Chapter 1

Introduction

Current research in Internet of Things (IoT) and sensor technology has significantly impacted multiple aspects of human life, particularly enhancing monitoring and control applications. The current state of Micro Electro Mechanical System (MEMS) and Integrated Circuit (IC) technology has enabled the integration of highly portable sensors, wireless communication, and signal processing into a unified system. This innovation has given rise to a transformative revolution in water pollution assessment and monitoring. The ready availability of low-power, portable, and cost-effective sensors for measuring parameters like pH and turbidity has simplified the task of assessing the quality of river water and identifying pollution levels. This has led to the development of Smart Sensing for Water Pollution (SSWP), an emerging paradigm that enables both quantitative and qualitative analysis of water pollution. SSWP empowers relevant authorities to proactively take preventive measures in response to the data generated [3].

The Internet of Things (IoT) is effectively utilized in the assessment and monitoring of water pollution. This involves a network of physical IoT devices equipped with sensors capable of exchanging data via the Internet. These IoT devices can measure key water parameters such as pH, Dissolved Oxygen (DO), turbidity, Nitrates (NO3), Biochemical Oxygen Demand (BOD), Fecal Coliforms (FC), and more, enabling them

to estimate water quality. Subsequently, they efficiently transmit this water quality data to a central base station through a communication network. However, IoT devices do have limitations in terms of storage, processing capacity, and power resources. In our research, we harnessed IoT devices to accurately identify water pollution levels within the constraints of their capabilities [4].

Water stands as the paramount natural environmental resource, playing a vital role in the sustenance of human beings and aquatic life. While we can endure without food for several months, our existence cannot persist for more than three days without water. Our planet boasts an array of water bodies, including lakes, streams, rivers, reservoirs, ponds, and more, all serving as repositories of this precious resource. The quality of water is determined by its fitness for specific purposes, such as agriculture, industrial applications, and drinking. Across the globe, a significant population relies on river ecosystems, which encompass gently flowing water, supporting plant life, and facilitating interactions among animals, plants, and microorganisms, including bacteria. Unfortunately, the degradation of river water quality is primarily driven by human activities, encompassing industrial, agricultural, and other anthropogenic factors. The result of such contamination is termed water pollution, significantly impacting water quality. In the market, a wide range of cost-effective sensors is available for monitoring water pollution. These sensors encompass various types, including the Potential of Hydrogen (pH) sensor, Electrical Conductivity (EC) sensor, Ultrasonic sensor (Water Level sensor), Temperature sensor, and Turbidity sensor, among others. The pH sensor is used to determine the water's acidity or alkalinity. Meanwhile, the Electrical Conductivity (EC) sensor gauges the water's ability to conduct an electrical current. The Ultrasonic sensor, also known as the Water Level sensor, is employed to measure water levels, while the Temperature sensor provides information on the water's temperature. Lastly, the Turbidity sensor assesses the clarity of the water. Fig. 1.1 illustartes different sensors utilized for deciding the condition of water.

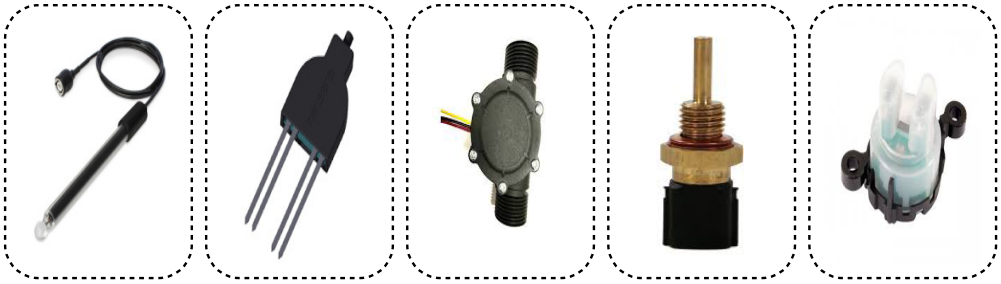

Figure 1.1: Sensors used for deciding water quality.

The Water Quality Index (WQI) is a concept employed to evaluate water pollution with regard to its impact on ecosystem health, human safety, and the suitability of the water for drinking. WQI incorporates quantitative data, which consists of numerical values, and translates this data into terms that are readily comprehensible to the general public, as shown in Table 1.1.

Table 1.1: Example: Healthy Drinking Water

WQI	$81-100$	$61-80$	$41-60$	$21-40$	$0-20$
Water Label	Excellent	Very Good	Fair	Bad	Very Bad

The WQI is designed to measure the level of pollution in bodies of water, including lakes, ponds, and rivers. This single, dimensionless metric has the unique capability of integrating various measurements with different units into a single numerical value. For example, parameters like Dissolved Oxygen (DO) are measured in Parts per Million (ppm), while turbidity is measured in Nephelometric Turbidity Units (NTU). The WQI assesses water pollution levels by combining measurements of specific parameters, serving as a vital and widely used tool for evaluating the quality of water in different bodies of water. WQIs play a crucial role in determining the suitability of water for various purposes such as irrigation, industrial use, bathing, cleaning, and competitive services. While numerous WQI methods exist, there remains a lack of a universally accepted, standardized approach that maintains its scientific validity. The WQI serves

as a criterion for assessing the condition of water concerning its diverse applications, which encompass industrial processes, recreational activities, drinking, agriculture, and the operation of hydroelectric power plants, as illustrated in Fig. 3.10.

Figure 1.2: Applications of water

Water Quality Indices (WQIs) play a crucial role in assessing the pollution levels in river water. They rely on the analysis of various water parameters, such as pH, Dissolved Oxygen, nitrates, Biochemical Oxygen Demand, Fecal Coliform, and more. This analysis employs two distinct methods: the laboratory-based approach and the sensor-based approach. In the laboratory-based method, water samples collected from the river undergo analysis to determine the river's pollution level. This process involves computing the WQI, which utilizes sub-indices and assigns weights to each parameter. However, this laboratory-based analysis typically yields results after a specific evaluation period, which can extend to several weeks. This delay poses a significant obstacle to real-time river monitoring.In our research, we have developed a mathematical model that takes multiple parameters as input and calculates the WQI. This WQI estimation process involves four key steps: parameter selection, weight assignment, sub-index computation, and aggregation of weights and sub-indices to generate the final WQI. Subsequently, we leverage the estimated WQI to automatically label the laboratory dataset.

No direct automated method is available for annotating sensory data, and manually annotating extensive sensory datasets is not feasible due to the significant time and effort required. To tackle this challenge, we leverage the labels from the labora-

tory dataset to automatically annotate the extensive sensory data instances, utilizing solely the GPS coordinates of the sensory data. However, this automated annotation process introduces the possibility of noisy labels within the dataset. To address this issue, we integrated a noise-handling loss function aimed at mitigating the potential impact of noisy labels. With the labeled sensory dataset in place, we proceeded to construct a classifier using a deep neural network. This classifier's purpose is to learn the relationship between input water parameters and the corresponding pollution label.

The demanding deep neural network requires substantial computational resources for WQI estimation. However, we are constrained by limited resources on our IoT devices. Hence, we adopt a knowledge distillation approach to train a compressed model. Initially, we obtain this compressed deep neural network through filter pruning. Subsequently, we conduct training for this compressed DNN. Following training, we implement the trained compressed DNN on our IoT devices for WQI estimation. This deployment on IoT devices allows us to attain the desired level of accuracy in detecting water pollution.

Furthermore, there is a necessity to transmit the estimated WQI (pollution data) from a river to the base station while minimizing energy consumption. To address this challenge, we employ a game theory-based approach. This approach calculates the optimal time duration for utilizing the appropriate spreading factor of the long-range network, thus reducing energy consumption during the data transfer from the river to the base station. The IoT devices in this scenario are the LoRa nodes, which have limited resources. The game theory-based approach not only accomplishes the transmission of WQI data with minimal energy usage but also ensures the successful and reliable data transmission.

1.1 Motivation of the Research Work

In this section, we delve into the motivation behind our research. The central objective in the realm of river water pollution assessment and monitoring is to ascertain the pollution level to determine its suitability for various applications such as bathing, drinking, irrigation, and more. This entails the utilization of machine learning and deep learning techniques to gauge water pollution levels. Machine learning techniques necessitate the use of statistical features like mean, median, and standard deviation, among others, to extract the relevant features essential for pollution level determination. In contrast, deep learning techniques exhibit the capability to automatically extract these features without relying on statistical measures. However, a key limitation when applying deep learning techniques is the need for an extensive set of accurately labeled data instances for training. Manually assigning labels to the vast amount of sensory data is simply impractical, rendering the use of deep learning techniques unfeasible without the employment of automatic annotation methods.

Regrettably, existing work often overlooks the critical aspect of automatic annotation. In the absence of an automated annotation mechanism, annotating a large dataset to enhance the learning technique's performance becomes a formidable task. However, adopting automatic annotation techniques introduces a significant number of noisy labels to the data. Thus, the necessity arises for noise handling mechanisms to mitigate the adverse effects of these noisy labels. This step is crucial for assigning accurate labels to the data before training, ultimately enhancing the performance of the constructed classifier. Notably, prior research has not incorporated these noise handling methodologies.

The river dataset encompasses a wide array of water parameters, including pH, electrical conductivity, dissolved oxygen, turbidity, and more, to serve as the basis for labeling the data instances. However, it's practically challenging to collect sensory values from all sensors simultaneously for each data instance. Artificially generating

data values for unrecorded instances during data collection is inadequate for accurately predicting water quality. Hence, the annotation process must rely on a select subset of parameters for labeling the data instances. This underscores the pressing need for a model that can effectively leverage a minimal number of parameters to assess water pollution.

Deep neural networks and machine learning techniques have found application in the realm of IoT to classify the quality label of given water samples. ML techniques typically necessitate domain expertise and the crafting of specific features. Consequently, we introduced a deep neural network, capable of not only automatically extracting key water features but also annotating data instances. However, DNN models tend to be parameter-intensive for water sample classification, making deployment on IoT devices with limited computing capacity challenging. Thus, there's a demand for a compressed DNN model, which, while compromising some accuracy, remains practical. To achieve an acceptable level of accuracy, we incorporated knowledge distillation, a concept that enhances the performance of the compressed DNN by leveraging the generalization capabilities of the standard DNN. With the application of knowledge distillation, we have successfully improved the performance of the compressed DNN. Subsequently, this compressed DNN is deployed on IoT devices to assess the level of water pollution. These IoT devices transmit water pollution data, gathered through sensors, to the nearest processing unit using a communication protocol.

The choice of a communication protocol significantly influences the energy consumption of IoT devices. Long Range Wide Area Network (LoRaWAN) stands out as an ideal option for IoT devices, offering extended-range communication with minimal energy consumption. The LoRaWAN architecture comprises LoRa Nodes (LNs), LoRa Gateways (LG), and a Network Server (NS). LoRa technology supports various Spreading Factors (SFs), ranging from 7 to 12, allowing for flexible long-range communication with low power consumption. These SFs exhibit distinct communication ranges and

varying power consumption during data transmission. Notably, none of the existing work has simultaneously prioritized energy efficiency, accuracy, and cost-effectiveness as its primary objective.

1.2 Contributions and Organization of the Thesis

In this thesis, we delve into the multifaceted challenges associated with the detection of river water pollution levels. We commence by addressing the task of ascertaining water pollution levels through the estimation of Water Quality Indices (WQI) using laboratory data and the automatic annotation of said data. To this end, we develop a mathematical model for WQI estimation and employ it to annotate the laboratory data, subsequently facilitating the annotation of extensive sensory data. The direct annotation of sensory data poses unique challenges, thus necessitating an automatic annotation process that, unfortunately, introduces noisy labels into the dataset. Subsequently, we proceed to construct a deep neural network (LSTM model) designed for the identification of river water pollution. However, we underscore the potential performance degradation of this classifier within the realm of deep learning due to the presence of noisy labels in the training dataset. To mitigate this challenge, we put forth a mechanism for handling noisy labels, particularly in cases where prior information regarding the noise concentration is unavailable. Furthermore, we address the challenge of accurately detecting water pollution levels, specifically the correct estimation of Water Quality Indices (WQI), while working with IoT devices that have constrained resources. To tackle this, we implement a knowledge distillation approach to train a condensed deep neural network (DNN). This streamlined DNN is then deployed on IoT devices, successfully achieving the desired level of accuracy in water pollution detection. In addition, we tackle the issue of transmitting pollution data from the river to the base station while minimizing energy consumption. To optimize this process, we introduce a game theory-based technique for estimating the optimal time duration for employ-

ing the appropriate spreading factor during data transmission. Specifically, this thesis investigates the following research problems:

- How to estimate the river water quality level using unlabelled limited lab data with high accuracy?
- How to do automatic annotation of the river water sensory data using lab data labels with high accuracy?
- How to construct a deep learning based classifier to predict a label for noisy sensory data and test the new instance of a sensory data?
- How to efficiently estimate WQI using limited processing of IoT devices with acceptable accuracy?
- How to transmit the pollution data from a river to the base station or remote host by consuming minimum energy?

The rest of the thesis is organized as follows:

Chapter 2: This chapter presents the preliminaries about the various techniques used in this thesis, including Internet of Things, knowledge distillation, long range communication technology and game theory. The chapter also presents state-of-the-art work covering the prior studies on river water pollution assessment, river water pollution monitoring and noisy labels in the dataset. Next, we have tabulated the comparative summary of the existing literature on the water quality assessment. Further, we have tabulated the comprehensive summary of the prior research work on the water quality monitoring. We concentrated on the different facets such as, statistical methods or deep learning methods for feature extraction, whether prototype designed, consideration of cost, energy-efficiency, and accuracy as primary goal, and communication techniques *etc.*. We review the existing work on assessing and monitoring of the condition of the river water. The research work on river water pollution has been studied and summarized.

Chapter 3: This chapter highlights the identified limitations of existing systems and delineates the principal contributions made by this work. Additionally, the chapter provides a comprehensive account of the methodology employed for the collection of the River Water dataset. This methodology encompasses the identification of specific sites of interest, the arrangement of a suitable vessel, the formulation of a spatio-temporal route map for data collection through sensors, and the subsequent uploading of collected data to the Cloud. The chapter also delves into a discussion on various wireless communication protocols utilized in IoT applications. It expounds upon the acquisition of two distinct datasets: the laboratory dataset collected offsite and the extensive real-time sensory dataset gathered online. Furthermore, it introduces the Hanna multi-parameter sensor, a critical component affixed to the boat for sensory data collection, and details the data pre-processing techniques integrated into our research. *We address the problem of identifying river water pollution level using lab data and annotate the limited lab data.* To solve this problem, this work proposes a mathematical model by incorporating the various selected six parameters DO, pH, BOD, FC, NO3, turbidity of unlabeled limited lab data to estimate WQI (or to identify river water pollution level). The estimation of Water Quality Indices (WQI) comprises a four-step process: parameter selection, weight assignment to chosen parameters, sub-index computation, and the aggregation of weights and sub-indices to estimate the final WQI. Subsequently, the chapter elucidates the automatic labeling of the lab data facilitated by the estimated WQI. Concluding this chapter is an extensive account of the experimental performance evaluation of the proposed approach, shedding light on its effectiveness and applicability.

Chapter 4: This chapter serves as the foundation for our work, commencing with the definition of terms and the establishment of their respective mathematical notations relevant to our research. It further delineates the existing limitations within current

systems and outlines the primary contributions made by this work. In this chapter, we consider the noisy labels in the sensory dataset of river water. Our target is to build a deep learning model that is robust towards the noisy labels in the dataset. *We address the problem of automatic annotation of huge sensory data and recognizing a water pollution level using deep learning model in the presence of noisy labels.* To solve the problem, this work proposes an automatic label transfer mechanism (ALT) which annotates the sensory data using limited labeled lab data and only GPS coordinates of sensory data. The automatic annotation process introduces noisy labels into the dataset, which prompts the proposal of a noise handling loss function. Furthermore, we propose a sensor-based deep learning approach, employing an LSTM model, for the assessment of water pollution using automated annotation. This deep learning approach consists of a single layer of LSTM unit with 25 cells, which prove to be sufficiently adept at autonomously extracting key water features. The performance evaluation of our system encompasses testing on five major Indian rivers, including Godavari, Ganges, Yamuna, Hindon, and Brahmaputra. Four key performance metrics, namely precision, recall, accuracy, and F1 score, are incorporated into the experiments for model evaluation. Precision measures the model's ability to predict the correct labels, while recall indicates the proportion of actual correct labels correctly identified. Concluding this chapter is a comprehensive presentation of the experimental performance evaluation of our proposed approach, demonstrating an impressive accuracy exceeding 90%, even in the presence of 20% noisy labels.

Chapter 5: This chapter lays the foundation for our research work, commencing with the comprehensive definition of key terms and the introduction of their corresponding mathematical notations. It goes on to elucidate the constraints and limitations inherent in existing systems while spotlighting the major contributions made by our work. Central to this chapter is the endeavor to address the challenge of *How to identify river water pollution level (estimate WQI) using resource-constrained IoT devices.* To tackle

this, we introduce a knowledge distillation approach for estimating river water pollution levels. This system leverages deep neural networks in conjunction with long-range communication technology to identify water pollution. The process begins with the development of a complex deep neural network, meticulously trained to recognize river water pollution. Subsequently, filter-level pruning is employed to create a compressed deep neural network. Furthermore, we utilize a knowledge distillation technique to train this compact deep neural network. During this training, two key losses come into play: the distillation loss and the combined loss of the student model. We present a mathematical formulation to minimize the student loss, thus achieving an acceptable level of recognition performance. Concluding this chapter is an in-depth exploration of the experimental performance evaluation of our proposed approach, demonstrating its effectiveness and viability in monitoring pollution data with commendable accuracy.

Chapter 6: This chapter presents the limitations of the existing systems and major contributions. In this chapter, we address the problem of *How to transmit the pollution data from river to the base station by consuming minimum energy?* To solve this problem, we propose a game theory based approach to estimate the time duartion for using the suitable spreading factor (virtual channel) of long range network to transfer the pollution data from river to the base station with minimum energy consumption. The game-theory based method ensures the successful transmission of data with acceptable accuracy by reducing energy. The system is implemented by developing an application on the IoT devices and various experiments are performed for validating the accuracy and energy efficiency of the system. Finally, this chapter presents the experimental performance evaluation of the proposed approach. We perform a real-world study to evaluate the feasibility and performance of the proposed approach.

Chapter 7: The main findings of the thesis are summarized in this chapter. We also cover some future directions of research. Our main contributions put into a global perspective in this chapter.

Some important technical reports, research papers and textbooks are listed in References. The publications of the research work presented in this thesis are listed in the List of Publications.