tend to be massive and for a long period of time. This will then tend to cause the swelling of the lake reservoir if the rate of precipitation is higher than the rate of evaporation.

Another research by (Yang et al., 2022) investigated the lake changes in different geomorphological zones of Central East Asia in the past 50 years, they effectively assessed the drivers of lake change in the context of climate change, and forecasted the trend of lake change under the climate background in the future. It also gave us a good basis for understanding the climatic and environmental responses of arid regions to global warming. This provided a good basis to understand the response of the climate and ecological environment in arid areas to global warming.

The lake area in Central Asia has grown by 41% over the past five decades from 17,442 km<sup>2</sup> to 24,541 km<sup>2</sup>. There have been large differences in the effects of lake change and their influence factors across Central Asian geomorphological zones. In the different geomorphological zones of Central East Asia, there are significant differences between lake changes and their influence factors. In the Tibetan Plateau and high mountain-basin zones, the area of lakes expanded owing to increased precipitation, and increased glacier, permafrost, and snow melt water caused by warming (which contributed to 66%), which led to the water content of mountains to increase. The main driver for changes in the low mountain zone lakes was changes in precipitation (which contributed to 87%).

Increased global greenhouse gas intensity has led to an increase in surface ocean temperatures, resulting in increased evapotranspiration and higher atmospheric water content. This allows more water vapor to pass through the atmosphere into arid regions, leading to an increase of cloud water resources and rainfall which leads to a rise in mountain water supplies as well as improved regional water cycles. As a result, the atmospheric circulation has enabled more water vapors to penetrate arid regions leading to increased cloud water resources and precipitation which in turn increases mountain water reserves and enhanced regional water cycles. These are the drivers that triggered lake expansion in Central East Asia.

Future climate projections based on the RCP4.5 and RCP8.5 scenarios, together with historical geological evidence, suggest that continued increases in temperature of 4.0 to 7.8 C and increased precipitation of 1.07 to 1.29 mm per day could lead to further expansion of lakes in Central Asia. Nevertheless, the increase in temperatures and precipitation is likely to lead to a larger frequency of severe weather events which could seriously affect the environmental stability of desert areas; these issues have yet to be adequately addressed.

### **2.2.2 Sedimentation**

According to (Nausheen, et al., 2021) the process of sedimentation starts when there is movement of runoff generated either by rainfall, snowmelt or wind. Upon sediments being deposited in the reservoir, the reservoir acts as a storage area since the speed is very slow and almost constant. Increased depth of flow and a decrease in the velocity flow causes a reduced capacity of the reservoir's transportation of sediment which then leads to the settling of the bulk of incoming sediments. Sedimentation can also be caused by human beings when they practice deforestation and Urbanization. When humans cut down trees, they leave the land without wind breakers and vegetation cover that slow speed of wind or even speed of water and hold soil particles together, this then leads to the soil and rock particles being swept by water and wind then deposited to the lake reservoir leading to decrease of lake depth and increase in the lake surface area

Sedimentation causes effects such as clogging of the gills in fish, decrease in the reservoir depth, and turbidity of water which blocks light from penetrating to the bottom of the water reservoir. When there are high temperatures for a long period of time the rocks and soil get weakened and disintegrated and when it rains the disintegrated particles are carried away by rainfall and deposited in the reservoirs such as lakes among others.

### **2.2.3 Deforestation**

This is the cutting down of trees without replacing them. Research by (Butler, 2019) found that Forests play a big role by reducing the speed of the wind, reducing the speed of surface runoff, and absorbing water in an area. These are crucial in mitigating floods. In an open area where there are no trees or the tree numbers are limited there tends to be room for erosion since the wind speed and water runoff will not be reduced and soil particles will be loosely held together. To be able to control floods and erosion people should plant more trees.

In research by (Lawrence et al., 2022) they discussed that forests are important and ignoring biophysical influences on local climate means ignoring local self-interest, a potent motivator to achieve global climate goals and enhance forest conservation. Because physical impacts in one location might cancel out effects in another, the biogeochemical influence of forests tends to exceed the biophysical effect at the global scale. However, at the local level, biophysical consequences can be very significant and huge. Although the importance of forests in preserving vital habitat for biodiversity is widely acknowledged, recent studies on extinction also demonstrate the importance of forests in preserving vital climates for biodiversity. Extinction is caused by variations in maximum temperature rather than average temperature.

In the tropics, deforestation is linked to year-round increases in the maximum daily temperature and in the upper latitudes during the summer. Of course, in tropical, mid-latitude, and boreal forests, deforestation also raises average daytime temperatures. Even at the mid-and high-latitudes, the biophysical effects of forests help moderate local and regional temperature extremes, making exceptionally hot days far more common after deforestation. About one-third of the current increase in the intensity of the warmest days of the year at a specific region can be explained by historical deforestation. Moreover, it has doubled or quadrupled the frequency and severity of hot, dry summers. Extreme temperature rises brought on by lost forests locally are similar in size to those brought on by 0.5°C of global warming. Anywhere on Earth, during the hottest months of the year, forests provide localized cooling, enhancing the resiliency of rural areas, urban areas, and conservation zones. In order to adapt to a warming planet, forests are essential.

Additionally, forests reduce the risk of drought brought on by extremely high temperatures. Trees dissipate heat and transfer moisture to the atmosphere under drought circumstances because of their deep roots, high water use efficiency, and rough surface. Apart



from this direct cooling, ET from forests can also affect cloud formation, improving albedo and possibly encouraging precipitation. As temperatures rise, woods produce more BVOCs and organic aerosols, which can increase the effects of albedo directly or indirectly (via cloud formation). It has been noted that at the mid-latitudes, this negative feedback on temperature offsets abnormal heat episodes.

The forest absorbs and soaks rainfall brought about by a heavy downpour while holding soils together and releasing water at slow intervals. This regulating feature of vegetation particularly forest can help reduce destructive flood cycles that can occur when forests fall. When forest cover is reduced or done away with completely, runoff rapidly flows into rivers and streams, it elevates river levels and submerging villages, cities, and agricultural lands to flooding, during the rainy season.

#### **2.2.4 Topography and Land Use Practices**

Land Use is how human beings utilize the land. There are activities that human beings tend to do to their advantage, but these actions can lead to problems for the environment and bring about disasters such as floods, and drought, among others. As (Avashia & Garg, 2020) found that human actions such as deforestation for urbanization purposes and the development of infrastructure on land that was initially used for other purposes. In the future when the lake would expand it would then find other inhibitors on its land and hence leading to the distraction of the developments in the area.

Topography, on the other hand, is the shape of the earth: the hills and valleys. (Hamid et al., 2020) explained in their research that when it rains the water collects as surface runoff and then flows gradually from the highest to the lowest points. The lowest point is mainly a reservoir such as a lake and all the deposits carried out by the water are deposited there, this tends to increase the reservoir content as all the sediments and water are deposited there.

Numerous recent studies have addressed the issue of climate change and the mountain biomes, and they have examined potential adaptation measures that protected area managers could implement. Regretfully, practical implementation of scientific suggestions for topographic environment management is frequently hampered by their theoretical nature or imprecise definition. Many times, even well-organized proposals have scant scientific support. Consequently, these suggestions may be hard to implement, exclusive to a given setting, or predicated on a small number of case studies. However, environments that are heterogeneous enable for species to adapt to varying climates. Because of the diversity of climates created by heterogeneity in elevation, aspect, and slope, species are able to make minor spatial adaptations to follow climatic conditions that suit them.

In other words, avalanches and other natural disturbances like them are particularly common in mountainous landscapes. These disturbances have the power to eliminate inertia from a system—such as long-lived, non-reproductive individuals—and promote the accelerated establishment of new species and structures, which helps species populations adapt to changing climatic conditions more quickly. The variety of environmental variables can reveal how resilient and adaptable mountain biomes are to changes in their surroundings. High topographic heterogeneity can cause temperature or precipitation to fluctuate, changing the climate. Furthermore, mountain biome regions demonstrate how great topographic heterogeneity can result in a large degree of environmental diversity because of variations in temperature, precipitation, sun exposure, and wind.

When the research area's station average temperatures are broken down by month, the Hamidiye station recorded the highest average temperature of 11.25 °C. Sutluk and the neighboring areas recorded the lowest monthly average temperature values (10.1 °C). The winter temperature tolerances in the centers around the north and west were found to have

significantly decreased. The region's topography, dryness, and continentality are all linked to these declines. The temperature tolerances from S to N decrease when the monthly average temperature is taken into account. The temperature significantly drops as a result of the elevation gain because it approaches the north.

### **2.2.5 Factors Leading to the Expansion of Lake Baringo**

Lake Baringo has been expanding due to floods. From the theories discussed above this can be attributed to; Climate change which is a prolonged long season without rainfall characterized by high temperatures which weaken the rock particles and hence disintegration, when the rainy season arrives the disintegrated particles are then carried away by the water, and deposited in the lake since the lake is the lowest point, the deposits are deposited on the bottom of the lake while water is above the deposits, this leads to the rising of the water levels due to the displacement of the area which was initially covered by water.

(Mulama & Ondieki, 2023) In order to show the change in the area of Lake Nakuru, the results were able to provide a smooth line graph using different classification techniques. Spatial distribution maps of the lake showed different images of the lake in each of the classification used, though the graph of the area of the shapefile was assumed to be more accurate because it's whereby the lake is calculated before any classification is performed on the image and is used to correlate the area of the lake with the other methods. Although no one knows the cause of the lake's change over time. In order to observe this, change a line graph was used and it showed that the area of Lake Nakuru had increased between the years 2010 and 2018.

Like Lake Baringo, Lake Nakuru experienced similar change according to (Mulama & Ondieki, 2023). The change in agricultural use practices has been significant, with the built-up area increasing by almost 400% during the last 30 years and these changes have had a major

impact on soil loss as vegetation and forest were destroyed resulting in increased susceptibility to erosion. The results of this classification have also shown that from 1990 to 2018, the area of Lake Nakuru increased by 50%. They modelled factors despite assumptions being made on the soil erodibility factor and Lake Surface (LS) factor the results obtained.

The research recommended that, in order to produce the best results from Lake Nakuru area, both satellite image and field survey data should be used. In this case study, the use of Landsat satellite imagery on the Earth Explorer website proved to be useful and successful, and proved to be appropriate for the mapping and monitoring of the lake area. Correlation graphs showed that the best way of classifying an image would be to use a maximum likelihood, as this yielded a coefficient of 0.93. Since the lake showed an expansion from the results of the line graphs. It's prudent to monitor it which will help predict the future change of Lake Nakuru.

## **2.3 Empirical Review**

Based on the theoretical review there have been numerous studies done to show and monitor how climate change, sedimentation, deforestation and topography, and land use have been affecting change thus leading to Lake Baringo's expansion. There have been different approaches to the research with some researchers focusing on Remote sensing and GIS methods to monitor the trend of the expansion of the lake. Machine learning has also been used to try and predict the extent of lake flooding.

### **2.3.1 Remote Sensing and GIS Methods**

(Wang et al., 2020) They conducted their research Using Remote Sensing and GIS to predict the changes in Land use and Land Cover in the area for the year 2030. The authors classified two Landsat satellite images for the Kathmandu district of Nepal in the years 1990 and 2010 into seven classes; Forests, Shrub land, Grass land, Agricultural, Barren area,

water body and Built-up area. The research was done in three phases: Data Acquisition and Preparation, LULC classification, and the prediction of LULC classification using the Cellular Automata-Markov model. The CA-Markov model is a combination of the cellular automata (Change of pixel value based on a neighboring pixel) model and the Markov (Future states depend on the present state) model. The model predicted transitions (two way) among available land use types better than regression models. The inputs which were fed into the model were: the Digital Elevation Model (DEM), transition probability matrix, Road Infrastructure, and the LULC data of 1990 and 2010. Kappa statistics was used to enhance the accuracy of the output.

(Muhammad et al., 2022) conducted research on Cellular Automata-Artificial Neural Networks (CA-ANN) to predict the changes in the LULC for the years 2030, 2040, and 2050 in Linyi in China. The researchers conducted their research in three bits; Generation of LULC maps of the ten-year interval through classification, running the independent variables to the CA-ANN model to predict the LULC for the year 2020 for validation purposes, and the predicted LULC for the years 2030, 2040, and 2050. Classification of the satellite images was done for the years;1990,2000,2010 and 2020 and from the classification change detection was determined. The CA-ANN model was then run with the independent variables being DEM, slope, and distance from the road to predict the LULC for the year 2020 which was used for validation against the created LULC map for 2020. The model was then used to predict the LULC for the years 2030, 2040, and 2050.

In this research (Soltani et al., 2020), They demonstrated the implementation of the new strategy for the case study of Lake Gregory in Australia by using satellite data and a stochastic approach to forecast changes in lake surface area (LS) regions. High resolution satellite photos from Landsat 5, 7, and 8 were taken on a monthly time frame on clear days. To create the lake surface maps, ENVI 5.3 software analysis was used, along with the normalized difference

vegetation index (NDVI) and modified normalized difference water index (MNDWI) indices. A decision tree was used to separate satellite pictures into water and non-water categories. The monthly area of the Lake was calculated using ArcGIS 10.3 software. According to the general trend data, the LS decreased gradually between 2004 and 2019.

The relationship between these variations and regional precipitation was ascertained using monthly temperature (T) and precipitation (P) measurements from the TRMM satellite. In order to predict lake surface (LS) variations, they created a novel generalized group method of data handling (GGMDH), in which the LS time-series database was taken from the satellite imagery. Precipitation and three distinct scenarios were defined for downscaling, based on forecasts of climate change, in order to forecast the LS for the 2020–2060 timeframe. When compared to stochastic models integrated with preprocessing scenarios, the GGMDH performs better in long-term logistic regression forecasting than the stochastic model. The outcome demonstrated that, for the forecasting stage, GGMDH is the most effective model when it comes to simulating lake surface, with an  $R^2$  (%) = 94.16 and an RMSE = 8.77 for the phase of predicting. The projected surface area of Lake Gregory varied over time, ranging from 226 by a limit of 0.008 km<sup>2</sup>.

### **2.3.1.1 Landsat satellite Images**

The program started in 1967 and initially, it was referred to as Earth Resource Technology Satellite Program. The satellite was launched in series that is 1, 2, 3, 4, 5,6,7,8. The satellites were unmanned and were designed to acquire data on earth resources in a systematic, repetitive moderate resolution on a multispectral basis.

**TABLE 1: Landsat Satellite Images Characteristics**

<b>Characteristic</b>	<b>Satellites</b>	<b>Property</b>
Altitude	Landsat (1-3)	900km
	Landsat (4, 5, 7, 8)	705km
Orbit	All	Sun synchronous
Revisit period	Landsat (1-3)	18 days
	Landsat (4, 5, 7, 8)	16 days
Period/Revolution	Landsat (1-3)	103 mins
	Landsat (4, 5, 7, 8)	99 mins
Inclination	Landsat (1-3)	9 <sup>0</sup> to the normal
	Landsat (4, 5, 7, 8)	8.2 <sup>0</sup> to the normal

Sun synchronous; designed that they pass equator every morning at the same time but different places; Landsat 1-3(9:42am) and Landsat 4,5,7,8 (9:45am). Revisit period; after some period, the satellite will go back to the same position. Period/Revolution; time satellite takes to go round the earth to produce an orbit. Inclination; the way the satellite is inclined to the normal. All satellites 1, 2,3,4,3,5,7,8 have a swath (area covered by a satellite when it moves along a straight line) of (185×185) km.

Landsat satellite images have numerous sensors based on the satellite type such as; Multispectral Scanner System (MSS), Return Beam Vidicon (RBV), Enhanced Thematic Mapper plus (ETM+), Thematic Mapper (TM), Enhanced Thematic Mapper (ETM), Operational Land Imager (OLI) and Thermal Infrared Sensors. Each scene of the satellite image covers 170km by 185km Below is a summary of satellite characteristics with the exception of panchromatic bands. (EOSDA LandViewer: Find And Download Satellite Imagery, 2023)

**TABLE 2: Summary of Landsat Satellite Images Properties**

<b>Satellite</b>	<b>Launch</b>	<b>Decommission</b>	<b>Sensors</b>	<b>Resolution(m)</b>	<b>Return Period(days)</b>
Landsat 1	23 <sup>rd</sup> July, 1972	January 6, 1978	MSS/RBV	79/80	18
Landsat 2	22 <sup>nd</sup> January, 1975	July 27, 1983	MSS/RBV	79/80	18
Landsat 3	5 <sup>th</sup> March, 1978	September 7, 1983	MSS/RBV	79/80	18
Landsat 4	16 <sup>th</sup> July, 1982	June 15, 2001	MSS/TM	82/30	18
Landsat 5	1 <sup>st</sup> March, 1984	2013	MSS/TM	82/30	16
Landsat 6	5 <sup>th</sup> October, 1993	Orbit not achieved	ETM	-	-
Landsat 7	15 <sup>th</sup> April, 1999	In use	ETM+	30	16
Landsat 8	11 <sup>th</sup> February, 2013	In use	OLI/TIRS	30/100	16

### **Limitations of Landsat Satellite Images**

Other than having their pros the Images have their disadvantages which must be looked into when analyzing them. The limitations are as listed below; Spatial Resolution: Landsat images have a moderate spatial resolution (30 meters for Landsat 5, 7, and 8). This means that smaller objects or fine details cannot be captured accurately. High-resolution applications, such as detailed urban planning or monitoring small-scale land changes, require data from sensors with much higher spatial resolution. Temporal Resolution: Landsat satellites have a revisit time



of 16 days. This means that a specific area on Earth is imaged approximately every two weeks. For applications requiring more frequent updates, such as monitoring rapidly changing environmental conditions, this temporal resolution might not be sufficient. Spectral Resolution: Landsat satellites capture data in several spectral bands (visible, infrared, thermal, among others). However, the number of spectral bands is limited compared to hyperspectral sensors. Hyperspectral sensors capture a much wider range of wavelengths, allowing for more detailed analysis of the Earth's surface characteristics. Atmospheric Interference: The Earth's atmosphere can scatter, absorb, and reflect light, which can distort satellite imagery. Although there are correction methods to mitigate atmospheric effects, they are not always perfect and can introduce errors, especially in areas with significant atmospheric pollution or variable atmospheric conditions.

Cloud Cover: Cloud cover can significantly affect the quality and usability of satellite images. Landsat data is often impacted by clouds, especially in regions with frequent cloud cover. Clouds block the view of the Earth's surface, making it challenging to obtain cloud-free images for analysis. Data Cost: While Landsat data is freely available to the public, accessing and processing large volumes of data for extensive studies can incur significant computational costs and require substantial data storage resources. Limited Nighttime Imaging: Landsat satellites are primarily designed for daytime imaging. While thermal bands can capture nighttime data, the spatial and spectral resolutions for nighttime imagery are limited compared to daytime observations. Limited Radiometric Resolution: Landsat data has a limited radiometric resolution (typically 8 or 16 bits per pixel), which means that subtle differences in reflectance or brightness might not be accurately represented in the images.

Despite these limitations, Landsat data remains one of the most widely used and valuable sources of Earth observation data due to its long historical record, global coverage,

and accessibility. Researchers often mitigate these limitations through careful data preprocessing, analysis, and integration with other data sources to enhance the accuracy and applicability of their studies.

In research by (Acharya et al., 2016), Water was identified using Landsat imagery. Water, according to the researchers, is a crucial component of every ecosystem. Water identification was crucial for several scientific approximations and for tackling social problems. Numerous techniques had been established, and fresh ideas were being investigated. Water body identification has long been facilitated by Landsat imagery. The quality of imagery sensed by Landsat 8 had improved because of the addition of new OLI sensors, which have high sensitivity linked to spectral resolution and enhanced signal-to-noise ratios because of the radiometric performance. But given the increased quality and volume of data, it was essential to investigate appropriate and workable techniques for identifying water that make the most of better photos while requiring the least amount of data input.

In the research, the Researchers used a Landsat 8 OLI image and solely its original OLI reflectance bands to apply and investigate the efficacy of a JDT in distinguishing between water and non-water bodies. To train (70%) and validate (30%) the model, stratified randomly-sampled pixels were labeled using expert judgments and a computerized topography map from the National Geographical Information Institute, Korea. The JDT model has an AUC value of 0.999, a kappa statistic of 0.9966, and accurately identified 99.83% of cases. Maps showing binary water and non-water were created using the model. Similarly, using a confusion matrix and relevant information, water and non-water maps based on five alternative approaches were created and cross-compared. Density slicing, NDWI, MNDWI, ML, SVM, and JDT have overall accuracy rates of 99.35%, 98.92%, 98.43%, 99.28%, 99.41%, and 99.15%.

In a similar vein, the kappa coefficients for density slicing, NDWI, MNDWI, ML, SVM, and JDT were, respectively, 0.9870, 0.9785, 0.9687, 0.9856, 0.9883, and 0.9830. With the exception of MNDWI, which incorrectly categorized a large number of non-water objects, nearly all approaches showed good accuracy based on these statistics and visual interpretation. Water in agricultural areas should be identified with great care, taking into account seasonal variations. In the same way that spatial resolution was crucial to the pixel-by-pixel categorization of multispectral images, it should be carefully taken into account when identifying smaller bodies of water. In the current study site, water bodies were recognized with high accuracy using one band (density slicing) or two bands (NDWI and MNDWI), however there were a lot of mistaken pixels. In contrast, ML and SVM also demonstrated high accuracy but used each band used as input for categorization. On the other hand, JDT had significantly fewer misclassified bodies and only used four OLI bands. Overall, the OLI imagery improvement shown good accuracy for different techniques, such as JDT, to identify water features. The deep blue band was shown to have the third-highest information gain in the decision tree, confirming the band's significance for identifying water.

In order to better evaluate the capabilities of different methods based on improved OLI imagery, future study addressing this water identification method may make use of data from other regions of Korea, such as from complex watersheds or flooded areas. In addition to more explanatory factors and sensor pictures, new techniques for identifying water could be utilized to evaluate and contrast the accuracy. That will enable for a more thorough comprehension of JDT classification. If training data are carefully chosen, analyses like this could also be helpful in other disciplines where binary classification challenges exist.

### **2.3.1.2 Land Use Land Cover Classification**

Earth's land mapping has historically been divided into two categories: land cover and land usage. Even though these two ideas are employed interchangeably in numerous studies and are sometimes misinterpreted, they are distinct when understood correctly. As per (Vali et al., 2020) "Land use is characterized by the arrangements, activities, and inputs by people to produce, change, or maintain a certain land cover type," whereas "Land cover is the observed (bio) physical cover on the Earth's surface." The definition states that land cover and land use are closely related, making a simultaneous classification of both nearly necessary. Consequently, the "land use and land cover" (LULC) classification is now regarded as a more inclusive term that encompasses this relationship in its whole in current research.

The grouping of different portions of an image can be done in two different methods those are Supervised Classification and Unsupervised classification (Mehmood et al., 2022). Supervised classification is whereby the number of groups to be generated is determined by the user or rather training data is usually labelled but in unsupervised classification, the number of groups to be generated is determined by the machine or rather the training data is usually unlabeled. Classification helps in creating different classes as the land is used on the ground, for example; Any water body such as river, lake, dam, well, sea or ocean(Water), Any piece of land that human beings practice farming of crops such as wheat and maize (Agriculture), Any piece of land that have plants that are not Agricultural land (Vegetation), and Any land that is not tilted and does not have plants or is not built in (Bare land ), among others. Classification mainly is achieved because the images have different reflectance and image pixels with the same reflectance as those of training dataset are classified to the same classes to come up with the different Land Use Land Cover zones. The accuracy of the classification can also be

computed to guide the level of accuracy of the classified image for purposes of confidence of the final results.

An essential technique for comprehending the spatial distribution of land features and the attributes that are linked with them is LULC categorization. For a variety of land management, natural resource management, and urban planning applications, accurate LULC classification is crucial. Land Use and Land Cover (LULC) classification involves categorizing the Earth's surface into different classes based on its physical and human characteristics. Determining appropriate classes for LULC classification is a crucial step in the process, as it defines the scope and granularity of the analysis. Here are the steps to determine classes for LULC classification: Define the Purpose of the Classification: Determine the specific goal of the LULC classification. Are you interested in urban planning, environmental monitoring, agricultural analysis, or something else? The purpose will guide your choice of classes.

Understand the Study Area: Familiarize yourself with the study area. Consider its natural features, human activities, and environmental concerns. Understanding the unique characteristics of the area will help you define relevant classes. Review Existing Classification Systems: Look for existing classification systems relevant to your study area. For example, the Corine Land Cover classification system in Europe or the National Land Cover Database (NLCD) in the United States. These systems provide predefined classes that might be suitable for your study. Consult Stakeholders and Experts: Engage with local communities, experts, and stakeholders. They often have valuable insights into the area's land use and cover. Their input can help identify specific classes that are culturally or economically significant.

Consider Temporal Aspects: Think about whether your classification will be a one-time snapshot or if it needs to capture changes over time. Classes might need to reflect seasonal variations or changes due to urbanization, deforestation, or agricultural practices. Balance

Specificity and Generalization: Strike a balance between detailed classes and broader categories. Overly specific classes might result in too many classes, making analysis complex, while overly generalized classes might not capture important distinctions. Utilize Remote Sensing Data: Analyze the spectral bands available in your remote sensing data. Different land cover types reflect and emit specific wavelengths. Utilize this information to identify distinct spectral signatures, which can inform your class definitions.

Consider Hierarchical Classification: Implement a hierarchical classification system if the area has a wide variety of land cover types. This approach involves general classes at the top level (e.g., forest, agriculture) and more specific classes at lower levels (for example deciduous forest, corn fields). Validate Classes: Validate the chosen classes using ground truth data or field surveys. Ground truthing involves verifying the classes by physically visiting sample locations and comparing the observed land cover with the classified data. Iterative Process: LULC classification is often an iterative process. You might need to refine classes based on the initial classification results and feedback from stakeholders or validation efforts.

It should be Remembered that the choice of classes should align with the study's objectives and the available data's resolution and spectral capabilities. Flexibility and a good understanding of the study area are key to determining appropriate and meaningful LULC classes.

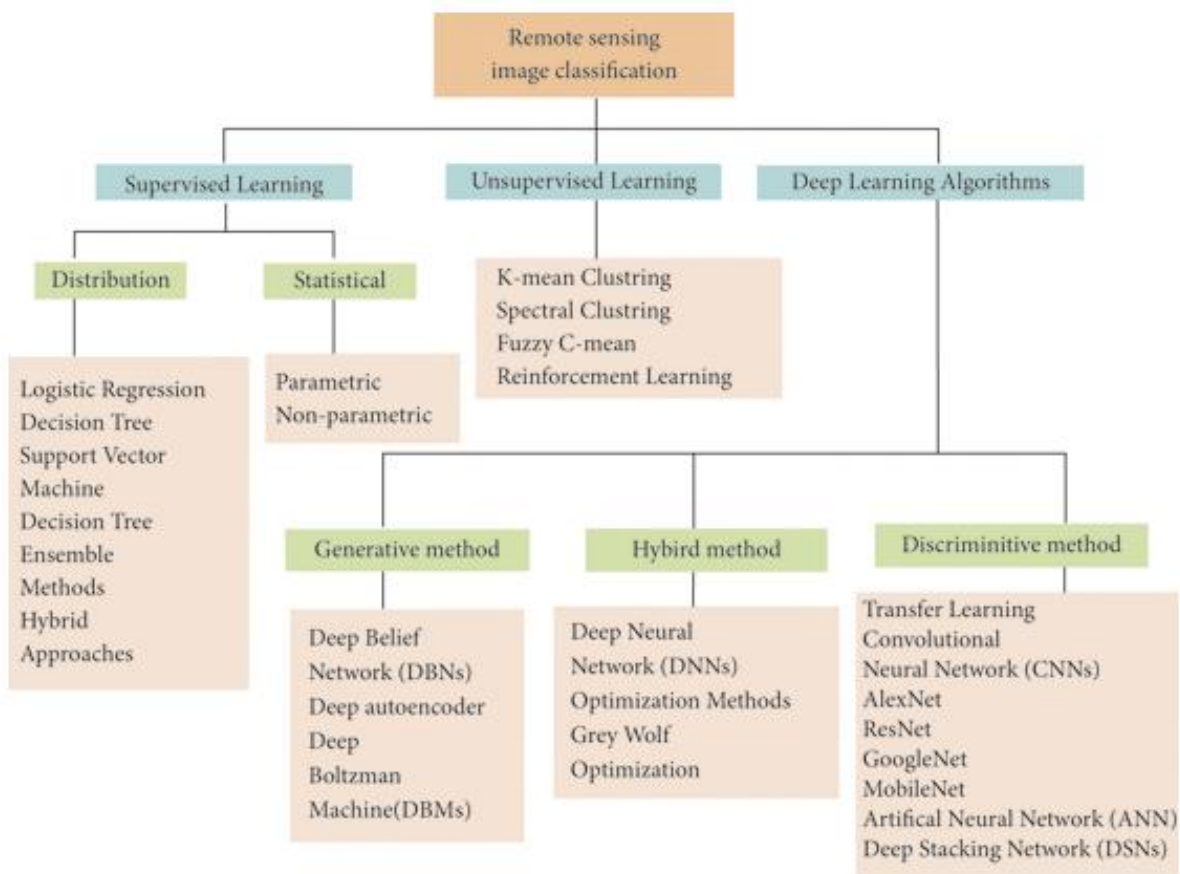
### **2.3.1.3 Land Use Land Cover Classification Methods**

Four main phases are often included in conventional supervised LULC machine learning pipelines: pre-processing, feature engineering, classifier training, and post-processing. These phases could each consist of a number of smaller tasks. To define standalone sub-problems that can be studied independently and have solutions or models that can be incorporated into a LULC pipeline to accomplish the targeted classification/segmentation, it is

helpful to break down the entire process into its sub-tasks, with an explicit statement of their assumptions. The demand for research using deep learning approaches to address these sub-problems has increased over the past few years due to the growing popularity of deep learning as a very strong tool in solving various sorts of AI challenges.

Machine learning algorithms are used to create models that can be used to classify the images. There are different methods of Remote Sensing classification as summarized by the chart below.

**FIGURE 2: Image Classification Methods**



Supervised Classification Algorithms for LULC- Among these classifiers are SVM, Random Forest, CART, and Naive Bayes. The standard process for categorization is: First gather practice data. Put together features with properties that hold the predicted values in numeric form for the predictors and a property that stores the known class name. Secondly

launch a classifier and adjust its settings as needed. Thirdly is to Use the training data to train the classifier. Fourth is to categorize a group of pictures or features. Lastly is to calculate the classification error using data from an independent validation.

A Feature Collection comprising attributes that store predictor factors and the class label is what makes up the training data. Class labels ought to be sequential, with numbers beginning at zero. if required. Numerical predictors ought to be used. Data for training and/or validation might originate from many different sources. To interactively get training data, you. A specific set of parameters is supplied for the SVM. Optimal parameters are unknown in the absence of prior knowledge about the physical nature of the prediction problem. When using the Random Forest, you have to describe the number of trees to be used.

Unsupervised Classification Algorithms for LULC-The Clutterer package manages clustering, often known as unsupervised classification. Currently, these algorithms are based on their corresponding Weka algorithms. Classifiers and clutterers are used in the same way. The typical clustering workflow is: To locate clusters, compile features with numerical attributes. The second is to get a clutter remover started. Adjust its settings as required. The third is to utilize the training data to instruct the clutterer. The fourth is to use the clutterer on a feature collection or image, and lastly to give the clusters names.

The feature collection that will be sent into the clutterer is the training data. A Clutterer does not have an input class value, in contrast to classifiers. The train and apply steps' data should have the same number of values as classifiers. When a trained clutterer is applied to a picture or table, each pixel or feature is given an integer cluster ID.

Choosing between supervised and unsupervised classification methods for Land Use and Land Cover (LULC) mapping depends on the availability of labelled training data, the



complexity of the study area, and the goals of your analysis. Here's a comparison to help you decide which method is more suitable for your specific scenario:

### **Supervised Classification:**

**When to Use: Availability of Training Data:** Supervised classification requires labelled training data, where you already know the land cover classes for certain sample pixels. If you have reliable training data, supervised classification methods can be highly accurate. **Specific Classes:** When you need to classify into specific, predefined land cover classes (e.g., urban, forest, agriculture) and you have representative training samples for each class.

**Pros: High Accuracy:** With proper training data, supervised methods often yield high accuracy because they learn from labelled examples. **Specific Classes:** You can classify the land cover into specific and meaningful categories. **Controlled Output:** You have control over the classes you want to identify.

**Cons: Dependent on Training Data:** The accuracy heavily relies on the quality and representativeness of the training data. **Time-Consuming:** Collecting and labelling training data can be time-consuming and labor-intensive.

### **Unsupervised Classification:**

**When to Use: Exploratory Analysis:** When you don't have labelled training data and you want the algorithm to explore the data and identify natural groupings (clusters) on its own.

**Unknown Classes:** When you are not sure about the number or nature of the land cover classes in your study area.

**Pros: No Training Data Required:** Unsupervised methods do not require labelled training data, making them useful when such data is scarce or unavailable. **Data Exploration:**

Unsupervised methods can reveal hidden patterns or groupings within the data, aiding in exploratory analysis.

Cons: Interpretation Challenge: Unsupervised methods provide clusters of pixels, but these clusters might not always correspond to meaningful land cover classes. Interpretation of the clusters can be challenging and requires additional domain knowledge. Less Specific: Unsupervised methods might group pixels in a way that doesn't align with specific land cover categories of interest.

### **How to Choose between Supervised and Unsupervised classification:**

Availability of Training Data: If you have high-quality, representative training data for the specific land cover classes you want to map, consider supervised classification methods.

Data Exploration and Discovery: If your goal is to explore the data, identify natural patterns, or discover unknown land cover classes, unsupervised methods can be more appropriate.

Hybrid Approaches: Sometimes, a combination of both methods (semi-supervised or hybrid methods) can be beneficial. You can use unsupervised methods to create initial clusters and then assign meaningful labels to those clusters using limited training data.

Consider Project Goals: Think about the project goals and the level of accuracy required. If precision and specific class identification are critical, supervised methods might be more suitable. If you're conducting a broad-scale analysis without specific class requirements, unsupervised methods might suffice.

Ultimately, the choice between supervised and unsupervised methods should align with your study objectives, available resources, and the nature of the land cover classes in your study area. Consider the advantages and limitations of each approach before making a decision.

### 2.3.2 Machine Learning Methods

Research on prediction of the deformation in lake Urmia area of Iran using and ensemble machine learning model of Convolution Neural Networks (CNN), Multi-layer Perceptron (MLP), and Long Short Term Memory(LSTM) by (Radman et al.,2021). On their research they used satellite images to monitor and identify areas of great subsidence. Data from satellite images together with rainfall, groundwater, and lake variation were used to assess the relationship between land deformation and environmental parameters. The environmental parameters were then used as inputs of the model and ground deformation as a target to be predicted. The accuracy of the independent models was low than the accuracy of the ensemble model of the three algorithms hence the prediction of the magnitude of deformation was of high accuracy.

Another research by (Park et al.,2022) on development of a model with high accuracy in the prediction of rapidly fluctuating water levels by evaluating LSTM and GRU methods at Hangang in Han River, South Korea. Hydrological data together with meteorological data were used as inputs of the models in order to get a higher accuracy of the water level prediction. The correlation between collected hydrological, water level and meteorological data was analyzed and used as input into the models to determine the priority of data to be trained. The results showed that GRU had higher accuracy than LSTM. The accuracy was also higher when multivariate data was used other than univariate data.

According to the research (Wu et al., 2020) it was determined that transformer machine learning models can also be used in the time series prediction. In their research they explained how Time series transformer models can be used with both Univariate and multivariate data to produce results of higher accuracy than other existing time series models in the prediction of Influenza Like Illness. Transformer models are composed of two layers an Encoder and a

Decoder. The encoder entails input, positional encoding, and a stack of four identical encoder layers. The input layer maps the input data (time-series) to dimension vector through a fully connected network. Positional encoding layer on the other hand is composed of cosine and sine functions used to encode information in a sequential manner from the data (time series) by addition of input vector to a positional encoding vector. The output from the result vector is then used as input to the 4 encoding layers. Each of the four layers is composed of 2 sublayers, namely; Self attention Sublayer and Fully connected feed -forward sublayer. The sublayers are then followed by a normalization layer. The encoder output is then fed to the decoder. The Decoder on the other hand has an Input layer, 4 identical decoder layers and output layer. The decoder input layer maps the input to a dimensional vector. The decoder has a sublayer that applies self-attention mechanisms over the encoder output. The output layer then maps the output of the last decoder layer to target time-sequence. Transformers models employs masking (look ahead) and one position offset between the decoder input and target output in the decoder that ensures prediction of a time series data point only depends on previous data points.

Another research (Xu et al.,2019) expounds how the work was done in four steps which include; Collected data from needed departments, and divided it into water level data and environmental data. Built the ARIMA model, verified the data was stationary and got the parameters through AIC, then trained the model. The water level could then be predicted and the residual calculated. Used Relational Neural Network model to predict residual, normalized the environment data and residual data, then constructed the training data, and got an RNN to predict the residual sequence. Predicted future water level. The relationship between the predicted value of the residual, the predicted value of the water level, and the actual value could be determined through the RNN network, and they could predict the water level for the next 30 days.

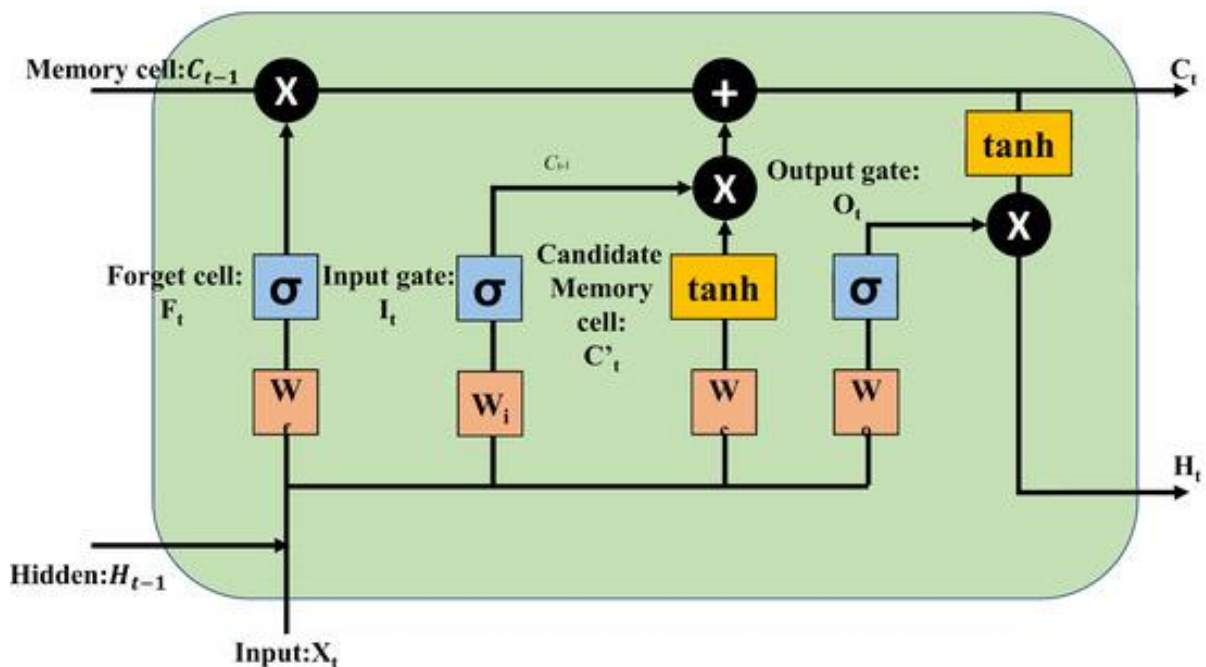
According to (Lim & Zohren, 2021) Recent advances in processing power and data availability have made deep neural network architectures quite successful at forecasting difficulties in a variety of disciplines. They examined the primary architectures for time-series forecasting in this research, emphasizing the essential components of neural network architecture. They discussed how they might be expanded for use in multi-horizon forecasting and looked at how they combine temporal information for one-step-ahead forecasts. They also discussed the current trend of hybrid deep learning models, which beat over pure techniques in both categories by combining statistics and deep learning components. Lastly, they outlined two ways—methods in interpretability and counterfactual prediction—that deep learning might be expanded upon to enhance decision support over time.

While several deep learning models have been created for time-series forecasting, there are still certain restrictions. First off, forecasting datasets with missing observations or ones that arrive at random intervals is challenging for deep neural networks since they usually require time series to be discretized at regular intervals. Neural ordinary differential equations had been used to conduct some basic research on continuous-time models; however, further effort was required to expand this work to datasets with complicated inputs (such as static variables) and to assess them against current models. Furthermore, time series frequently exhibit a hierarchical structure with logical groups within trajectories, as in the case of retail forecasting, where shared trends may have an impact on product sales within a certain region. Therefore, the creation of structures that specifically taking into consideration such hierarchies could be a worthwhile area of study and lead to better forecasting results than current univariate or multivariate models.

### 2.3.3 LSTM and GRU

LSTM is a form of an improved Recurrent Neural Network that had a vanishing gradient problem solved (Wu, et al., 2023). It processes and analyses time-series data by extracting saved information by selecting and combining selected information with subsequently input time-series data. It has ability to predict a fragmented sequence of data and passing deviations backward to predict the sequence dynamically. LSTM has a hidden state(H) and memory state(C) in its output and it has three gates to process the inputs differently.

**FIGURE 3: LSTM Algorithm Structure**



LSTM regulates the internal state to enable information from earlier moments to be retained and filtered. Information to be forgotten in the internal state of the preceding moment  $C_{t-1}$  is controlled by the forget gate  $F_t$ . Gate of input  $I_t$  chooses and stores the input data memory and calculates the amount saved to the cell state  $C_t$  at the given instant  $X_t$ . The output gate's primary function is to regulate the amount of data that the internal state must currently output to the external state.  $H_{t-1}$  and  $X_t$  are the inputs used by the forget gate.

The output is controlled by the sigmoid activation function ( $\sigma$ ), which ranges from 0 to 1. A "0" means that no information can pass at all, whereas a "1" means that all information passes. This permits control over the internal network of the gate. The hidden state is updated and candidate cell state values are determined using the tanh activation function.

In a study by (Song et al., 2020), A new hybrid model for estimating the daily oil rate of a volcanic reservoir was proposed. It is an LSTM neural network optimized by PSO. Engineers and decision-makers can use the model to obtain well performance data ahead of time, which could help them manage and modify reservoir development plans more effectively. It was questioned if the suggested model could predict the daily oil rate of the two situations. Concurrently, the suggested model was contrasted with the conventional approaches.

The following were the main conclusions: The suggested model demonstrated a potential approach to time-series oil rate prediction, as demonstrated by its good performance on oil rate data from the Xinjiang oilfield and the simulator. It may be used as a substitute instrument to provide a trustworthy prediction. The suggested model performed better than ARIMA, conventional neural networks, and decline curve analysis. It correctly depicted the intricate pattern of fluctuation in the oil rate under various conditions. Furthermore, when the input was predicted to have never existed during the sequence problem's training, its generalization ability was higher. The performance of the LSTM model depended heavily on the parameter selection. It is important to choose the training epoch carefully in order to prevent over-fitting and insufficient training. At the expense of longer computation times, the hidden layer could somewhat increase the LSTM's accuracy.

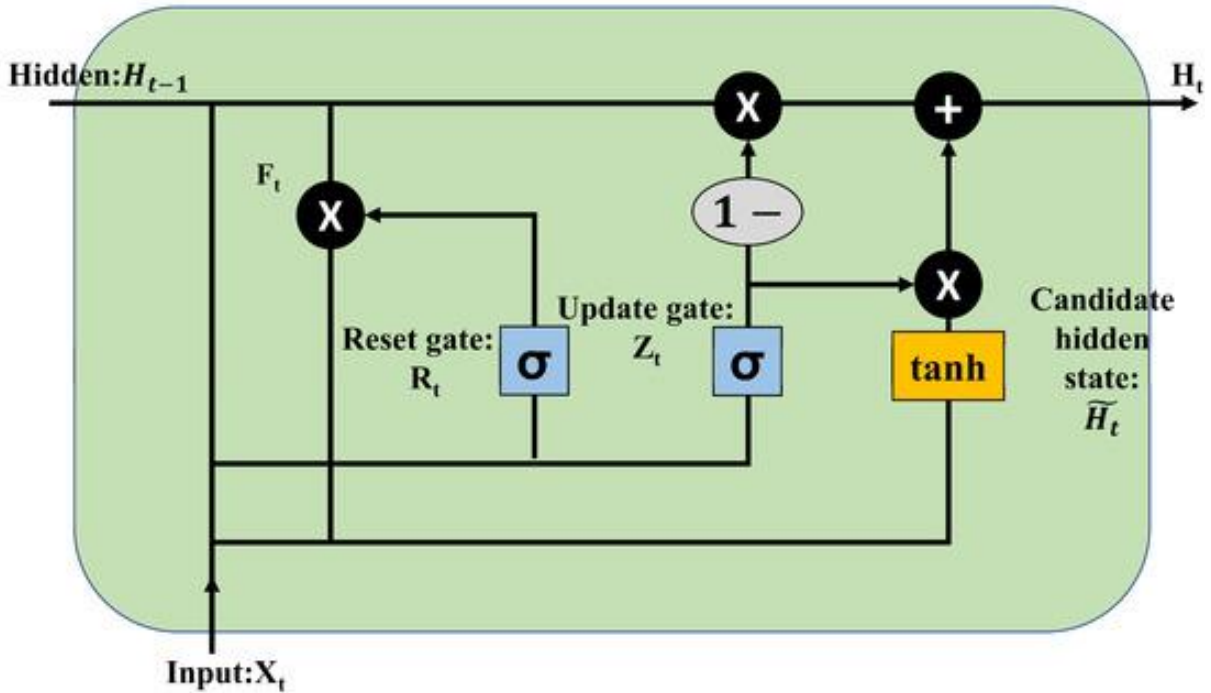
They would look into combining LSTM and convolution neural networks in the future to manage the production prediction of several wells that is dependent on time and space. This was one instance where how a prediction model was created using LSTM.

An overview of LSTM topologies designed to forecast nonlinear time series behavior was given in the publication (Lindemann et al., 2021). Numerous architectures are in use today and have been used to a variety of fields, including autonomous systems, manufacturing, and image processing. Within the parameters of this work, the methods were classified and assessed in relation to predetermined attributes. The following is a summary of the main findings: LSTM that use optimized cell state representations—like attention-based and hierarchical LSTM—perform better while processing multidimensional input. Grid and cross-modal LSTM are examples of LSTM with interacting cell states that can jointly and accurately predict many quantities. LSTM Seq2Seq can forecast multi-step forward prediction with reduced error propagation for many parameters. Additionally, partially conditioned Seq2Seq LSTM provide the best fit for modeling both short- and long-term dependencies.

In contrast, GRU is a more efficient variant of LSTM with a simpler training procedure. It has just two gates, as opposed to LSTM's three; the Update gate (which controls changes in the state of hidden units over time through a particular gated structure) and the Reset gate (which merges LSTM's memory cells and hidden layer states).  $Z_t$  and  $R_t$ , respectively, represent the update and reset gates.  $Z_t$  determines how much of  $H_{t-1}$  is kept in  $H_t$ , while  $R_t$  determines how much of the current candidate set  $H_t$  will be written into  $H_{t-1}$ .



FIGURE 4: GRU Algorithm Structure



The input  $X_t$  in the current moment and the state in the preceding moment  $H_{t-1}$  are used to determine the gating state. The mapping of data between 0 and 1 represents the gating signal. The reset gate is "reset" as a coefficient of the previous moment  $H_{t-1}$  once the gating signal is achieved, yielding the candidate hidden state  $\bar{H}_t$ . To obtain a result between 0 and 1, the sigmoid activation function is employed, and the tanh function is used by the activation function to calculate a potential hidden state. The reset gate, update gate, and weight matrix for determining the candidate's hidden state are denoted by the letters  $W_r$ ,  $W_z$ , and  $W_h$ . are used to determine the candidate's hidden state. They include the reset gate, update gate, and weight matrix.

(Xu et al., 2021) In order to address the issue of time series prediction of water quality data, they developed a novel model in this work called FM-GRU. By using the FM model to solve the common issue of missing data in the water quality monitoring dataset and obtain the potential high-dimensional feature interactive information in the time series, FM-GRU technically realized the effective integration of factorization machine and seq2seq framework.

This information was then processed as input to the GRU model. The classic time series model's long-range information loss issue is resolved by the dual-Attention component of the framework, which allows the FM-GRU model to continue having a positive impact on time series analysis over an extended period of time. It was evident in the experimental section that FM-GRU offers glaringly better performance than several popular time series forecasting algorithms. Natural language processing (NLP) and time series forecasting (TSF) are currently two of the most important research topics in both academia and business. This paper suggested a novel approach and strategy for resolving issues with TSF, NLP, and river water quality. Novel and practical solutions have been suggested by eco-environmental governance and planning.

(Mahjoub et al., 2022), This research introduces three deep neural networks: LSTM, GRU, and Drop-GRU, and presents an energy management technique based on the forecasting process. These methods' primary goal was to predict and manage load usage. In order to maximize ability prediction, the power consumption prediction techniques first processed the input data, carried out efficient feature extraction, and then constructed the proper network structure. Lastly, a comparison analysis of the suggested algorithms is carried out. After a set of power load data was used to evaluate and execute these three strategies, the Drop-GRU outperformed the GRU and the LSTM. More precisely, the project's use of the GRU technique to anticipate energy usage over a specified horizon based on past consumption was ideal. readings, enabling them to anticipate consumption peaks and foresee the best course of action for load shedding. The goal of the research will be to create hybrid models with even greater accuracy and speed in the future. By including additional external aspects like holiday and weather data, we may further enhance these outcomes. Therefore, it is simple to identify high consumption points that surpass the permitted consumption level and subsequently safeguard the electrical grid based on the forecast findings and external data on production capacity.

The research by (Cho et al., 2022) was based on LSTM and GRU. Basically worldwide, flood damage was found to be growing, and if floods could be forecasted, the financial and human costs associated with them could be minimized. Data on water levels is a crucial factor in floods, and this study suggested a water level prediction model that makes use of a gated recurrent unit (GRU) and long short-term memory (LSTM). Meteorological data, such as temperature, humidity, precipitation, and upstream and downstream water level, were employed as variables in the input data. When trials were conducted using alternative model structures and different input data formats, the best results were obtained when the Automated Synoptic Observing System (ASOS) meteorological data and the LSTM–GRU-based model were included in the input data. The experiment's mean squared error (MSE) value was 3.92 as a result. The top result across all cases was the mean absolute error (MAE) value of 2.22, and the Nash–Sutcliffe coefficient of efficiency (NSE) value of 0.942. The test results also revealed the study area's historical highest water level of 3552.38 cm, as well as the lowest possible maximum water level error of 55.49. The performance difference according to the input data composition and the time series prediction model was confirmed through this paper. We intend to develop a flood risk management system in a later study that can anticipate floods and evacuate in advance based on the anticipated water level.

## **2.4 Knowledge Gaps**

In research, a "knowledge gap" is an area or topic where knowledge, information, or data is scarce or inadequate within a specific field. Finding knowledge gaps is an essential component of research because it allows scientists to assess what has been researched, what is known, and what areas require additional study. In this study some of the identified knowledge gaps were as summarized by the table below;

**TABLE 3: Summary of the Knowledge Gaps**

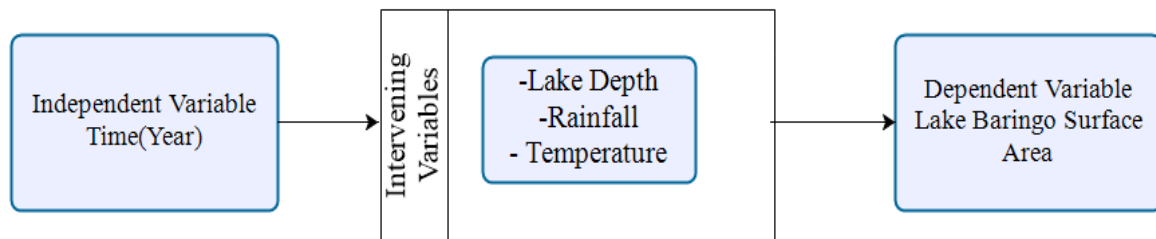
Researcher	Study Focus	Findings
(Xu et al.,2019)	A Model for Predicting Water Level Using ARIMA-RNN.	Created an ARIMA-RNN model to predict the rise in water level
(Wang et al., 2020)	Used Remote Sensing and GIS to analyze and detect land use and land cover change in Nepal's Kathmandu district and predicted future change.	Generated LULC maps between 1990-2020 then used the CA-Markov model to predict the future LULC changes in 2030
(Radman et al.,2021)	utilized deep learning and InSAR to model and forecast subsidence over the nearby region of Lake Urmia, Iran.	Created an ensemble model of Convolution Neural Networks(CNN), Multi-layer Preception(MLP), and Long Short Term Memory(LSTM) to predict subsidence in the adjacent areas of the lake.
(Muhammad et al., 2022)	A case study of Linyi-China was presented that illustrates the use of the QGIS MOLUSCE plugin and remote sensing big data for spatiotemporal change analysis and prediction of future changes in land use and cover.	used Cellular Automata-Artificial Neural Networks(CA-ANN) to predict the changes in the LULC for the years 2030, 2040, and 2050 in Linyi China.
(Park et al.,2022)	Creation of Deep Learning Models to Increase the Precision of Time Series Predictions of Water Levels Using Multivariate Hydrological Data.	Compared GRU and LSTM algorithms in the prediction of the multivariate dataset in the prediction of water levels of a river

<b>Knowledge Gap</b>	<b>Strategy Used to Address Knowledge Gap</b>
<p>Used RNN which has a vanishing gradient problem and dwelt only in water level Prediction.</p>	<p>Compared LSTM and GRU which solve the vanishing gradient problem and dwell in the lake expansion prediction.</p>
<p>The focus was on all the changes on the land but not on the lake area change. Used an already existing model.</p>	<p>Generated the maps of specific years for the Lake Baringo Area, analyzed the change in the area, and created a lake expansion prediction time-series model</p>
<p>Compare the 3 individual models and the ensemble model to come up with a prediction model of higher accuracy.</p>	<p>Compare the accuracy of LSTM and GRU then created a model with the algorithm with better accuracy.</p>
<p>Used an already existing model, the Focus was on all the LULC changes.</p>	<p>Create a time series model that was able to forecast the area of the Lake expansion Accurately. LULC maps generated served as validation of the available dataset.</p>
<p>Used Multivariate hydrological datasets in the comparison of the algorithms.</p>	<p>Compare the 2 algorithms and develop a model with the best accuracy out of the two.</p>

## 2.5 Conceptual Framework

The conceptual framework highlights the relationship between the available variables in the dataset. Based on the previous studies done on this topic, the study proposes the following conceptual Framework.

**FIGURE 5: Conceptual Framework**



Intervening variables, also known as mediator variables, are variables that come between the independent variable (the variable that is being manipulated or controlled) and the dependent variable (the variable being measured) in a cause-and-effect relationship. These variables explain how or why the independent variable influences the dependent variable. Here, the intervening variable serves as a mediator between the independent and dependent variables, meaning that changes in the independent variable affect the intervening variable, which, in turn, affects the dependent variable.

### 2.5.1 Characteristics of Intervening Variables:

**Mediating Role:** Intervening variables explain the process or mechanism through which the independent variable affects the dependent variable. They provide insights into the underlying causal relationship. **Partial Effect:** Intervening variables can either enhance or diminish the effect of the independent variable on the dependent variable. They can amplify or reduce the strength of the relationship between the two variables. **Complex Relationships:** In real-world situations, cause-and-effect relationships are often complex. Intervening variables

help researchers understand the nuanced interactions between variables. Mediation Analysis: Researchers use various statistical techniques, such as mediation analysis, to explore and quantify the role of intervening variables. Mediation analysis helps determine if the relationship between the independent and dependent variables is mediated by the intervening variable.

## 2.6 Operationalization of Variables

**TABLE 4: Operationalization of Variables**

Variable	Indicator	Values
Time	Date	Day-Month-Year
Lake Level	Surface Height	Meters
Rainfall	Amount	Millimeters
Temperature	Value	Degrees Celsius
Lake Surface Area	Area	Km <sup>2</sup>

## **2.7 Summary**

Flooding of Lake Baringo is a threat to the community in the area and many destructions occur during the flooding. There is a need for research to be done to alert people early of future flooding and to help reduce the adverse effects that would be experienced in case it occurs. Many researchers have done work along the same line but there have been some gaps that exist in the previous research and there has not been a prediction model that has been developed to help predict the potential flooding area. The purpose of this study was to map the changes that have occurred in the Lake Baringo area due to the expansion of the lake and develop a forecasting model that would help to predict the area that is likely to be affected in the future and to help solve or reduce the existing gaps in the previous studies. This chapter looked at the relevant theories, past research done nearly with the topic, the relevant conceptual frameworks of the study conducted. It also talked about the two algorithms and how they differ from each other and examples of past models created using the two algorithms.



## **CHAPTER THREE: RESEARCH METHODOLOGY**

### **3.1 Introduction**

This chapter aims at discussing research design and how the four objectives will be achieved. It outlines current methodologies as well as the proposed model that will be implemented to achieve the desired model. Available datasets, manipulation of datasets, and development of output are also discussed in detail.

### **3.2 Current Methodological Approaches**

Climate change is the current term being discussed by many researchers in present times to explain the change in Earth's weather patterns. Researchers are putting their best foot forward to determine the adverse effects of climate change and how it can be reduced or even mitigated to reduce the effects. Remote Sensing, GIS, and Machine Learning methods have been used to conduct the study by visualizing the previous and current changes in the earth's features as a result of climate change. Machine Learning has been integrated with GIS and Remote Sensing methods to predict future changes based on previous changes.

### **3.3 Research Design**

The main objective of this research was to develop a time series model that would be used in forecasting Lake Baringo areal change. The research design used in this study was Longitudinal research design where the data used was secondary data that had been collected over a long period of time(1985 – 2023) and it was used to study changes over time. The study had two parts of datasets which included; Landsat Images data (This is secondary data from United States Geological Survey. It is a form of database that stores satellite data from the previous years 1969 to date and can be accessed through scripting) and the tabular time series data(from the DAHITI website and Chirps website, this also was accessed through the

Knowledge Discovery in Databases procedure). The Landsat satellite data is composed of different bands and was acquired based on the required band combinations that is (Red band, Green band and Blue band- since these are the bands which bring out the true colors of features as they are, for example vegetation to appear green, water to appear blue and land not interfered by man to remain brown. The manipulation and analysis of the Landsat images was done by use of script in the Google Earth Engine platform, each of the four scripts(for the image years) consisted of five main components as expounded below;

Landsat Surface Reflectance picture collection: This is the initial stage of data access and acquisition, initiated by a function that calls a stack or sequence of images from the picture Collections to create a composite image. In GEE, each of them had a unique ID and Image Collection. The Landsat 7 and 8 collection was called and composited in this instance. The GEE can also be used to derive an image collection from individual photos or an image merger from pre-existing collections. To restrict the acquired image to only the selected location and period of interest, a filter function was programmed in. A function to mask out the clouds using a pixel Quality Assurance cloud band value was used because the images suffered from a heavy cloud cover.

Study area and sampling selection - The user's preferred designated shape (vector) was then trimmed from the image. In GEE, vector data can be obtained in four different ways: by importing an existing vector shape via Google Fusion Table, by drawing manually, by using an existing vector dataset, or by uploading a shapefile directly to the personal Asset folder. In this investigation, the GEE platform was used to import the vector data from the Google Fusion Table.

Supervised classification using MLA: For the purpose of classifying land covers, this study employed a pixel-based supervised classification technique using a machine learning

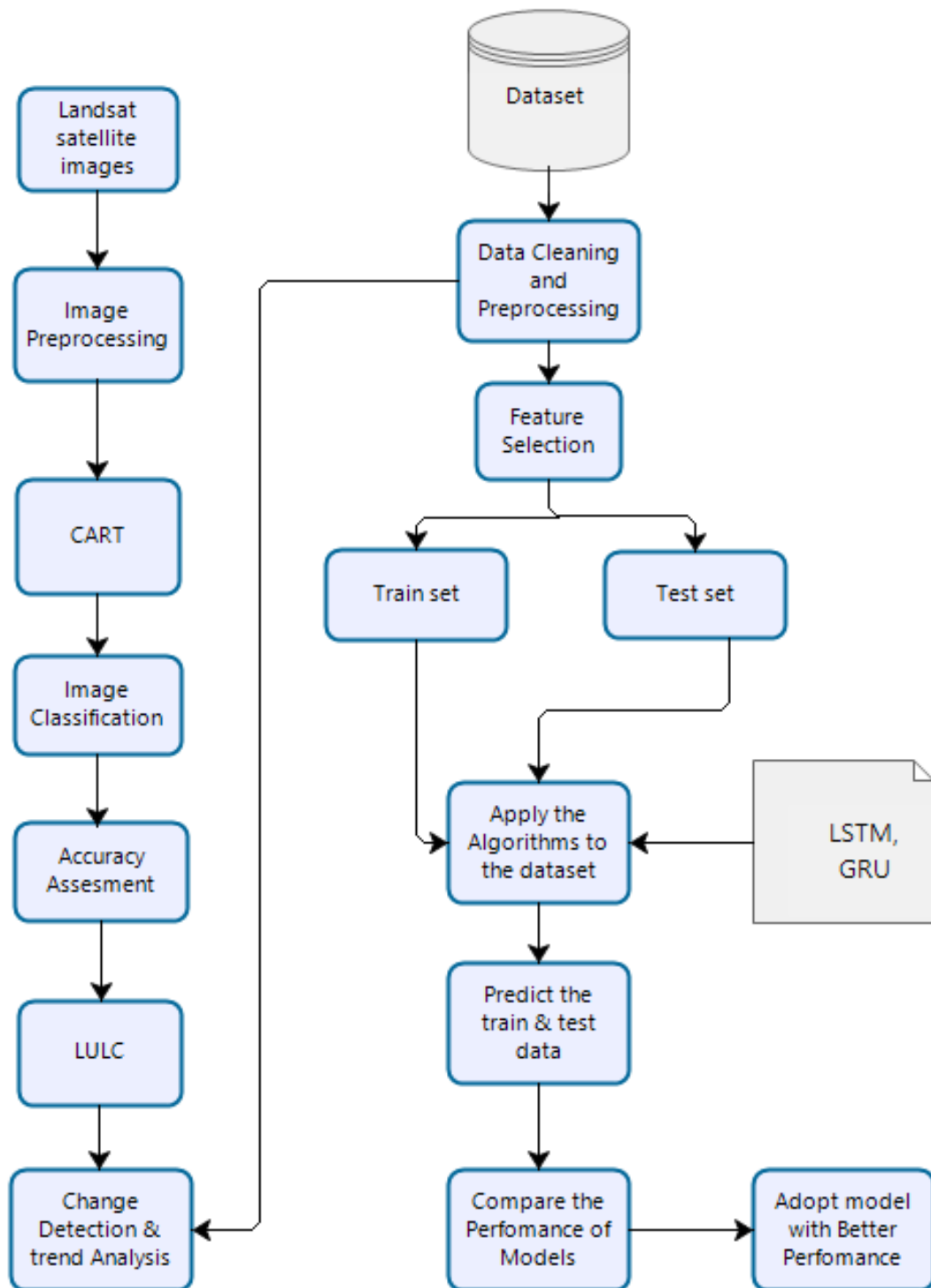
algorithm. Four sets of training and testing polygons were found by analyzing the composite images (based on the photos and Google Earth,) for three classes (Vegetation, Water, Bare-land/Built-up) as Feature Collection utilizing the Geometry Tools and Import in the years 2002, 2009, 2016, and 2023. Due to the manual selection of these samples based on the accessible references, there were some biases in the process. By repeatedly running the script, the quantity of good samples was experimentally tested in order to obtain satisfactory statistical and visual findings. The Classification and Regression Trees (CART classifier) within the Earth Engine platform were then trained using the samples (API Reference | Google Earth Engine, n.d.). The classification result was then shown using a map function. The quantity of training and testing pixels as well as the classifier's complexity may impact processing time and result delivery, therefore extreme vigilance is required.

Validation and assessment: Using Google Earth for the relevant year and existing images, ground truth data or testing samples were gathered for this study as previously mentioned in the part above. Confusion matrices were utilized to evaluate the performance of supervised classifiers, and the introduction and allocation of "testing" data to the classifier is required to obtain a true validation accuracy. To make using GEE easier and more visually appealing, the previously gathered samples have been programmed to be randomly divided into training and testing groups of 80:20 percent. The script was designed to hold out data for testing, evaluate the Confusion Matrix for this withheld validation data, and then apply the classifier to the testing data. The validation outcomes were subsequently printed in the Code Editor's right-hand console or exported as a table with properties saved on Google Drive.

Exporting results: Giving a long-term record of the output gain from the analysis, exporting results might have a variety of uses. The export of pertinent analytical findings, including tables, charts, photos, and maps, is the last stage of the GEE processing process.

The Figure below illustrates the procedures followed to achieve the end results.

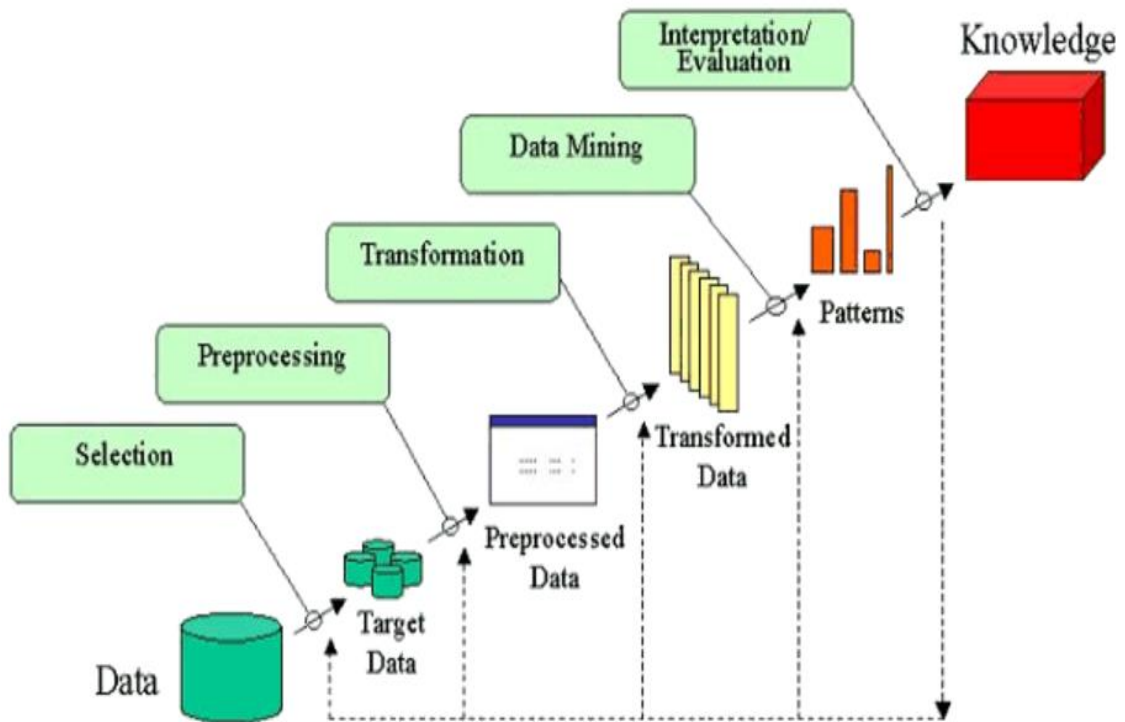
**FIGURE 6: Methodology Workflow**



### 3.3.1 Dataset

All datasets used was secondary dataset. The data entailed Landsat satellite images for the years;2002,2009,2016 and 2023(temporal resolution of seven years). These Landsat images were retrieved from the United States Geological Survey (USGS) website via the google earth engine. Another dataset that was important in the accomplishment of the task were the rainfall, temperature, Lake Baringo Depth and Lake Baringo area over the years from April 1984 to August 2023(at a monthly temporal resolution), the dataset was from the DAHITI website (Herrnegger et al., 2021) In order to offer water level time series of inland waters, the Deutsches Geodätisches Forschungsinstitut der Technischen Universität München (DGFI-TUM) created the Database for Hydrological Time Series of Inland Waters (DAHITI) in 2013. Nowadays, DAHITI offers a wide range of hydrological data on wetlands, rivers, lakes, and reservoirs that are obtained from satellite data, such as optical remote sensing imaging and multi-mission satellite altimetry. After a quick registration process, the user community has free access to all goods.

**FIGURE 7: Knowledge Discovery in Databases (DBD, 2020)**



(Jodha, 2023) Knowledge Discovery in Databases (KDD) is the process of discovering useful patterns, trends, correlations, or knowledge from large amounts of data. It involves several steps and techniques that allow researchers, data scientists, and analysts to uncover valuable insights from complex datasets. The KDD process typically includes the following stages: Data Selection: Identifying the data to be analyzed based on the research questions or objectives. Gathering relevant data from various sources, including databases, text files, and web sources. Data Preprocessing: Cleaning the data to handle missing values, outliers, and inconsistencies. Transforming the data into a suitable format for analysis. Integrating data from multiple sources if necessary. Data Transformation: Reducing data dimensionality through techniques like Principal Component Analysis (PCA). Discretizing continuous variables or normalizing data to make it suitable for mining algorithms. Feature engineering, creating new variables derived from existing ones to enhance analysis.

Data Mining: Applying various data mining techniques to discover patterns in the data. Common data mining methods include clustering, classification, regression, association rule mining, and anomaly detection. Using algorithms like decision trees, neural networks, support vector machines, and k-means clustering. Pattern Evaluation: Assessing the discovered patterns to identify those that are interesting, valid, and potentially useful. Utilizing metrics such as accuracy, precision, recall, and F1-score for evaluation. Eliminating redundant or irrelevant patterns. Knowledge Presentation: Presenting the discovered knowledge in a comprehensible format to users or stakeholders. Using visualization techniques like charts, graphs, and dashboards to represent patterns and trends. Summarizing findings in reports, presentations, or interactive interfaces. Knowledge Utilization: Integrating the discovered knowledge into decision-making processes. Applying the insights to solve specific problems, make predictions, or optimize processes. Iteratively refining the models and patterns based on feedback and new data.

### **3.3.2 Image Preprocessing and Data Preparation**

Data pre-processing is a crucial component of LULC classification since it influences the accuracy of the final classification outputs, according to the research of (Darem et al., 2023). Prior to LULC classification, data pre-processing is a technique used to maximize the accuracy of the classification outcomes. It facilitates the improvement of the data to yield better outcomes. To achieve more precise classifications, data pre-processing involves a number of steps, including spatial filtering, data improvement, and radiometric adjustments. In order to account for offset effects and sensor gain, which might have a negative impact on the LULC classification findings, radiometric corrections are made. The two main radiometric adjustments employed in LULC classification are equalization of the histogram and adjustment of the signal-to-noise ratio. By adjusting the signal-to-noise ratio, the variation is reduced and the output Image data preparation was done in Google Earth engine. Tabular dataset



preparation on the other hand included data cleaning to remove all the cells that had null values, outliers and other inaccuracies, data integration involved combining of dataset that was from 2021 to 2023 with that was from 1984 to 2020, and data transformation tasks involved formatting the date to the required format and saving the data values to float in order for proper application of the algorithms. This was carried out in Python program.

signal distribution in the image is more uniform. The process of histogram equalization involves spreading the pixel values throughout the dataset to account for the non-uniform brightness distribution in areas with different levels of illumination.

To rectify the geometric and radiometric errors that arise during capture, scanning, and transmission, the RS data must be pre-processed. Clipping maps, image registration, georeferencing, radiometric correction, image enhancement, camping bands, and layout maps are some of the pre-processing procedures. By removing noise sources and enhancing image quality, these actions produce a dataset that is more trustworthy.

Here is a description of the steps:

To extract particular areas of interest from the original photos, clipping maps are utilized. By removing unnecessary background information, it makes it possible to analyze photos more precisely. Registration of images is the practice of lining up images from several locations so that comparable elements can be contrasted. It is also capable of aligning the coordinate systems of two photos that were taken at the same spot (Ramdani et al., 2021).

### **3.3.3 Image Classification**

The Landsat images were then classified using Classification and Regression Trees (CART) to get the different classes of land use such as; Water, Vegetation, and Bare land/Built-up area. The process involved backdating the google earth software images then selecting the

training samples for the relevant years. Training samples collected were 100 for each of the classes considering the different characteristics such as color.

CART classifier was chosen based on the research by (Wahap & Shafri, 2020) Geographic big data is currently receiving a lot of attention and is being highlighted worldwide. Google Earth Engine (GEE) is thought to be the leading platform for big data processing in remote sensing and geographic information systems. There are currently few or no studies on the use of this platform to examine changes in land cover and use through time in Malaysia. The goal was to classify the Klang Valley area from Landsat composites of three different years (1988-2003-2018) using several Machine Learning Algorithms (MLA) in order to assess the viability of GEE as a free cloud-based platform. Using commercial software, the best categorization results were then imported and subjected to additional processing to measure changes over time.

Despite the great accuracy of the classification findings, CART displayed the best accuracy with Comparison of 1988, 2003, and 2018 to 94.71%, 97.72%, and 96.57% in relation to RF and SVM. A small number of incorrectly classified pixels were found as a result of the annual composited pictures' compilation without crop phenological stages being taken into account. This led to the incorrect classification of agricultural land as bare land and urban areas. Because they may have an impact on the classification outcome and subsequent analysis, the initial selection and composition of the data needed to be planned and organized before processing. Nevertheless, with little assistance from humans, GEE processed several datasets quickly and with high performance in terms of time and complexity. Overall, GEE demonstrated dependability in achieving the study's goals of classifying and measuring the land use/cover of the investigated area in order to assess the viability of the GEE. The accuracy of

the classification was then computed to know the level of the output confidence. Carried out in Python and Google Earth Engine.

### **3.3.4 Data Analysis**

The LULC maps for the different years were compared to see the positive or negative changes in the area over the years and a summary of the change analyzed. Analysis of rainfall, temperature, water level and lake area were also computed to determine the behavioral change, this was done using the tableau software. The data used for the study was available in raw format in an Excel spreadsheet. It was imported into the TensorFlow environment for preprocessing and analysis using Python programming language. The input/ predictor variable used to construct the GRU model was the lake Surface area. The variable was fed into the input layer which then feeds into the single hidden layer.

Each hidden layer with determined neurons through experimentation and turning feeds into a single output neuron that carries the decision of the variable, which is the surface area of the lake. The Adam activation function was used. Decisions must be taken to divide the dataset into training, and test ratio. Data samples of 369 observations are randomly mixed and 75% of them are used for training while 25% are used for testing. The training epoch was set to 200 with a no batch size. The GRU training performs continuously and terminates when the validation error failed to decrease during the validation process. The model was evaluated using the MSE and RMSE score evaluation metric since it was a regression model the values were less than 0.5 and proved to be accurate and reliable enough to solve the given nature of problem.

### **3.3.5 Modelling**

After the data preparation phase the tabular dataset only had two columns which were; Date and Lake Surface Area in Km<sup>2</sup>. The data was then split to train and test set then the model

was created by applying the two algorithms (LSTM and GRU) on the dataset and defining all the constants and the requirements, the created models were then applied through fitting. The prediction of both the train and test set was then done. The accuracy and the efficiency of the two models was determined using the mean squared error and the root mean squared error. The graph of the actual, train and test set data was plotted to check the behavior. The loss and validation curve were then plotted to see the reaction of the model on the training data and performance of the models on the unseen data.

### **3.3.6 Model Evaluation**

The model performance was determined by the use of Mean Squared Error and the Root Mean Squared Error. MSE is calculated by taking the average of the squared differences between the predicted values and the actual values. MSE gives more weight to large errors due to the squaring operation. Therefore, outliers or large errors have a significant impact on MSE. RMSE is simply the square root of MSE. It is a more interpretable metric because it is in the same units as the target variable. RMSE provides a measure of how spread out the residuals (the differences between predicted and actual values) are. It penalizes large errors more significantly than small errors due to the square root operation.

### **3.3.7 Deploy the model**

This is the step where the model is moved from the mining environment to the scoring environment. Depending on the development environment, the process may be hard or easy. The hard part comes from the fact that some development may have been done in a special environment where the software can't run anywhere else and the miner has to move it and recode it in another programming language to make it possible for it to run.

### **3.3.8 Assess the results**

In this step, actual results are compared against expected results. The actual measure of data mining is the value of actions taken as a result of the data mining. Having a measure of lift can help one choose a model and use of these models can help one choose how to apply the results from the models. In addition to using lift, it is also important to do a measure at the field.

### **3.3.9 Begin Again**

The completion of a data mining project raises more questions without answers. This is true since there come new relationships that were not thought of before and this creates a new question to answer hence starting of data mining process again.

## **CHAPTER FOUR: DATA ANALYSIS FINDINGS AND DISCUSSIONS**

### **4.1 Introduction**

The main findings of this study are presented in this chapter. The chapter begins by declaring the results obtained from the summary descriptive statistics. The analysis of dataset and the characteristics of various components will also be discussed here. The chapter will present results based on various validation methods. The performance of the most significant time series algorithm will also be discussed.

### **4.2 Descriptive Statistics**

The sample size of 369 was drawn covering a Thirty nine-year period of between April 1984 and August 2023 in which 75% of the sample was applied in the development of the predictive model, while the 25% was used for validation. The dataset had columns of Lake Surface Area (Km<sup>2</sup>), water level (m) and Rainfall(mm). The obtained observations were described in terms of their frequency that is the number of observations in each characteristic, under the same date of observation. The distribution is shown in the table below;

**FIGURE 8: Summary of the Dataset Used**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 369 entries, 0 to 368
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Date                   369 non-null   object
1   Surface_Area(Km2)     369 non-null   float64
2   Water_Level(m)        369 non-null   float64
3   Rainfall(mm)          360 non-null   float64
dtypes: float64(3), object(1)
memory usage: 11.7+ KB
```

In [3]: `Data.describe()`

Out[3]:

	Surface_Area(Km2)	Water_Level(m)	Rainfall(mm)
<b>count</b>	369.000000	369.000000	360.000000
<b>mean</b>	172.152358	976.240393	85.142889
<b>std</b>	31.998955	3.132684	60.672630
<b>min</b>	110.750000	970.884000	0.090000
<b>25%</b>	133.860000	972.387000	40.107500
<b>50%</b>	185.780000	977.316000	75.510000
<b>75%</b>	194.200000	978.559000	116.740000
<b>max</b>	230.800000	982.304000	321.460000

## 4.3 Research Findings

### 4.3.1 Objective one Findings (Factors leading to spatial-temporal (2002, 2009, 2016, and 2023) change in the Lake Baringo region)

The spatial- temporal changes in the Lake Baringo region can be attributed to a number of factors that relate to climate change, below is the characteristics of the LULC;

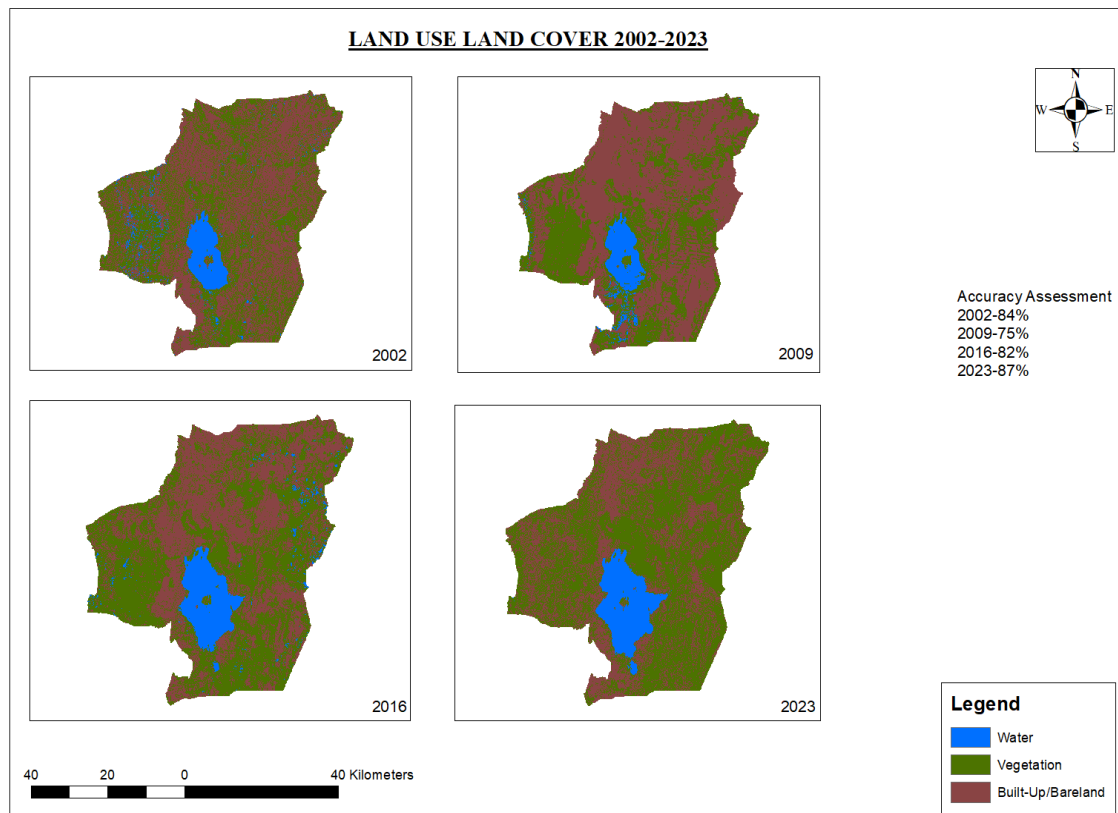
The lake area has been increasing over the years as displayed by the maps (Increase in precipitation led to increase in sedimentation or erosion carried by the rivers and deposited at the bottom of the lake hence the lake becoming shallow and expanding, erosion is brought

about by expansion and contraction of rocks and soils which disintegrate when coming into contact with water). Vegetation has been Increasing also over the years, this can be attributed to rainfall and climate change (The Landsat satellite images were of the same time period that is they were of different years but same month of June hence there was increased rainfall that led to increase in vegetation growth). Bare land/ Built-up area has been on the decrease (this is inversely proportional to vegetation behavior that is when vegetation increases Bare land/Built-up decreases and vice versa)

Accuracy assessment is a crucial step in evaluating the quality and reliability of land use/land cover (LULC) classification results. It helps in understanding the level of agreement between the classified map and the ground truth data. The accuracies of the images were computed as displayed in figure 9.



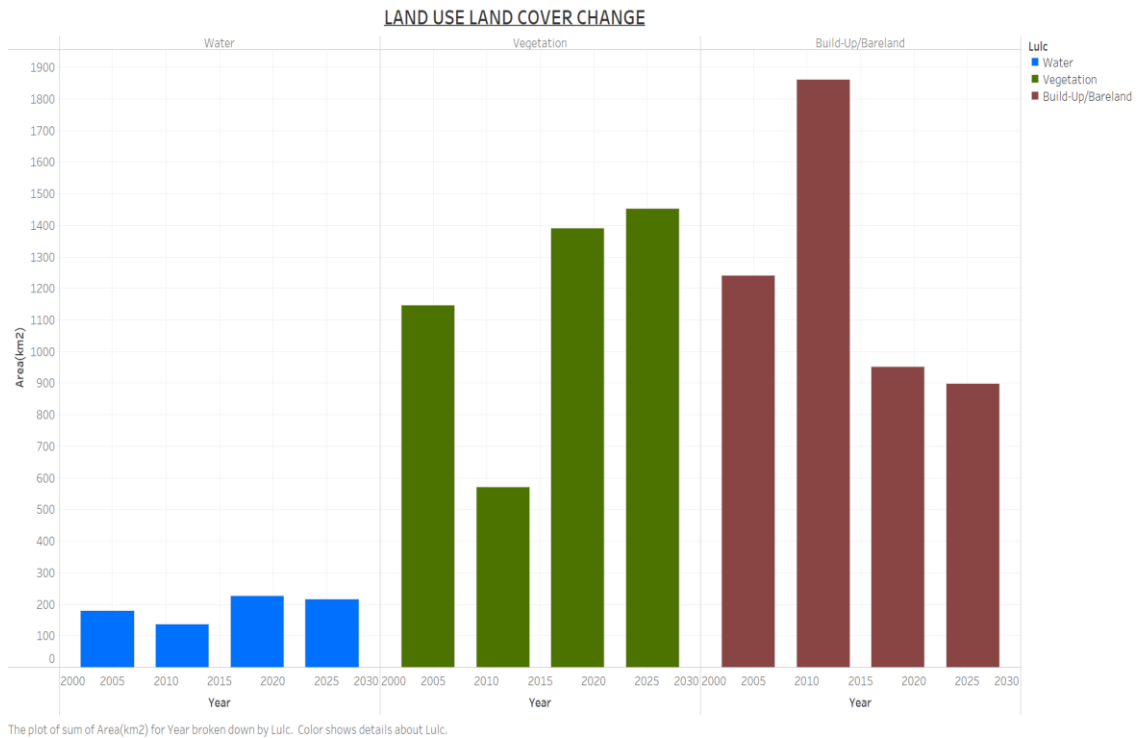
**FIGURE 9: Spatial-temporal changes in LULC in the Lake Baringo and Neighbouring Area**



As seen in figure 10 the year 2009 is special simply because it was characterized by severe drought. In a report by the government of Kenya (Kenya Post-Disaster Needs Assessment (PDNA), 2008) An analysis of the length, intensity, and spatial characteristics of the drought using data from the Intergovernmental Authority for Development (IGAD) Climate Prediction and Applications Centre (ICPAC) and the Kenya Meteorological Service revealed evidence of a drought that occurred in Kenya from 2008 to 2011, with varying intensities across geographies and time. This period of persistently low rainfall constituted a drought in the following ways: a meteorological drought due to lower-than-normal precipitation duration and intensities at different times; an agricultural drought due to insufficient soil moisture to meet the needs of the nation's various crops; a hydrological drought due to deficiencies in surface and groundwater supplies over time; a socio-economic drought.

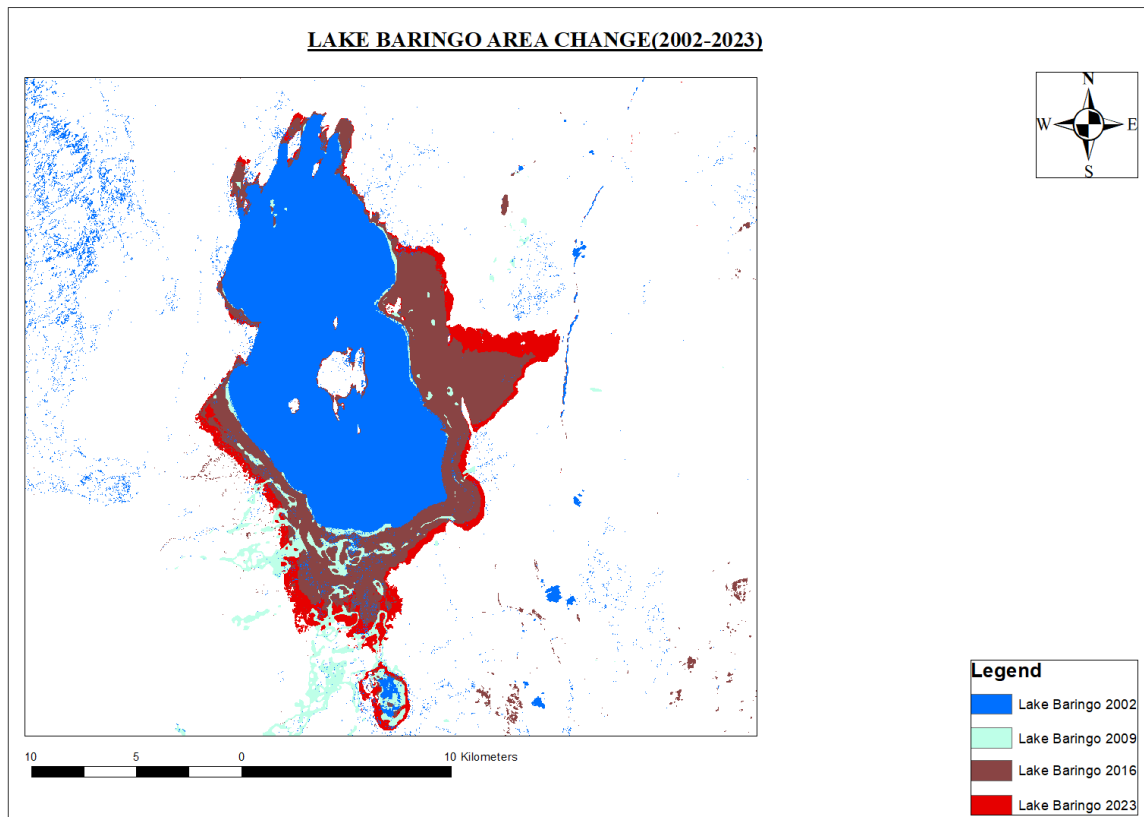
Severe water shortages have an impact on people's health, happiness, and standard of living in places all around the nation. These results led to the consideration of a drought period from 2008 to 2011 during the exercise and analysis.

**FIGURE 10: LULC Change Summary**



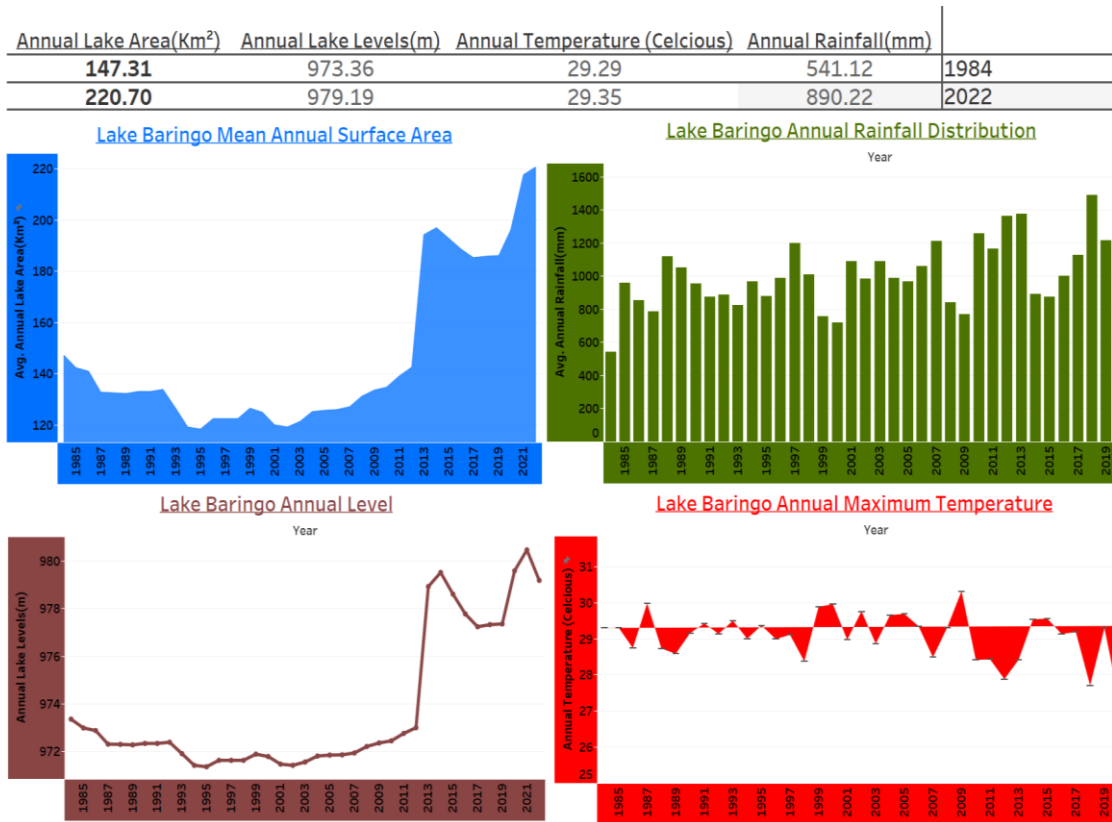
There was no doubt that lake Baringo had expanded over the years, Figure 9 below is a clear indication of the statement, The blue color shows the size of lake Baringo in the year 2002, the average size in that year was 117Km<sup>2</sup> .The cyan color shows the lake size in the year 2009 which was 142Km<sup>2</sup> .The brown color shows the lake area in the year 2016 with an average size of 189Km<sup>2</sup> .Lastly the red color shows Lake Baringo in the current state which is the year 2023 up to the month of August with an average area of 214 Km<sup>2</sup> .

**FIGURE 11: Visualization of Lake Baringo area change (2002 to 2023)**



From the above diagram it can be seen that the growth of Lake Baringo occurs mostly on the Southern and the Eastern side. This is mainly because many rivers that feed Lake Baringo are located in these areas hence the numerous erosion particles are dumped on these sides. Over the years as displayed below the lake Surface area has been increasing tremendously from 2003 to date. Rainfall has been fluctuating but from 2010 it has been on a high rate. Lake level started rising steadily from 2012 to date. Temperature has been fluctuating over the years but the change is on the duration of the high temperatures.

**FIGURE 12: Characteristics of lake Baringo Surface area, Rainfall, Lake Level and Temperature**



The Lake Baringo Surface area has been on the rise due to increased rainfall, The rainfall due to its increased magnitude has carried too much soil deposits which led to dumping in the lake reservoir, this in turn led to continuous rising of the lake level. Increased rainfall in terms of LULC change led to increased vegetation cover in the lake Baringo and its environs. The increase in vegetation absorbed the land that was bare and hence the reduced barehand. Increased lake surface area reduced some of the area occupied by humans and some of the area that was initially bare. Across the LULC change the year between 2009 to 2014 was characterized by a major drought which led to reduced water cover, reduced vegetation cover and increased barehand cover.

#### **4.3.2 Objective two Findings (To develop a time-series model (LSTM/GRU) for forecasting the area growth of Lake Baringo.**

When creating the model, the distribution of the data had to be understood first, Outliers had to be removed and others transformed, missing values had to be removed or to be filled. The unimportant features on the dataset had to be removed. The dataset was then split into two that is train set (75%) and test set(25%). The test set was to be used at a later stage during the testing of the developed models. The LSTM and GRU algorithms were then used to create the models as displayed in the model summaries below.

**FIGURE 13: LSTM Model Summary**

```
Single LSTM with hidden Dense...
Model: "sequential"

Layer (type)                Output Shape                Param #
=====
lstm (LSTM)                  (None, 80)                  26240
dense (Dense)                 (None, 40)                  3240
dense_1 (Dense)              (None, 1)                   41
=====
Total params: 29,521
Trainable params: 29,521
Non-trainable params: 0
Train...
```

**FIGURE 14: GRU Model Summary**

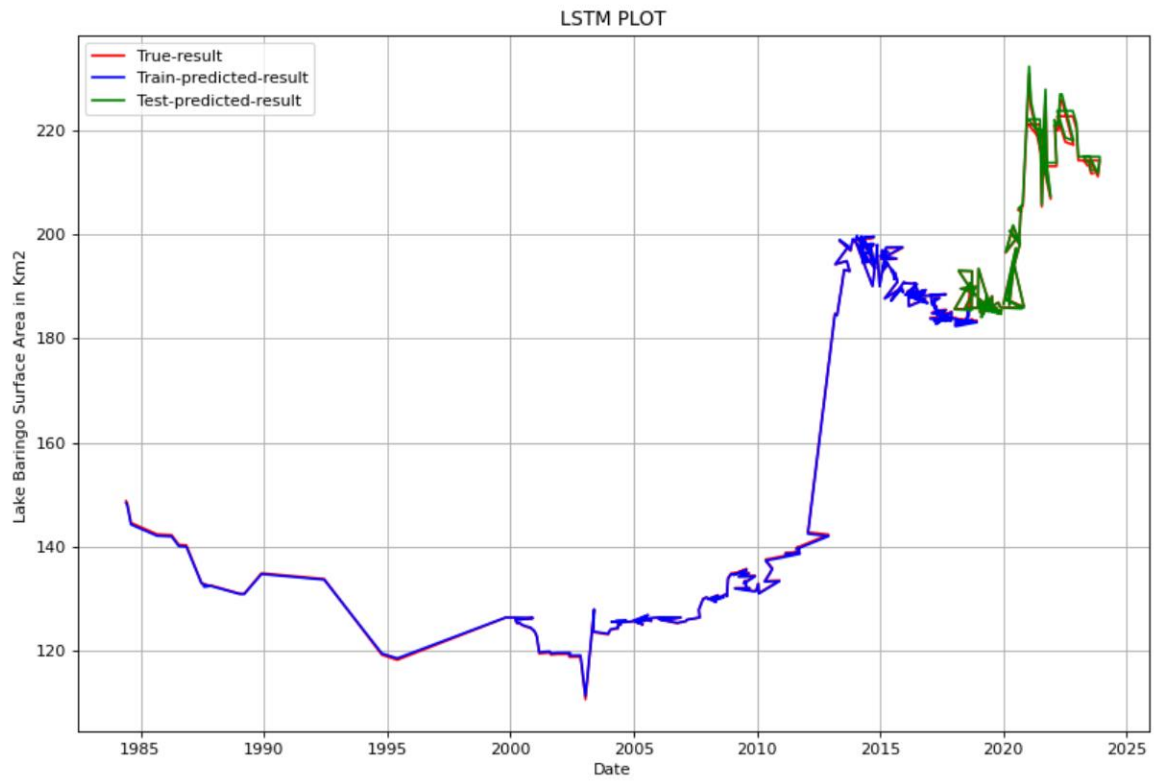
```
Single GRU with hidden Dense...
Model: "sequential_1"

Layer (type)                Output Shape                Param #
=====
gru (GRU)                    (None, 80)                  19920
dense_2 (Dense)              (None, 40)                  3240
dense_3 (Dense)              (None, 1)                   41
=====
Total params: 23,201
Trainable params: 23,201
Non-trainable params: 0
Train...
```

To see the relationship between the true data, predicted trained data and the predicted test data for both the models plots were created to visualize the outcome and as displayed

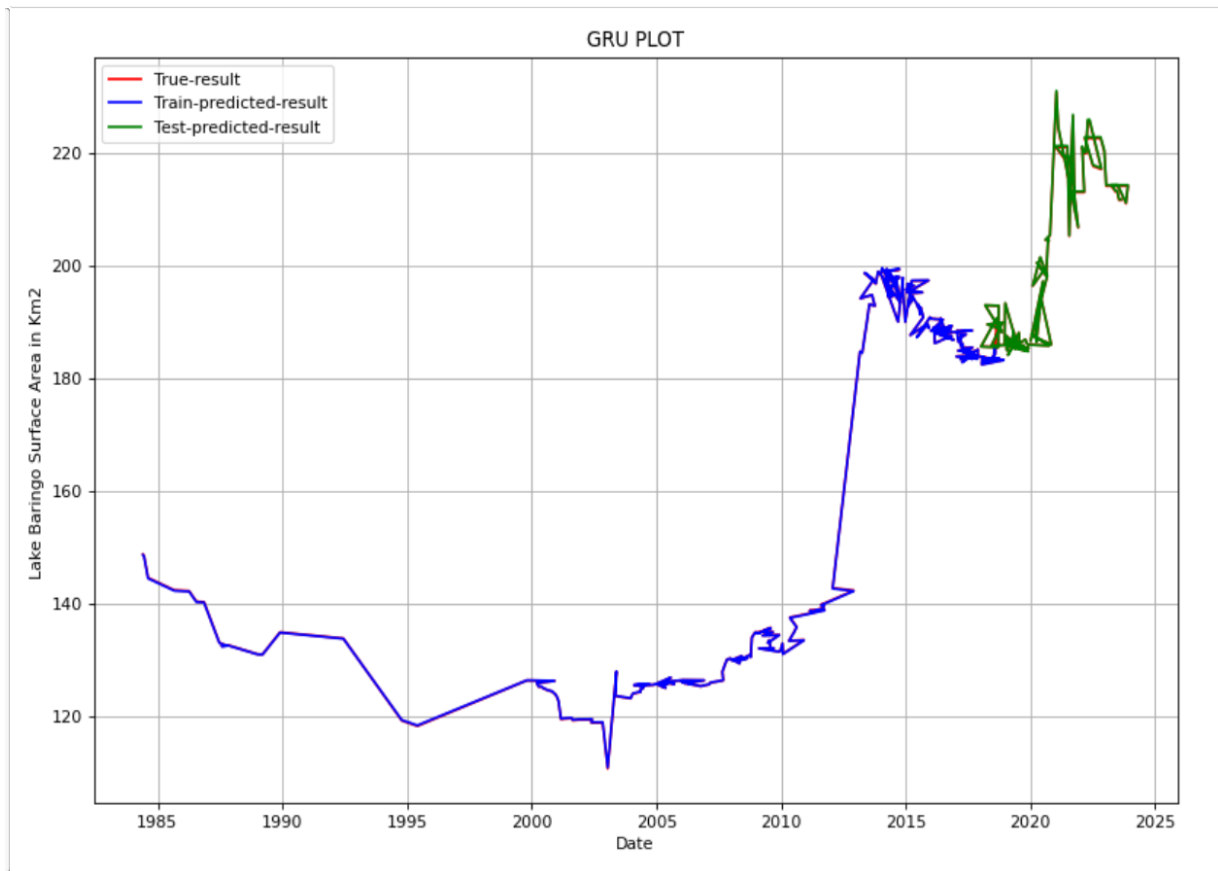
below, it could be seen that it was almost impossible to determine the difference between the three in GRU model.

**FIGURE 15: Baseline dataset against train and test Predicted Data by LSTM Model**





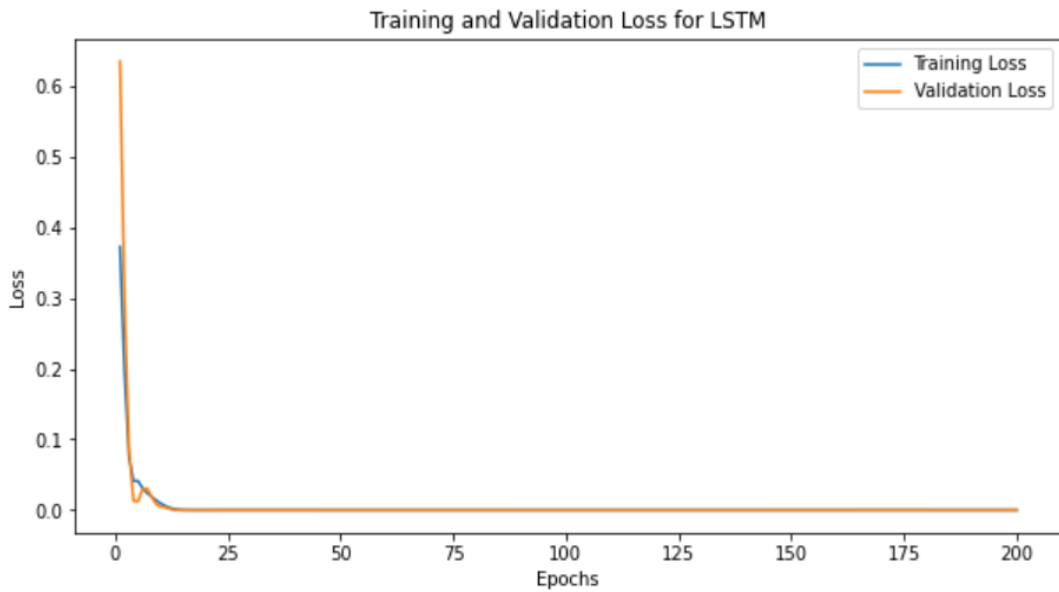
**FIGURE 16: Baseline Dataset Against Train and Test Predicted Data by GRU Model**



For time series data, the plot of true values and predicted values against time.

Interpretation: Visualization of the predicted values follow the same patterns over time as the true values. Differences can indicate the model's ability to capture trends and seasonality.

**FIGURE 17: Training and Validation Loss Curves for LSTM Model**



**FIGURE 18: Training and Validation Loss Curves for GRU Model**



Training Loss Curve: Training Loss: During the training process, the model's parameters (weights and biases) are adjusted to minimize a loss function. The training loss represents a measure of how well the model is performing on the training data. It quantifies the

difference between the model's predictions and the actual target values in the training set.

**Training Loss Curve:** This curve shows how the training loss changes over epochs or iterations during the training process. An epoch is one complete pass through the entire training dataset. As training progresses, the model tries to minimize the training loss. In the absence of overfitting or underfitting, the training loss should generally decrease over epochs.

**Validation Loss Curve: Validation Data:** A portion of the dataset, called the validation set, is set aside and not used during training. After each epoch, the model's performance is evaluated on this validation set. The validation set helps assess how well the model generalizes to unseen data.

**Validation Loss:** Similar to training loss, validation loss measures how well the model is performing, but it does so on the validation dataset. It indicates how well the model is likely to perform on new, unseen data.

**Validation Loss Curve:** This curve shows how the validation loss changes over epochs. It provides insights into the model's generalization performance. Ideally, both training and validation loss should decrease simultaneously. However, if the training loss continues to decrease while the validation loss starts increasing, it might indicate overfitting. Overfitting occurs when the model learns to memorize the training data instead of generalizing to new, unseen data.

**Interpreting the Curves:**

- Decreasing Training Loss:** Indicates that the model is learning from the data. A steady decrease signifies effective learning.
- Decreasing Validation Loss:** Shows that the model is generalizing well to unseen data. This is a positive sign of the model's performance.
- Increasing Validation Loss while Training Loss Decreases:** Suggests overfitting. The model is becoming too specialized to the training data and may not perform well on new data.

Monitoring these loss curves is crucial during the training process. Techniques like early stopping can be employed, where training is halted when the validation loss starts to increase, preventing overfitting and ensuring the model generalizes well to new data.

The training performance was impressive as shown on the above training and validation loss curve. The models did learn the data well. the models were not overfit (performed well on training data but poorly on unseen data) nor underfit (performed poorly on both training and unseen data).

#### **4.3.3 Objective three Findings (To test and evaluate the developed model)**

The two time series algorithms were used to train the dataset and test their efficiency on the test data. This helped in the determination of the best algorithm to adopt when creating the model to be used in the forecasting of the lake growth. The efficient algorithm in accomplishment of the objective was one that; takes a short training time to train the model, takes a short inference time meaning it is able to predict within the short time possible, uses minimum numbers of parameters to represent complex patterns in the data. Efficiency of the machine learning algorithm also depends on the quality of accuracy.

**Mean Squared Error (MSE):** MSE measures the average of the squared differences between predicted and actual values. Interpretation: The smaller the MSE, the better the model is at predicting the target variable. A MSE of 0 means the model perfectly predicts the data. However, MSE is not in the original unit of the target variable, so it might be harder to interpret in real-world terms.

**Root Mean Squared Error (RMSE):** RMSE is the square root of MSE and is in the same unit as the target variable. Interpretation: RMSE provides a more interpretable measure of the average error. Like MSE, lower RMSE values indicate a better fit. It's useful for understanding the typical error magnitude in the same unit as the target variable. It penalizes large errors more significantly than small errors due to the square root operation.

**FIGURE 19: Time taken by LSTM to Predict Train and Test Data**

```
9/9 [=====] - 0s 14ms/step  
3/3 [=====] - 0s 23ms/step
```

**FIGURE 20: Accuracy Measure of the LSTM Model**

```
Train Score: 0.01 MSE  
Test Score: 0.10 MSE  
Train Score: 0.09 RMSE  
Test Score: 0.31 RMSE
```

**FIGURE 21: Time Taken by GRU to Predict Train and Test Data**

```
9/9 [=====] - 0s 10ms/step  
3/3 [=====] - 0s 15ms/step
```

**FIGURE 22: Accuracy Measures of GRU Model**

```
Train Score: 0.00 MSE  
Test Score: 0.02 MSE  
Train Score: 0.05 RMSE  
Test Score: 0.13 RMSE
```

The time taken by the GRU model to predict the train and test data was shorter than that taken by the LSTM model, this showed that GRU algorithm is more efficient than LSTM algorithm in the training of the dataset. Based on the Mean squared error and the Root mean squared error, GRU algorithm proved to have outperformed LSTM algorithm in the training of the dataset. GRU model was a better model in the prediction of lake area growth

**FIGURE 23: Actual Surface Area Vs GRU Model Predicted Surface Area**

Surface_Area(Km2)		Predicted_Value	
Date		Date	
2023-04-05	214.20	2023-04-05	214.319687
2023-05-31	213.18	2023-05-31	213.294327
2023-06-27	213.18	2023-06-27	213.294327
2023-07-24	211.67	2023-07-24	211.776306
2023-08-20	211.67	2023-08-20	211.776306

#### 4.4 Discussion of the Research Findings

This section discusses the results of this research, as presented in chapter 4.3. The findings are discussed in relation to the existing literature. Moreover, the limitations of the current research are addressed and possible solutions are offered.

From the maps prepared from the satellite images it could be seen that The lake area has been increasing due to the increase in rainfall and the consequent increase in sedimentation, This is due to the increase of erosion materials that are deposited in the floor of the reservoir by the surface runoff water. Vegetation increase in the area of Lake Baringo and its neighborhood can be attributed to increased rainfall or the rainy season when the image was captured, The Bare land and Built-up has been on the decline due to the increase in the vegetation cover,

There is variation in the year 2009 between the three classes since that year was characterized by major drought. Water cover in the area decreased by 15.7% between 2002 and 2009, then there was an increase of 100% between the year 2009 and 2016 and a minor decline of 8% between the years 2016 and 2023. Vegetation cover had a decrease of 47.3% between the years 2002 and 2009 like the water class, there was 141% increase in vegetation cover between the years 2009 and 2016 and between the years 2016 and 2023 there was a slight increase of 3.6%. On the Build-up/Bare land class there was a sharp increase of 45.8% between the years 2002 and 2009, then there was a decrease of 48.6% between the years 2009 and 2016 and a further decrease between the years 2016 and 2023 of 5.6%. From the analysis it can be seen that Lake Baringo has been increasing gradually from 2002 to 2023 with percentages of 13.12%, 40.15% and 13.09% in the years 2009, 2016 and 2023.

This study constructed and utilized a variant of Recurrent Neural Network Specifically Gated Recurrent Unit and Long Short-Term Memory to come up with a predictive models to show the area of Lake Baringo in the later dates. Due to the nature of the expected outcome being numerical, the study used time series algorithms which were then evaluated using the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE) metrics. Mean square error of 0 for training data and Root Mean Squared Error of 0.05 was achieved on the model developed using GRU, LSTM on the other hand achieved training score of 0.01(MSE) and 0.09(RMSE). MSE measures the average of the squared errors, giving higher weight to large errors. Consequently, MSE tends to penalize large errors more severely than small errors. A lower MSE value indicates better model performance. It is not in the same unit as the target variable, as it is squared. On another hand RMSE provides an interpretable measure of the average error magnitude. It is in the same unit as the target variable, making it easier to understand in the context of the problem. Like MSE, lower RMSE values indicate better model

performance. Based on the test data LSTM achieved 0.10 (MSE) and 0.31(RMSE) while GRU achieved 0.02 (MSE) and 0.13 (RMSE)

Both MSE and RMSE quantify the average discrepancy between predicted and actual values. RMSE is preferred when you want a metric that is in the same unit as the target variable, making it easier to communicate the error magnitude to stakeholders or to compare it with the scale of the data. However, if you want a metric that penalizes large errors more severely, you might choose MSE. The choice between MSE and RMSE depends on the specific context of the problem and the preference for interpretability in the same unit as the target variable. The values of the predicted test data were lower than the predicted train data because the train data had been seen before by the model while the predicted test data had not been seen before by the model, since the actual data was divided into train and test at the ratio of 75% to 25% respectively for both models.

According to the research (Cahuantzi et al., 2023) it is determined that by examining the intricacy of the string sequences that recurrent neural networks (RNNs) can memorize, they researched to better understand their design. Symbolic sequences with varying levels of complexity were created to mimic RNN training and examine parameter configurations in order to assess the network's capacity for inference and learning. They made a comparison between gated recurrent units (GRUs) and Long Short-Term Memory (LSTM) networks. They discovered that when the training period is limited, an increase in RNN depth does not always translate into an improvement in remembering ability. Their findings also suggested that two of the most crucial hyper-parameters to adjust are the learning rate and the number of units per layer. In general, GRUs performed better than LSTM networks on low-complexity sequences, whereas LSTMs performed better on high-complexity sequences.



These findings of this study corroborate with those of the (Azad et al., 2022) studies that suggested using the hybrid model of seasonal autoregressive integrated moving average and artificial neural networks, or SARIMA-ANN. The models were developed and tested using average monthly Reservoir Water Level (RWL) data from January 2004 to November 2020. Every model's performance was assessed using a number of model assessment criteria. The results demonstrated that, while taking into account all performance criteria for reservoir RWL prediction, the SARIMA-ANN hybrid model performed better than the other models (the individual SARIMA model and the individual ANN model). As a result, this study unequivocally demonstrated that the SARIMA-ANN hybrid model might be a practical choice for precisely predicting reservoir water level.

We can ascertain from the study (Hu et al., 2021) that the region encompassed by the four lakes in the Zhuonai Lake-Salt Lake the Zhuonai Lake, Kusai Lake, Heidonor Lake, and Salt Lake basin in Hoh Xil has undergone substantial alteration in the last thirty years. The areal parameters of four lakes were extracted in this work using remote sensing picture data collected between 1989 and 2018 using the Landsat thematic mapper, enhanced thematic mapper plus, and operational land imager. Climate change has caused the four lakes' combined area to grow by 18% over the last 30 years. Trend analysis was conducted using interpolated values derived from 28 meteorological stations within the basin. The variations in the basin's yearly lake evaporation were examined using a single-layer lake evaporation model. The yearly evaporation of lakes increased somewhat between 1989 and 1995, then by a significant decline between 1995 and 2018. The basin's yearly evaporation varied from 615.37 to 921.66 mm from 1989 to 2018, with a mean of 769.73 mm. To calculate how much precipitation and evaporation would alter lake volumes, a mass balance model was created. The four lakes are expanding because of the rise in precipitation and fall in yearly lake evaporation. The primary cause of the changes in the lake areas is evaporation from the lakes. This theory focused on climate

change, although it is also possible that the sedimentation case at Lake Baringo contributed to the case.

From the findings Lake Baringo has been expanding and the expansion can be attributed to factors such as; climate change (Decrease in the evaporation rate and the increase in Precipitation), sedimentation (Due to human factors such as deforestation there is a lot of erosion when there is a heavy downpour, the soil and rock particles are carried by the runoff water and deposited in the lake which makes Lake basin to be shallow and hence the expansion of the area. So long as the climate changes there might be further increase in the lake area. GRU algorithm also outperforms LSTM algorithm in the development of area prediction models.

## **CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS**

### **5.1 Introduction**

Predicting the lake area in future years is an important factor as it helps save on a lot of costs and lives. The main goal of this study was to develop a Time series forecasting model to help in the prediction of lake expansion. This section provides conclusion derived from the study and descriptions of how the study extended previous studies, models, or frameworks. This chapter presents the conclusions from the study findings, some of the limitations and challenges that the study experienced and the recommendations for policy areas of interest and future research.

### **5.2 Conclusions**

The conclusions of the study are drawn from the objectives since they were able to be achieved as discussed below;

#### **5.2.1 Objective One**

The following conclusion can be drawn to address the objective: to determine the factors leading to spatial-temporal (2002, 2009, 2016, and 2023) change in the Lake Baringo region. The research data both the classified satellite images, analyzed satellite data and the tabular data were studied and the similarity of the behavior of the output were used to determine the factors leading to spatial-temporal change in the region. The factors were identified as; Increased Rainfall in the Lake Baringo Region, Increased Lake level of Lake Baringo, Decrease or Increase of the different Land Use Land Cover zones in the Lake Baringo region and change of climatic factors in the area. It can therefore be concluded that these are the factors that should be looked into in order to prepare for further expansion of the lake area in the future.

### **5.2.2. Objective Two**

The second objective was to develop a time-series model for forecasting the area growth of Lake Baringo. This was achieved through understanding of the underlying problem, understanding of the data, preparation of the data, exploration of the data, development of the data mining model, evaluation of the built data mining model and deployment of the model as according to (Schneider et al., 2022).

It can also be concluded that the input variables used and the dataset were sufficient enough to come up with the predictive model. Training and validation loss curves are graphical representations used in machine learning to monitor the performance of a model during training. These curves are essential for understanding how well the model is learning from the data and whether it is overfitting or underfitting. The curves showed that the training of the data to form a model was according to the required standards.

### **5.2.3 Objective Three**

The third objective was to test and evaluate the developed model. The two algorithms were used to train the dataset and the evaluation was determined by checking the time taken by each of the algorithm to train the dataset whereby the model that took little time was found to be efficient, another comparison was based on the time taken by the model to predict an outcome whereby the model which predicted under a very short time was found to be efficient and the last way to compare the efficiency was the use of metrics which are mean squared error and root mean squared error, for a model to be considered better it should have the two values ranging to less than one. Based on the three parameters used to rank the better algorithm of the two it can be concluded that GRU algorithm performed better than LSTM algorithm since it took lesser time to train the dataset, it took less time to predict both the train and test datasets and it performed better than LSTM based on the performance metrics (MSE and RMSE).The

test was carried out using the 25% data set aside in the model development stage. Both models produced outputs but it can be concluded that the prediction value by GRU was of higher accuracy than that of LSTM.

### **5.3 Contributions of the study**

The study has significantly contributed to natural disaster management field by demonstrating how data mining can be applied in the Lake area prediction field for purposes of flood prediction with time series prediction models that consider time and the available variables that are related to surface area to produce accurate predictions. The study has also helped to visualize the lake growth over the years and to highlight features that lead to the Lake expansion, this is likely to improve the awareness and preparedness of the relevant parties on the likely occurrence of the floods caused by lake swelling and expansion as the measures shall be more target focused than general.

The study through the limitations of the data was able to show how relevant and crucial there is need to record data on a frequent basis since the data had some missing values and observations which are important on training of the machine learning time series algorithms in order to achieve accurate predictions from large amount of data. The study has also demonstrated how continuous improvement can be achieved in computer field through diving into the field of disaster management and other fields in general even to an extent of collaborating with researchers in other fields.

From this study researchers are able to appreciate the fact that python as a programming language is well equipped to perform data analysis from end to end. It is well equipped with libraries that are suitable to handle machine learning tasks like keras and TensorFlow. The libraries together with other python libraries can handle tasks from data preprocessing, modelling, evaluation and visualization of the data. This one stop shop eliminates the need to

migrate from one platform to the other in different stages of analysis hence improving the overall productivity. The results from this study alongside with the study findings of other researchers can help policy implementors to come up with strategies to improve on how to handle and plan for the future.

#### **5.4 Limitations of the Study**

The study was only limited Lake Baringo as a case study and not all the Kenyan Lakes. This was based on the data availability since there was readily available data for the case study, Interdisciplinary Nature was also another reason for Choosing of Lake Baringo since other researchers had done some analysis of the trend of the lake behavior and there was need for expansion and continuation in the same line, Longitudinal studies was another factor for choosing of the study area as there was need to monitor data over a long period of time, policy and practice implications was another reason for choosing the area of study since the policy makers and the inhibitors and lastly the lake expansion prediction in Kenya is a Unique and a rare field.

Another limitation of the study was the availability of data, getting data for this type of research and the analysis was time consuming and expanding the area would take a lot of time, The data also was limited in temporal perspective whereby it was limited to the years between 1984 and 2023 with other years having more monthly coverage than other years.

The performance analysis of the model was also simply limited to the mean squared error and the Root Mean Squared Error metrics as the study was a regression problem other than classification problem where the accuracy metric could be used to evaluate the performance of the model.

## **5.5 Recommendations for future research**

Number of recommendations are suggested in this section based on the limitations of the study. Analysis should be performed on a larger study area such as all lakes within rift valley this can be made possible by the availability of data that has been recorded on a regular period, the study focus was mainly on univariate time-series, future work should be done on multivariate time-series or even combination of two or more machine learning algorithms. Data in this area of disaster management is not readily available and other data have to be created from the already existing data, recommendation is for researchers in Kenya to have a way of generating data from satellite images and other form of data and make them readily available in order for efficient analysis and research.

The study came up with a model that could only visualize a value for the case of lake area prediction, future study should focus on a way of generating future area of the lake visually in form of maps or other forms of visualizations, future studies could also use other different time series algorithms to compare the output with the GRU developed model for Lake Baringo.

Lake Baringo has been growing rapidly over the years and in this study, it was discovered that the lake growth is mainly on the southern and Eastern parts. policy makers and stake holders could control and prevent further losses by zoning out the area which is likely to be occupied by the lake as early as possible through the use of analysis of the previous lake growth and a prediction for the future lake growth

## REFERENCES

- Adminusr (2019) Kenya Population and Housing Census Reports - Kenya National Bureau of Statistics.” Kenya National Bureau of Statistics, 21 Feb. 2020, [www.knbs.or.ke/2019-kenya-population-and-housing-census-reports/](http://www.knbs.or.ke/2019-kenya-population-and-housing-census-reports/).
- API Reference | Google Earth Engine. (2023). Google for Developers. Retrieved October 17, 2023, from <https://developers.google.com/earth-engine/apidocs>
- Avashia, V., & Garg, A. (2020). Implications of land use transitions and climate change on local flooding in urban areas: An assessment of 42 Indian cities. *Land Use Policy*.
- Azad, A. S., Sokkalingam, R., Daud, H., Adhikary, S. K., Khurshid, H., Mazlan, S. N. A., & Rabbani, M. B. A. (2022). Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study. *Sustainability*, 14(3), 1843. <https://doi.org/10.3390/su14031843>
- Baringo County (2023) County Trak Kenya. [countytrak.infotrakresearch.com/baringo-county/](http://countytrak.infotrakresearch.com/baringo-county/). Accessed 17 Oct. 2023.
- Butler, R. (2019). The Impact of Deforestation. From Mongabay: <https://rainforests.mongabay.com/09-consequences-of-deforestation.html>
- Cahuantzi, R., Chen, X., & Güttel, S. (2023). A Comparison of LSTM and GRU Networks for Learning Symbolic Sequences. *Lecture Notes in Networks and Systems*, 771–785. [https://doi.org/10.1007/978-3-031-37963-5\\_53](https://doi.org/10.1007/978-3-031-37963-5_53)
- Daniel Muia, M. G. (2021). Effects of Extreme Flooding of Lake Baringo on Livelihoods of Communities Lining around the Lake. *Applied Sociology*, 404-414.
- Darem, A. A., Alhashmi, A. A., Almadani, A. M., Alanazi, A. K., & Sutantra, G. A. (2023). Development of a map for land use and land cover classification of the Northern Border Region using remote sensing and GIS. *The Egyptian Journal of Remote*



- Sensing and Space Science, 26(2), 341–350.  
<https://doi.org/10.1016/j.ejrs.2023.04.005>
- DBD, R. (2020). KDD Process/Overview. Uregina.ca.  
[https://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1\\_kdd.html](https://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html)
- Dong kun, L., Kyungmin, K., & Haekyung, P. (2019). Prediction of Severe Drought Area Based on Random Forest: Using Satellite Image and Topography Data. *Water*, 705.
- EOSDA LandViewer: Find And Download Satellite Imagery. (2023, 4 17). From eos.com:  
<https://eos.com/find-satellite>
- Hamid, H. A., Wenlong, W., & Qiaomin, L. (2020). Environmental sensitivity of flash flood hazard using geospatial techniques. *Global Journal of Environmental Science and Management*, 31-46.
- Hernegger, M., Stecher, G., Schwatke, C., & Olang, L. (2021). Hydroclimatic analysis of rising water levels in the Great rift Valley Lakes of Kenya. *Journal of Hydrology: Regional Studies*, 36, 100857. <https://doi.org/10.1016/j.ejrh.2021.100857>
- Huda, J. J., Mohammed, H. A., Ghadah, H. M., & Qayssar, M. A. (2022). Monitoring and evaluation Al-Razzaza lake changes in Iraq using GIS and remote sensing technology. *The Egyptian Journal of Remote Sensing and Space Science*.
- Hu, Z., Tan, D., Wen, X., Chen, B., & Shen, D. (2021). Investigation of dynamic lake changes in Zhuonai Lake–Salt Lake Basin, Hoh Xil, using remote sensing images in response to climate change (1989–2018). *Journal of Water and Climate Change*.  
<https://doi.org/10.2166/wcc.2021.285>
- Jodha, R. (2023, June 12). KDD in Data Mining. *Scaler Topics*.  
<https://www.scaler.com/topics/kdd-in-data-mining/>
- Kenya Post-Disaster Needs Assessment (PDNA). (2008).  
<https://www.gfdr.org/sites/default/files/publication/pda-2011-kenya.pdf>

- Lake Baringo Flood Resilience Project. (2023, 02 01). From PlanAdapt: <https://www.plan-adapt.org/projects/lake-baringo-flood-resilience-project/>
- Lim, B., & Zohren, S. (2021). Time Series Forecasting With Deep Learning: A Survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200209. <https://doi.org/10.1098/rsta.2020.0209>
- Lindemann, B., Müller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. *Procedia CIRP*, 99, 650–655. <https://doi.org/10.1016/j.procir.2021.03.088>
- Mahjoub, S., Chrifi-Alaoui, L., Marhic, B., & Delahoche, L. (2022). Predicting Energy Consumption Using LSTM, Multi-Layer GRU and Drop-GRU Neural Networks. *Sensors*, 22(11), 4062. <https://doi.org/10.3390/s22114062>
- Manoj Chhetri 1, S. K.-G. (2020). Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan. *Remote Sensing*.
- Mehmood, M., Shahzad, A., Zafar, B., Shabbir, A., Ali, N., & Ahmad, A. (2022). Remote Sensing Image Classification: A Comprehensive Review and Applications. *Mathematical Problems in Engineering*, 1-24.
- Moskolai, W. R., Abdou, W., Dipanda, A., & Kolyang. (2021). Application of Deep Learning Architectures for Satellite Image Time Series Prediction: A Review. *Remote Sensing*, 13(23), 4822. <https://doi.org/10.3390/rs13234822>
- Muhammad, R., Zhang, W., Abbas, Z., Guo, F., & Gwiazdzinski, L. (2022). Spatiotemporal Change Analysis and Prediction of Future Land Use and Land Cover Changes Using QGIS MOLUSCE Plugin and Remote Sensing Big Data: A Case Study of Linyi, China. *Land*, 419.

- Muita, R. (2021). Assessment of Rising Water Levels of Rift Valley Lakes in Kenya: The Role of Meteorological Factors. *Environmental Sciences and Ecology: Current Research (ESECR)*, 1-9.
- Mulama, R. Y., & Ondieki, J. O. (2023). Assessment of Spatial Expansion of Rift Valley Lakes Using Satellite Data. *Advances in Remote Sensing*, 12(3), 88–98. <https://doi.org/10.4236/ars.2023.123005>
- Lawrence, D., Coe, M., Walker, W., Verchot, L., & Vandecar, K. (2022). The Unseen Effects of Deforestation: Biophysical Effects on Climate. *Frontiers in Forests and Global Change*, 5(756115). <https://doi.org/10.3389/ffgc.2022.756115>
- Nausheen, M., Ali Iqtadar, M., Sohail, A., Muhammad Ameer Nawaz, A., Ali, M., & Javid, K. (2021). Effects of climatic factors on the sedimentation trends of Tarbela Reservoir, Pakistan. *SN Applied Sciences*.
- Neo, W., Bradley, G., Xue, B., & Shawn, O. (2020, 1 22). Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case. From arXiv.org: <https://arxiv.org/abs/2001.08317v1>
- Nyakundi, V., Plal, E., & Kandu, K. (2023). The Risk of Flooding to Architecture and Infrastructure amidst a Changing Climate in Lake Baringo, Kenya. *American Journal of Climate Change*, 80-99.
- Omweno, J., Opiyo, s., Omondi, A., & Zablou, W. (2021). Natural and anthropogenic changes threatening the ecological and limnological integrity of Lake Baringo, Kenya: A Review. *PAN AFRICA SCIENCE JOURNAL*, 103-121.
- Park, K., Jung, Y., Seong, Y., & Lee, S. (2022). Development of Deep Learning Models to Improve the Accuracy of Water Levels Time Series Prediction through Multivariate Hydrological Data. *Water*, 469.

- Radman, A., Akhoondzadeh, M., & Hosseiny, B. (2021). Integrating InSAR and deep-learning for modeling and predicting subsidence over the adjacent area of Lake Urmia, Iran. *GIScience & Remote Sensing*, 1413-1433.
- Ramdani, F., Setiawan, B., Rusydi, A., & Furqon, M. (2021, March 9). An Artificial Neural Network Approach to Predict the Future Land Use Land Cover of Great Malang Region, Indonesia. <https://doi.org/10.20944/preprints202103.0247.v1>
- Read, J. S. (2022). Process-Guided Deep Learning Predictions of Lake Water Temperature. *Water Resources Research*, 9173-9190.
- Schneider, J., Seidel, S., Basalla, M., & vom Brocke, J. (2022). Reuse, Reduce, Support: Design Principles for Green Data Mining. *Business & Information Systems Engineering*. <https://doi.org/10.1007/s12599-022-00780-w>
- Soltani, K., Amiri, A., Zeynoddin, M., Ebtehaj, I., Bahram Gharabaghi, & Hossein Bonakdari. (2020). Forecasting monthly fluctuations of lake surface areas using remote sensing techniques and novel machine learning methods. *Theoretical and Applied Climatology*, 143(1-2), 713–735. <https://doi.org/10.1007/s00704-020-03419-6>
- Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L., & Cheng, Z. (2020). Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *Journal of Petroleum Science and Engineering*, 186, 106682. <https://doi.org/10.1016/j.petrol.2019.106682>
- Tabari, H. (2020). Author Correction: Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*.
- Vali, A., Comai, S., & Matteucci, M. (2020). Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sensing*, 12(15), 2495. <https://doi.org/10.3390/rs12152495>

- Wahap, N. A., & Shafri, H. Z. M. (2020). Utilization of Google Earth Engine (GEE) for land cover monitoring over Klang Valley, Malaysia. *IOP Conference Series: Earth and Environmental Science*, 540, 012003. <https://doi.org/10.1088/1755-1315/540/1/012003>
- Wang, S., Gebru, B., Lamchin, M., Kayastha, R., & Lee, W.-K. (2020). Land Use and Land Cover Change Detection and Prediction in the Kathmandu District of Nepal Using Remote Sensing and GIS. *Sustainability*, 3925.
- Will, D. (2022, 8 29). Floods and droughts could cost the global economy \$5.6 trillion by 2050, report says. From Fortune: <https://fortune.com/2022/08/29/climate-change-costs-floods-droughts-could-cost-global-economy-5-trillion-by-2050/>
- Wu, Z., Lu, C., Sun, Q., Lu, W., He, X., Qin, T., . . . Wu, C. (2023). Predicting Groundwater Level Based on Machine Learning: A Case Study of the Hebei Plain. *Water*, 823.
- Xu, G., Cheng, Y., Liu, F., Ping, P., & Sun, J. (2019). A Water Level Prediction Model Based on ARIMA-RNN. *IEEE Xplore*.
- Xu, J., Wang, K., Lin, C., Xiao, L., Huang, X., & Zhang, Y. (2021b). FM-GRU: A Time Series Prediction Method for Water Quality Based on seq2seq Framework. *Water*, 13(8), 1031. <https://doi.org/10.3390/w13081031>
- Yang, Y., Wu, J., Miao, Y., Wang, X., Lan, X., & Zhang, Z. (2022). Lake Changes during the Past Five Decades in Central East Asia: Links with Climate Change and Climate Future Forecasting. *Water*, 14(22), 3661. <https://doi.org/10.3390/w14223661>
- Zhu, W., Zhao, S., Qiu, Z., He, N., Li, Y., Zou, Z., & Yang, F. (2022). Monitoring and Analysis of Water Level–Water Storage Capacity Changes in Ngoring Lake Based on Multisource Remote Sensing Data. *Water*, 14(14), 2272–2272. <https://doi.org/10.3390/w14142272>

## APPENDICES

### Appendix 1: Budget and Resources

The budget for the research is tabulated below

**TABLE 5: Estimated budget of the Study**

<b>Serial</b>	<b>Item</b>	<b>Cost (Ksh)</b>
1	laptop	80000/=
2	Stationery; Hard disk, Notebook, Pens	10000/=
3	Printing	5000/=
4	Binding	6000/=
5	Data Expenses	25000/=
<b>Total</b>		<b>126000/=</b>

Appendix 2: Schedule

FIGURE 24: Schedule of Work

