A LONG SHORT-TERM MEMORY (LSTM) NETWORK MODEL FOR PREDICTING WATER CONSUMPTION IN RESIDENTIAL PROPERTIES USING SMART WATER METER DATA

BY

STANLEY MWAURA KAMAU

REG NO: 12/01296

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MASTER OF SCIENCE IN DATA AMAYLTICS IN THE SCHOOL OF TECHONOGY

AT KCA UNIVERSITY

2023

DECLARATION

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

Student Name: Stanley M Kamau Reg No: 12/01296

Sign: _____ Date: 5/10/2023

I do hereby confirm that I have examined the master's dissertation of

Stanley Mwaura Kamau

And have certified that all revisions that the dissertation panel and examiners recommended have been adequately addressed.

Date: 5/10/2023

X

Dr. Lucy Waruguru
mburul@kcau.ac.ke

Dr. Lucy Waruguru

A LONG SHORT-TERM MEMORY (LSTM) NETWORK MODEL FOR PREDICTING WATER CONSUMPTION IN RESIDENTIAL PROPERTIES USING SMART WATER METER DATA

ABSTRACT

Rapid urbanization in Kenya and the subsequent population increase have caused a severe imbalance between water demand and water availability. This imbalance poses serious challenges in managing water consumption in urban areas. Furthermore, water leakages and variable human activity generate non-linear patterns in domestic water consumption data which make traditional linear time series models such as autoregressive integrated moving average (ARIMA) ineffective. Using a case study research design with Nairobi City, the author developed a novel Long Short-Term Memory (LSTM) network model for predicting water demand through deep learning of smart water meters data. The model uses high frequency non-linear time series data collected between January and December 2022 from smart sensors within an Internet of Things (IoT) framework, alongside other information such as timestamp and temperature. Nine different variables were constructed from the study data and used to train and validate the LSTM network model for smart water meter data management. The model was then evaluated using root mean square error (RMSE) and the correlation coefficient. Although significant variation was observed in the daily and monthly patterns of domestic water consumption, the model outcomes were relatively accurate. LSTM generated values that mirrored observed values more closely than the ARIMA model. Evaluation metrics also indicated that LSTM had lower prediction errors. It is expected that the developed model will be generalizable for estimating future water consumption in other urban households in Kenya and other regions. The study is limited by a small sample dataset of 320 households and the lack of socio economic and demographic factors to determine water consumption. A more extensive study with multiple influencing factors is recommended to assist water authorities and service providers to properly distribute water, identify leakages, and take corrective actions to prevent degradation of the ecological environment.

Keywords: Smart water meters, domestic water consumption, LSTM network model, residential

ACKNOWLEDGEMENT

I thank the Almighty God for giving me the wisdom, courage and determination to execute this project. I also thank my family for their continuous support and perseverance as I spent much of my time doing the project work.

LIST OF ACRONYMS AND ABBREVIATIONS

CC	Correlation coefficient
DDD	Data-Driven Decision-making
EDA	Exploratory Data Analysis
IOT	Internet of Things
LSTM	Long Short-Term Memory
RMSE	Root mean square error.
RNN	Recurrent neural network
SWM	Smart water meter
WSN	Wireless sensor network

LIST OF DEFINITIONS

Smart water meter: Smart sensor combined with a communication system to collect and transmit real-time high frequency information on water usage.

Water demand: A measurement of total amount of water that end users utilize in a system.

End User: An individual user or an apartment that is linked with water usage in a system.

Water consumption: The utilization of water in a system, usually measured in liters.

Accumulated water consumption: The total amount of water that passes through the water meter during a specified time interval (e.g., liters/ month)

Individual/ Isolated metering: The measurement and monitoring of the water consumption per end user.

Deep learning: A sub-field of artificial intelligence that uses artificial neural networks to learn previous patterns and make future predictions.

Recurrent neural network (RNN): a category of deep learning artificial neural networks that is designed to process sequential data, such as time series or natural language processing data in order to generate patterns and insight.

Long Short-Term Memory (LSTM): An RNN architecture that is specifically designed to analyze sequential data.

Data-driven Decision-making: A system in which decision-making is highly influenced by data analysis as an alternative to personal experience.

Wireless sensor network: a cluster of electronic devices which communicate information about a specific event or phenomena remotely using wireless connections.

Internet of Things: A worldwide network of inter-connected objects which can be uniquely identified and addressed using standardized communication protocols.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1Background

Water is a critical resource for sustainability. Nevertheless, the number of fresh-water sources in Kenya, like in many African countries, is ranked among the lowest globally. Water availability has reduced by two-thirds in the past four decades and is projected to further reduce by 50% before 2050 (He et al., 2021). This dire situation has left millions of residents without access to reliable and clean water. The Kenyan governments faces many obstacles in ensuring sustainable access to quality water to meet growing demands. While most cities in the developed world have a water supply rate of more than 80%, the penetration rate in Nairobi and other developing world cities is less than half and variable depending on the extent of urbanization (Kim et al., 2022).

The complexity of water supply in Nairobi and many other African cities is further compounded by the aging infrastructures for water distribution that are prone to leaks and breakdowns. According to Water Services Regulatory Board, 52 percent of Kenya's water utility companies collectively lose an equivalent of 2 out of every 5 liters of treated and pumped water (Newsroom, 2020). This impacts water availability, especially in the urban areas. The estimated annual loss of between 8 and 27 billion Kenya Shillings also causes consumers to be billed more in a bid to recover the losses. Absence of real-time detection mechanisms aggravates this situation and leads to unavoidable depletion of water resources. Because of the need to establish an efficient water supply infrastructure, city administrations and researchers must increase efforts to ensure sustainable water supply through smart innovations.

1.1.1 Smart technologies

The concept of "smart cities" has recently gained wide acceptability. More and more of these smart cities are sprouting in Africa. A report on smart cities by Pelikh (2022) rated the "Silicon Valley of Africa" in Rwanda and Konza City in Kenya as the most popular African smart cities. The smart city is now classified as the future emerging market that will dominate the digital economy. Concurrently, smart technologies are now commonplace for monitoring water demand and supply following the advances in information and communications technology (ICT), the 4th industrial revolution, and Internet of Things (IoT, Shah, 2017).

IoT is a network that interconnects physical devices which are embedded with sensors, software and connectivity mechanisms to allow exchange of real-time high frequency data over the internet (Owen, 2023). IoT devices are usually installed in the pipeline, water meters, and other sections of the distribution channel to collect data on water flow, consumption, pressure and temperature. Useful insights can then be generated about consumption patterns and general state of the water supply infrastructure. Proper analysis of the smart water meter generated data can generate insight for confirming leakages, water usage costs and other details that are customer oriented (Kim et al., 2022). While adoption of IoT-enabled smart water meters has been rapid, modern analytics models of water consumption have not been widely adopted for effective management of water data in many urban areas.

1.1.2 Residential water consumption

Water usage patterns for residential areas are non-linear. Models that can properly reflect non-linear time series characteristics are increasingly being preferred for predicting water usage. For example, a recent study found that it is important to forecast future water demand for expanding existing water supply systems to generate

an optimum water consumption infrastructure based on accurate prediction (Velasco et al., 2022). The study compared output from artificial neural network (ANN) models with conventional ARIMA models to predict potential demand for water. The authors found out that the ANN models performed significantly better than the ARIMA models. These traditional models have low levels of accuracy for predicting customer specific usage of water in residential areas. Evidence from other previous studies shows that when water consumption is predicted using a single variable, errors in prediction can arise (Piasecki, et al., 2018). Artificial neural networks (ANN) and multiple regression models have been applied to predict water consumption in different cities of nations such as USA, Japan, Poland and Korea. A review of the models applied showed that when factors such as historical water consumption details and other day-to-day variables were applied the models yielded high precision levels (Krishnan et al., 2022). It is therefore necessary to consider multiple variables when modelling domestic water consumption for urban areas in the developing world using smart water meter data.

1.2Problem Statement

Most African countries have insufficient water reserves, and Kenya is no exception. According to recent statistics, Nairobi is among the cities in Africa that have the most unstable water supply (Makoni, 2021). Additionally, Nairobi experiences significant wastage of water because of unidentified leakages and illegal connections. The Kenyan government faces significant challenges in efficient management of water resources (Oiro et al., 2020). These challenges include the methods for accessing water, efficient water usage, and how to effectively determine consumption fees (Mulwa et al., 2021). Traditional techniques that have been adopted for estimating future water consumption generate minimal insights and make it difficult to anticipate and respond to variations in local demand.

The aging infrastructure for water distribution in Kenya is prone to leaks and breakdowns. The lack of effective monitoring and predictive maintenance causes repair efforts to be reactive and expensive. IoT-enabled devices such as smart water meters generate massive real-time high-frequency data on water consumption for each user. Analysis of such massive data using machine learning techniques can potentially ease the detection of leakages and errors in water consumption and advise measures for specific users or groups of users as well as for the entire water supply infrastructure. However, existing models of water management using artificial neural networks, recurrent neural networks, random forest, and support vector regressions, have focused on waste-water treatment and smart irrigation (Gaya et al., 2017; Phasinam, 2022; Roshni et al., 2022; Zanfei et al., 2022). Liu (2022) developed an IoT and data-analytics-based intelligent water management system using time series forecasting with hardware-based wireless sensor network monitoring model for measuring data in short-distance transmission of the agricultural environment. This study showed that such models can produce accurate results by using different parameters. However, the study applied only 78 statistical data attributes. In Kenya, only one study has applied IoT and machine learning in precision agriculture to manage crop variety, crop performance and soil quality (Micheni, Machii&Murumba, 2022).

Finally, there is limited knowledge on how water is consumed in residential homes within an urban setting and using real-world data. Given the unique nature of water supply and water consumption in Kenya, the increasing population in urban areas and the unexpected droughts, the lack of contextual models is a significant gap in sustainable water management. This study has addressed the identified research gaps by establishing an LSTM network model of water management using the IoT-based smart water meter. A trained deep learning model can be used for future real time prediction.

Such a model can forecast future water consumptions using new data captured by IoT-based smart water meters and identify anomalies, detect leakage, optimize water distribution to provide decision support for water management strategies.

1.3Objectives

1.3.1 General objective:

The main objective of this study was to establish an LSTM network model for predicting water consumption in residential properties using smart water meter data. This was accomplished through the specific objectives that have been listed below.

1.3.2 Specific objectives

i. To assess the factors influencing variation in residential water consumption.

ii. To design and develop an LSTM network model for processing time series data to predict water consumption and detect anomalies.

iii. To test and validate the developed model.

In addressing the above specific objectives, the current research examined the following questions in detail:

1.3.3Research questions

i. Which factors influence variation in residential water consumption?

ii. How can an LSTM network model be used with time series data to predict water consumption and detect anomalies?

iii. What is the effectiveness and applicability of the developed solution?

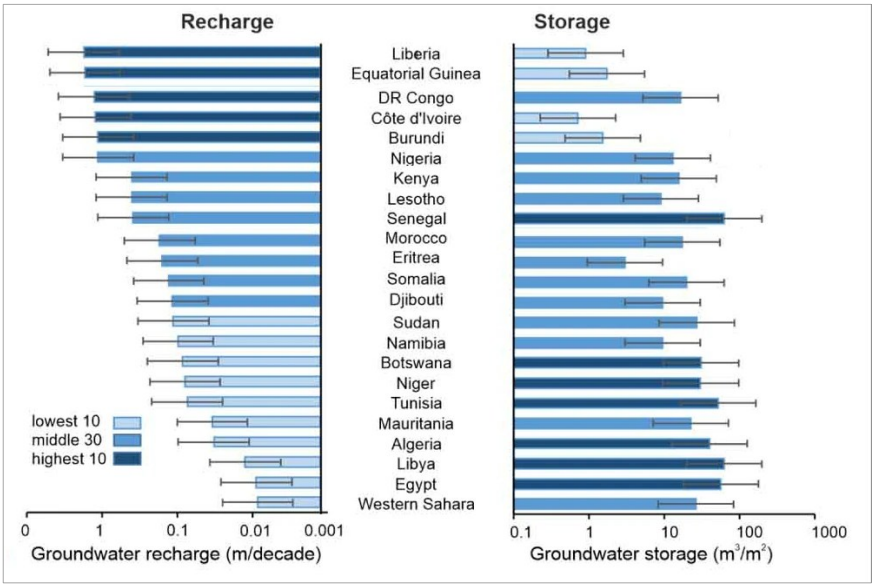
1.4Motivation of the Study

Water is precious for the life of individuals, agriculture, and industries. It is therefore necessary to effectively predict water consumption patterns and detect

anomalies in water usage to inform interventions. Although there is constant demand for water in residential properties, especially in urban areas, it is difficult for water utility companies to track how the water is being used or paid for(Mulwa, Li &Fangninou, 2021). Unexpected and prolonged droughts add to the challenge of water consumption management and create the need for models which apply intelligent algorithms on historical data to predict future consumption. According to recent statistics, also shown in Figure 1.1, Nairobi is among the cities in Africa that have the most unstable water supply (Makoni, 2021).

FIGURE 1.1:

Statistics Of Rainfall Recharge Per Annum for Selected African Countries (Source: Makoni, 2021)



The need for sustainability in water supply is well documented in the literature, and many initiatives have come up to address the problems resulting from declining quantity and quality of water. Sustainable Development Goal (SDG) number 6 seeks to ensure availability and management of clean water and sanitation for all as an incentive

for development. There is evidently a need for more efficient tracking and management of the water use and supply, especially in urban area municipalities. A possible solution is to utilize IoT-enabled devices such as smart water meters to generate massive real-time high-frequency data on water consumption for each user. Analysis of such massive data using machine learning techniques can potentially ease the detection of leakages and errors in water consumption and advise measures for specific users or groups of users as well as for the entire water supply infrastructure.

LSTM models are ideal for analyzing smart water meter data due to the temporal and sequential nature of such data. Patterns in smart water meter generated data emerge over time and can vary per day, week, or season. smart water meter data is also sequential, where each reading is built upon previous readings which can be used to predict future readings. However, the strengths of LSTM network models have been rarely utilized to effectively capture and model the intricate patterns and dependencies that are inherent in data generated by IoT-enabled smart water meters.

LSTM network models contain memory cells that retain information for long time periods. This enables LSTMs to capture long term contexts and dependencies that affect water consumption, such as sudden changes in water usage within a residential property. Additionally, these models can effectively handle time lags that exist in water consumption data. For example, variation in a day's water consumption can affect consumption for the subsequent days. LSTMs learn these lag patterns and make predictions based on the learned information. Furthermore, LSTMs are able to process irregular records of water consumption data from smart water meters which result from sensor errors, maintenance and irregularities in meter readings. These irregular timesteps and missing data can be smoothly incorporated into LSTM network models. This ability also makes LSTMs capable of adapting to changes in smart water meter-

generated data due to changes in household behavior, infrastructure, or policy. LSTMs are resilient to noise in data and will still generate meaningful predictions with irregular input.

Finally, since IoT-enabled smart water meters generate massive consumption data at regular time intervals, the ability to process and analyze the large volumes of data is critical. LSTM network models are able to efficiently process large data sets and generate accurate predictions.

1.5 Significance of the Study

Results of this study will benefit various stakeholders by generating different outcomes. First, water utility companies can be more efficient in their operations by using accurate predictions. Water utility companies are increasingly adopting IoT-enabled infrastructures where networked smart water meters transmit and receive real time data on water consumption. Proper analysis of the generated data can generate insight for confirming leakages, water usage costs and other details that are customer oriented. This will allow these utility companies to optimize the supply and distribution of water, reduce wastage and ensure more sustainable use. The result will be improved customer service and the planning of the water infrastructure, as well as cost saving.

Accurate predictions of water consumption patterns for residential properties can aid consumers and households to make informed decisions concerning the usage of water. By accurately determining water consumption levels and empowering the consumers to proactively manage their consumption, potential behavioral change as well as transparency can be achieved. The use of consumer feedback to encourage consumption conservation has been widely applied and evaluated in other sectors but have not been documented to a similar extent in managing water consumption. The

provision of more frequent and detailed consumption information and feedback has been found in several studies to result in significant water savings, whereas others are critical arguing that there is little evidence at the time whether high frequency feedback is effective in reducing consumption in the water sector or not. Nevertheless, smart metering offers an improved possibility to supply end users with feedback can potentially promote water savings. In the current digital age, information and feedback in real-time or near-real time could easily be provided. This will lead to a reduction in water bills and more responsible consumption practices. The depletion of water tables can then be controlled and managed.

Urban planners and policy makers can better understand the demand for water through accurate predictions of consumption. Data from the research can be made available to different decision-makers and stakeholders upon request and in real time. This will allow all key stakeholders to make more relevant decisions that can help with intervention measures regarding usage and pricing. The insight from this study will inform decision making on regulations for different water zones, infrastructure development and the development of policies to manage water in the face of rapid population growth and changes in consumption patterns. Furthermore, the results of this study will inform emergency services to prepare for potential water related incidents, such as peaks in demand, leakage and interruptions in water supply.

Research of this study will allow environmental organizations which are mandated to conserve water to leverage on the results and increase awareness about water consumption patterns and the benefits of managing water efficiently. This will result in initiatives which promote sustainable water use for individuals, households, and communities. Environmental agencies can also monitor water stress levels and

potential overuse using the prediction data from this study. This will help to maintain the ecological balance and ensure availability of adequate water in future.

Future researchers can benefit from the generated model as a basis to improve it for better predictions. Smart water meter technology providers will also leverage on the study results to refine and develop more advanced smart water meters for increased predictive accuracy.

1.6 Structure of the Research

The subsequent chapters will proceed as follows: Chapter 2 is named Literature Review. This chapter will review the literature associated with smart water meter, machine learning techniques and IoT. It will also look into potential variables that influence water consumption. It is in this chapter that a conceptual framework is proposed using the review of existing literature as a basis for the formulation of the study hypotheses and identification of potential study variables. Chapter 3 is named Methodology. This chapter will highlight the methodology and strategy for achieving the study objectives by introducing the study design, data, and data analysis. It also describes the ethical considerations that have been made throughout the collection and analysis of data, and the presentation and interpretation of the study results. Chapter 4 is named Results and Discussion. This chapter presents the key findings of the research. The results include outcomes from the descriptive analysis of study variables, as well as the correlation statistics. Furthermore, the model outcomes are presented and validated using different metrics. These results are interpreted to introduce context and establish their perspective. Chapter 5 is named Conclusion. This is the final chapter of this research. The chapter revisits the key results and highlights the contributions to the existing literature. It then discusses the limitations identified and uses these limitations as a basis to recommend future research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Water is an important component that helps to sustain life and ensure the survival of all living things. Nevertheless, a challenge to this is the constant depletion of water among the population of cities that is growing quickly in the modern era because more people are migrating from rural to urban regions. Many people around the world are now living in urban areas. The rapid population growth and density have caused an increased demand in infrastructure and services, including water (He et al., 2021). Most cities in the developed world have a water supply rate of more than 80%, as opposed to developing world cities where the penetration rate is less than half and depends on the degree of city development (Kim et al., 2022). In ensuring water security and sustainable water supply, the Kenyan government meets complex and interconnected challenges at the local and regional levels. Strategies are needed to ensure that water is always available in residential houses, especially those in urban areas. At the same time, there should be a reduction of the unused water that is being pumped into the network of these homes. Precision in the prediction of water supply using traditional models can therefore be very challenging leading to persistent water scarcity in most areas.

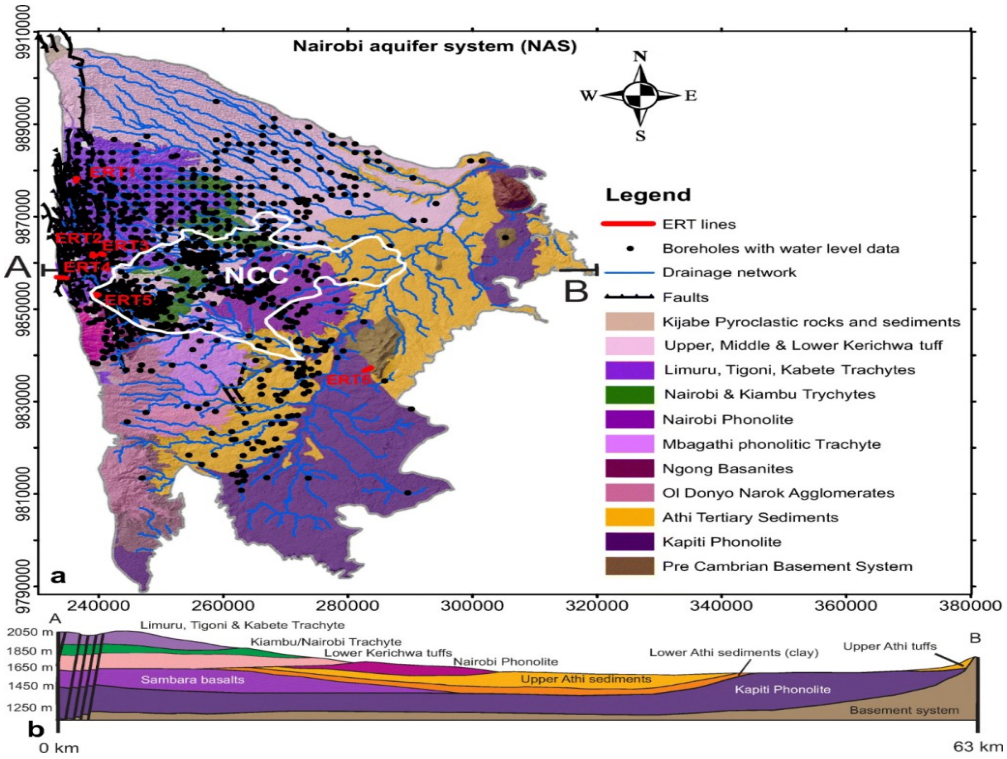
For many countries and urban areas, conserving water and ensuring adequate supply are increasingly being prioritized as the gap between demand by the growing population and the ever-reducing water availability widens. Water scarcity is a serious threat, especially for urban cities in developing countries. Increased digging of boreholes and water wells to increase supply has led to falling water tables and

depleting aquifer (Oiro et al. (2020). Nairobi has a population of more than six million residents, and the Nairobi Aquifer System (NAS) in Figure 2.1 is the largest area of water demand in Kenya with constantly depleting water tables.

FIGURE 2.1

A) Nairobi Aquifer System (Nas) with Nairobi City Council (ncc).

B) Geological Cross Section of the Nas from West to East (source: oiro et al., 2020)



The critical nature of sustainability in water supply has been well documented, and many initiatives have come up to address the problems the result from declining quantity and quality of water. Sustainable Development Goal (SDG) 6 seeks to ensure availability and management of clean water and sanitation for all as an incentive for development. Many researchers agree that short-term prediction of future water consumption is important for ensuring the efficiency of water management (Kim et al.,

2022). Parchin and his co-authors emphasize that water-supply systems must be operated on the basis on foreseen demand of water in the short-term future in order to increase efficiency in the water supply (Pacchin et al., 2019). Other studies have found out that a large-scale water supply management system can be more relevant for accurate prediction of water supply, especially in residential properties (Gagliardi et al., 2018).

However, many countries in Africa including Kenya are still largely relying on analogue and manual water-reading systems for tracking and managing water supplies and water use (Amankwaa et al., 2023). Because of the importance of establishing an efficient water supply infrastructure, city administrations and researchers have increased effort to increase sustainability through creation of smart innovations, including new digital tools and technologies that have emerged in the modern age. For example, the Internet of Things (IoT) is a new technology that cities are blending into infrastructure and services, including water supply and consumption monitoring and management (Shah, 2017). The concept of "smart cities" has recently become popular across the globe. In Africa, more and more of these smart cities are growing. A report on smart cities by Pelikh (2022) rated the "Silicon Valley of Africa" in Rwanda and Konza City in Kenya as the smart cities with highest attraction. Smart cities are now being regarded as a future emerging market that will soon be central to the digital economy.

Water readings have been historically collected by water company representatives using a pen and paper. Nowadays the readings are collected using mobile phones or a hand-held electronic device called a meter reader. This newer method was sufficient a few years ago, but the rapid population growth has resulted in more construction of residential areas and industries which are unsustainable for human data collection.

Manual water reading is labor-intensive and often generates a lot of errors. There is time delay, and sometimes the companies make estimates from previous historical readings. There is evidently a need for more efficient tracking and management of the water use and supply, especially in urban area municipalities.

Smart technologies are increasingly being introduced into the water-supply field following the advances in information and communications technology (ICT), the 4th industrial revolution, and IoT. Over the past few years, there has been a deliberate move towards smart metering within the Internet of Things (IoT) infrastructure. IoT is a network that interconnects physical devices which are embedded with sensors, software and connectivity mechanisms to allow exchange of real-time data over the internet (Owen, 2023). In the context of smart water metering, the IoT devices are usually installed in the pipeline, water meters, and other sections of the distribution channel to collect real-time high-frequency data on water flow, consumption, pressure and temperature. Useful insights can then be generated about consumption patterns and general state of the water supply infrastructure.

In Kenya, the Kenya Power and Lightning Company (KPLC) has installed networked smart meters throughout the country which are used to record electricity consumptions and determine pricing. Advances in electricity usage have also resulted in smart water metering to monitor and manage water consumption. Smart water meters are increasingly being applied to increase efficiency in water management (Di Nardo et al., 2021). The IoT digital technology makes it possible for smart water meters to collect a large number of high-frequency data using smart sensors so that it can be used for monitoring and measuring performance and usage patterns for technical systems such as water distribution systems. IoT-enabled smart water meters record details of water consumption more effectively than the traditional water supply and metering

infrastructure. Consequently, water utility companies are increasingly adopting IoT-enabled infrastructures where networked smart water meters transmit and receive real time data on water consumption (Söderberg& Dahlström, 2017). Proper analysis of the generated data can generate insight for confirming leakages, water usage costs and other details that are customer oriented (Kim et al., 2022). The smart sensors feed information dashboards with large amounts of data for decision making. Despite the rapid adoption of IoT-enabled smart water meters and the massive data generated from these meters, modern analytics models of water consumption have not been widely adopted for effective management of water in many urban areas.

Detailed analysis of high frequency real-time data generated by networked smart water meters within an IoT infrastructure can improve insights about water consumption, water leaks and anomalies.

2.2 Smart Cities Concept

Although water is an essential resource that is consumed by everybody, urban areas make up the largest proportion of the water consumption ecosystem. This means that urban users place a special demand on the need to efficiently manage water and create smart cities. The concept of smart cities was first introduced in the 1990s and focused on the technical perspective of a city that majorly incorporates novel information and communication technologies within infrastructures and services (Krishnan et al., 2022). Several authors have tried to define the "smart city" concept, but there is no unanimously agreed upon definition of the smart city. Nevertheless, according to an earlier definition by Hancke et al. (2012), a smart city is formed through the integration of its infrastructure and services in an intelligent manner to form a coherent unit. High levels of sustainability and efficiency can be assured using Internet of Things for monitoring and control. By using sensors, a lot of real-world data

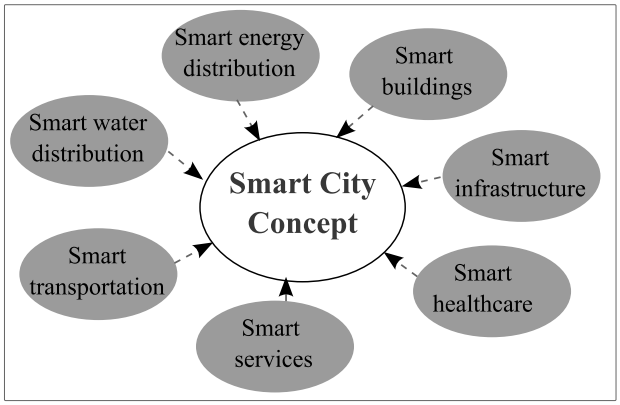
is captured and integrated into a computing platform. The collected data from the sensors become “smart” when complex analytics, modelling, optimization, and visualization are included and applied with the intention of assisting to improve operational decisions.

The concept of a smart city involves a strategic decision basis that targets sustainable development, economic growth and an increased quality of life for citizens. Because of the high consumption levels for resources such as energy, water, etc. in cities and other urban areas in most countries, there is need to regulate through innovations such as smart city. The concept of “smart city” is defined as an ecosystem where infrastructure and services have been integrated in a smart (intelligent) way to form a single coherent unit (Söderberg& Dahlström, 2017). Smart cities are being widely adopted since they are increasingly being recognized. The smart city has been pinpointed as an emerging market feature expected to drive the digital economy (Humayun et al., 2022). Real-world data is collected using sensors and combined into a digital platform that performs complex modeling, analytics and optimization to make the collected data “smart”. The output is visualized using end user applications to make improved decisions. Smart cities therefore help in developing sustainable decision-making, development and economic growth for improved quality of life.

The water sector is relevant in the smart-city concept as shown in Figure 2.2. Smart buildings and smart service provision can include a smart water distribution component in a connected system that gathers high-frequency data on end-user consumption using smart sensors.

FIGURE *2.2*

Smart City Concept



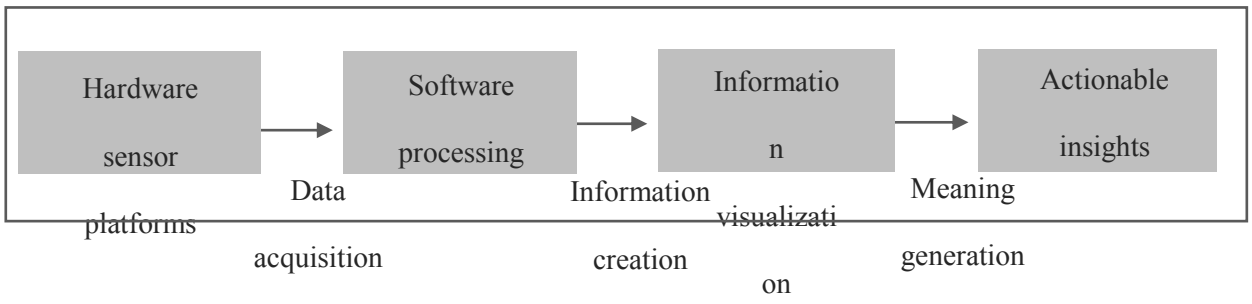
2.3 Internet of Things (IoT)

The definition of IoT varies depending on the context. The general definition given by Ashton is a "worldwide network of interconnected objects which are uniquely addressable based on standard communication protocols" (Ashton, 1999). Within the computing environment, IoT extends computing capabilities and networks connections to other electronic objects to enable them to generate and send data with minimal human intervention. Improvement of sensor networks in areas such as scalability, energy efficiency, and reliability has been important in rapidly expanding products and services within the IoT environment. Developments within manufacturing have made it possible for small-scale-computation units to be embedded into small objects. This, combined with the increase in computer economics to increase computational power and lower costs of electronic devices, has increased popularity of IoT. There is more capability for data collection, storage and communication both from in-house data stores and cloud-based services (Micheni et al., 2022).

Collecting data from multiple sources is not valuable in itself; this is only one part of the IoT concept. The application of automation and analytics to the large amounts of collected data using well defined algorithms is what provides useful information and creates value. Figure 2.3 shows the steps in the data access process that facilitate data collection up to generation of actionable insight.

FIGURE 2.3

A process flowchart of the IoT environment



2.4 Smart Water Sensors and Meters

2.4.1 Design of smart-water meter reading infrastructure

At the household level, a smart sensor is a connected device that uses the internet and other ICT advancements to communicate over a large network. Smart sensors usually communicate with end nodes (called sinks) which forward the communication to dedicated servers or a cloud service. Such networks are usually called wireless sensor networks (WSN). Information about water flow in the WSN is obtained using the smart-water meter, the smart sensor that collects water usage information, and a communication channel for transmitting the information in real or near-real time (e.g., every 30 minutes).

Collection and communication of water meter readings can be designed in several ways: The first is to send the readings to a gateway device or concentrator using dedicated serial lines. This technique requires the water meters to have a universal serial bus (USB), an RS485 connector or any other suitable connector (Yan & Gang, 2019). Wired systems can offer higher security and less interference, but the resources (time, money) used to install makes this approach not practical. It is also possible for the user to initiate reading using a mobile device and some app, but this is unsuitable for frequent readings, autonomous system, or real time monitoring.

A second option is to establish a dedicated serial line using wireless technologies where the WSN end node is linked to both the water meter and the gateway device. Bit-rate sensors used with this approach utilize the ZigBee requirement with integrated IEEE 802.15.4 standard for low-data networks. The approach is a low cost and energy-efficient alternative for WSNs with a short range (10-100m). The detection range can be extended for the gateway device using some network topology such as star, mesh, etc. (Funk, 2018). Various communication technologies are applicable for smart-water meter systems and depend on network topologies and infrastructures such as wireless-packet data services (WPDS), remote frequency (RF), broadband technology, and satellite. The satellite technology is expensive because of complexity and need for regular maintenance.

A design schema that uses the IEEE 802.15.4 standard for wireless smart-water meter reading basically comprises of a lower-level network to connect the water meters, actuators, and data concentrators, and a higher-level network dedicated for visualization, report creation and managing information. Other suggested alternatives are hybrid water-meter reading systems which apply robust and stable technologies like ZigBee. However, the problems with the hybrid options include complexity due to the

hierarchical-network architecture, high power consumption and high maintenance requirements.

Another technological option that facilitates direct data communication over power lines is Broadband over Power line (BPL). BPL technology can be combined with smart-water meter systems to achieve bi-directional broadband communication (Funck, 2018). Nevertheless, the requirement of additional components (e.g., collector, concentrator, master station) makes this an expensive alternative. Fewer cables are used in this option, but the use of an open physical later introduces security risks.

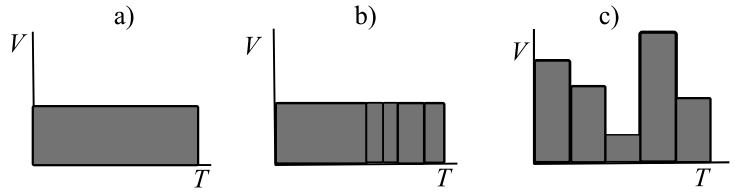
2.4.2 Smart water meters

Different water meters exist (e.g., electromagnetic meters, velocity meters, displacement meters, etc.) based on different physical characteristics of the water flow. The most used meters are the displacement meters which use the movement of water to record the water flow (Kiran et al, 2022). Three techniques are available for the mechanical water-flow recording as shown in Figure 2.4:

1) Accumulation: this is the oldest and simplest technique. The accumulated usage is recorded for a certain resolution of intervals is recorded and communicated in a single meter read, with no data being stored in-between.
2) Pulse: a pulse is triggered by a quantum of water passing through the pulse water meter. Pulse meters record both the pulse and the pulse timestamp
3) Interval: an interval meter consistently monitors the flow of water moving through the meter at a specific time interval.

FIGURE 2.4:

Accumulation meter (a), pulse meter (b) and interval meter (c)



Unlike the traditional water meters which transmit low resolution data, such as single data-points per annum, the smart water meter yields high resolution data at spatial and temporal scales.

Over the past few years, several projects have been conducted in Asian countries (e.g., Korea, Japan), China, USA and European countries (France, Malta, Spain and Italy to support development and implementation of smart water meters. Pilot studies have discovered a lot of savings in the water consumed when the smart meter was being used. However, many African countries including Kenya have experienced very slow adoption of smart water meters. Safaricom, the largest telecommunications service provider in Kenya has ventured into the smart metering domain by installing water ATMs in Mathare that allow water to be purchased and paid directly through their mobile money payment system (Guma & Wiig, 2022). The uptake of this IoT technology for water management is nevertheless extremely limited in spite of the potential benefits.

From the research angle, very little research has been done about the benefits of analyzing IoT-enabled smart water meter data in Kenya to forecast consumption, and how best to provide an appropriate modeling solution for the local context. The data that can be extracted from smart water applications is unique such that it is difficult to

generalize information based on the studies that have been done in the developed countries (Krishnan et al., 2022). It is therefore necessary to develop a new model or algorithm that can be applied to provide solutions for water consumption management in Kenya. Modeling options include artificial intelligence, deep learning and other machine learning techniques that can be applied in designing a smart water management system for sustainable water usage from natural resources. Statistical analysis is also needed to develop an effective and efficient localized smart water management model.

2.5 Machine Learning Approaches for Smart Water Data Management

Different machine learning algorithms have been applied for effective water management. The challenges experienced with predicting water consumption and the general management of water resources have seen widespread use of machine learning algorithms across a variety of areas. These methods use algorithms and data analytic tools to glean insights from smart water meter data and create precise forecasts. The recent rise of smart cities is causing research interest to expand to effective water management in the urban water distribution system. A study used Principal Component Analysis (PCA) and Random Forest algorithms to build an urban water management system in China (Gao et al., 2020). The authors used mean absolute error and root mean square error to validate their model. Other studies have incorporated more complex machine learning models for water management and prediction which this chapter will explore in the sections that follow below.

2.5.1 Naïve predictors

Models of naïve prediction, which are also called persistence models, tend to follow a kind of heuristic that have a direct and straight-forward approach. In many instances, naïve predictor models are used as a baseline when making predictions

alongside other more complex models. Naïve models have been used as baseline models in cases such as the forecasting of solar radiation, power generation and health outcomes (Takeyosi, 2016; Kaushik et al., 2020; Kühnert et al., 2021). Naïve predictors can be utilized on the basis of a heuristic that is meant to generate a form of expert system. Specifically, a prediction profile is usually derived through computations of hourly averages for the final days of the week that experience similar consumption patterns. An example of a naïve predictor algorithm is as follows:

Algorithm1: Pseudocode for naïve water management and prediction model

```
result: Prediction of water flow and consumption for next 24 hours
if next_day is public_holiday or if next_day is Saturday then
|   prediction = hourly average for last tSaturday;
else if next_day is Sunday then
|   prediction = hourly average for last tSunday;
else
|   prediction = hourly average for last tsimilar week_day;
End
```

Predicting water consumption in urban areas requires accurate, unbiased and interpretable tools. The limitation with naïve predictors is their inability to meet these three requirements. Such methods are simplistic, but they suffer from decreased accuracy of predictions and increased bias. Furthermore, their results are sometimes difficult to interpret.

2.5.2 Vector autoregressive (VAR) model

VAR models are among the most common machine-learning techniques used for multi-variate time-series forecasting. In addition to predicting water profiles, VARs have been widely applied in aviation industry, the prediction of brain functions and many other areas (Kühnert et al., 2021). VARs operate under the fundamental principle

that each variable at a certain timepoint t can be predicted using all variables at timepoint t-1. This means that the VAR model assumes that 1) time series follow a stochastic process and 2) the dependency between any two variables is linear. The formal description of a first-order VAR model is:

$$x_t = \beta_0 + CX_{t-1} + \varepsilon_t$$

where *C* is an *x* by *x* matrix that represents the coefficients of the VAR model, and ε_t is a sequential order of independent and normally distributed errors with a mean of zero and a co-variance that is the sum of errors (Kühnert et al., 2021).

VAR models improve the insight of generated outcome to a large extent. However, the complexity of these models make their interpretation difficult.

2.5.3 Support vector machine

The popular machine-learning approach known as Support Vector Machine (SVM) is utilized for categorization problems. The ideal hyperplane that divides several classes in the feature space is what SVM seeks to identify. Sentiment analysis, spam identification, and false news detection are a few examples of water consumption categorization and prediction problems where it has been effectively used (Styawati et al., 2022).

SVM has demonstrated strong performance when working with high-dimensional data and is capable of tackling both linear and non-linear classification issues. Widely used in a variety of water consumption categorization and prediction applications, SVM is regarded as a potent machine learning method. Its widespread use results from its capacity to identify the best hyperplane that optimally divides various classes in the feature space. The adaptability of the SVM approach in processing high-dimensional

data is one of its main benefits. In linear and non-linear water consumption prediction tasks where the input data may contain a variety of characteristics, this makes it very helpful. SVM can efficiently handle these complicated datasets and spot patterns that other algorithms might miss. Additionally, SVM is not just applicable to linear classification issues. According to Zhang et al.(2023), the radial basis function (RBF) is one of the several kernel functions that may be used in SVM to tackle non-linear classification Due to its adaptability, SVM can capture complex correlations between data and classes, making a wide range of water consumption categorization and prediction problems viable for it. SVM has been effectively used in the field of sentiment analysis to categorize text data into positive, negative, or neutral feelings. SVM may be taught to correctly estimate the sentiment of fresh, unseen text inputs training on huge datasets containing labeled sentiments. Numerous industries, including customer feedback research, brand reputation management, and social media monitoring, can benefit from this (Wu et al.,2022). SVM has also been used to discriminate between harmful and authentic emails in spam detection.

SVM can successfully categorize erratic and non-linear time series data and patterns to detect probable anomalies such as those caused by irregular smart meters, broken piles and leakages and other forms of anomalies. This lessens the incorrect predictions that can be made due to errors in the incoming data.

SVM models suffer from two key limitations. The first limitation is the difficulty in choosing the appropriate kernel function and input space to solve the practical problem. Proper selection of kernel function and input space is necessary to improve the accuracy of prediction and the efficient of computation for any model. Nevertheless, identification of the input space efficiently and effectively remains a research problem for analysts using the SVM approach. The second limitation of the

SVM model concerns the efficient and effective optimization of SVM model parameters (Wang & Shi, 2023). It is important to optimize a model to allow efficient computation in the high dimensional feature space, since SVM imposes a lot of computational resources. However, this process remains unclear and creates a lot of ambiguity when analyzing data using SVMs.

2.5.4 Random Forest

An ensemble learning technique called Random Forest uses many decision trees to provide predictions. Dueto its capability to handle complicated data structures, manage missing values, and offer feature priority rankings, it has been extensively employed in water consumption categorization and prediction problems. According toZarei. (2023), numerous urban analysis fields, including electricity distribution, water consumption tracking and diagnosis, consumer segmentation, and fraud detection, have used Random Forest. The field of water consumption categorization and prediction problems has seen tremendous growth in the use of Random Forest, a potent ensemble learning technique.

To produce precise and dependable forecasts, it integrates the predictions of many decision trees. Random Forest resilience in handling complicated data structures is one of its main features. It is useful for water consumption categorization problems that entail complex interactions between factors since it can handle data with a lot of characteristics and variables. Random Forest can capture the intricate relationships and interactions between variables, producing predictions that are more accurate (Meghana, 2023). Furthermore, the datasets missing values may be handled via Random Forest. This is very helpful in water consumption categorization data where it is frequent to have missing data. The prediction model is not harmed by partial data since Random Forest deftly imputes missing values or irregular incoming data. Ranking the relevance

of features is another advantage of Random Forest. This is essential in water consumption categorization assignments as it aids academics and industry professionals in determining the traits that have the most predictive power. It is possible to get important insights by studying the feature importance rankings, which will improve knowledge and interpretation of the underlying water consumption processes.

Several urban domains have effectively used Random Forest (Nyanjara et al., 2023). Based on a collection of symptoms and a person's medical history, it has been used to diagnose illnesses by estimating the chance of developing particular conditions. Random Forest can help in early detection and treatment by identifying significant predictors by studying vast volumes of patient data. However, the main limitation of the Random Forest approach is the high computation requirements of this machine learning technique, especially with large data sets.

2.5.5 Neural networks

Recent years have seen a huge increase in the popularity of neural networks, particularly deep learning models, because of their capacity to identify intricate patterns from vast volumes of data. For water consumption categorization and prediction tasks, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are often employed architectures (Wang & Zhao, 2023). While RNNs are better suited for sequence data processing tasks like sentiment analysis and stock market prediction, CNNs excel at classifying images and texts. CNNs have demonstrated to be quite good at tasks like text and picture categorization. Convolutional layers allow CNNs to automatically extract features from unprocessed data and recognize intricate patterns that are essential for water consumption categorization tasks. CNNs are capable of

recognizing minute characteristics and spatial correlations in pictures, which enables precise categorization and recognition in the context of image classification. According to Mulumba et al. (2023), RNNs, on the other hand, excel in analyzing sequential data, which makes them the perfect choice for projects like traffic-speed analysis and stock market forecasting. Because of their exceptional capacity to remember previous inputs, RNNs can recognize temporal connections and patterns within sequences.

Baek et al (2020) applied convolution Neural Network (CNN) in Korea to simulate water levels and quality in the Nakdong-river basin. CNN has also been used in combination with U-net in an integrated deep-learning automatic detection model (Saraiva et al. (2020). U-net applies the Tensor Flow Python library and the Google cloud platform to train the images for detecting leakage in the water pipes. A different study used an IoT model with a deployed sensor to monitor water quality, turbidity and temperature (Jan et al., 2021). Some other authors built an ensemble-learning machine learning model which combined support vector regressions (SVR), adaptive neuron-fuzzy inference system (ANFIS), multivariate linear regression (MLR) and Artificial Neural Network (ANN). The model was used to forecast infiltration of a water irrigation system using factors such as in-flow rate, infiltration opportunity time and cross sectional area of infiltration (Sayari et al., 2021).

Further experiments have employed the IoT system with different smart sensors for measuring water supply, flow, and pH using various undisclosed machine learning techniques (Shah, 2017). The model was used to monitor the water storage system and ensure uniform supply to all regions in the network. However, the model was found to be too expensive to be used for real-time water supply systems. Another study developed a machine learning model to carry out suspended sediment yield (SSY)

prediction in Godavari River Basin in India. They combined the genetic algorithm (GA) with ANN to form GA-ANN (Yadav et al., 2022).

By considering neural networks, this class of machine learning approaches can comprehend the context of data learned over a period of time through deep learning analysis. This makes categorization more accurate and sophisticated. To anticipate future market behavior, RNNs may evaluate previous consumption patterns and extract significant patterns. This helps analysts estimate how consumers will behave in the future. Neural networks are a useful tool in a variety of urban application domains because of their adaptability and versatility. From Ye's et al. (2023) perspective, their power to handle complicated patterns and learn from big datasets has enabled them to make substantial strides in industries like healthcare, marketing, and finance. In activities ranging from illness diagnosis and patient prediction to consumer behavior analysis and fraud detection, neural networks have been used. The way water consumption categorization and prediction problems are conducted has undergone a radical change because of their capacity to reveal hidden patterns and make precise forecasts.

RNN is a category of Artificial Neural Networks that allows learning of long term dependencies that is useful when the network needs to retain information over long time periods. This means that the approach allows handling of successive sequence of events in which the understanding of each even is based on previous events. Furthermore, the deeper the RNN, the longer the memory period and consequently better capabilities can be achieved. However, RNN has its limitations because of the vanishing gradient problem due to its architecture restriction to long term memory capabilities.

2.5.6 Naïve bayes

Based on Bayes theorem, Naive Bayes is a statistical machine learning method. The computation is made simpler by supposing that the characteristics are conditionally independent provided the class label (Siregaret al., 2023). In water consumption classification applications including anomalies detection, pricing evaluation, and water flow categorization, naive Bayes has been used. It is renowned for its effectiveness, simplicity, and strong performance on tasks involving text classification. One clear benefit of Naive Bayes is its simplicity.

It is an appealing option for applications where interpretability is essential because it is reasonably simple to comprehend and apply. Additionally, Naive Bayes performs superbly in problems requiring text categorization. It has a long history of successful application in fields including spam detection, sentiment analysis, and document categorization (Ding et al., 2022). To distinguish between legitimate emails and spam, Naive Bayes may assess the content and structure of emails.

Naive Bayes can effectively categorize emails by considering the frequency of particular words or phrases, assisting in the maintenance of clutter-free inboxes free of unwelcome communications. Like this, Naive Bayes sentiment analysis may examine the sentiment included in text, such as postings on social media or client evaluations. Naive Bayes can categorize the sentiment of the text by considering the frequency of specific words or phrases linked to positive or negative sentiment, giving companies insights into client opinions and preferences (Yilmaz et al., 2022). Another use of Naive Bayes that has been successful is document classification. Naive Bayes can categorize documents into predetermined categories, including news items, academic papers, or legal documents, by looking at their content and structure. This capacity

offers effective information organization and retrieval, assisting with activities like document management and information retrieval.

2.5.7 Regression

A straightforward and popular machine learning approach is the regression modelling or applications of linear regression. When the objective variable is binary or ordinal, it may also be utilized for water consumption categorization problems, while its main usage is for continuous variable prediction. In social science research, Auerbach (2022) observed that linear regression has been used to forecast outcomes including human consumption and lifestyle activity patterns, income levels, and life satisfaction. Recent research has demonstrated that linear regression may produce findings that are easy to understand and can include significant social aspects in the prediction process. In social sciences studies, linear regression has proven to be more useful than other ML models in predicting outcomes like educational attainment, income level, and life happiness. As a result, it is the best. Researchers may learn a lot about the connections between various social factors and the projected results by using linear regression to examine pertinent components and their effects on the target variable (López &Arboleya,2022). Its interpretability further increases its usefulness by allowing researchers to comprehend the variables influencing the predictions and confirm the findings.

The appropriateness and interpretability of linear regression in water consumption categorization tasks was demonstrated in research by James et al. (2023) that used linear regression to predict water consumption inclination based on social media data. Based on social media data, this study used linear regression to predict water consumption inclination, demonstrating how this algorithm may effectively capture key water consumption aspects and offer insightful data. Linear regression effectively

predicted political orientations by considering a variety of variables and patterns in the social media data, revealing insight on the impact of social media on political attitudes. Linear regression is a useful tool for water consumption categorization tasks as well as regression problems due to its clarity and interpretability. Its use in social science research has made it easier to anticipate different outcomes, and recent studies have shown how successful it is in capturing key social aspects (López &Arboleya, 2022).

Linear regression is a fundamental and adaptable approach for comprehending and forecasting the urban water consumption phenomena as machine learning develops further. Notably, various machine learning techniques have been used for water consumption categorization and prediction problems. Even though each method has advantages and disadvantages, it has been demonstrated that linear regression is a good option for urban water consumption categorization tasks due to its interpretability and capacity to capture significant aspects (James et al., 2023). However, the job, data qualities, and intended results ultimately determine which method is used.

2.5.8 Long short-term memory (lstm) networks

Long Short-Term Memory (LSTM) networks are a special prediction approach that has resulted in significant improvements for time series prediction using data such as water consumption data. LSTM is a special type of recurrent neural network (RNN) that is designed to allow retention of information for longer periods of times. The LSTM precision has been achieved over the recent years using special deep learning algorithms that are capable of predicting non-linear water demand in real time. LSTM is among the methods that are potentially useful for water suppliers, particularly those in urban areas where many variables exist that affect water consumption and make it difficult to predict this consumption in the short and long terms.

Indeed, it has been discovered in studies done in the recent past that LSTM network models can easily outperform other methods since they are able to quickly integrate additional information like pump error, smart meter defects, and unnatural patterns such as vacations and holidays. High levels of water leakages, contamination, and deterioration of water pipes, especially in agriculture, have caused research interest in using machine learning tools for water management.

Deep learning models such as the Long Short-Term Memory (LSTM) network model provide good alternatives for implementing IoT-based smart water meter. Such models allow collection and analysis of massive datasets for smart water meter on individual and collective water consumption within an Internet of Things (IoT) framework (Kim et al., 2022). An LSTM network enhances the predictive capabilities of such a system by processing time-series data and making accurate forecasts.

The LSTM network model can maintain a constant error that allows this model to recursively learn through both time and layers. Additionally, the LSTM utilizes special type of cells named “gated cells” which store information in a different manner as compared to the standard recurrent neural network and allows the reading from them. Each cell can make a decision on its own regarding the information while closing and opening these cells to execute those decisions. The LSTMs architecture comprises of chains that allows the holding of information over long time periods to solve problems that a standard RNN may not be able to solve.

2.6 Data-driven decision support

The Data-Information-Knowledge-Wisdom (DIKW) ladder first presented by Ackoff in 1989 is a well-known model for understanding the value of data collection (Ackof, 1989). Data refers to the discrete facts and figures which are observed. They

are not organized or processed to generate useful meaning. Information refers to data that has been processed to add value. Knowledge combines data and information, and expert opinions, skills and experiences to generate a valuable output for decision making. Knowledge can also be described as combination of information with capability and understanding to generate "actionable information". This actionable information is the cornerstone of data-driven decision making (DDD).

DDD applies to the decision made from evidence in data as opposed to decisions from personal experiences and intuitions. DDD opens more opportunities for making better decisions (Osman&Elragal, 2021). However, DDD does not include wisdom. Improvement of the DDD through IoT has significantly increased availability of information to decision making entities in different sectors. Organizational learning, scaling, employee training, and ICT proficiency are variables which affect the ability of an organization to successfully incorporate DDD.

2.7 Factors influencing water management and consumption in cities.

Several factors have been studied by researchers in the field of water consumption and management in cities and other urban areas within the IoT-enabled smart water meter domain. One study in Sweden found out that the time-of-day patterns was significantly correlated with water consumption in residential properties. Households experienced peak usage of water was determined to be during off-peak hours (Söderberg&Dahlström, 2017). Another study in the Pune city of India found out that previous usage patterns of residents affected the consumption of water through the pipe-based water supply (Singla &Bendigiri, 2019). This study worked with four regression models where the authors also assessed other factors that can affect the value of a residential property. A similar study by Hargreaves et al. (2019) in the Greater Southeast of England has established that household size heavily influenced the average

water consumption. The study was seeking to understand the impacts of urban planning on viability of alternative water supply systems.

More recently, the influence of community factors in water consumption and management has been studied in Shanghai, China (Han et al., 2021). The authors discovered that some apartments rented by several tenants were the main consumers of residential water. The study showed that average consumption per household must be considered when modeling water management and consumption for cities. Other authors investigated the future water scarcity problem at the global level and shoed that seasonality is a factor influencing water usage and water management (He et al., 2021). During the dry season water tends to be used more as residents 'water plants and dust or wash more.

In a study of factors affecting water consumption in Bangladesh, Hasan et al. (2019) found out that the quality of water is directly related to the consumption. They concluded that the higher the quality of water, the more the consumption. Residents with poor quality water used less water, and those with safe drinking water in their supply system were found to use much more water. Furthermore, Söderberg and Dahlström (2017) and several other authors have found out the size of the apartment of house, which is equivalent to the number of rooms or the general house size in square meters can affect consumption. This is because more space requires more cleaning water. Dream Civil (2022) further notes that an increase of pressure in the pipeline can cause water supply to rise and automatically increase demand for water.

More recently, a review of Smart Water resource management using machine learning has shown that there are very limited studies which have studied the water management in urban areas. This means that the factors which can potentially influence

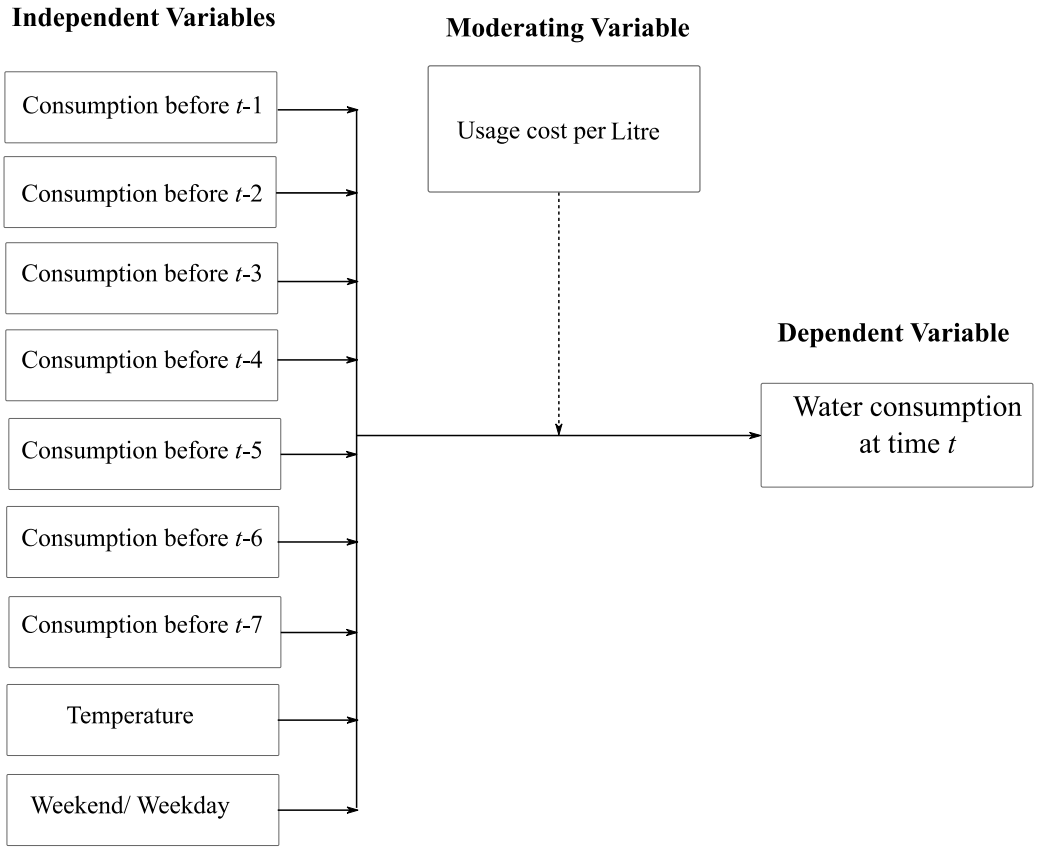
fluctuations in water consumption have not been fully identified. Furthermore, the precision of using the identified factors to accurately predict consumption levels in the urban residential areas has not been ascertained. One of the reasons is because of the limitations in generalizing the inference obtained from previous studies. Most of the existing studies have been done in developed countries (Krishnan, 2022). As a result, further analysis of the factors is needed for expanded domains of application geographies, which include developing countries such as Kenya, and the cities within these developing countries..

2.8 Conceptual framework

This current research has identified the conceptual framework shown in Figure 2.5 using variables identified from the literature review.

FIGURE 2.5

Conceptual Framework



To use the study variables, they need to be in a measurable manner. Table 2.1 illustrates the indicators and values that will represent each study variable. In order to understand the water consumption per liter for the current day, t, which is treated as the dependent variable, a number of associated variables have been identified. One of these, namely, Usage cost per liter, represents the moderating variable. This is because the cost is a limiting factor that uniformly affects all urban residents and consumers of the water resource. Cost is an impediment that allows rational usage while preventing wastage of water. At the same time, the determination of values for the dependent variable, which are controlled for using the identified independent variable, can allow water service providers and other water authorities to regulate and determine the

pricing of water as based on peak and off-peak times. In total, the conceptual model has nine independent variables, which include 1) the water consumption for previous 1 day, t-1, 2) the water consumption for previous 2 days, t-2, 3) the water consumption for previous 3 days, t-3, 4) the water consumption for previous 4 days, t-4, 5) the water consumption for previous 5 days, t-5, 6) the water consumption for previous 6 days, t-6, 7) the water consumption for previous 7 days, t-7, 8) temperature, 8) water pressure, and 9) water consumption on weekend versus weekday. Table 2.1 provides information about how these variables have been operationalized.

TABLE 2.1

Operational Definition of Variables

Variable	**Abbreviation**	**Description**	**Values**
Dependent variable	C_t	The water consumption per liter for current day, *t*	Continuous/ number
Moderating variable	UC	Usage cost per liter	
Independent variable	C_{t-1}	The water consumption for previous 1 day, *t*-1	Continuous/ number
	C_{t-2}	The water consumption for previous 2 days, t-2	Continuous/ number
	C_{t-3}	The water consumption for previous 3 days, t-3	Continuous/ number
	C_{t-4}	The water consumption for previous 4 days, t-4	Continuous/ number
	C_{t-5}	The water consumption for previous 5 days, t-5	Continuous/ number
	C_{t-6}	The water consumption for previous 6 days, t-6	Continuous/

			number
	C_{t-7}	The water consumption for previous 7 days, t-7	Continuous/ number
	T	Temperature	Continuous/ number
	P	Water pressure	Continuous/ number
	Wk/Wn	Water consumption on weekend versus weekday	Continuous/ number

2.9 Knowledge gaps

This study will contribute to the knowledge gap that exists with previous studies that have been concluded on the smart water meter and management of water consumption. First is the model of smart water meter application in a major city in Kenya, a developing country. Most studies have concentrated on cities in the developed world. Secondly, LTSM network models for consumption data generated by IoT-enabled smart water meters are rare. Finally, the study contributes on how water is consumed in residential homes within an urban setting and using real-world data.

2.10 Conclusion

Based on the review of exiting literature, several gaps were established which have been addressed in this study. Results of the study are expected to help Nairobi City Water and Sewerage Company and other water and sanitation companies in Kenya to understand the savings that can be gained by adopting the smart water meter technology. Furthermore, the findings of this study are expected to contribute to the body of knowledge in the field of domestic household-level water supply and its applications.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter provides a description and justification of the research design, research methods and analysis methods. The case study and population are described and the techniques for constructing an LSTM network model using IoT-enabled smart water meter data are presented. The testing and validating processes are then described. The chapter concludes with ethical considerations.

3.2Research Design

Due to the nature of the research problem and availability of a mixed dataset, a case study research design was adopted. Case study research design is bounded by the phenomena of interest exiting within an individual, organization or group (Hancock et al., 2021). The research design allows researchers to collect different kinds of data about the case and provides the opportunity to obtain an in-depth look at the case study of interest to understand the inner workings and interactions that are used to generate insights. It provides a uses an extensive knowledge of the bounded unit and allows one to examine the case and learn from it. In this manner, it is then possible to apply the principles and lessons learned on a specific case to other scenarios and situations in order to transfer knowledge, as opposed to the generalization of inference that is dine in many quantitative studies.

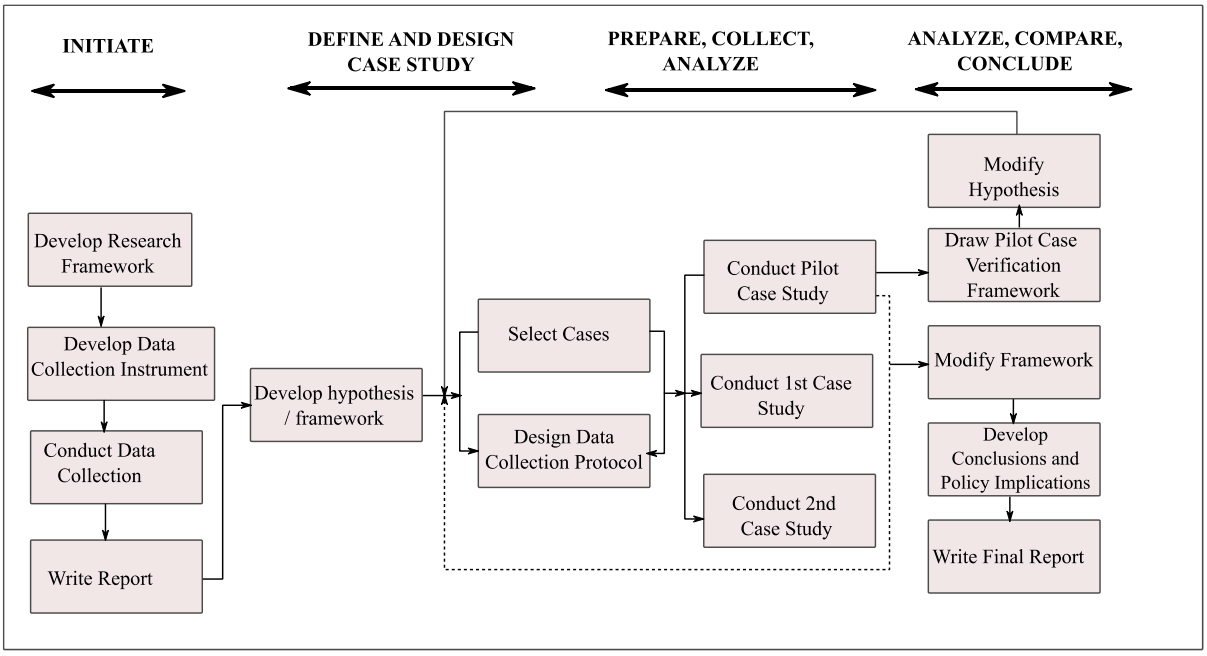
Through critical and extensive observation of the study data, case study research is capable of delivering useful information in real time and can also discover patterns and general relationships which may guide further analysis and potentialize its results.

As such, case study research plays an important role in a typical data analysis workflow.

Case study research design uses a group of statistical techniques that aim to explore, describe and summarize the nature of data within a bounded case study in order to guarantee its objectivity and interoperability (Garg et al., 2023). In this research, the use of Nairobi case study allowed for possible errors outliers and relationships of variables to be identified, as well as an in-depth analysis of data using graphical illustrations and summaries. The aim of the analysis was to increase knowledge in the observed phenomena under observation and also to extend the comparison of findings. The steps of case study research design used in this study are shown in Figure 3.1.

FIGURE 3.1

Steps In Case Study Research Design



3.2.1 Data collection

This is usually a key step in case study research. Smart water meters are capable of transmitting and receiving real time data on water consumption for every resident or water consumer, which are then used to confirm leakages, consumption patterns and consumption fees alongside other details. The analysis in this study used secondary water consumption data generated by IoT-enabled smart water meters. The dataset is a subset of bulk data generated from many smart water meters located in 320 residential properties within several estates in Nairobi city that are managed by Maji Pay, a local smart water meter service provider. Each of the smart water meters collect water consumption data at an interval of every 15minutes. In addition to the smart water meters, several other sensors within the IoT framework collect accompanying descriptive data, such as 1) battery voltage data which indicates the state of the smart water meter and is important for water management, 2) timestamp in date and time, 3) average temperature for each day, and 4) water pressure.

The frequency of the other sensors in collecting data ranges from every 30 minutes, hourly, and daily. The data set therefore has each of the individual meter readings as well as the consolidated readings for each hour of the observation period. The consumption data from the IoT-enabled smart water meters is high frequency data and adequate for a time series analysis using LSTM network model for the household level. The data consist of water readings from a subset of smart water meters for each day during the period of January 2022 to December 2022. Table 3.1 shows a header sample of the data generated by different sensors in the IoT framework.

TABLE1.1

Sample Data Generated by Different Iot-Enabled Smart Devices In The Period June 2022

Time	ManagedObjectid	TypeM	Series	Unit	Value
2022-06-30T07:15:00.000	83008	Pulse1	P1	L	0
2022-06-30T07:15:00.000	61285911	Pulse1_Total	T1	L	11
2022-06-30T07:15:00.000	21967995	Switch2	SW2	State	0
2022-06-30T07:15:00.000	2506	Pulse1_Total	T1	L	531079
2022-06-30T07:15:00.000	61285911	Pulse1	P1	L	0
2022-06-30T07:15:00.000	21967995	Pulse1_Total	T1	L	3357040
2022-06-30T07:15:00.000	2506	Pulse1	P1	L	0
2022-06-30T07:15:00.000	65438040	Pulse1_Total	T1	L	13149680
2022-06-30T07:15:00.000	21967995	Pulse1	P1	L	60
2022-06-30T07:15:00.000	61487282	Pulse1_Total	T1	L	31
2022-06-30T07:15:00.000	65438040	Pulse1	P1	L	10
2022-06-30T07:15:00.000	61201153	Pulse1_Total	T1	L	75220
2022-06-30T07:15:00.000	61487282	Pulse1	P1	L	0
2022-06-30T07:15:00.000	61201153	Pulse1	P1	L	0
2022-06-30T07:15:00.000	324588	Pulse1_Total	T1	L	391963
2022-06-30T07:15:00.0	61144943	Pulse1_Total	T1	L	26592

00					
2022-06-30T07:15:00.000	324588	Pulse1	P1	L	0
2022-06-30T07:15:00.000	65087778	Pulse1_Total	T1	L	11395800
2022-06-30T07:15:00.000	61144943	Pulse1	P1	L	0
2022-06-30T07:15:00.000	21264412	Pulse1_Total	T1	L	2626551
2022-06-30T07:15:00.000	65087778	Pulse1	P1	L	20
2022-06-30T07:15:00.000	21264412	Pulse1	P1	L	0
2022-06-30T07:15:00.000	22607088	Switch2	SW2	State	0
2022-06-30T07:15:00.000	458	Pulse1_Total	T1	L	146085
2022-06-30T07:15:00.000	22607088	Pulse1_Total	T1	L	40146482
2022-06-30T07:15:00.000	65090110	Pulse1_Total	T1	L	13968890
2022-06-30T07:15:00.000	458	Pulse1	P1	L	0
2022-06-30T07:15:00.000	22607088	Pulse1	P1	L	0
2022-06-30T07:15:00.000	65090110	Pulse1	P1	L	20
2022-06-30T07:15:00.000	46058019	Switch2	SW2	State	0
2022-06-30T07:15:00.000	66214272	Pulse1_Total	T1	L	33
2022-06-30T07:15:00.000	621763	Pulse1_Total	T1	L	116204
2022-06-30T07:15:00.000	46058019	Pulse1_Total	T1	L	31957753
2022-06-30T07:15:00.0	66214272	Pulse1	P1	L	0

00					
2022-06-30T07:15:00.000	621763	Pulse1	P1	L	5
2022-06-30T07:15:00.000	46058019	Pulse1	P1	L	290
2022-06-30T07:15:00.000	65088496	Pulse1_Total	T1	L	7886600
2022-06-30T07:15:00.000	61183262	Pulse1_Total	T1	L	23
2022-06-30T07:15:00.000	96372	Pulse1_Total	T1	L	151603
2022-06-30T07:15:00.000	65088496	Pulse1	P1	L	0
2022-06-30T07:15:00.000	61183262	Pulse1	P1	L	0
2022-06-30T07:15:00.000	96372	Pulse1	P1	L	0

In the above table the columns can be described as follows:

- “Time” represents the timestamp using the local time, as captured at 15-minute time intervals, 30-minute intervals, hourly and daily depending on the type of smart meter that is capturing the data.
- "ManagedObjectID" represents the ID of smart meters which are managed within the residential properties in the study are. Each is a unique number that allows the tracking and monitoring of the smart meter within the internet of things framework.
- "TypeM" indicates the type of measurement that is being recorded (such as water flow, battery voltage, average temperature, etc.)
- "Series" is the sub-type of the measurement being recorded in “TypeM” above (such as volt, min, median, etc.). For example the smart meter embedded in a battery will record a reading in voltage. The first pulse value of a smart water

meter can be recorded as P1 while the total value collected within a 1-hour interval is labelled as T1.

- "Unit" represents the unit of measurement, which can be a measurement of water flow in liters, the state of a switch as off (0) or on (1), among other units of information.
- "Value" is the actual value that is registered. All the recorded values in this field are quantities.

The entire dataset made up of 99,370,800 records was collated and cleaned in the preprocessing steps that have been discussed below.

3.2.2 Data preprocessing

To generate reliable results and accurate interpretation of the data to be collected, the data was cleaned and processed. This step involved a series of steps where data was cleaned, normalized, and pre-processed in preparation for model training to ensure the quality and accuracy of the analysis. A deep understanding of the data structures and potential flaws of the data were considered during the analysis. The data was therefore assessed and validated before proper analysis could be done. This included detecting faulty sensor data, outliers, and other obvious flaws such as unreasonable consumption.

3.2.3 Feature selection

After preprocessing the collected data, the next activity involved feature selection, a machine-learning process meant to identify important features of the dataset to improve LSTM model performance. Filter methods in R programming language were used to identify the properties of feature through univariate statistics. Additionally, the study used information gain to compute the reduction in entropy from a transformation of the original dataset. Feature selection then identified the

information gain of each feature in the context of the target variable. The feature selection was to a great extent informed by correlation results from the exploratory data analysis.

3.2.4 LSTM model development

The key deliverable is an LSTM model of smart water meter-generated water consumption and water management data. The model was developed using a deep learning RNN approach that is more suitable for time series data. A typical LSTM unit has 4 components: 1) cell state–C_t, 2) input gate–i_t, 3) output gate–o_t, and 4) forget gate–$f_t$as depicted in Figure 3.3.

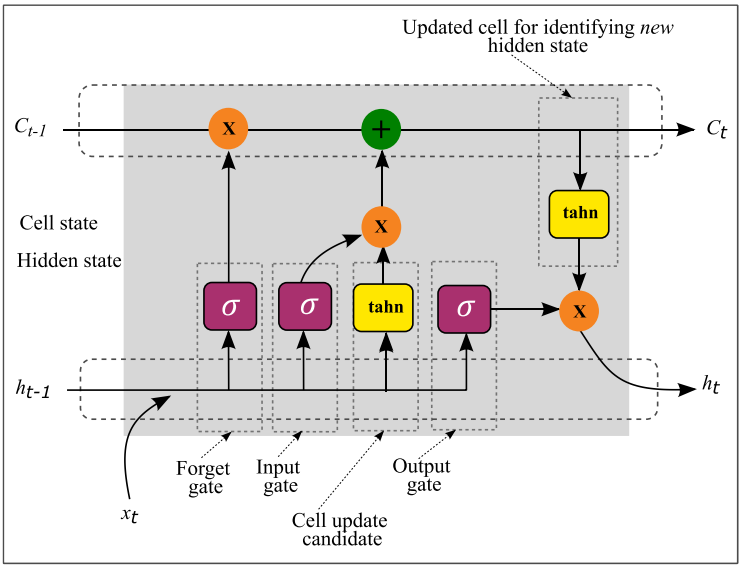
Cell states exchange information with each other. Each cell state is responsible for transmitting historical information (C_{t-1}) and current updated information (C_t) to the subsequent cell via forget, input and output gates in order to update the information. The forget gate is charged with receiving new inputs (x_t) and past hidden states (h_{t-1}) which are used to determine the information to be communicated to the cell state. At the same time, the input gate determines which information should be updated among the various series of new information and subsequently create a new information value ($\bar{C}_t$) using an activation function (tahn). The output gate thereafter identify the information that will form the output. The LSTM equation is as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1,} x_t\right] + b_f\right),$$
$$i_t = \sigma\left(W_i \cdot \left[h_{t-1,} x_t\right] + b_i\right),$$
$$o_t = \sigma\left(W_o \cdot \left[h_{t-1,} x_t\right] + b_o\right),$$
$$\tilde{C}_t = tahn\left(W_C \cdot \left[h_{t-1,} x_t\right] + b_C\right),$$
$$h_t = o_t \cdot tahn(C_t),$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

where W_f,W_i, W_o and W_C represent the weights of the forget gate, input gate, output gate and cell state respectively (Kim et al., 2022). The structure of the LSTM is as shown in Figure 3.3.

FIGURE 3.3

LSTM network model structure



To test suitability of the LSTM model for predicting water consumption in residential properties, performance of the model was compared against the traditional ARIMA model.

3.2.5 Model training

A model is a simplified representation of reality, and it is therefore important that a model is sufficiently trained using data from the real life setting in order to calibrate it with reality. Usually, the training time is a significantly longer period and uses a larger proportion of the study data set. The pre-processed dataset was randomly split into

model training (70%) and validation data (30%). The training dataset was then used to train both the traditional ARIMA model and the LSTM network model on historical data so as to allow the models to learn patterns, relations and anomalies that exist in the data. The model parameters were adjusted accordingly, and this process applied optimization approaches to minimize the prediction errors. The ARIMA model was trained using the historical water consumption data and the periodicity of these data. Periodicity refers to the sequences of the observation values which have been captured at equal time intervals such as 15 minutes, 30 minutes, one hour or daily. The LSTM model was trained using both the periodicity and other external factors that are shown in the conceptual model.

3.2.6 Model evaluation

Evaluation involves confirming the extent to which the developed model effectively represents the phenomenon that is being observed. This is a very critical process since it validates the usefulness of a model and its applicability within the real-world setting. During the model evaluation stage, a comparison was made on the prediction results for data observed by both ARIMA model and LSTM model in order to select the better-performing model as the final model for adoption. The metrics to be used were correlation coefficient (CC) and the root mean standard error (RMSE). The CC measure is a normalized measure of covariance that was used to assess the linear correlation between the predicted values and observed values of the consumption dataset. The CC value is a ratio with values between 1- and +1, where values closer to -1 indicate negative correlation, values closer to +1 indicate positive correlation and values closer to zero indicate weak or no correlation. The ratio is presented as:

$$CC = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}}_i)^2}}$$

where $\bar{y}$ is the observed mean value and $\bar{\hat{y}}$is the predicted mean.

The metric, Root Mean Square Error (RMSE) is one of the key accuracy metrics that can be used to ascertain the accuracy of a model for predicting a particular response to regression problems. The RMSE calculates the square-root value for the residual variance and can be used to calculate how well the model fits to the study data, or how close the observed data gets to the values that have been predicted by the model (Kühnert et al., 2021). In this study, the RMSE metric was used to indicate inherent errors in the predicted and observed values. A lower RMSE value indicated an increase in performance. The RMSE equation is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}$$

where n is the number of observation points, where y_i is the observed value and $\hat{y}_i$ is the predicted value.

3.3 Ethical considerations

Research ethics and ethical behavior are concerned with the moral aspects of a scientific study and are the key pillars of a civilized society. The focus on ethical behavior is indispensable in many research fields, especially those involving analysis of data. In practice, ethics applies to everything that would include, affect, transform, or influence upon individuals, communities or living creatures. The domain of research ethics mainly deals with analysis of what is good for society and individuals. Applications of ethics to fields as diverse as medicine and environmental protection are now well-established, and data ethics is developing into a distinct branch of applied ethics. The researcher in this study was honest and accurately reflected the work done

using appropriate methods and techniques for data analysis and interpretation. Decision about the research design and the method of presenting results were justified to develop trustworthy research. The study revealed all the research limitations and challenges in implementing any of the research methods.

Finally, regarding the study participants, all responses and data collected about individual participants were treated with confidence. The researcher sought permission from National Commission for Science, Technology and Innovation (NACOSTI) before collecting the data, and all participants were informed about the use of the data that they provided. The processing of data obstructed all information that could be linked to a single participant.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the outcomes of applying the LSTM network model to smart water meter time series data related to residential properties water consumption. It delves into the insights gained from modelling the interactions among the variables and explores the potential of the model for accurate forecasting. Through a rigorous analysis of the results and findings, this chapter seeks to contribute to the growing body of knowledge water consumption and urban water distribution research and aid in the design of more targeted interventions, policy formulation, and resource allocation. The subsequent sections of this chapter will provide a detailed account of the observed results, model estimation, validation, and interpretations.

4.2 Descriptive Analysis of Domestic Water Consumption Trends

Table 4.1 below provides a concise overview of various critical factors related to residential water consumption and its associated conditions across the observation period of January to December 2022. It presents both the cumulative values for the entire period and the values observed during each quarter of the observation year 2022. The factors under consideration include water flow, temperature readings, battery voltage and water pressure.

The data within the table unveils distinct patterns and variations across the quarterly time periods, highlighting potential relationships between these factors and the variation in water consumption. Moreover, the distribution of temperature and water

pressure measures also varies across the observation period. These variations may contribute to the differences observed in water consumption levels and underscore the importance of effective water distribution interventions.

TABLE 4.1

Descriptive Analysis of Water Consumption and Related Factors

Factor	Total Quarterly Values in year 2022				Total Across the year 2022
	Jan-Mar	Apr-Jun	Jul-Sep	Oct-Dec	
Average water flow in liters (SD, Min - Max)	1,681,865 (830,919, 763,288 – 1,429,632)	1,221,409(795,652, 598,361 – 1,983,989)	1,380,010 (234,812, 984,786 – 2,150, 465)	1,746,068 (510,919, 798,511 – 2,127,989)	1,431,364 (906,327, 598,361 – 2,150, 465)
Average water pressure (SD, Min - Max)	45 (15.45, 39 – 53)	46(17.61, 36 – 51)	51(18.41, 42 – 57)	49 (12.78, 40 – 52)	47 (16.63, 36– 57)
Average temperature (SD, Min - Max)	26.14 (3.78, 27.78 – 24.49)	20.62 (3.52, 16.09 – 22.38)	18.22 (3.41, 17.11 – 20.87)	26.55 (6.78, 26.13 – 29.39)	20.55 (8.79, 16.09 – 29.39)
Average battery voltage (SD, Min - Max)	3.63 (0.042, 3.62 – 3.64)	3.63 (0.071, 3.62 – 3.64)	3.62 (0.052, 3.61 – 3.64)	3.64 (0.081, 3.61 – 3.67)	3.63 (0.097, 3.61 – 3.64

Based on the results shown in Table 4.1 above the following can be observed:

Water Consumption: the average water consumption measured as the water flow was very variable across the observed quarters of the observation year, 2022. The fourth quarter (October – December 2022) had the highest average consumption of 1,746,068 liters and the second quarter (April – June 2022) had the lowest consumption of 1,221,409 liters. Nevertheless, deviation of consumption during this second quarter was quite high at 795,652, as compared to the fourth quarter that registered a standard deviation of 510,919 liters. This showed that there was high variation in the daily consumption rate within the months of April, May and June 2022. The average yearly consumption rate was 1,431,364. The deviation rate was also high at 906,327, indicating high variability in consumption across the observation months.

Water pressure: The water pressure was also relatively variable for the months observed. The months between January and March of 2022 (first quarter) had the lowest average pressure reading of 45pounds per square inch (PSI). The third quarter of July to September had the highest pressure reading of 51 PSI, but this month also had the highest standard deviation (SD = 18.41 PSI). Cumulatively, there was an average of 47 PSI as the average pressure reading recorded for the months of 2022.

Average temperature: Moderate temperature was observed throughout the year, with only slight variation in average temperature being observed during the second (April to May) and third (June to August) quarters. The lowest average temperature was registered at 18.22 degrees centigrade during the third quarter. These are usually the coldest months of the year. The highest average temperature (at 26.55 degrees centigrade) was experienced in the fourth quarter of 2022. Low standard deviation

values for the temperature ranging between 3.41 and 6.78 degrees centigrade suggested some level of stability in daily values recorded.

Battery voltage: Battery voltage values measured in V unit was also relatively stable, suggesting that the smart device batteries were in optimal condition, and they could efficiently manage the smart water meter data. The lowest voltage recorded was during the third quarter of the year at 3.63 V (SD = 0.052 V). This quarter also recorded the lowest average temperature values. The highest voltage was recorded at 3.64 during the fourth quarter of 2022. This quarter had the highest average temperature values. This observation indicated the possibility of a linear relationship between the average temperature and the average battery voltage for the IoT-enabled smart devices.

In general, the values for all variables were highly varied with contrasting averages over the observation yearly quarters. While the standard deviation was not higher than the mean and suggested that the data was not highly skewed, high standard deviation values were observed, especially with the domestic water consumption rates.

4.3 Correlation Analysis of the Study Variables

Table 4.2 below provides a correlation matrix for all the variables that were included in the current study. The correlation value is a ratio with values between 1- and +1, where values closer to -1 indicate negative correlation, values closer to +1 indicate positive correlation and values closer to zero indicate weak or no correlation(Kühnert et al., 2021). This research assessed the correlation among a set of 12 different study variables, of which one was the dependent variable (i.e., the water consumption for the current day, Ct), one was a moderating variable (i.e., the usage cost, UC), and the remaining variables were independent variables. All variables with a

positive correlation value greater than 5 have been highlighted in gold while the variables with a negative correlation value higher than 5 are highlighted in blue.

In agreement with the study hypothesis and previous studies, all the independent variables had some level of relationship to the dependent variable (C_t), with a correlation value greater than 0.3. The relationship between the independent variables themselves was not very high since no two independent variables had a correlation value higher than 5.5. The test of multicollinearity using variance inflation factors indicated the absence of the risk since the largest value was 2.18 which was far below the cut off mark of 10.

TABLE 4.2

Correlation matrix of study variables

Variable		1	2	3	4	5	6	7	8	9	10	11	12
1	C_t	1											
2	UC	.342**	1										
3	C_{t-1}	-.684**	-.512**	1									
4	C_{t-2}	.561**	-.471*	.542**	1								
5	C_{t-3}	.405*	-.319*	.375*	.402**	1							

6	C_{t-4}	.421**	-.308**	.292*	.371**	.408**	1							
7	C_{t-5}	.409**	-.328**	.369**	.450**	.415**	.422**	1						
8	C_{t-6}	.411**	.298*	.308**	.384**	.392**	.402**	.419**	1					
9	C_{t-7}	.385*	-.408**	.289**	.297**	.305**	.311**	.318**	.348**	1				
10	T	-.527*	.287*	-.306**	-.208**	-.451*	.328*	.444	-.548**	-.511**	1			
11	P	.432***	.271	.477**	.458**	.385*	.451**	.285**	.524*	.425*	.512*	1		
12	Wk/Wn	.433*	.278*	.333**	.578**	.473**	.527**	.472**	.258**	.358**	.378**	.222**	1	

Notes: * Correlation is statistically significant at the 0.05 level (2-tailed); ** Correlation is statistically significant at the 0.01 level (2-tailed); *** Correlation is statistically significant at the 0.001 level (2-tailed)

C_t = Water consumption per liter for current day, t; UC = Usage cost per liter; C_{t-1} = The water consumption for previous 1 day, t-1; C_{t-2} = The water consumption for previous 2 days, t-2; C_{t-3} = The water consumption for previous 3 days, t-3; C_{t-4} = The water consumption for previous 4 days, t-4; C_{t-5} = The water consumption for previous 5 days, t-5; C_{t-6} = The water consumption for previous 6 days, t-6; C_{t-7} = The water consumption for previous 7 days, t-7; T = Temperature; P = Water pressure; Wk/Wn = Water consumption on weekend versus weekday

The highest level of correlation existed between the water consumption (C_t) and the previous day consumption, and this relationship was positive (r = .684). This indicated that an increase in the previous day consumption was likely to increase the consumption of the current day. Similarly, all historical consumption patterns, up to day 7 prior to the observation day were positively correlated to the day's consumption and were statistically significant at the 0.01 level. However, the strength of correlation decreased with the increase in number of days from the current consumption day.

The highest negative correlation existed between temperature and water consumption for the previous 6 days (C_{t-6}) at r = -0.548. Other significantly high relationships existed between usage cost (UC) and water consumption for previous 1 day (C_{t-1}), as well as between temperature and water consumption for the current day, at r = -0.512 and r = -0.527 respectively. Indeed, the temperature was negatively correlated with most of the other independent variables, even though this correlation was weak and was not statistically significant for two variables, usage cost and water consumption for previous five days (C_{t-5}).

The moderating variable, usage cost (UC) had a negative correlation with the consumption at the current day as well as with all previous consumption values except the water consumption for previous six days (C_{t-6}). It is worth noting, however, that this last correlation was not statistically significant. The correlation between usage cost and water pressure was also not statistically significant. Usage cost had a weak positive correlation with the water consumption for weekday versus weekend (r = 0.278), but this correlation was statistically significant at the level of 0.05.Only two of the variables under observation, i.e., water consumption on the current day (C_t) and water pressure had a correlation that was statistically significant at the 0.001 level. This was a positive relationship of r = 0.432.

4.4 Graphical Depiction of the Average Trends in Data

Figure 4.1 shows the pattern of domestic water consumption what was observed for each day of the week (Figure 4.1A) and for each day of the month (Figure 4.1B). The average consumption in the residential properties was fairly uniform for most weekdays, with between 70 and 80 liters per day being consumed per household. Saturday experienced a surge in demand and registered more than double the daily usage. On Sunday there was a dip in demand to return to the consistency that characterized the weekdays. In contrast, the usage per day of the month was highly varied and there were a lot of fluctuations. The first four days of the month had a progressive upward increase in water consumption. On day 5 of the month there was a sudden drop and subsequent fluctuation until the middle of the month when the consumption levels started to increase steadily again, up to a peak of 200 liters per day on day 19. Afterwards the pattern was variable until day 27 when consumption dropped sharply until the last day of the month when the lowest level was recorded. From the general observation, the domestic water usage increased between the middle of month and towards the end of the month, then steadily dropped in the last three days of the month.

FIGURE 4.1:

Average domestic consumption for each day of the week (A) and day of the month (B)

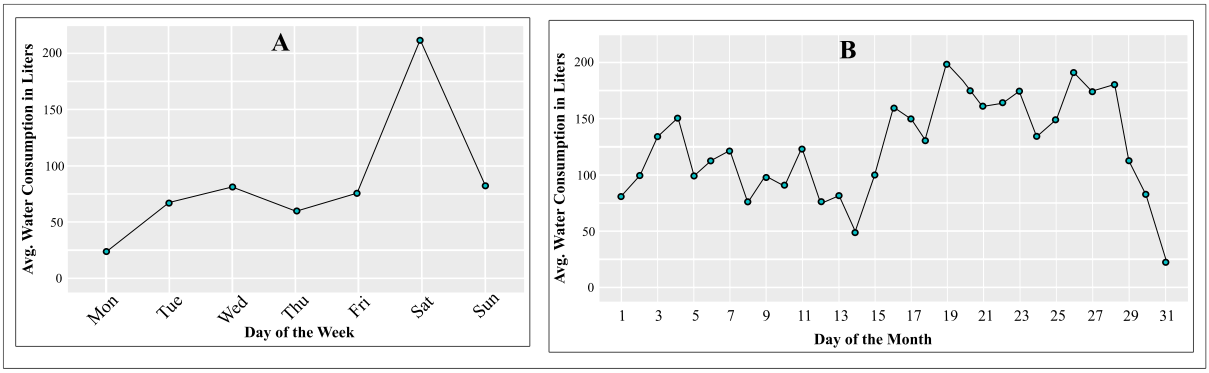


Figure 4.2 shows the trends in domestic water consumption for each month of the observation year (6A) against the average temperature that had been observed for each moth (6B). A distinct pattern was observed in monthly usage of water in the residential properties. Specifically, the first four months of the year had the highest usage, but the consumption level decreased slightly in each subsequent month. There was a sudden decrease in water consumption between the months of June and July of 2022. In general, consumption was lowest in June, July and August as compared to the remaining months of the observation period. These three months also experienced the lowest average temperatures in the year (between 19 and 22 degrees centigrade), unlike the earlier months which had much higher temperatures of up to 28 degrees centigrade. A steady increase in water consumption was observed in the subsequent months, but there was a sudden drop in the consumption pattern in the month of December 2022. This month was also the warmest month of the year with an average temperature of 28 degrees centigrade.

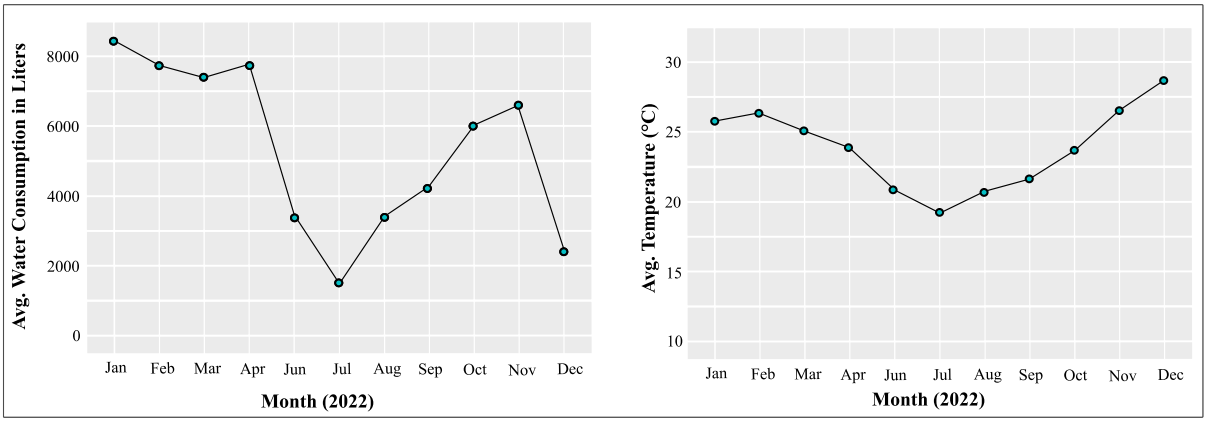


Figure 4.2: Average monthly water consumption (A) and temperature (B)

4.5 Prediction of Domestic Water Consumption Patterns in Training Data for LSTM

Figure 4.1 shows results from the prediction of consumption patterns in the training data for the LSTM network model as a time series. The broken green lines show the observed consumption patterns, and the red lines show the predicted patterns. Comparison of the two lines shows that the consumption patterns are almost similar during the training phase. Aside from the months of April and in mid-October where the LSTM predicted values that were much higher than the values that had been observed, the rest of the observation period saw consistency between predicted and observed values. Performance metrics portrayed a high value of 94.3% of the correlation coefficient with an RMSE value of 0.22, both signifying good model performance with the training dataset. Table 4.3 portrays the result of the validation tests using the correlation coefficient and RMSE metrics.

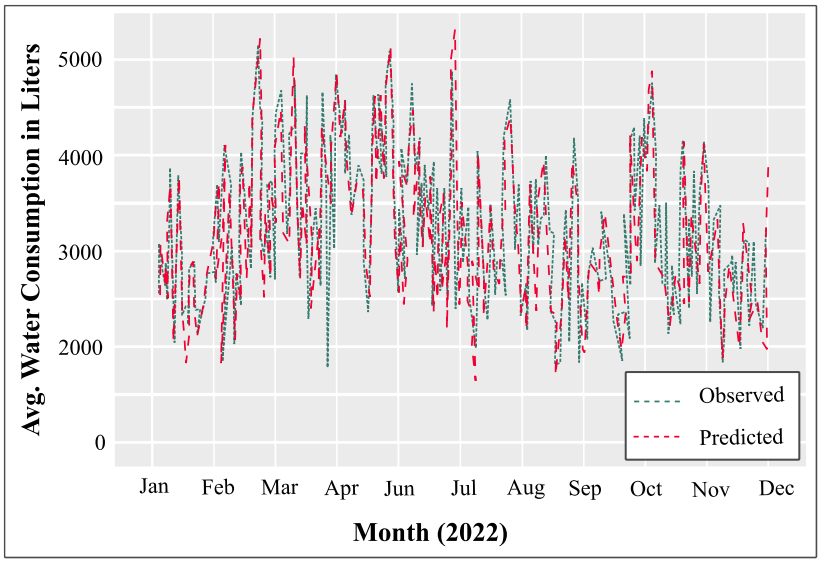
TABLE 4.3

Performance metrics of LSTM network model using the training dataset

Correlation Coefficient	RMSE
94.341%	0.223

FIGURE 4.3

Time series results of the predicted and observed consumption patterns by the LSTM network model.



4.6 Comparison of LSTM and ARIMA Models using the Validation Dataset

The LSTM model was thereafter compared with the ARIMA model using the validation dataset. This was done in order to test suitability of the LSTM model for predicting water consumption in residential properties in comparison with a conventional model in line with the case study research approach that has been adopted for this study. Significant variation was observed in the generated outcomes from the two models. Table 4.4 below shows the performance metric for ARIMA and LSTM. Based on the correlation coefficient values, LSTM outperformed ARIMA by more than 20%. The root means square error (RMSE) value echoed this performance by portraying a much lower value for LSTM (i.e., 3.61), which showed that the LSTM predicted with far less errors than ARIMA which had an RMSE value of 14.32. Figure 4.4 shows the consumption values predicted by the models against the observed values.

TABLE 4.4

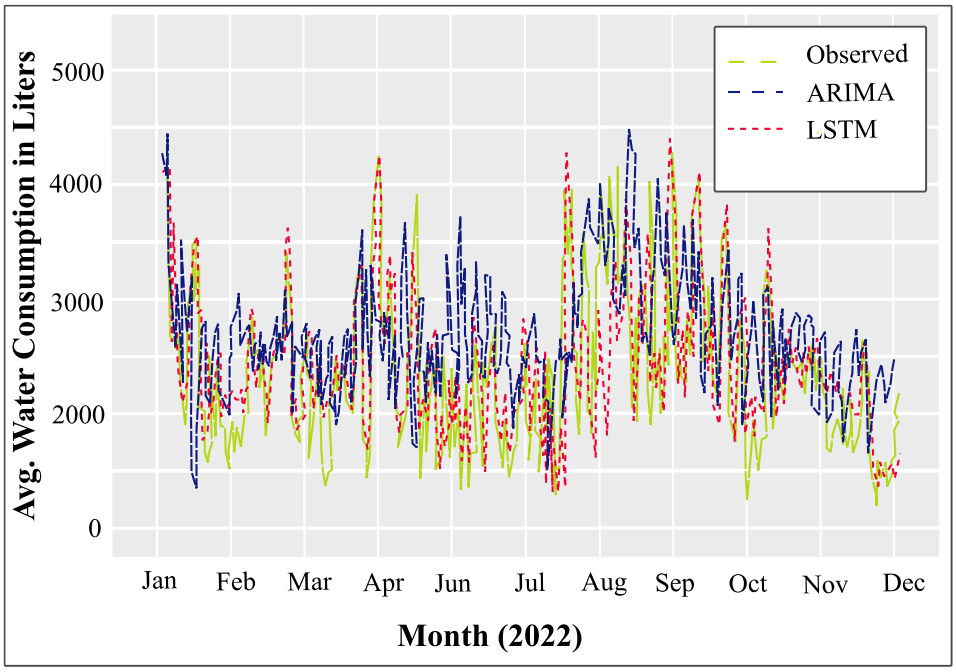
Performance metrics of ARIMA and LSTM models using the validation dataset

Model	Correlation Coefficient	RMSE
ARIMA	93.87%	3.61
LSTM	71.43%	14.32

The time series results shown in Figure 4.4 are presented using a broken green line for observed values, a broken blue line for ARIMA model output, and broken red lines for LSTM prediction using the validation data. Results indicate high consistency between the observed pattern and what was predicted by the LSTM network model. However, the ARIMA model over-estimated the observed values in several instances. This was more so within the months of February, August and September. The performance deviation echoed the metrics that had been observed in Table 4.2.

FIGURE 4.4

Time series output of predicted values from ARIMA and LSTM models against the observed values



4.7 Discussion of Results

It can be acknowledged that smart water meters have the potential to provide useful information to both operational managers as well as to urban water consumer, and that it is important to ensure that there are benefits for both associated with the deployment of smart meters and sensors at the apartment level. By accurately predicting water consumption levels and empowering the consumers to proactively manage their consumption, potential behavioral change as well as transparency can be achieved (Söderberg & Dahlström, 2017. The use of consumer feedback to encourage consumption conservation has been widely applied and evaluated in other sectors but have not been documented to a similar extent in managing water consumption. The provision of more frequent and detailed consumption information and feedback has been found in several studies to result in significant water savings, whereas others are critical arguing that there is little evidence at the time whether high frequency feedback is effective in reducing consumption in the water sector or not. Nevertheless, smart metering offers an improved possibility to supply end users with feedback can potentially promote water savings. In the current digital age, information and feedback in real-time or near-real time could easily be provided. One example is through the availing of individual digital page to consumers to monitor their own consumption levels in real time.

The modern use of artificial intelligence allows the anticipation of urban residential consumption output and environmental implications to make accurate predictions. Such technologies are becoming popular due to their ease of use, adaptability, and utilization of historical data to predict future energy usage patterns under limits. Smart water metering when used together with precise analytics models of end used consumption patterns can allow access to more detailed information on urban household water

consumption and how the water is being consumed much more precisely that ever before. The information needs to be made available to a number of stakeholders and be simple enough to be understandable in order to make the insight meaningful. How to interpret the information generated from the analysis is not always straight forward, and this section discusses the result of the analysis and its interpretation as combined with other important features and findings.

In cognizance of the severe mismatch in demand and supply of water in urban residential areas of most cities and urban areas in Kenya, this research adopted Nairobi city as a case study to construct a prediction model for water demand. The baseline results that were provided in the descriptive study were meant as a first orientation of the data. The outcome from the descriptive analysis provides a frame of reference to further analysis and other studies. The average daily water consumption in the residential properties was fairly uniform for most weekdays, with between 70 and 80 liters per day being consumed per household. Noteworthy is that this consumption rate was highly dynamic and fluctuated across the observation days. Saturday experienced a surge in demand and registered more than double the daily usage. On Sunday there was a dip in demand to return to the consistency that characterized the weekdays. Similar patterns have been observed in previous studies where water consumption has been found to fluctuate as based on the day of usage (Söderberg & Dahlström, 2017; He et al., 2021).

Observations over a longer period of time than one week did not reveal stabilization of the water consumption patterns over time. In contrast, the usage per day of the month was highly varied and there were a lot of fluctuations. The first four days of the month had a progressive upward increase in water consumption. On day 5 of the month there was a sudden drop and subsequent fluctuation until the middle of the month when the

consumption levels started to increase steadily again, up to a peak of 200 liters per day on day 19. Afterwards the pattern was variable until day 27 when consumption dropped sharply until the last day of the month when the lowest level was recorded. From the general observation, the domestic water usage increased between the middle of month and towards the end of the month, then steadily dropped in the last three days of the month. Studies such as Kim et al., (2022) have similarly conducted exploratory data analysis over a data set of water consumption in Korea and discovered similar results. In particular, the authors found out that earlier days of consumption registered high usage of water which reduced as the month progressed, then the ending days of the month experienced increased water consumption rates again. The authors attributed these fluctuations to the mobility of residents which sometimes caused them to be away for work or on vacation and therefore caused less usage of water (Kim et al., 2022). Such information can be useful in determining flexible cost schedules that can take peak and off-peak hours into account.

A correlation analysis of the study variables showed a good level of relationship between all the independent variables with the dependent variable. This was more so with the variable representing water consumption for the previous 1 day that had a strong positive correlation with the dependent variable (i.e., $r = 0.684$). This showed that historical water consumption patterns can be used to determine the subsequent usage of water in urban residential properties. However, the observed positive correlation became weaker decreased as the number of days from the current consumption day increased until by day 7 it was relatively weak at $r = 0.385$. This implied that historical patterns can be useful, but only up to a certain number of days prior to the current consumption day. Similar results have been observed by Krishnan et al (2022) where the authors found a significant relationship of historical usage with the

water consumption patterns in an extensive review of previous water consumption studies.

Although some level of relationship existed among the independent variables in the study, the variance inflation factors indicated the absence of serious multicollinearity which can affect the analysis. A deep learning LSTM network model was developed for the IoT-enabled smart water meter data using a split dataset of 70 percent training and 30 percent validation while controlling for factors such as previous consumption and daily temperature. Initial assessment of the study data revealed water consumption patterns that differ widely from previous observations in Swedish, German, and Korean cities (see e.g., Söderberg & Dahlström, 2017; Kühnert et al.,2021; Kim et al., 2022). Three key contextual differences inherent in African cities account for this variation. Firstly, the population density of most African cities is much higher than for cities in the developed world. For example, Germany has a density of 4,126 people per km2, while Nairobi has 6,317 people per km2. Secondly, authorities in the developed world countries prescribe maximum limits on house occupancy based on the house size in square meters or the number of bedrooms, but most African governments place no such constraints. Thirdly, while most developed world cities have daycare centers, African households predominantly employ household nannies.

Finally, lack of preventive maintenance of the water supply infrastructure causes susceptibility of the water system to leaks. Cumulatively, these aspects result in non-uniform consumption patterns. It was evident in this study, for example, that water usage reduced substantially during certain days of the week, such as on Sundays. This can be attributed to individuals who spend time away from home. However, Saturday is generally a day for household cleaning and may explain why water usage is highest on this day. The month of December is mostly spent on vacation since schools close

during this month and many families move to the rural areas to spend the Christmas season with the extended family. This explains the low water usage during this month. Nevertheless, high variability in the consumption patterns limits the applicability of models that have been developed for monitoring water use in the Western world.

Overall, the developed model showed high accuracy levels for both the training and validation datasets, and it also outperformed the ARIMA model that has been conventionally used for water predictions. Results show that the LSTM model can maintain consistent performance even upon addition of new data. As such the data generated from the LSTM model can provide insight on the supply adjustments to be made to different areas of the city, the pricing variations, and any potential water loss due to leakages.

CHAPTER FIVE

CONCLUSIONS

5.1 Introduction

The importance of forecasting the daily, short term and future water consumption patterns in urban residential households cannot be overstated. There has been a gradually increasing need to determine this consumption, even in the face of many variables which interact to make the prediction of water consumption a complex task. In seeking to address this problem, this chapter revisits the objectives of the study and the key findings, while also presenting the implications of findings and the entire research to the management of smart water meter data. presents the outcomes of applying the LSTM network model to smart water meter time series data related to residential properties water consumption. It delves into the insights gained from modelling the interactions. The chapter further reflects upon the contributions made by the research to the field of domestic water consumption estimation and elucidates the broader significance of employing LSTM models for predicting this consumption.

5.2 Key Results from the Study Objectives

This study set out with the main objective of establishing an LSTM network model, a deep learning model used to predict water consumption in residential properties using smart water meter data. Based on this main objective, the current study has achieved the following results from its specific objectives:

5.2.1 Specific Objective 1: To assess the factors influencing variation in residential water consumption.

To achieve the first specific objective of assessing the factors influencing variation in residential water consumption, this study conducted extensive literature

review of current research and identified several influencing factors. Among these factors were the time-of-day patterns which were found to be significantly correlated with water consumption in residential properties. Households experienced peak usage of water was determined to be during off-peak hours (Söderberg&Dahlström, 2017). Similarly, usage patterns were found to affect consumption of water through the pipe-based water supply in the Pune city of India (Singla &Bendigiri, 2019).

With respect to how the water flows can affect water consumption, it was found out that household size heavily influenced the average water consumption in the Greater Southeast of England (Hargreaves et al., 2019). Additionally, that the influence of community factors in water consumption and management was found to be influential in Shanghai, China (Han et al., 2021). The authors discovered that some apartments rented by several tenants were the main consumers of residential water. Further assessment of the existing literature revealed that seasonality is a factor influencing water usage and water management (He et al., 2021). During the dry season water tends to be used more as residents' water plants and dust or wash more. Similarly, during the rainy and cold season less water tends to be used. Additional findings revealed that an increase of pressure in the pipeline can cause water supply to rise and automatically increase demand for water (Dream Civil, 2022). This meant that water pressure was an important factor to consider when determining the water usage patterns for residential areas.

Using the variables that had been identified from the literature, this study formulated a set of hypotheses that were the basis for analysis of the study dataset. Descriptive and correlation of the dependent and independent variables showed results that echoed the previous findings indicated above and led the current study to confirm the hypotheses. Given the consistency of the observed results from this objective when

compared with the results from later stages, it can be concluded that the study objective has been sufficiently achieved.

5.2.2 Specific Objective 2: To design and develop an LSTM network model for processing time series data to predict water consumption and detect anomalies.

The second specific objective was addressed mathematically in the formulation of the LSTM network model for analysis of time series data from the smart water meters. The data used to build the LSTM network model comprised bulk data generated from many smart water meters located in 320 residential properties within several estates in Nairobi city that are managed by MajiPay, a local smart water meter service provider. The smart water meters collect water consumption data at 15-minute intervals. Several other sensors within the IoT framework collect accompanying descriptive data, such as 1) battery voltage data which indicates the state of the smart water meter and is important for water management, 2) timestamp in date and time, 3) average temperature for each day, and 4) water pressure. The consumption data from the IoT-enabled smart water meters was high frequency data and was considered as adequate for a time series analysis using LSTM network model for the household level. The data consisted of water readings from a subset of smart water meters for each day during the period of January 2022 to December 2022.

Results from the model showed that although significant variation in daily, hourly, weekly and monthly consumption patterns had been observed, the LSTM model was relatively accurate in predicting the consumption levels for the urban residential households. In agreement with the study hypothesis and previous studies, all the independent variables had some level of relationship to the dependent variable (Ct), with a correlation value greater than 0.3. The relationship between the independent variables themselves was not very high since no two independent variables had a

correlation value higher than 5.5. Nevertheless, the test of multicollinearity using variance inflation factors indicated the absence of the risk since the largest value was 2.18 which was far below the cut off mark of 10.

5.2.3 Specific Objective 3: To test and validate the developed model.

The observation data from the smart water meters was split randomly into 70 percent test and 30 percent validation data sets. The training data was used to train the LSTM model until a high level of accuracy was observed. The correlation coefficient for LSTM on the training dataset was found to be 94.341 with RMSE value of 0.223. Aside from the months of April and in mid-October where the LSTM predicted values that were much higher than the values that had been observed, the rest of the observation period saw consistency between predicted and observed values. Performance metrics portrayed a high value of 94.3% of the correlation coefficient with an RMSE value of 0.22, both signifying good model performance with the training dataset.

During the testing phase of the LSTM model, results from this model were compared with the results that had been generated by the traditional autoregressive integrated moving average (ARIMA) model. A profound difference in model performance was observed, with the LSTM outperforming the ARIMA by more than 20%. The prediction values from LSTM closely matched the observed values, whereas the ARIMA predictions were overshot on several occasions. The root mean square error (RMSE) value echoed this performance by portraying a much lower value for LSTM (i.e., 3.61), which showed that the LSTM predicted with far less errors than the errors generated by ARIMA which had an RMSE value of 14.32. This confirmed the stability of the trained LSTM model for predicting water consumption patterns in urban residential properties.

5.3 Key Contributions of the Current Research

This research sought to make significant contributions to the existing research, and in this regard the study has to a great extent achieved this goal. One of the primary contributions of this project lies in its adoption of a multivariate approach. Traditional univariate time series analysis for instance as conducted by (Muthukumar et al, 2019) often overlooks the interconnectedness of various contributing factors, leading to limited accuracy in forecasting. By incorporating a diverse set of independent variables, the project acknowledges the interplay between previous consumption patterns, daily temperature and water pressure resulting in a more holistic understanding of how these factors jointly influence water consumption in urban residential areas.

There are very few studies that have addressed domestic water consumption in urban areas using a multivariate approach. Furthermore, even the few studies that exist have been conducted in urban areas of Europe, USA and other parts of the developed world (Kühnert, 2021). This research project has undertaken a multivariate analysis of time series water consumption data within a developed world country, Kenya, while also considering other factors such as previous day consumption, previous week consumption, average pressure, average monthly temperature and water pressure. The incorporation of these multiple variables that can potentially influence domestic water consumption patterns underscores the project's practical relevance. By integrating these factors into the analysis, the research recognizes the multifaceted nature of domestic water consumption. This expands the scope of the forecast beyond mere flow factors, and it acknowledges the importance of historical data and environmental influence which can potentially affect consumption levels.

Another strength and important contribution of the current research is with regards to the technical details of the study data. The smart meters which were used to

measure water consumption employ a pulse sensor. This is a valuable addition to knowledge, especially in view of the level of sophistication of the pulse sensor as compared to accumulation sensors that have commonly been used in the past (Söderberg & Dahlström, 2017). When data contains accumulated consumption, it is practically impossible to pinpoint the water usage to a certain end user consumption activity, such as flushing the toilet or taking a shower because of the aggregation. For example, concurrent activities such when one takes a shower when a washing machine is running, and sequential activities such as showering and breakfasting in the morning, are all aggregated into single, hourly readings of the consumption volume. Therefore, several analysis approaches and techniques that have been applied in the past have not been sufficient. The use of pulse readings provides a novel method for identifying water consumption patterns within accumulated consumption meter data. Even if the resolution is higher, there would be no possibility to pinpoint human activity that is attributable to the unit of analysis uses accumulated consumption. In dealing with this problem, the current study has used pulse meter data where a pulse is triggered by a quantum of water passing through the pulse water meter. Pulse meters record both the pulse and the pulse timestamp. This ensures that fine-level details about water consumption have been accounted for.

Another important contribution of this study to the existing literature and body of knowledge is the application of case study research using a large number of apartments under study and over an extended observation time frame studied than what has been observed in most of previous studies. Case study research design is a powerful analysis method that allows the examination of deeper understanding of the observation data, as well as the main factors describing the data set. These characteristics can easily be understood, and patterns of consumption can be discovered through in-depth data

analysis. The focus on who is using the water (e.g., by observing at the house level), as well as information regarding when and how water is being used can be made possible through exploratory data analysis. This study has applied a new method for conducting analysis with the existing data set from 320 residential properties within several estates in Nairobi city. To the researcher's knowledge, there are limited studies to date that have used case study data analysis techniques to discover these patterns. Within other fields and observation domains, case study research has been shown to establish a fundamental understanding of the content, structure and weaknesses of the data when the objectives of analyzing data are not fully understood. However, once data becomes transformed into information and further on to knowledge, it becomes possible to establish the ability to ask and answer more context relevant questions with a lower risk of misinterpretation. More in-depth analyzes can be performed when a clear objective has been presented.

The utilization of an LSTM model adds another layer of sophistication to the analysis. This model considers not only the lagged values of previous consumption but also the lagged values of all the other independent variables. This captures the potential feedback loops and interactions that can occur among these variables. The LSTM model's ability to model -dependencies and interactions among multiple time series provides a valuable contribution to the field of domestic and urban water consumption forecasting. This is because it will address the challenges of modeling complex systems with interconnected variables which is especially crucial in regions with complex and dynamic water consumption dynamics. The project demonstrates the potential of LSTM models to provide deeper insights and more accurate predictions in a variety of urban-related consumption contexts beyond water use and by considering multiple factors that influence water flow and consumption, the LSTM model provides more robust and

reliable forecasts. Indeed, a comparison of the developed LSTM model with the traditional autoregressive integrated moving average (ARIMA) model showed that the LSTM outperformed the ARIMA in terms of prediction accuracy for both the training and validation data sets.

The findings of this study can guide different stakeholders, such as water utility companies and service providers, domestic water consumers, urban planners and policy makers to understand the dynamics of domestic water consumption. This will lead to more efficient efforts in allocating the water resource, detecting leakages and other anomalies which cause consumption to deviate from the normal observed patterns, determination of water prices, and the making of other informed decisions about resource allocation and intervention prioritization. Furthermore, the LSTM model developed in this study is an important contribution to the literature on urban domestic water consumption. The patterns observed here can inform future researchers on how to improve their future models to increase prediction accuracy.

For all urban systems, water consumption management is crucial. Data must therefore be made available to different decision-makers and stakeholders upon request and in real time. This will allow all key stakeholders to make more relevant decisions that can help with intervention measures regarding usage and pricing. In order to gather, analyze and share water consumption data for a variety of purposes, this research work has proposed a conceptual framework comprising one dependent variable, one moderating variable and nine independent variables which can be put into practice. The conceptual framework has been operationalized in the literature review section, and it has also been applied with real time data to develop a Long short-term memory (LSTM) network model for predicting non-linear water consumption patterns. Data gathered continuously from several sources and processed into a common format

were made available under controlled access in order to achieve the study objectives. The primary components of the analysis included exploratory data analysis to discover trends in the consumption data, a predictive analysis using the LSTM network model, and a validation analysis which used the validation data to evaluate both the LSTM model and the more conventional autoregressive integrated moving average (ARIMA) model.

The study's key advantage is that it allows the observation of patterns over non-linear consumption data within residential houses in an urban setting, and it is one out of few studies made in Kenya. This assists the with understanding the level of consistency, standardization, and data sharing in the water management sector, and thus enables key stakeholders to focus more on data analysis rather than data retrieval and manipulation. Results from this study can be extended into a technology that will undergo rigorous testing to be implemented in numerous connected instances. With this method, the most recent data as well as past records can be constantly accessible for immediate inquiries and deeper analysis among all stakeholders within the water consumption ecosystem.

5.4 Limitations of the Current Research

This research has made important contributions, it also has several limitations. One of the limitations of this study is that findings are usually influenced by the geographic area for which the data has been collected. Water consumption patterns can vary widely across different regions due to ecological, climatic, and socio-economic factors. The lack of available sufficient time to conduct this study, as well as the limited knowledge regarding the house owners and tenants involved in the study when conducting analysis made it impossible to incorporate socio economic and demographic factors about the water consumption study. This is because of the level of anonymity

that is required when processing the study data within ethical considerations. It is possible, therefore, that high mobility among Nairobi residents has caused lower consumption patterns to be observed than what can be observed elsewhere, or it has also caused a lot of fluctuations in weekday versus weekend usage patterns. Therefore, the results obtained which were based in Nairobi region might not be directly applicable to other regions without proper validation. Nevertheless, it is possible to argue that the factors that have been presented in this study can be deemed as being sufficiently representative of observations that can be expected from within the specific geographic region of interest. Therefore, if not applying the results obtained in this study within a different geographic context, the results observed in this study can be reasonably regarded to be relevant and reliable. The same argument can be extended to the climate factors that can affect water consumption within the urban setting.

Secondly, although the study incorporated a comprehensive set of independent variables, there may be other potential covariates that could influence urban consumption levels. Socioeconomic factors, population mobility, and land use patterns are examples of variables that were not included but may play a significant role in water consumption dynamics. The exclusion of these variables may have resulted in an incomplete model. While our study identifies relationships between variables, it does not establish causality or the direction of influence. For instance, while we observe a correlation between previous day consumption and the current day consumption, we cannot definitively determine whether increased historical consumption patterns directly lead to increased current consumption levels. Addressing causality requires experimental designs or sophisticated causal inference methods.

Thirdly, the study employed a relatively small number of residential households to monitor urban water consumption. The residential properties that were observed

were only 320 households, which is a very small proportion of the urban residential population in Nairobi. The consumption patterns that were observed in this sample, and their causal factors, may therefore not be representative. As smart water meters become more available in Kenyan households, a more comprehensive analysis is needed at the citywide, or even countrywide level to predict water demand.

Finally, this research has not provided solutions or answers regarding the behavioral question of why we consume water in a particular manner. Instead, the study has concentrated on answering the question about the manner in which water is consumed within the urban residential context. Socioeconomic and sociodemographic factors can avail important additional information that can improve our knowledge and understanding in order to increase insight. However, it is not particularly critical when determining how and when water consumption happens, or the application areas that can yield such information.

5.5 Recommendations for Future Research

In view of the limitations that have been identified from the current research, this study makes several recommendations. First, future research can consider analysis of water consumption in other cities such as Kisumu, Nakuru and Mombasa, as well as the consumption patterns in rural areas. This study only used data depicting the consumption patterns in one city, Nairobi. Additional insight from the expansion of this domain of observation on to other geographic urban areas will increase the generalizability if inference in using the LSTM for predicting consumption patterns in Kenya. It is also worthwhile to comparatively assess the consumption levels in different cities within the same study. The incorporation of region-specific data and factors will

ensure accurate and tailored predictions for each location which are due to the fact that water consumption dynamics vary across regions due to factors like local climate, and socioeconomic aspects of individuals in different localities.

Secondly, this research used a set of nine independent variables for its research. These included 1) the water consumption for previous 1 day, t-1, 2) the water consumption for previous 2 days, t-2, 3) the water consumption for previous 3 days, t-3, 4) the water consumption for previous 4 days, t-4, 5) the water consumption for previous 5 days, t-5, 6) the water consumption for previous 6 days, t-6, 7) the water consumption for previous 7 days, t-7, 8) temperature, 8) water pressure, and 9) water consumption on weekend versus weekday. Future research can include more variables that are potentially influential, such as socioeconomic factors, population mobility, and land use patterns into the analysis of domestic water consumption in urban areas. Although the current study incorporated a comprehensive set of independent variables, these additional covariates will yield a more stable model that can more accurately explain the variation in water consumption and can make more accurate predictions. Inclusion of these variables will assure the researchers of a more complete model that can establish both causality and the direction of influence.

Thirdly, future studies can apply large observation data sets to monitor urban water consumption in residential households to increase the representativeness of the observations. Availability of smart water meters within an Internet of Things (IoT) framework offers many possibilities for the assessment of large datasets to improve real-time analysis and the generation of accurate predictions for Kenyan households. Finally, this study compared two models and discovered substantial differences. The main model that the study proposes is the Long short-term memory (LSTM) network model for predicting non-linear water consumption patterns, which was compared

against the more conventional autoregressive integrated moving average (ARIMA) that is more suitable for linear time series data. The LSTM was found to be more accurate. More models can be compared, including the bidirectional LSTM model which has been found in other studies to yield relatively good predictive performance.

REFERENCES

Amankwaa, G., Heeks, R., & Browne, A. L. (2023). *Smartening up: User experience with smart water metering infrastructure in an African city*. Utilities Policy, 80, 101478.

Ashton, K. (1999). *That 'internet of things' thing. 2009*. URL http://www.rfidjournal.com/articles/view?4986.

Baek, S.S.; Pyo, J.; Chun, J.A. *Prediction of water level and water quality using a CNN-LSTM combined deep learning approach*. Water 2020, 12, 3399.

DreamCivil (29 May 2022) 11 *Factors Affecting Water Demand*. Available at https://dreamcivil.com/factors-affecting-water-demand/

Di Nardo, A., Boccelli, D. L., Herrera, M., Creaco, E., Cominola, A., Sitzenfrei, R., & Taormina, R. (2021). *Smart urban water networks: Solutions, trends and challenges. Water*, 13(4), 501.

Funck, J. H. (2018) *Synchronous Data Acquisition with Wireless Sensor Networks*. Universitätsverlag der TU Berlin

Gagliardi, F., Alvisi, S., Franchini, M., &Guidorzi, M. (2017*). A comparison between pattern-based and neural network short-term water demand forecasting models. Water Science and Technology*: Water Supply, 17(5), 1426-1435.

Gao, X.; Zeng, W.; Shen, Y.; Guo, Z.; Yang, J.; Cheng, X.; Hua, Q.; Yu, K. *Integrated Deep Neural Networks-Based Complex System for UrbanWater Management. Complexity* 2020, 2020, 8848324.

Garg, S., Mitra, S., Yu, T., Gadhia, Y., &Kashettiwar, A. (2023, June). *Reinforced Approximate Exploratory Data Analysis. In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 6, pp. 7660-7669).

Gaya, M. S., Zango, M. U., Yusuf, L. A., Mustapha, M., Muhammad, B., Sani, A., ... & Khairi, M. T. M. (2017). *Estimation of turbidity in water treatment plant using Hammerstein-Wiener and neural network technique*. Indonesian Journal of Electrical Engineering and Computer Science, 5(3), 666-672.

Guma, P. K., & Wiig, A. (2022*). Smartness Beyond the Network: Water ATMs and Disruptions from below in Mathare Valley, Nairobi. Journal of Urban Technology*, 29(4), 41-61.

Hancock, D. R., Algozzine, B., & Lim, J. H. (2021). *Doing case study research: A practical guide for beginning researchers*.

Hasan, M. B., Driessen, P. P., Majumder, S., Zoomers, A., & van Laerhoven, F. (2019). *Factors affecting consumption of water from a newly introduced safe drinking water system: the case of managed aquifer recharge (MAR) systems in Bangladesh*. Water, 11(12), 2459.

Han, S., Zhou, J., Liu, Z., Zhang, L., & Huang, X. (2021). *Influence of Community Factors on Water Saving in a Mega City after Implementing the Progressive Price Schemes*. Water, 13(8), 1097.

Hargreaves, A. J., Farmani, R., Ward, S., & Butler, D. (2019). *Modelling the future impacts of urban spatial planning on the viability of alternative water supply*. Water Research, 162, 200-213.

He, C., Liu, Z., Wu, J., Pan, X., Fang, Z., Li, J., & Bryan, B. A. (2021). *Future global urban water scarcity and potential solutions. Nature Communications*, 12(1), 4667.

Humayun, M., Alsaqer, M. S., &Jhanjhi, N. (2022). *Energy optimization for smart cities using iot.* Applied Artificial Intelligence, 36(1), 2037255.
Kato, T. (2016). *Prediction of photovoltaic power generation output and network operation*. In Integration of Distributed Energy Resources in Power Systems (pp. 77-108). Academic Press.

Kaushik, S., Choudhury, A., Sheron, P. K., Dasgupta, N., Natarajan, S., Pickett, L. A., & Dutt, V. (2020*). AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures*. Frontiers in big data, 3, 4.

Kim, J., Lee, H., Lee, M., Han, H., Kim, D., & Kim, H. S. (2022). *Development of a Deep Learning-Based Prediction Model for Water Consumption at the Household Level*. Water, 14(9), 1512.

Kiran, K., Sharma, A. K., & Van Staden, R. (2022). *Development of an intelligent urban water network system*. Water, 14(9), 1320.

Krishnan, S. R., Nallakaruppan, M. K., Chengoden, R., Koppu, S., Iyapparaja, M., Sadhasivam, J., & Sethuraman, S. (2022). *Smart Water Resource Management Using Artificial Intelligence—A Review. Sustainability*, 14(20), 13384

Kühnert, C., Gonuguntla, N. M., Krieg, H., Nowak, D., & Thomas, J. A. (2021). *Application of LSTM networks for water demand prediction in optimal pump control*. Water, 13(5), 644.

Meyer, B. E., Jacobs, H. E., &Ilemobade, A. (2020*). Extracting household water use event characteristics from rudimentary data. Journal of Water Supply*: Research and Technology-Aqua, 69(4), 387-397.

Micheni, E., Machii, J., &Murumba, J. (2022, May). *Internet of Things, Big Data Analytics, and Deep Learning for Sustainable Precision Agriculture*. In 2022 IST-Africa Conference (IST-Africa) (pp. 1-12). IEEE.

Mulwa, F., Li, Z., &Fangninou, F. F. (2021). *Water scarcity in Kenya: current status, challenges and future solutions*. Open Access Library Journal, 8(1), 1-15.

Oiro, S., Comte, J. C., Soulsby, C., MacDonald, A., & Mwakamba, C. (2020). *Depletion of groundwater resources under rapid urbanisation in Africa: recent and future trends in the Nairobi Aquifer System, Kenya.* Hydrogeology Journal, 28, 2635-2656.

Osman, S. A. M., &Elragal, A. (2021). *Smart cities and big data analytics: a data-driven decision-making use case*. Smart Cities, 4(1), 286-313.

Owen, D. L. (2023). *Smart water management*. River.Pacchin, E., Gagliardi, F., Alvisi, S., & Franchini, M. (2019). *A comparison of short-term water demand forecasting models*. Water resources management, 33, 1481-1497.

Pelikh, K.(2022) Smart Cities*: The best smart cities in Africa*. Available at https://www.o-city.com/blog/the-best-smart-cities-in-africa

Phasinam, K., Kassanuk, T., Shinde, P. P., Thakar, C. M., Sharma, D. K., Mohiddin, M. K., & Rahmani, A. W. (2022). *Application of IoT and cloud computing in automation of agriculture irrigation*. Journal of Food Quality, 2022, 1-8.

Piasecki, A., Jurasz, J., & Kaźmierczak, B. (2018*). Forecasting daily water consumption*: a case study in Torun, Poland. PeriodicaPolytechnica Civil Engineering, 62(3), 818-824.

Salehi, M. (2022). *Global water shortage and potable water safety; Today's concern and tomorrow's crisis*. Environment International, 158, 106936.

Saraiva, M.; Protas, É.; Salgado, M.; Souza, C. *Automatic mapping of center pivot irrigation systems from satellite images using deep learning*. Remote Sens. 2020, 12, 558.

Sayari, S.; Mahdavi-Meymand, A.; Zounemat-Kermani, M. *Irrigation water infiltration modeling using machine learning*. Comput. Electron. Agric. 2021, 180, 105921.

Shah, J. (2017*). An internet of things-based model for smart water distribution with quality monitoring*. Int. J. Innov. Res. Sci. Eng. Technol, 6(3), 3446-3451.

Singla, H. K., &Bendigiri, P. (2019). *Factors affecting rentals of residential apartments in Pune, India*: An empirical investigation. International Journal of Housing Markets and Analysis, 12(6), 1028-1054.

Söderberg, A., & Dahlström, P. (2017). *Turning smart water meter data into useful information*: a case study on rental apartments in södertälje. Lund University Press

Velasco, L. C. P., Granados, A. R. B., Ortega, J. M. A., &Pagtalunan, K. V. D. (2018). *Performance analysis of artificial neural networks training algorithms and transfer functions for medium-term water consumption forecasting*. International Journal of Advanced Computer Science and Applications, 9(4).

Yan, Z., & Gang, H. (2019, October). *Design of intelligent water metering system for agricultural water based on nb-io*t. In 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) (pp. 1665-1669). IEEE.

Zanfei, A., Brentan, B. M., Menapace, A., Righetti, M., & Herrera, M. (2022). *Graph convolutional recurrent neural networks for water demand forecasting*. Water Resources Research, 58(7)

APPENDICES

Appendix 1

Project Schedule

Activity	Apr	M	Ju	Ju	Aug	Septem	Octob
Develop project idea							
Developresearchquestions							
Write research proposal							
Conduct literature review							
Present proposal							
Collect research data							
Develop Study model							
Conduct data analysis							
Conduct model validation							
Compile research results							
Prepare project dissertation							
Final project defense							

APPENDIX 2

Project Budget

No.	Item	Unit	Price	Total
1	Computer/ hard disk/ operating	1	60,000	60,000
2	Internet connectivity/ data	5	3,000	15,000
3	Travel cost to-from KCA University	15	500	7,500
4	Field data expenses	1	2,000	2,000
5	Final dissertation preparation (i.e., printing, binding, etc.)	600	15	9,000
6	Flash drive	1	1,500	1,500
7	Miscellaneous costs	1		10,000
	Grand Total			**105,000**