Data Compression for Flood Detection System Using Huffman Coding

Airi Kokuryo[1], Kohei Inoda[2], Akihito Kohiga[3], and Takahiro Koita[4]

Department of Information Systems Design, Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu 610-0394, JAPAN
[1] cguh1034@mail4.doshisha.ac.jp

[2] cgug1019@mail4.doshisha.ac.jp

[3] akohiga@mail.doshisha.ac.jp

[4] tkoita@mail.doshisha.ac.jp

Abstract

This study addresses the challenge of transmitting large volumes of sensor data for flood detection using LoRaWAN, a low-power wide-area network with payload limitations. We propose Huffman-based data compression methods that dynamically adapt to seasonal variations and incorporate flood risk bias. By analyzing temperature and humidity data from different regions in Japan over four seasons, we organized both season-specific and all-season Huffman trees with and without flood-risk-based symbol weighting. Evaluation results show that the flood-risk-weighted all-season Huffman model consistently achieves the higher compression ratio, outperforming conventional and seasonal methods. This approach enables efficient, selective data transmission during high-risk periods, contributing to practical early flood warning systems in remote environments.

1. Introduction

The occurrence of extreme precipitation and the risks of flooding have intensified with global warming. As the atmosphere warms, its water-holding capacity increases, and regions with higher water availability are more likely to experience severe flooding[1]. In Japan, annual average rainfall is on the rise, and Tokyo recorded its highest-ever average temperature in 2023.

In response to this growing threat, we are developing an early flood detection system for use in forests and mountainous regions, utilizing LoRaWAN and IoT-based sensors. However, these areas are often inaccessible and lack power infrastructure, necessitating the use of Low-Power Wide-Area Networks (LPWAN). Additionally, forecasting flood events requires the efficient transmission of large volumes of sensor data.

We utilize LoRaWAN as a basement of the sensor system. LoRaWAN is a LPWAN protocol designed for enabling long-range communication between battery-powered IoT devices. It operates on unlicensed radio frequencies and is optimized for low power consumption, making it ideal for applications where devices must operate autonomously over long periods without frequent battery replacements. Despite its long communication range and low power usage, LoRaWAN has a payload limitation of a maximum of 11 bytes, which poses challenges for transmitting large or frequent sensor readings[2]. To address this challenge, we propose a Huffman coding-based data compression method specifically optimized for flood detection scenarios.

2. Proposed Method

This section presents our proposed compression method for transmitting sensor data over LoRaWAN in flood detection scenarios. Sensor data, including temperature and humidity, is transmitted only when the Flood Probability (FP) exceeds 90%. Details on the computation of FP are provided in the following.

Flood Probability (FP):
FP is calculated using weather data collected from 2023 to 2024 across eight different regions in Japan, covering all four seasons. Each data point includes temperature (°C), relative humidity (%), and precipitation (mm). To model flood risk, we apply logistic regression with precipitation samples that are equal to or greater than 10 mm. Our flood risk model is trained to learn the relationship between temperature and humidity under high-precipitation conditions. The resulting coefficients and intercept are then used to represent the probability of flooding given a specific temperature and humidity.

Variant of Huffman Coding:
We propose two kinds of Huffman coding methods tailored to seasonal variations and weighting strategies for data with a Flood Probability (FP) of 90% or higher when making the Huffman tree.

I. Seasonal Huffman Coding:
Seasonal Huffman coding involves dividing the dataset into four temporal segments corresponding to the four seasons: Spring(March-May), Summer(June-August), Fall(September-November), and Winter(December-February). For each segment, a separate Huffman tree is constructed based on the symbol frequencies observed during that specific season. This seasonal segmentation exploits the natural variation in environmental parameters—such as temperature, humidity, and rainfall—that occurs throughout the year. Since symbol distributions shift significantly with the seasons, using season-specific Huffman trees enables more accurate probability modeling, thereby enhancing compression efficiency.

II. All-Season Huffman Coding:
All-season Huffman coding, by contrast, constructs a single Huffman tree using an equal number of samples from each season. This approach attempts to generalize across seasonal fluctuations, producing a tree that reflects the average symbol distribution across the entire year.

For both methods, we evaluate two variants: 1. using weighted symbol frequencies that reflect flood risk relevance, 2. using raw (unweighted) frequencies. This yields four total configurations: seasonal-unweighted, seasonal-weighted, all-season-unweighted, and all-season-weighted. These are compared in terms of compression performance under high-risk flood conditions.

3. Evaluation Method

We conducted a comprehensive evaluation of various compression strategies using temperature and humidity sensor data collected during rainy periods across different seasons. While abundant data with significant rainfall was available for Spring and Summer of 2024, relevant Fall and Winter data was scarce. Therefore, our comparative analysis primarily focuses on the Spring and Summer datasets.

We tested on six different methods: Seasonal-Unweighted Huffman Coding, Seasonal-Weighted Huffman Coding, All-Season-Unweighted Huffman Coding, All-Season-Weighted Huffman Coding, No Compression (Raw Data in a binary form), and ZIP Compression (General-Purpose Algorithm).

Each method was evaluated on its compression size, across four meteorological seasons—Spring (March–May), Summer (June–August), Fall (September–November), and Winter (December–February).

4. Results and Discussion

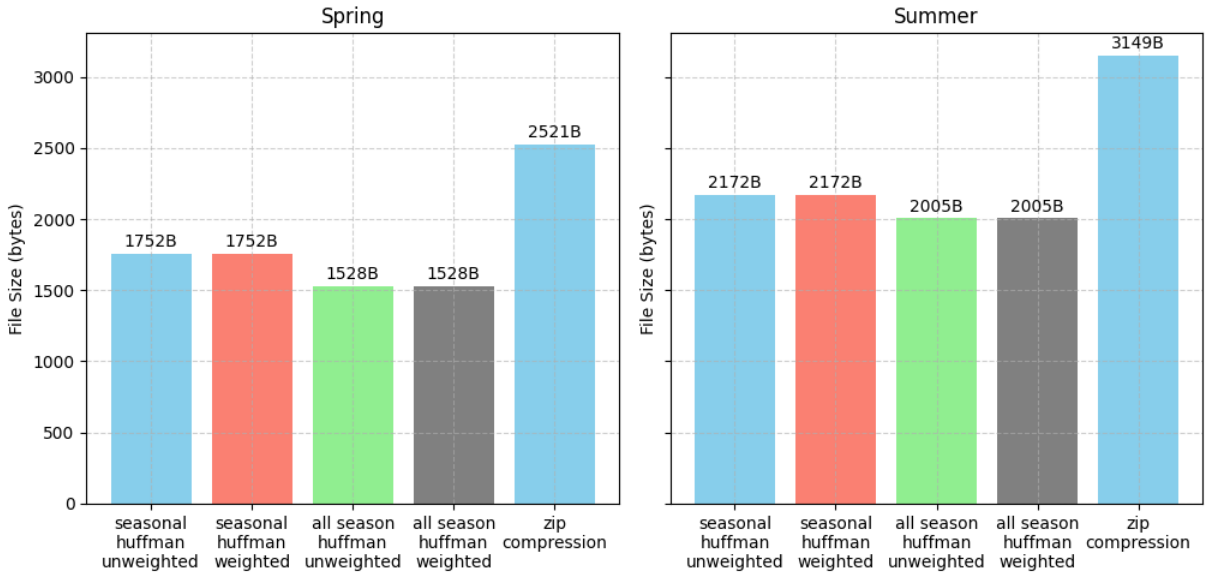

Figure 1: Results of Comparison: Spring and Summer

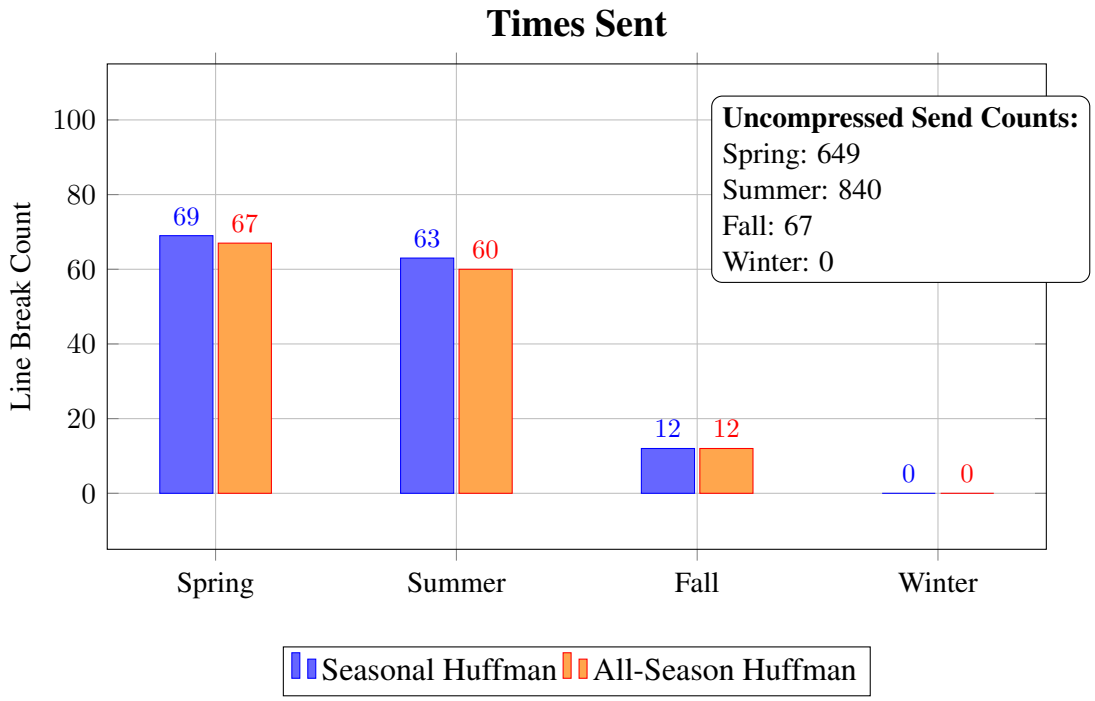

Figure 1 is the graph that compares the file size after compression. According to Figure 1, the flood-risk-weighted All-Season Huffman method consistently achieved the highest compression ratio among all tested methods. In particular, incorporating flood-risk-based symbol weighting—emphasizing data with a flood probability of 90% or higher—significantly enhanced the performance of the All-Season Huffman approach. In contrast, the same weighting had minimal effect on the Seasonal Huffman method.

The relative sizes of compressed data (from largest to smallest) were:

$$\text{No Compression} \gg \text{All-Season Huffman (Unweighted)} > \text{ZIP} > \text{Seasonal Huffman (Weighted} \approx \text{Unweighted)} > \text{All-Season Huffman (Weighted)}$$

Additionally, the bar graph shows that the number of transmissions in summer is lower than in spring, excluding the uncompressed method, which transmits data every time a new value is received. This indicates that the compression method effectively reduced the number of transmissions, particularly for periods with frequent data collection, making it more efficient in high-traffic scenarios.

These results suggest that weighting symbol frequency by flood risk and using a unified All-Season Huffman model achieves the most efficient compression. While seasonal segmentation seems intuitive, its practical benefit is limited, highlighting the greater impact of risk-aware symbol prioritization.

5. Conclusion

In this paper, we proposed and evaluated a flood-aware Huffman coding strategy tailored for LoRaWAN-based environmental monitoring systems. Our findings reveal that weighting symbol frequencies by flood probability is more impactful than season-specific modeling in achieving high compression efficiency. The flood-risk-weighted all-season Huffman method not only achieved the best compression ratios across multiple seasons but also required maintaining just a single tree structure—simplifying deployment and reducing overhead. These results suggest that risk-aware data prioritization is a key enabler for real-time, low-power flood detection in resource-constrained networks. Future work will explore energy consumption optimization.

References

[1] H. Tabari, "Climate change impact on flood and extreme precipitation increases with water availability," *Scientific Reports*, vol. 10, no. 1, 2020. doi: 10.1038/s41598-020-70816-2

[2] J. D. C. Silva, J. J. P. C. Rodrigues, A. M. Alberti, P. Solic, and A. L. L. Aquino, "LoRaWAN - A low power WAN protocol for Internet of Things: A review and opportunities," in *Proc. of the 2017 2nd Int. Multidisciplinary Conf. on Computer and Energy Science (SpliTech)*, 2017.

[3] D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proceedings of the IRE*, vol. 40, no. 9, 1952. doi: 10.1109/JRPROC.1952.273898

[4] Japan Meteorological Agency, "Historical Weather Data Download: Update History, How to Use This Page, FAQ, CSV File Format," *JMA — Past Weather Data Download*. `https://www.data.jma.go.jp/risk/obsdl/index.php`. [Accessed: May 6, 2025].