**INTRODUCTION**

Access to water is a critical component of human lives and is now considered a basic human right. Access to clean water is also one of the 17 Sustainable Development Goals (SDG) set up by the United Nations in 2015 to achieve a better future for all [1].Specifically, the sixth goal, which is to ensure and sustain the availability of water and sanitation to all [2]. Potable water can also be linked to the third SDG goal \_ good health and well-being, as contaminated water can be a transmission medium for diseases such as cholera, typhoid, and diarrhoea, which are jointly the highest cause of mortality (especially children) in developing nations of Africa and Asia [3]. Water is also important in agriculture and food production. Recent statistics shows that about 10% of the world population is malnourished, with developing countries being hit the hardest, with starvation resulting in about 45% of infant mortality [5]. Ensuring global food security is thus of utmost importance. Food security has been recognized as a critical requirement, hence its inclusion as one of the SDG (goal 2), with specific focus on ending hunger, by promoting sustainable agriculture and improving food distribution. Food production and agriculture in general rely heavily on water, both for irrigation and for animal consumption. It is thus pertinent to ensure the availability and sustainable management of water \_t for agricultural use.

There are several sources of water for both drinking and irrigation use, including rivers, streams, rain, and groundwater (accessed through wells and boreholes). The nature and characteristics of a source of water are often critical factors that influence the constituents of water samples obtained therein. Beyond natural factors, chemical wastes from human activities such as mining, crude oil extraction, and industrial wastes, most often end up in streams, rivers, and other sources of water, changing the nature and properties of these waters. These waters then end up in homes or farms, where they are used for domestic purposes, drank, fed to livestock, or used to water crops. Consuming this type of water can have dire health consequences or result in death. It is therefore paramount that a proper process be put in place to ensure end to- end monitoring of the water right from the source to its last point of use. At each monitoring point, samples of water need to be collected to assess the quality or ``fitness for use'' for human (and animal) consumption, irrigation and domestic (or industrial) uses.

Several models have been developed to assess water quality, all of which consider various parameters, including chemical (such as hydrogen potential (pH), calcium, oxygen, sulphate levels etc.), microbial (such as E. coli, rotaviruses, Entamoeba etc.), and physical (temperature and clarity). These models produce a unit metric, known as the Water Quality Index (WQI), as output. Globally, different guidelines have been adapted for calculating WQI. For instance, in parts of Europe, the British Columbia Water Quality Index (BCWQI) and the Scottish Research Development Department (SRDD) are used, while in North America, the Canadian Council of Ministers of the Environment Water Quality Index (CCMEWQI) and National Sanitation Foundation Water Quality Index (NSFWQI) are predominant. In Asia, specifically India, the Bureau of Indian Standards (BIS) is prominent, while in Africa, notable standards include the South African National Standard for drinking water (SANS 241-1) and the Kenya Bureau of Standards (KEBS). A number of these models have been reviewed in [6]. It is important to note that many of these national standards are mostly local adaptations of the standards defined by the World Health Organization (WHO) [7]. This work is based on the South African and WHO standards.

Indeed, measuring water parameters for diverse water samples can be a laborious and daunting task, as it often involves adhering to a stringent set of rules in collecting the water samples, maintaining set conditions during transportation to the test laboratories, following standard methodologies in analysing the samples, and generally ensuring quality control. Some of these processes (and corresponding guidelines) are given in [8], [9]. The output of these processes indicates if the water sample is potable or non-potable. In this work, we propose a Cyber-physical network architecture for real-time monitoring of water parameters across a city and an alternative model based on machine learning to determine potability of water samples. Like [10]\_[13] [14], our work also only focuses on the physical and chemical parameters of water, while ignoring the biological. This is because our model is meant to be sensor based (in the context of the Internet of Things), and to our knowledge, there are no physical sensors for measuring biological parameters, such as the presence of E. coli in water. We do not trivialize the importance of microbial water parameters, and our proposed model can indeed be adapted to consider these parameters by simply incorporating suitable physical sensors (if available) or virtual / soft sensors, such as the one proposed in [15] into our model.

Figure 1 gives a high-level depiction of our proposed architecture which is built upon 4 layers. The constituent components of this architecture are described as follows

:1) **Sensing Layer:** As depicted in the figure, the sensing layer interacts directly with the water samples in a river, stream, dam etc. to measure water parameters. It is built into a vertical pole tagged ``sensor probe'' and consists of numerous sensors bundled together. These sensors might include pH, conductivity, turbidity, temperature, residual chlorine etc., similar to those offered by Libelium [16]. All telemetry data measured by these sensors are sent to the Fog Nodes (FNs), wired or wirelessly, via the sending unit. In scenarios where installing sensors in water source(s) is extremely difficult or when the required sensors are not readily available, water parameter readings can be collected from the associated water treatment plants.

2) **Edge Layer:** This layer consists of low-end processing devices (edge modules), such as single board computers (e.g., Raspberry Pi or Nvidia Jetson), or microcontrollers (e.g. Arduino, ESP32). These devices act as i.) data pre-processing units, responsible for the collection, aggregation, filtration, and shaping of data received from the sensing layer; ii) network gateway to ``ferry'' telemetry data to the FNs, through 3G/4G/5G cellular or other low powered long-range network solutions.

3) **Fog / Cloud Layer:**

\_ Fog Nodes (FNs): these are small sized distributed cloud computing nodes that bring computing and storage closer to the data source, thus reducing latency resulting from transmission delay to/from the remote Cloud [17]. The FN is responsible for classification of water samples using machine learning models such as the ones proposed in this work. Due to the limited computing power at the Fog (compared to the Cloud), only the most influential parameters need to be considered when classifying water samples. This can be beneficial as less sensors would be required (since not all parameters are being measured) and by extension

lower computing resources would be needed for the classification process. Furthermore, resource management, scheduling etc. can also be carried out on FNs. When long term storage and/or advanced computations are required, which are beyond the Fog's capacity, data are forwarded to the Cloud data centre.

\_ Cloud Data Centre: The Cloud is a remote high performance computing infrastructure, which provides computing on demand [18]. In our system, the Cloud serves as a data warehouse as well as a platform for performing advanced data analytics, dash boarding, and hosting for relevant services and software.

4) **Application Layer:** serves as an interface between users (water management authorities, end users / customers, other stakeholders) and software / services running in the Cloud. Relevant software for water parameter monitoring are hosted at this layer and made available to users through mobile and web platforms.

The water monitoring network proposed in this work is to be deployed in the City of Cape Town in Western Cape, South Africa, with the intention of monitoring water parameters in water storage dams and/or water treatment plants across the city. Data gathered by the monitoring network are then passed through Machine Learning (ML) models to determine their suitability for consumption or irrigation purposes. The specific contributions of this work can be summarized as Follows V

1) Build a network for real-time collection and monitoring of water quality across water storage dams in the city of Cape Town. This network takes into consideration the unique geographical features of Cape Town, such as mountains and elevations that might obstruct radio frequency propagation.

2) Curate ample sized datasets on drinking and irrigation water that can be used to train (and test) machine learning models to automatically determine the ``fitness for use'' of a sample of water for drinking and/or irrigation purposes.

3) Build models that determine the most critical parameters that influence the accuracy of machine learning models in analysing water for drinking or irrigation.

Regarding the order of this paper, following this introductory section, is a review of related works in section II. Section III discusses our methodology for building the city-wide water monitoring network, while section IV presents the datasets duration process and machine learning models considered for determining quality of water samples. Implementation processes and obtained results from our experiments are discussed in sections V and VI respectively. Section VII discusses the economic viability of our proposed solution, while section VIII concludes the paper and gives insights into future directions.