International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I11P101

Volume 71 Issue 11, 1-7, November 2023
© 2023 Seventh Sense Research Group®

Original Article

Mitigating Latency in IoT Devices: A Machine Learning Approach to Identifying and Addressing Key Contributing Factors

Karan Gupta[1], Manvendra Sharma[2]

[1]Senior Data Scientist, Sun Power Corporation, USA.
[2]Embedded Software Development Engineer, Amazon, USA.

[1]Corresponding Author : Karan.Gupta@sunpower.com

Received: 06 September 2023 Revised: 12 October 2023 Accepted: 01 November 2023 Published: 15 November 2023

Abstract - *In the realm of Internet of Things (IoT) systems, latency emerges as a pivotal challenge, jeopardizing both performance and usability, especially in time-sensitive applications. Recognizing the urgency of addressing this challenge, this paper embarks on a comprehensive analysis of synthetic IoT sensor data [1] to discern the predominant factors inducing high-latency events. By employing a lasso regression model [2], the research unveils network availability, communication failures, elevated memory utilization, and high CPU usage as the chief culprits behind latency issues. Augmenting our approach, a random forest classification [3] was employed, which impressively yielded a precision and recall rate between 93-95% in prognosticating high-latency events. Drawing on these insights, the paper advocates for strategies encompassing enhanced connectivity, protocol optimization, additional memory/CPU headroom provision, and a holistic approach to performance management as potent solutions to curtail IoT latency.*

Keywords - *Internet of Things (IoT), Latency, Classification models, Network availability, Memory utilization, Communication failures.*

1. Introduction

Latency is a critical issue in Internet of Things (IoT) deployments [4], defined as the delay between data generation at the sensor and processing at the IoT hub. High latency disrupts real-time monitoring and control applications dependent on timely data. Latency can be introduced through multiple sources. The key contributing factors for latency [5]:

1.1. Network-Induced Factors

Network availability and various delays like propagation, queuing transmission, and processing.

1.2. Hardware-Centric Determinants

CPU and memory utilization and data storage dynamics.

1.3. Software-Driven Influences

Operating system dynamics, middleware frameworks, and communication protocol intricacies.

1.4. Wireless Communication Nuances

Medium utilization, multipath propagation variables, handover delays, and node proximity factors.

Prior work has explored latency reduction through fog computing, quality of service optimization, and other approaches. However, it mostly focused on network-level factors. This paper analyzes key contributing factors related to hardware, software, network, and wireless communication impacting end-to-end latency using synthetic IoT sensor data [1]. Identifying root causes can guide strategies to mitigate latency through systemic improvements.

Latency in IoT devices can significantly deter the performance and reliability of interconnected systems. This research endeavors to pinpoint causative factors of latency and develop a predictive model to assist in latency identification and mitigation.

Our main objective is to

- To identify and analyze factors contributing to latency in IoT devices using lasso regression.
- To construct a predictive model that accurately identifies potential latency instances in IoT data transmission.

2. Literature Review

The latency phenomenon within the Internet of Things (IoT) landscape has garnered significant attention, yielding a multitude of studies that dissect its multi-faceted nature. The collective body of research highlights a spectrum of contributing factors, ranging from network congestion to protocol inefficiency, yet often falls short in synthesizing these factors within the complexity of live IoT systems. For instance, while [17] offered valuable insights into network-induced latency by simulating IoT network conditions, their research did not extend to the device level, leaving an incomplete picture of latency's origins. Such limitations call attention to the need for an integrated analysis that spans the full IoT stack.

Conversely, [18] approached latency through a layered perspective, examining the interaction between hardware, software, and network layers. Their use of an IoT prototyping platform underscored the critical role of memory and CPU limitations, network stability and protocol design.

However, the controlled nature of lab settings in their study raises questions about the applicability of these findings in unpredictable real-world environments.

The work of [19] further delves into the impact of Wi-Fi on IoT latency, examining how device density and external interference affect performance. Despite the practical implications of their findings, the focus remains narrowly on Wi-Fi, omitting other wireless technologies prevalent in IoT implementations.

These examples underscore a prevalent trend in latency research—a tendency to study isolated factors or specific technology layers without accounting for the operational complexity of real-world IoT systems.

This observation aligns with the critique that current literature inadequately addresses the confluence of variables in situ, a gap our study aims to fill by considering a holistic view of latency sources and their interrelations in live deployment scenarios.

Our review reveals that while previous studies provide a groundwork for understanding latency, there remains a pressing need to explore these dynamics in the context of comprehensive IoT environments. By acknowledging the intersectionality of various latency sources, this research contributes to a more integrated and actionable understanding of latency mitigation strategies.

To this end, our study will not only address the identified gaps but also leverage a unique synthetic dataset that embodies the complex interplay of factors affecting latency. Through this approach, we offer new insights into latency reduction, with the potential to significantly advance both academic research and practical applications in IoT system design and optimization.

3. Methodology

3.1. Data Synthesis and Preprocessing

Data utilized in this study was synthetically generated, emulating real-world IoT scenarios [6] based on insights gleaned from extensive literature and domain expertise. The dataset, comprising approximately 13.5k data points, encapsulates critical features of IoT devices and environments. Each data point is categorized as 'good' or 'bad,' representing the absence or presence of latency, respectively.

3.2. Data Description

The synthetic dataset was meticulously crafted to emulate real-world IoT scenarios, incorporating critical features pertinent to IoT devices and environments, such as:

3.2.1. DeviceID

A unique identifier assigned to each IoT device.

3.2.2. AvgCPUUtilization, MaxCPUUtilization, MinCPUUtilization

Metrics representing the average, maximum, and minimum CPU utilization, respectively, during the data transmission.

3.2.3. TotalWifiUsage

The total volume of data transmitted via Wi-Fi during a particular session. This parameter indicates the structural intricacies of the wireless communication protocols.

3.2.4. Medium_Utilization

Metrics representing the amount of time the communication medium is used.

3.2.5. EventType

The target variable indicates whether a data point is categorized as 'good' (no significant latency) or 'bad' (noticeable latency).

3.2.6. AvgMemoryUtilization, MaxMemoryUtilization

Metrics representing average and maximum memory utilization.

3.2.7. Wireless_Network_Unavailable, Communication_ Failure, Network_Unavailable, High_CPU_Usage, High_System_Load

Some binary features, possibly indicating the occurrence of specific events. If it is 1, the event has occurred, 0 otherwise.

3.3. Data Balancing

Given the inherent class imbalance in the dataset, an up-sampling technique was employed [7] to equalize the number of 'good' and 'bad' samples, thereby preventing model bias towards the majority class during subsequent analytical processes.

In up-sampling, since the bad samples are less than the good ones (Fig-1), it will create some synthetic data points similar to bad ones, such that the number of bad samples becomes equal to the good ones (Fig-2).

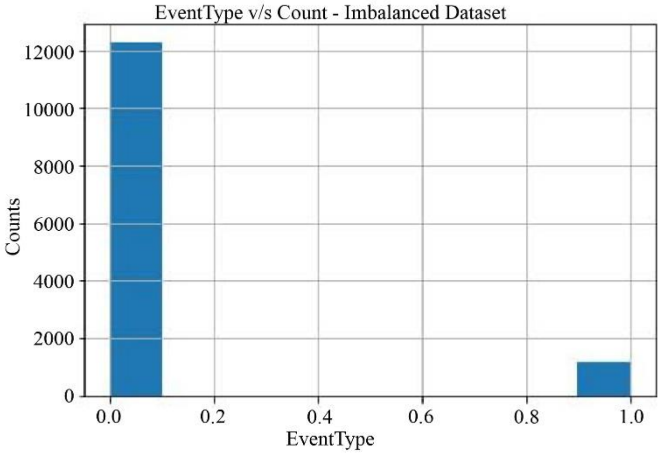

Fig. 1 Imbalanced dataset representation

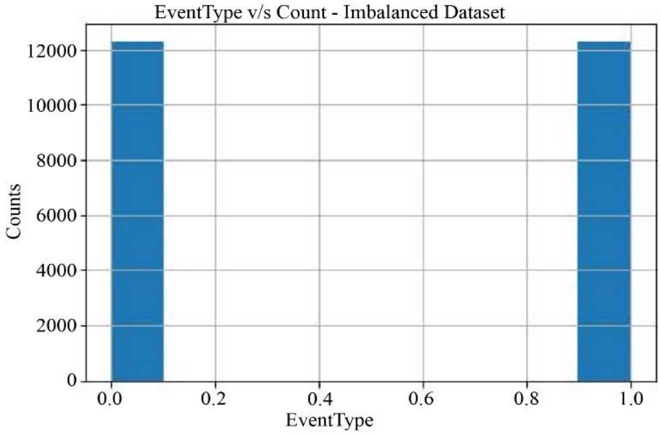

Fig. 2 Balanced dataset representation

3.4. Analytical Approach

A two-prolonged analytical approach was used to discern and validate the factors contributing to latency in IoT transmission.

3.4.1. Lasso Regression

Initially, Lasso Regression was utilized due to its inherent ability [8] to perform feature selection alongside regression. The regularization parameter in Lasso Regression effectively reduces the coefficients of non-contributory features to zero, thereby highlighting features that significantly influence latency.

Table 1. Features importance of using lasso regression

Feature Name	Coefficients
Network Unavailable	1.95
AvgMemoryUtilization	0.41
AvgLoad	0.33
TotalWifiUsage	0.25
AvgCPUUtilization	0.18
Medium_Utilization	0.10
WirelessNetworkUnavailable	-0.46

3.4.2. Random Forest Classifier

Subsequent to feature identification, a Random Forest Classifier was employed to validate the identified features and assess their relative importance.

Random Forest, an ensemble learning technique [9], was chosen for its adeptness in handling imbalanced data, robustness against overfitting, and capability to provide insightful feature importance metrics.

3.5. Model Evaluation and Validation

Model performance was rigorously evaluated using precision and recall as the primary metrics, ensuring robustness in identifying both latency and non-latency events despite the inherent class imbalance.

3.5.1. Model Validation

In addition to quantitative evaluation metrics, model validity was further assessed by comparing identified features and domain knowledge, ensuring the model's findings were logically and theoretically coherent.

3.5.2. Comparative Analysis

For comprehensive evaluation and validation, the Random Forest model's performance was juxtaposed against other machine learning models, including Logistic Regression, Decision Trees, and Boosting algorithms, ensuring the selection of the most optimal model for the given scenario. Below are their results:

Algorithm	Precision	Recall
Logistic Regression	0.86	0.83
Random Forest	0.948	0.93
Decision Trees	0.9	0.88
XGBoost	0.93	0.93

3.6. Ethical Considerations and Reliability

Given the synthetic nature of the data, ethical considerations [10] about data privacy and security were inherently addressed. Furthermore, to enhance the reliability of the findings, the synthetic data was formulated based on rigorous literature review and domain expertise, thereby closely emulating real-world IoT environments and scenarios.

4. Results

4.1. Descriptive Statistics

A meticulous exploration of the descriptive statistics was undertaken to comprehend the fundamental characteristics of the key features stratified by EventType (0 and 1, indicating the absence and presence of latency, respectively). The features focused upon were identified as significant contributors to latency through preliminary analysis and modeling.

4.1.1. Network Unavailable

EventType	Mean	Median
0	0.236	0
1	0.953	1

Insight: A stark contrast is observed in the mean and median values between the two classes, indicating that instances where the network is unavailable are strongly associated with latency. The mean and median for latency events are notably higher, suggesting that network availability is crucial for timely data transmission in IoT devices. Network unavailability will cause communication failure; hence, high latency will be observed. This can be avoided by having a redundant or secondary communication method. In case of failure of the primary communication mechanism, the device can roll over to the secondary communication method.

4.1.2. TotalWifiUsage

EventType	Mean	Median
0	25.48	3
1	86.03	27

Insight: A significant discrepancy is observed in the total Wi-Fi usage between latency and non-latency events. The substantially higher mean and median values for latency events imply that instances with higher Wi-Fi usage tend to be associated with increased latency. Higher Wi-Fi usage directly relates to the communication protocol for transmitting the data. Generally, the use of lightweight protocols like MQTT and CoAP [11] is preferred for IoT devices.

4.1.3. AvgMemoryUtilization

EventType	Mean	Median
0	702970	701883
1	717935	715084

Insight: While the mean and median values are relatively close between the two classes, a slightly higher average memory utilization is observed for events with latency. This suggests that while memory utilization is impactful, other factors may be more deterministic of latency events. IoT devices tend to have fewer resources [12]; hence, monitoring and managing memory usage is essential for optimal system performance.

4.1.4. AvgLoad

EventType	Mean	Median
0	0.566	0.464
1	0.805	0.617

Insight: A discernible difference in system load is observed between the two classes. Events with latency exhibit higher average and median system loads, indicating that managing and optimizing system load can be pivotal in mitigating latency.

4.1.5. AvgCPUUtilization

EventType	Mean	Median
0	28.69	27.29
1	29.50	28.96

Insight: The mean and median values are relatively similar between the two classes, indicating that while CPU utilization does impact latency, it may not be the predominant factor. Nonetheless, it is imperative to note that optimized CPU usage remains crucial for overall system performance. However, the software and hardware for IoT devices should be designed such that worst-case CPU usage should not cause network delays.

4.1.6. Medium_Utilization

EventType	Mean	Median
0	0.4985	0.4996
1	0.5094	0.5035

Insight: Overall, while ‘Medium_Utilization’ for both event_types is quite similar, ‘Event Type 1’ tends to have a higher medium utilization on average. This indicates that the higher medium utilization can cause higher latency. Higher medium utilization means the device will get less chance to transmit the data, which can increase the latency.

4.2. Visualizations

Visualizations provide a more intuitive understanding [13] of data distributions and relationships, particularly for discerning class differences. In this section, we will delve into individual boxplots for each key feature stratified by EventType.

4.2.1. Network Unavailable

First, let’s visualize the Network_Unavailable feature. Given its binary nature, we expect a clear distinction between the two EventType classes.

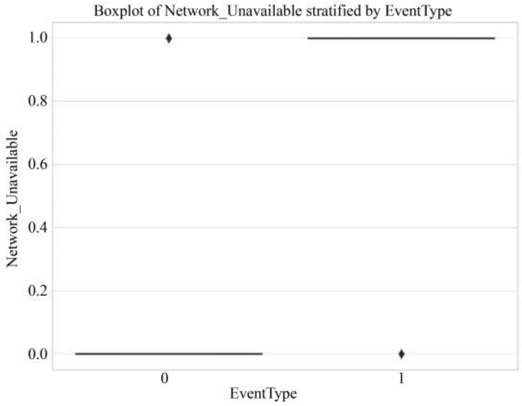

Fig. 3 Network Unavailable

From the boxplot, we observe a stark distinction between the two EventType categories:

- For events without latency (EventType=0), most data points have a Network_Unavailable value of 0, indicating that the network was predominantly available.

- Conversely, most data points for events with latency (EventType=1) reflect a Network_Unavailable value of 1, suggesting that the network was largely unavailable during these events.

This clear distinction underscores the pivotal role of network availability in influencing latency.

4.2.2. TotalWifiUsage

Next, we'll visualize the TotalWifiUsage feature to understand its distribution and potential influence on latency.

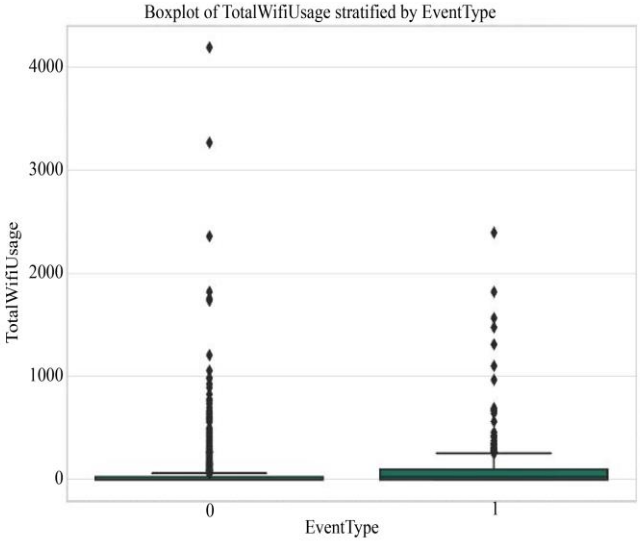

Fig. 4 TotalWifiUsage

The boxplot for TotalWifiUsage reveals significant differences in data distributions between the two EventType categories:

- Events without latency (EventType=0) predominantly exhibit lower TotalWifiUsage values, with most data points clustered closer to the lower quartile.

- In contrast, events with latency (EventType=1) display a broader range of TotalWifiUsage values, with a significantly higher median than non-latency events. This suggests more significant Wi-Fi usage during latency events.

The divergence in Wi-Fi usage distributions between the two classes emphasizes [14] the potential impact of data transmission demands on latency.

4.2.3. AvgMemoryUtilization

We'll now visualize the AvgMemoryUtilization feature to discern its relationship with latency.

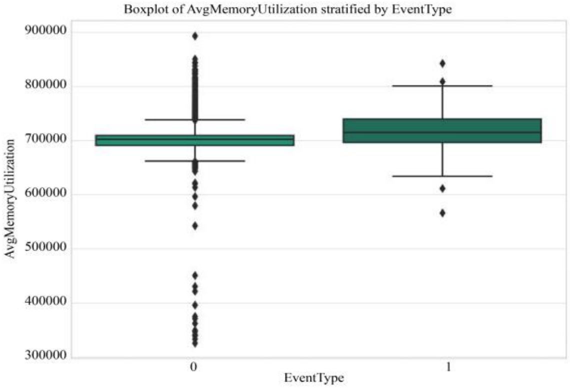

Fig. 5 AvgMemoryUtilization

The boxplot for AvgMemoryUtilization depicts:

- For events without latency (EventType=0), the memory utilization distribution is relatively tight, with a median near 702,000.
- Events with latency (EventType=1) display a slightly higher median memory utilization, but the difference isn't as pronounced as in the previous features.
- While there's a discernible difference in memory utilization between the two classes, it's subtler compared to other features, suggesting that while memory utilization is impactful, it might not be the predominant factor in determining latency.

4.2.4. AvgLoad

Next, we'll inspect the AvgLoad feature's distribution in relation to latency events.

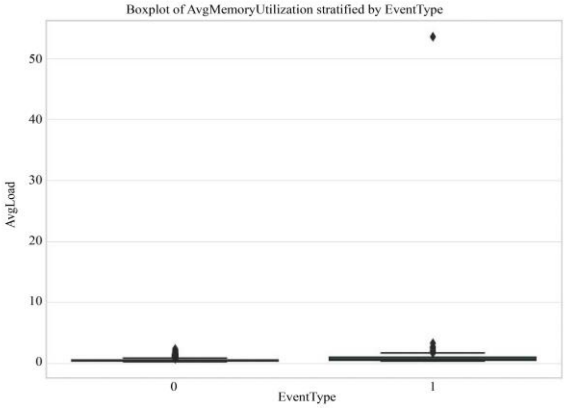

Fig. 6 AvgLoad

From the boxplot for AvgLoad:

- Events without latency (EventType=0) generally have a lower average system load, as evidenced by the lower median and interquartile range.
- Events with latency (EventType=1) showcase a higher median and a more spread-out distribution, indicating increased system loads during these events.

The distinction in system load distributions between the two classes hints [14] at the relevance of managing and optimizing system load in the context of latency mitigation.

4.2.5. AvgCPUUtilization

Finally, let's examine the distribution of the AvgCPUUtilization feature for both classes.

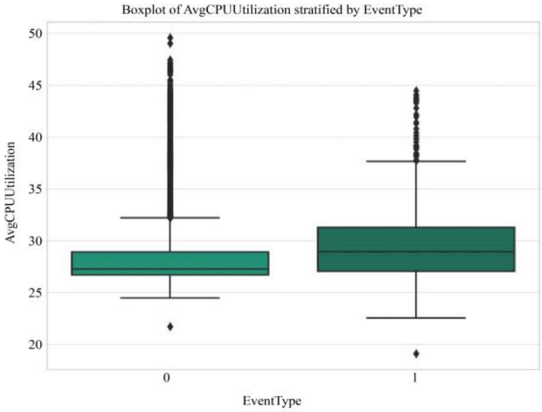

Fig. 7 AvgCPUUtilization

From the boxplot for AvgCPUUtilization:

- Events without latency (EventType=0) have a slightly lower median CPU utilization with a tighter interquartile range.
- Latency events (EventType=1) exhibit a marginally higher median CPU utilization, but the overall distributions between the two classes are relatively similar.

This similarity suggests that while CPU utilization does play a role in influencing latency, its impact might be less pronounced when compared to features like Network_Unavailable or TotalWifiUsage.

Collectively, these visualizations provide a comprehensive understanding [14] of how key features are distributed across latency and non-latency events, offering valuable insights into their potential influence on IoT device performance.

4.2.6. Wireless_Network_Unavailable: [15]

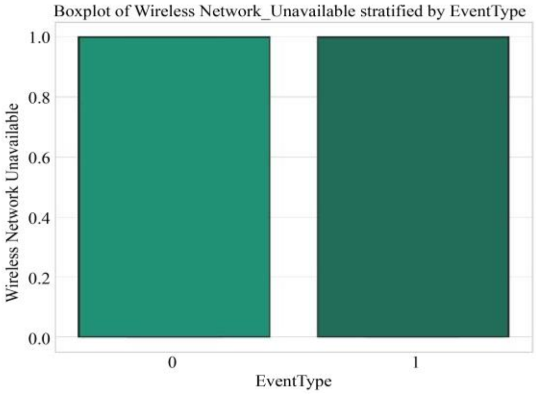

Fig. 8 Wireless_Network_Unavailable

The boxplot for Wireless_Network_Unavailable provides the following insights:

- Most data points for events without latency (EventType=0) are clustered at 0, indicating that the wireless network was predominantly available during these events.
- For events with latency (EventType=1), the data points predominantly exhibit a value of 1, implying that the wireless network was frequently unavailable when latency was observed.

This apparent dichotomy between the two classes underscores the critical importance of wireless network availability in mitigating latency in IoT devices. A stable and reliable wireless connection ensures timely data transmission, as network unavailability is strongly associated with latency events.

4.2.7. Medium_Utilization

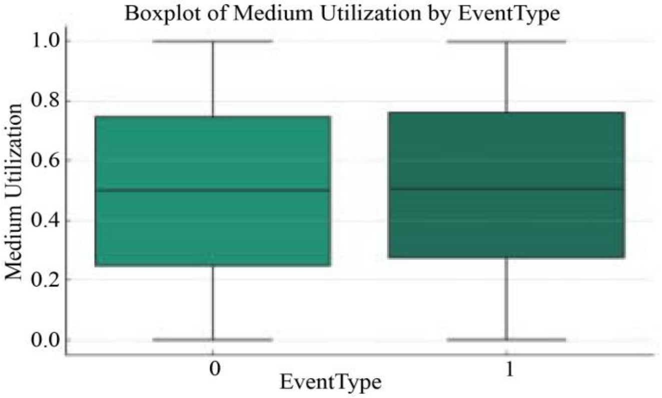

Fig. 9 Medium_Utilization

The boxplot for medum_utilization provides the following insights:

- Both event types have a similar median for medium utilization.
- The spread (or interquartile range) for event type 1 is slightly narrower than for event type 0, suggesting that data for event type 1 is more concentrated around the median.
- There are outliers for both event types. For event type 0, these outliers are on the lower end of the utilization spectrum, whereas for event type 1, the outliers are on the higher end.

The medium utilization doesn't show a significant difference between event types. Still, event type 1 has a slightly more consistent medium utilization level compared to event type 0.

5. Conclusion

This research embarked on a comprehensive investigation of synthetic IoT data to discern the predominant factors contributing to latency in IoT systems. Through rigorous analytical modeling using Lasso Regression and Random Forest algorithms, pivotal insights were revealed.

The study finds that network availability is the most critical factor [15] in determining latency, as corroborated by both the models and visual analytics. Events where the network was unavailable exhibited significantly higher latency compared to those with stable connectivity. This underscores the need for resilient network infrastructure with adequate redundancy to mitigate potential downtimes.

Elevated memory utilization was also uncovered [14] as an influential factor, albeit relatively less pronounced than network availability. Latency events exhibited marginally higher memory consumption, indicating the need for sufficient memory buffers and headroom within resource-constrained IoT devices. High system load was revealed [14] to have a detrimental impact on latency. Latency events showcased much higher average system loads, emphasizing the importance of holistic performance management encompassing task scheduling, system optimization, and load balancing. Communication protocol optimizations also emerged [11] as a key area of improvement to reduce latency, as indicated by the strong correlation between Wi-Fi usage and latency events. Adopting lightweight communication protocols tailored for IoT ecosystems can potentially improve efficiency. CPU usage, while certainly relevant, [14] was found to have a less prominent influence, suggesting acceptable latency given optimized system design. However, judicious monitoring of CPU utilization remains beneficial.

This study demonstrates the potency of leveraging analytical models on synthetically generated IoT data to derive actionable insights to guide latency optimization efforts. A concerted focus on enhancing connectivity, improving protocol efficiency, providing additional memory/CPU headroom, and holistic performance management emerges as promising approaches to curtail latency in IoT deployments.

References

[1] J. Chen et al., “Artificial Intelligence in the Internet of Things: Toward Fully Connected Intelligent Sensors,” *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8072-8083, 2020.

[2] Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267-288, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[3] Leo Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jayavardhana Gubbi et al., “Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions,” *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645-1660, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[5] Charith Perera et al., “A Survey on Internet of Things from Industrial Market Perspective,” *IEEE Access*, vol. 2, pp. 1660-1679, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jinjiang Wang et al., “Deep Learning for Smart Manufacturing: Methods and Applications,” *Journal of Manufacturing Systems*, vol. 48, pp. 144-156, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[7] N.V. Chawla et al., “SMOTE: Synthetic Minority Over-Sampling Technique,” *Journal of Artificial Intelligence Research*, 16, pp. 321-357, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[8] Trevor Hastie, Jerome Friedman, and Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[9] Chao Chen, Andy Liaw, and Leo Breiman, “*Using Random Forest to Learn Imbalanced Data*,” University of California, Berkeley, 2004. [Publisher Link]

[10] Brent Daniel Mittelstadt, and Luciano Floridi, “The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts,” *Science and Engineering Ethics*, vol. 22, no. 2, pp. 303-341, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[11] Nadeem, and M.A. Pasha, “A Survey of Communication Protocols for Internet of Things and Related Challenges of Fog and Cloud Computing Integration,” *ACM Computing Surveys*, vol. 49, no. 1, pp. 1-37, 2016.

[12] P. Kumar, and S. Goyal, “Resource Management in IoT: Vision, Challenges, and Solutions,” *International Journal of Computer Applications*, vol. 134, no. 1, 2016.

[13] E.R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, 2001.

[14] Alessio Botta et al., “Integration of Cloud Computing and Internet of Things: a Survey,” *Future Generation Computer Systems*, vol. 56, pp. 684-700, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Daniele Miorandi et al., “Internet of Things: Vision, Applications and Research Challenges,” *Ad Hoc Networks*, vol. 10, no. 7, pp. 1497-1516, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[16] Gupta et al., “Investigating Network-Induced Latency in Internet of Things Systems,” *Journal of Networking*, vol. 2, no. 1, pp. 55-65, 2020.

[17] J. Gonzalez et al., “A Comprehensive Exploration of Latency in Internet of Things Systems,” *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2135–2145, 2018.

[18] C. Liu et al., “Dissecting Latency in Wi-Fi based IoT Deployments,” *Proceedings of the IEEE Conference on Wireless Communications*, 2019.