Integrating Path-Loss Modeling and Machine Learning for Localisation within LoRaWANs

By

AlaaAllah Ahmed ElSabaa

A thesis submitted as part of the requirements
for the degree of Doctor of Philosophy
at the College of Engineering
Birmingham City University, Birmingham, UK

November 2025

Supervisor(s) Prof. Wenyan Wu and Dr. Florimond Gueniat

DECLARATION

I confirm that the work contained in this PhD project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed .... Date ....

AlaaAllah Ahmed ElSabaa

"And that man attains only what he has striven for; and that surely his pursuit will eventually be seen; then he is repaid for it with the fullest repayment; and that to your Lord is the finality."

— Quran [Ch 53, verse 39-42]

ABSTRACT

Internet of Things (IoT) technologies and applications have surged across various environments; urban and remote environments have amplified the need for energy-efficient and accurate localisation techniques. Low-Power Wide-Area Networks (LPWANs) operate independently for extended periods, making understanding the communication channel between end nodes and gateways crucial for network range, performance, and profitability. This research addresses the challenge of node localisation in LoRaWAN-based IoT networks by proposing a data-driven, GPS-independent framework utilising Long Short-Term Memory (LSTM) neural networks, constrained to single-gateway, sector-level classification across three angular sectors.

The study begins with the design, deployment, and analysis of a real-world LoRaWAN network across a defined urban area in Birmingham, UK. Through extensive field experimentation, empirical path-loss models were derived to characterise the radio environment, capturing spatial signal attenuation under diverse conditions. The Birmingham-specific regression model yielded a path-loss exponent of $n = 2.52$ and intercept $B = 130.85$ dB, outperforming classical models (Okumura-Hata and COST-231) with a mean absolute error of 9.93 dB. These models formed the basis for generating synthetic Received Signal Strength (RSS) values, augmenting the original dataset and addressing the imbalance in sector-wise data distribution. Statistical validation confirmed KDE divergence values below 1.71×10^{-4} between real and synthetic distributions, ensuring the fidelity of synthesised samples for training.

Subsequently, a deep learning localisation model was developed using LSTM networks, trained solely on time-series RSS (Received Signal Strength) and SNR (Signal-to-Noise Ratio) values of sequentially received packets. The proposed LSTM framework achieved an overall classification accuracy of 72.1% (single split) and $68.8\% \pm 2.2\%$ across five-fold cross-validation, with balanced accuracies of 0.72–0.80 per sector. The architecture outperformed a CNN–LSTM hybrid (65% accuracy) and static baseline models; including SVM (30.9%), kNN (27.0%), Logistic Regression (30.0%), and HMM (46.1%). Statistical validation via McNemar's test confirmed the significance of improvement over the CNN–LSTM baseline ($\chi^2 = 9.26$, $p = 0.0023$). Noise-injection experiments demonstrated accuracy stability at approximately 58% under Gaussian RSSI perturbations of up to ±5 dB, confirming resilience to real-world signal fluctuations.

This study presents a unified methodology for coarse-grained, single-gateway localisation using only sequential RSSI and SNR measurements, without relying on dense infrastructure or GPS hardware, providing an extensible foundation for next-generation IoT and smart-city deployments.

ACKNOWLEDGMENTS

First and Foremost, all praise to Almighty Allah for blessing this journey, and bestowing me with endurance to complete this PhD study.

The PhD journey is often described as a marathon—long, exhausting, and demanding far more persistence than speed. Every mile brings new challenges, and there are moments when the finish line feels impossibly far away. Yet this race was made achievable through the encouragement and belief of those who paced me, lifted me, and kept me moving forward when my strength alone was not enough. My sincere gratitude goes to all those who have been part of my experience.

I would like to express my deep sense of gratitude to my supervisors. Prof. Wenyan Wu as my Director of Studies (DoS) for providing her insights, guidance, and academic support during my journey. Dr Florimond Gueniat as my second supervisor for his deep knowledge, insightful discussions and continuous support. His ability to gently guide me back to clarity during overwhelming times, played a crucial role in helping me regain focus and continue my work with confidence.

My heartfelt thanks go to my family, my mom, Prof. Eman ElMallah for her confidence and unconditional love. Her support has been the foundation upon which this entire journey was built, and who never once let me believe that giving up was an option. To my husband, Eng Yasser Rashwan, whose understanding and unwavering support gave me strength during the most challenging moments, who held me through the late nights when the tears came easier than the words and celebrated with me through the victories. To my little toddler, Malika, who brought joy, laughter, and perspective to every difficult day, your presence reminded me of what truly matters. To the little one on the way, you have already given me more motivation to finish this thesis than any deadline ever could. Thank you for defending this thesis with me.

I would also like to extend my gratitude to my former supervisor, Prof. Michael Ward, whose guidance during the early stages of this work was invaluable. His support contributed meaningfully to the foundation of this research.

My thanks go to the Doctorate Research College (DRC), namely to Prof. Peter Larkham, Ms. Sue Witton, Ms. Faheemah Yilmaz and Mr. Lee Lawrence, for their administrative and technical assistance for this research and for always being helpful. I would like to give a special acknowledgement to the PhD committee who will invest time to assess this research and evaluate the PhD defence.

To the friends who became family throughout this journey — thank you for the laughter, the late nights, board games, coffee and cold fast-food and the completely unsolicited but deeply appreciated advice. You made this PhD feel less like a solitary endeavour and more like a shared adventure. Dr. Fatma Ahmed, Dr. Haitham Hassan, Mr. Ian McDonald, Dr. Bassant Rateb, Dr. Fatima AbdAllah, Dr. Diana Haidar, Dr. Sara Hassan and Dr. Antonio Nehme, thank you for your help and encouragement throughout the PhD journey. You made my days brighter during the darkest times.

My appreciation extends to my family and friends including Mrs Mariam Zaki, Mrs Dina Atef, Eng. Abdelrahman ElSabaa. Thank you for supporting me, especially when you pretended to listen when I explained my research for the hundredth time.

Finally, I would like to thank my examiners for their thorough and constructive engagement with my work — their feedback has made this thesis considerably stronger, and

their questions during the viva were a reminder that I am not the only one interested in this work.

Contents

Page

List of Figures

List of Tables

List of Abbreviations

ADR	**A**daptive **D**ata **R**ate
AWGN	**A**dditive **W**hite **G**aussian **N**oise
BER	**B**it **E**rror **R**ate
BiLSTM	**Bi**directional **L**ong **S**hort-**T**erm **M**emory
BW	**B**and**w**idth
CI	**C**onfidence **I**nterval
CNN	**C**onvolutional **N**eural **N**etwork
CNN-LSTM	**C**onvolutional **N**eural **N**etwork – **L**ong **S**hort-**T**erm **M**emory
CR	**C**oding **R**ate
CSS	**C**hirp **S**pread **S**pectrum
CSV	**C**omma-**S**eparated **V**alues
DB	**D**ata**B**ase
DBC	**D**istribution-**B**ased **C**lustering
DR	**D**ata **R**ate
EUI	**E**xtended **U**nique **I**dentifier
FCN	**F**ully-**C**onnected **N**etwork
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
FSPL	**F**ree-**S**pace **P**ath **L**oss
GFLOPs	**G**iga **F**loating-point **O**perations per **S**econd

GNSS	**G**lobal **N**avigation **S**atellite **S**ystem
GPS	**G**lobal **P**ositioning **S**ystem
GRU	**G**ated **R**ecurrent **U**nit
HMM	**H**idden **M**arkov **M**odel
IoT	**I**nternet **o**f **T**hings
ISM	**I**ndustrial, **S**cientific and **M**edical band
kNN	**k**-**N**earest **N**eighbours
KDE	**K**ernel **D**ensity **E**stimation
LoRa	**L**ong **R**ange
LoRaWAN	**L**ong **R**ange **W**ide **A**rea **N**etwork
LOS	**L**ine-**o**f-**S**ight
LPWAN	**L**ow-**P**ower **W**ide-**A**rea **N**etwork
LPP	**L**ow **P**ower **P**ayload
LSTM	**L**ong **S**hort-**T**erm **M**emory
MAD	**M**edian **A**bsolute **D**eviation
MAC	**M**edium **A**ccess **C**ontrol
MAE	**M**ean **A**bsolute **E**rror
MSE	**M**ean **S**quared **E**rror
NLOS	**N**on-**L**ine-**o**f-**S**ight
OTAA	**O**ver-**T**he-**A**ir **A**ctivation
PLE	**P**ath-**L**oss **E**xponent
PL	**P**ath **L**oss
ReLU	**Re**ctified **L**inear **U**nit

RMSE	**R**oot **M**ean **S**quare **E**rror
RMSProp	**R**oot **M**ean **S**quare **Prop**agation optimizer
RSSI	**R**eceived **S**ignal **S**trength **I**ndicator
SNR	**S**ignal-**t**o-**N**oise **R**atio
SSE	**S**um of **S**quared **E**rrors
SGDM	**S**tochastic **G**radient **D**escent with **M**omentum
SVM	**S**upport **V**ector **M**achine
TDoA	**T**ime **D**ifference **o**f **A**rrival
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
TTN	**T**he **T**hings **N**etwork
Tx	**T**ransmit
Rx	**R**eceive
UDR	**U**ser **D**ata **R**ecord
WSN	**W**ireless **S**ensor **N**etwork

List of Publications

[Published]

1. ElSabaa, AlaaAllah A., Michael Ward, and Wenyan Wu. "Hybrid Localization techniques in LoRa-based WSN." 2019 25th International Conference on Automation and Computing (ICAC). IEEE, 2019. Link to Article: https://ieeexplore.ieee.org/abstract/document/8895196

2. ElSabaa, AlaaAllah, Florimond Guéniat, Wenyan Wu, and Michael Ward. "Enhanced Data-Driven LoRa LP-WAN Channel Model in Birmingham." In 2022 IEEE World AI IoT Congress (AIIoT), pp. 766-772. IEEE, 2022. Link to Article: https://ieeexplore.ieee.org/abstract/document/9817253

3. ElSabaa, AlaaAllah, Florimond Guéniat, and Wenyan Wu. "Data-driven single gateway location estimation using improved channel model and LSTM: A. ElSabaa et al." Telecommunication Systems 89.2 (2026): 59. Link to Article: https://link.springer.com/article/10.1007/s11235-026-01429-9

Chapter 1

Introduction

With the exponential growth of Internet of Things (IoT) devices connectivity on the road of digitalising everyday life and the expectation of the market growth for manufacturing IoTs from the value of $326.1 billion in 2021 and to reach $1742.8 billion by 2030 (Okokpujie and Tartibu, 2024), existing communication technologies are faced with new challenges to fulfill the life quality promised. Those challenges include assured connectivity with low-power demand to increase the devices' lifetime, long range, low data rate and cost-effectiveness.

Low-powered wireless area networks (LP-WANs), often associated with IoT and Wireless Sensor Networks (WSNs), are designed to operate efficiently with limited energy resources. Technologies such as Sigfox (UnaBiz, 2021) and Long Range (LoRa) (Alliance, 2015) offering optimised device cost, higher coverage when compared to LTE and low-power utilisation, has facilitated the maturation of Smart Cities. These networks encompass a wide range of applications, from environmental monitoring (Nor, Zaman, and Mubdi, 2017) and healthcare to smart cities (Mdhaffar et al., 2017) and industrial automation. With their abilities to communicate at lower power levels and lower transmission rates less than 150 kb/s over a radius that can reach up to several kilometers (Popli, Jha, and Jain, 2018), making them an attractive solution for event-based IoT applications.

This thesis addresses the challenge of infrastructure-independent, data-driven localisation using signal-based features and machine learning, particularly in LoRaWAN networks.

1.1 Research Motivation and Question

Accurate location estimation is a critical challenge in Low-Power Wide-Area Networks (LPWANs), particularly for IoT applications such as environmental monitoring, asset tracking, and smart agriculture. This ease and accessibility of using electronics has created a large demand for electronic gadgets, making the Internet of Things (IoT) a large area of interest. The economic benefits of IoT in additive and advanced manufacturing, on the other hand, outweigh the hurdles. Improved efficiencies, lower prices, and higher quality contribute to increasing competitiveness (Okokpujie and Tartibu, 2024).

Traditional localisation methods, including Global Navigation Satellite Systems (GNSS), are often power-intensive and unreliable in obstructed environments such as urban canyons and dense forests. LoRaWAN, a leading LPWAN technology, offers long-range, energy-efficient communication, but its inherent limitations —such as signal attenuation, multipath fading, and network constraints— pose challenges for precise localisation.

It is important to note that the term "localisation" in this thesis refers specifically to sector-level classification; the prediction of the angular quadrant in which a LoRaWAN node resides relative to a single gateway, rather than the estimation of precise Cartesian coordinates. This distinction is deliberate: given the constraints of a single-gateway deployment and the inherent variability of LoRaWAN signal propagation, coordinate-level positioning is neither feasible nor the objective of this work. Instead, coarse-grained, region-level inference from sequential signal features is pursued as a practical and energy-efficient alternative to GPS-dependent or infrastructure-intensive positioning systems.

Recent advancements in machine learning have demonstrated significant potential in improving location accuracy by leveraging signal-based features such as Received Signal Strength Indicator (RSSI), Time Difference of Arrival (TDoA), and other propagation characteristics. However, existing models often suffer from data sparsity, environmental variability, and limited generalisation when applied across different terrains and deployment scenarios.

Despite significant advancements in LoRaWAN localisation, several gaps remain unaddressed. Firstly, most studies are confined to simulated environments. Real-world data-driven localisation models are scarce, resulting in insufficient validation of their practical performance. Secondly, while LSTM-based approaches show promise, hybrid models combining other neural architectures (like CNN) and adaptive path loss estimation remain underexplored. Finally, few models explicitly account for noise in real-world LoRaWAN signals, despite its significant impact on localisation accuracy.

This research aims to develop a robust and scalable machine learning data-based localisation framework for LoRaWAN networks using real-world experimental data. The research also focuses on maintaining low-power consumption through exploring the limitations of utilising a single gateway for locating the node. By leveraging supervised and unsupervised learning techniques, the study seeks to improve localisation accuracy while addressing key challenges such as environmental dynamics, hardware variability, and computational efficiency. The outcomes of this research will enhance the feasibility of LoRaWAN-based localisation for large-scale IoT deployments in outdoor environments.

This research formulates the core problem as: **Can temporal signal features (e.g., RSSI, SNR) from LoRaWAN devices be used to accurately infer spatial zones using machine learning, without relying on GPS or external anchors?**

To address this question, it is divided into the following sub-questions:

1. What are the key limitations of existing LoRaWAN-based localisation techniques in outdoor environments, and how do environmental factors impact localisation accuracy?

2. How can machine learning models be optimized to improve localisation accuracy in LoRaWAN networks while maintaining computational efficiency?

3. What are the most relevant signal features (e.g., RSSI, TDoA, SNR) that contribute to accurate location estimation using LoRaWAN?

4. How can experimental data collection and preprocessing be designed to enhance the robustness and generalisability of machine learning-based localisation models?

5. Can a hybrid machine learning approach (e.g., combining supervised learning with probabilistic models) improve the performance of LoRaWAN localisation compared to traditional empirical models?

6. What are the trade-offs between localisation accuracy, power consumption, and scalability in real-world LoRaWAN deployments?

1.2 Aim and Objectives

The main aim of this work is to develop a data-driven, GPS-independent localisation framework for LoRaWAN devices using LSTM-based learning models.

1. Experimentally characterise signal propagation in a real-world LoRaWAN deployment.

2. Develop empirical and analytical path loss models to describe the propagation environment.

3. Synthesize realistic RSS data to augment underrepresented regions in the dataset.

4. Design and implement LSTM models for classification-based localisation.

5. Evaluate the system's performance under varying environmental and spatial conditions.

1.3 Scope and Delimitation

This study focuses on urban and campus-scale outdoor environments where LoRaWAN coverage is available. Localisation is limited to 120 angle sector-based classification, representing coarse-grained location prediction rather than exact coordinates. The proposed methodology uses only RSSI, SNR, and derived features, deliberately excluding GPS and other external references to maintain generalisability and low power consumption.

1.4 Methodology

1.4.1 Research Design and Methodological Approach

This thesis adopts an **experimental, data-driven research approach**, combining real-world empirical data collection, physical signal modelling, and supervised machine learning within a unified investigative framework. Rather than relying on simulation or theoretical derivation alone, the research is grounded in measurements collected from a live LoRaWAN deployment in Birmingham City Centre, ensuring that all modelling assumptions and learning outcomes are validated against real propagation conditions. This positions the work within the applied engineering research tradition, where scientific rigour is demonstrated through experimental reproducibility, statistical validation, and quantitative performance evaluation rather than purely through mathematical proof. The methodology integrates three complementary research activities that are deliberately sequenced but mutually dependent: (1) **experimental data collection**, in which a real-world LoRaWAN testbed is designed, deployed, and used to gather empirical RSSI and SNR measurements across a defined urban environment; (2) **empirical propagation modelling**, in which the collected data is used to derive and validate a Birmingham-specific path-loss model that characterises the physical signal environment; and (3) **supervised sequential learning**, in which the validated propagation model and collected measurements are integrated into an LSTM-based classification framework for GPS-free sector-level localisation. Each activity informs and enables the next, forming a coherent research pipeline that progresses from physical observation through mathematical modelling to data-driven inference. The specific steps within this pipeline are detailed below. The complete workflow is summarized in Figure 1.1. The numbered steps below elaborate each phase of this workflow in sequence, from the initial literature review through to the final integration and evaluation of the proposed localisation framework, mapping directly onto the chapters that follow.

1. Literature Review of LP-WAN localisation Techniques: through conducting an in-depth survey of existing localisation methods within LP-WANs, focusing on technologies such as LoRaWAN and identify commonly used methodologies, their limitations, and opportunities for enhancement. Establish a baseline understanding of the strengths and weaknesses in current approaches.

2. Selection of Suitable Technologies and Techniques: Extract and justify a suitable combination of technology, methodology, and technique based on findings from the literature review. Focus on technologies with the most promise for GPS-free localisation in outdoor LP-WAN environments.

3. Framework Proposal and Conceptualization: Further investigate the selected approach and propose a novel conceptual framework for outdoor localisation using LP-WANs. Most importantly, define how the framework integrates empirical modeling with machine learning for enhanced accuracy.

4. Preliminary Simulation Using MATLAB: Perform initial simulations to validate the feasibility of LoRaWAN-based localisation models using MATLAB. This aims to assess the basic performance metrics and behaviors under simulated environments.

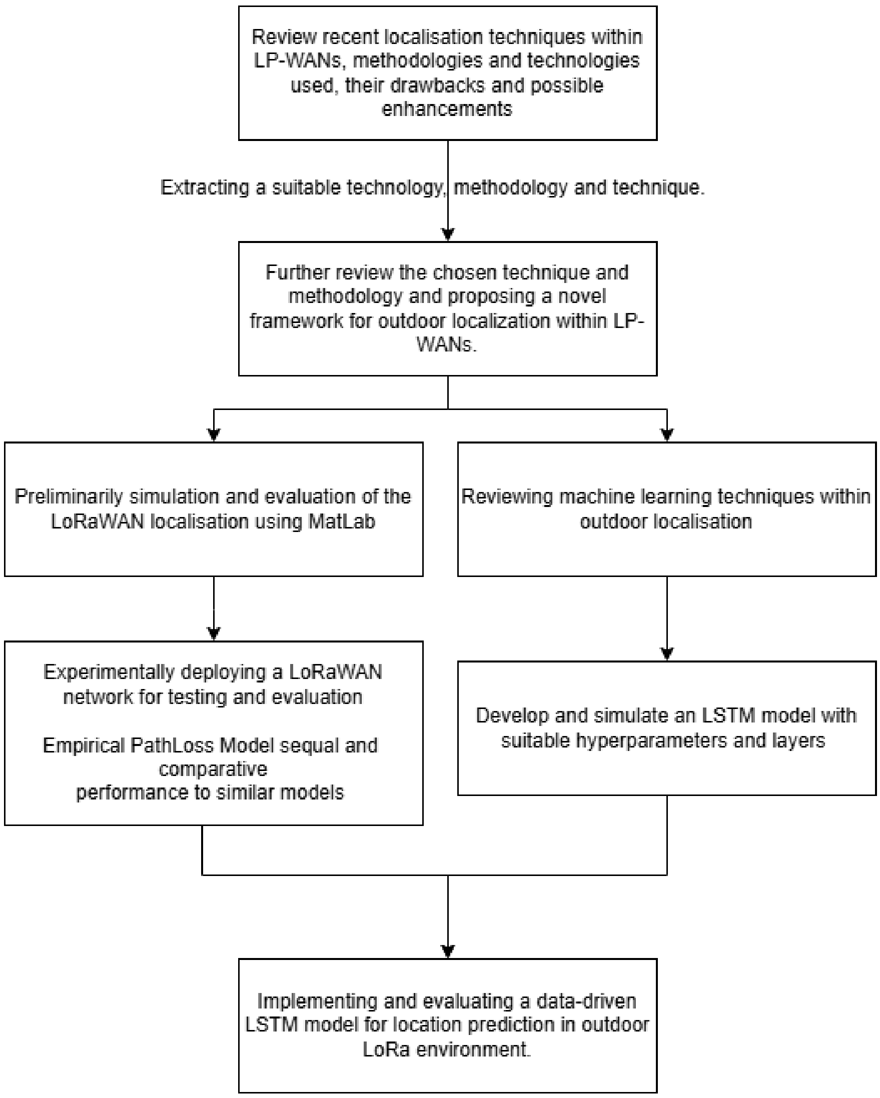


Figure 1.1: Overall research methodology framework

5. Review of Machine Learning Techniques for localisation: Explore state-of-the-art machine learning algorithms, particularly those used for spatial and temporal prediction tasks. Emphasize sequence-based models like LSTM due to their ability to model time series data effectively.

6. Empirical LoRaWAN Network Deployment: Design and deploy a real-world LoRaWAN network to collect signal data. Use the deployment to test model hypotheses and gather training and validation data for the machine learning model.

7. Development and Simulation of an LSTM Model: Build an LSTM-based architecture tailored for location prediction using collected time series data. Experiment with various hyperparameters and network depths to optimize model performance.

8. Path Loss Model Calibration and Comparison: Derive an empirical path loss model from the experimental data. Evaluate the model against existing standard path loss models to validate accuracy and applicability.

9. Final Implementation of the LSTM-Based localisation System: Integrate the empirical model and LSTM network into a unified GPS-free localisation solution. Evaluate system performance in terms of localisation accuracy, robustness, and scalability in the outdoor LoRa environment.

This structured methodology ensures a logical progression from theoretical foundation to experimental validation. Critically, the relationship between path-loss modelling and machine learning in this framework is not merely sequential but fundamentally interdependent. The empirically derived path-loss model serves two critical roles in supporting the learning pipeline: first, it provides physically grounded synthetic RSS values used to augment underrepresented spatial regions in the training dataset, ensuring that the LSTM is exposed to signal patterns consistent with real propagation behaviour rather than arbitrary noise; second, model-derived distance estimates and near/far-field regression parameters are incorporated directly as input features to the LSTM, enriching the temporal signal representation beyond raw RSSI and SNR values. In this way, the physical characterisation of the channel directly informs both the quality of training data and the discriminative power of the learned features — linking empirical modelling and machine learning under a unified framework for practical LoRaWAN localisation.

1.5 Contributions

- Deploying a simple operative single-gateway LoRa network within Birmingham while providing an unprecedented coverage map showing the estimated RSS for covered distances.

- Development of a well-established and published path loss model for Birmingham City-Centre, combining near-field and far-field regimes.

- A statistically validated RSS data synthesis technique to address spatial data imbalance.

- Proposing, testing and validating a data-driven LSTM framework for quadrant classification utilising a single gateway. LSTM was selected over static classifiers due to the inherently sequential nature of LoRaWAN packet reception, where signal strength and noise ratio vary temporally as environmental conditions change, making temporal dependency modelling essential. Furthermore, LSTM's gating mechanism allows it to retain relevant long-term signal patterns while discarding transient noise, which is particularly valuable in a single-gateway scenario.

- A comprehensive validation framework for single-gateway LoRaWAN localisation The framework was comprehensively assessed using classification accuracy, balanced accuracy per sector, macro F1-score, precision and recall, and statistical significance testing, going beyond the accuracy-only reporting typical of existing LoRaWAN localisation studies. The framework additionally incorporates spatial consistency analysis through angular error metrics, robustness assessment via controlled Gaussian noise injection across varying RSSI perturbation levels, and computational efficiency characterisation to confirm suitability for resource-constrained gateway deployment. This validation methodology is itself a transferable contribution, providing a rigorous and reproducible benchmarking approach for future single-gateway localisation research.

1.6 Thesis Outline

The documentation of the thesis is organised as follows: Chapter 1 introduces the research motivation, objectives, and contributions. Chapter 2 extensively reviews the state-of-art localisation techniques and methodologies used within LP-WANs, discussing their advantages and limitations. Moreover, it reviews the related machine learning techniques for localisation within LPWANs and LoRaWANs in particular.

Chapter 3 introduces the theoretical foundations and methodological framework underpinning this research. It explores key concepts in LoRaWAN communication, empirical and machine learning-based path loss modeling, and presents the rationale for adopting Long Short-Term Memory (LSTM) networks for time-series-based localisation.

Chapter 4 starts by simulating LoRaWAN to test its theoretical boundaries and build-up the need for experimental deployment. The experimental work is further discussed thoroughly, highlighting the various experiments and data collection process.

Chapter 5 discusses the processing of the data collected after experimentation, the outlier detection and elimination process, the initial regression and its comparative performance to various empirical path loss model to finally formulate a well-established path loss model for LoRaWAN propagation within Birmingham.

Chapter 6 presents the development and evaluation of the proposed data-driven Long Short-Term Memory (LSTM) localisation framework, which integrates the previously derived path loss features with sequential signal data to enable single-gateway, sector-level position estimation.

Chapter 7 concludes the developed methodologies in the work, highlights the advantages and their limitations and outlines the future work.

Chapter 2

Literature Review

2.1 Introduction

Before reviewing localisation techniques, it is useful to briefly orient the reader to the key signal characteristics of LoRaWAN that underpin the methods discussed throughout this chapter. LoRaWAN (Long Range Wide Area Network) is a low-power wide-area network protocol that employs Chirp Spread Spectrum (CSS) modulation, operating in sub-GHz ISM bands (868 MHz in Europe) to achieve communication ranges of up to 15 km in urban environments. Two signal-level parameters are of particular relevance to localisation: the Received Signal Strength Indicator (RSSI), which quantifies the power level of a received packet in dBm and serves as a proxy for transmitter-receiver distance, and the Signal-to-Noise Ratio (SNR), which captures the ratio of signal power to background noise and reflects link quality under varying propagation conditions. Both parameters are passively available at the gateway without additional hardware, making them the primary features exploited by the localisation approaches reviewed in this chapter. A fuller technical treatment of LoRaWAN fundamentals, modulation parameters, and propagation characteristics is provided in Chapter 3.

The rapid expansion of the Internet of Things (IoT) has fueled the deployment of Low-Power Wide-Area Networks (LPWANs), with LoRaWAN emerging as one of the most prominent protocols due to its long communication range, low power consumption, and scalability in dense environments. However, it was not originally designed with localisation in mind, and thus, several challenges arise when using it for geolocation purposes. These include high signal variability, limited bandwidth, and the absence of dedicated localisation infrastructure.

Traditional positioning systems such as the Global Positioning System (GPS) provide high accuracy but are often energy-intensive, cost-prohibitive for large-scale deployments, and unreliable in obstructed or indoor environments. Consequently, researchers have increasingly explored alternative, infrastructure-independent localisation methods tailored to Low Power Wide Area Networks (LPWANs), particularly LoRaWAN.

This chapter critically reviews the body of literature that forms the foundation of this research. It begins with a review of LoRaWAN and existing IoT localisation techniques, highlighting the unique characteristics and limitations of LoRa-based positioning. Following this, the chapter delves into empirical and experimental approaches to path loss modeling within IoT networks — an essential component for understanding signal behavior in real-world deployments. Next, the review explores how machine learning has been leveraged both for enhancing path loss models and for performing localisation directly from signal features. Given the increasing complexity of IoT environments and the availability of large-scale signal data, learning-based approaches — particularly those involving time-series models — offer a compelling alternative to traditional techniques.

Finally, the chapter presents a synthesis of the reviewed studies and identifies key gaps in the literature, particularly in the integration of signal modeling, data augmentation, and deep learning-based classification for GPS-independent localisation. These gaps establish the rationale and direction for the methodology and experimental design adopted in this thesis.

2.2 LoRaWAN and IoT localisation Techniques

Accurate localisation is a fundamental requirement in Internet of Things (IoT) networks, enabling a wide spectrum of applications including asset tracking, smart agriculture, industrial automation, and logistics. While technologies such as GPS offer high precision, their cost, energy consumption, and limited indoor performance make them impractical for many IoT scenarios. Consequently, research has increasingly focused on alternative localisation techniques that leverage the communication infrastructure itself — especially within Low Power Wide Area Networks (LPWANs) like LoRaWAN.

LoRaWAN-based localisation techniques are gaining increasing attention due to the inherent low-power and long-range capabilities of LoRaWAN networks. The primary challenge lies in achieving accurate localisation without relying on GPS. localisation in LPWANs is generally achieved using received signal strength indicator (RSSI), time difference of arrival (TDoA), angle of arrival (AoA), fingerprinting, or hybrid techniques. Each method has its advantages and limitations in terms of accuracy, power efficiency, and deployment complexity. LoRaWAN-based localisation techniques leverage a mix of these approaches to compensate for LoRa's limited bandwidth and susceptibility to interference. This section explores the current landscape of LoRaWAN-based IoT localisation techniques and categorizes them into core methodological classes.

Outdoor localisation is a critical requirement for many IoT applications, particularly in smart cities, environmental monitoring, and asset tracking. In (Magsi et al., 2023), authors conducted an experimental evaluation of trilateration-based outdoor localisation using LoRaWAN. Trilateration is a classical localisation technique that estimates the position of a device based on distance measurements from multiple known reference points. The research found that while trilateration is feasible, it suffers from accuracy degradation due to signal interference and environmental factors.

2.2.1 Coarse vs Fine-Grained localisation Techniques

localisation techniques for LoRaWAN can be broadly classified into coarse-grained and fine-grained approaches. Coarse-Grained Methods such as centroid localisation or zone-based RSSI classification provide approximate location estimates sufficient for region-level tracking. They are computationally lightweight and energy-efficient but limited in resolution (Coluccia and Fascista, 2019). While Fine-Grained Techniques like multilateration using TDoA, fingerprinting with RSSI or Channel State Information (CSI), and machine learning-based regression models deliver higher precision. However, they require dense infrastructure or intensive pre-processing, which may not always be feasible in real-world LPWAN deployments. Authors in (Pospisil, Fujdiak, and Mikhaylov, 2020) examined TDoA-based localisation using LoRaWAN as an effective technique. The work found that while TDoA provides better accuracy than RSSI, it requires synchronized gateways and higher network density to be effective. Moreover, work in (Y. Li et al., 2021) proposed a public LoRaWAN-based vehicle localisation system, highlighting the feasibility of tracking moving objects over LPWANs with adaptive positioning algorithms.

2.2.2 GPS-Free Localisation Approaches

Given the power and cost constraints of GPS modules, GPS-free localisation is a growing research focus for LPWAN-based IoT systems. These approaches include RSSI-Based Fingerprinting, that uses pre-collected signal strength patterns to map an unknown signal to a location. While simple, it suffers from low generalisability due to environmental changes. Another approach uses ML algorithms trained on RSSI or CSI data. It has shown promising results in adapting to non-linear signal environments and improving robustness without GPS (Ren et al., 2022). The work in (Moradbeikie et al., 2021) reviewed GNSS-free localisation methods for LoRaWAN, exploring RSSI fingerprinting, TDoA, and hybrid methods. The study found that a combination of RSSI-based fingerprinting and AI-enhanced models offers improved accuracy.

A hybrid strategy combining multiple weak signals and learning-based correction models—offers the most practical pathway forward for GPS-free, low-power localisation in LoRaWAN systems.

2.3 Path Loss Modeling within IoT Networks: Empirical and Experimental Approaches

Path loss modeling plays a foundational role in the design and analysis of wireless communication systems, particularly in LPWANs. Within the Internet of Things (IoT) ecosystem, accurate path loss models are essential for network planning, link budget estimation, and more critically, for node localisation without GPS. Path loss refers to the reduction in power density of an electromagnetic wave as it propagates through space. It is influenced by multiple factors, including distance, frequency, environmental conditions, and obstacles such as buildings or vegetation. Conventional free-space path loss models assume ideal line-of-sight (LOS) conditions, while more advanced models account for multipath fading, shadowing, and environmental clutter. Two main approaches have dominated recent literature: empirical models, derived from theoretical equations fine-tuned to field conditions, and experimental models, validated through direct field measurements.

2.3.1 Empirical Path Loss Models

Empirical models are commonly adapted from classical radio propagation theories, calibrated using statistical regression on collected signal data. One of the most widely applied models in IoT networks is the Log-Distance Path Loss Model, given by:

$$PL_d = PL_{d_0} + 10n \log_{10}(d/d_0) + X_\sigma \tag{2.1}$$

where, PL_{d_0} is the path loss at a reference distance d_0, n is the path loss exponent and X_σ is the Guassian random variable accounting for shadowing. Empirical models like Okumura-Hata, COST-231, and ITU-R have traditionally been used in cellular contexts, but their applicability to low-power IoT technologies like LoRaWAN is limited due to their low-frequency operation, varying antenna heights, and deployment in diverse urban/rural environments.

Recent work in (Moradbeikie et al., 2023), conducted an empirical RSSI-based localisation study in harsh harbor environments using LoRaWAN and proposed environment-aware enhancements to conventional PL models. In (Svertoka et al., 2024), researchers introduced a kNN-enhanced LoRaWAN localisation model based on real-time RSSI measurements and distance estimation in Industry 5.0 contexts. These studies underline the shift toward hybrid approaches that blend empirical measurements with algorithmic enhancements.

2.3.2 Experimental Path loss Modeling

While empirical models offer generalisability, experimental modeling grounded in field measurements has gained popularity for its realism and accuracy. These models directly derive propagation behavior from signal strength measurements collected across varying distances, terrains, and environments.

Early exploration of LoRa propagation was studied in many countries. In 2018, researchers in Germany, specifically Dortmund, tested different published path-loss models to study their capabilities against real signal measurements from a LoRa test deployment. They extended their work to include real-world measurements in a semi-urban area to validate the models. The research concluded that incorporating real-world elevation data enhances path loss estimations, making network planning more reliable (Linka et al., 2018). In (Petajajarvi et al., 2015), the authors explored the coverage capabilities of LoRa through real-world experiments conducted in Oulu, Finland. The study evaluated LoRa's communication range in different environments — on land (using a car-mounted node) and on water (using a boat-mounted node). Their findings showed that LoRa could achieve a communication range of over 15 km on land and close to 30 km on water using a node operating in the 868 MHz ISM band with 14 dBm transmit power and the maximum spreading factor. Moreover, a channel attenuation model was derived from the measurement data, which can be used to estimate path loss in environments similar to Oulu. In mining environments, (Bhavanam and Ragam, 2025) proposed a LoRa Sense framework, applying experimental modeling to optimize link reliability in open-cast mining operations. Their data-driven approach demonstrated that real-world terrain and topographical factors can cause substantial deviation from classical path loss trends. Such studies emphasize the importance of continuous benchmarking, especially in dynamic or densely populated deployments.

2.3.3 Comparative and Hybrid Approaches

Some studies advocate for hybrid models that blend theoretical predictions with experimental tuning. The work in (El Chall, Lahoud, and El Helou, 2019) compared various standard propagation models (Okumura-Hata, COST-231, etc.) with measured LoRaWAN data across urban and rural areas in Lebanon. They concluded that no single theoretical model suffices across all scenarios, advocating for local calibration using experimental data.

In summary, path loss modeling remains a critical component in the optimization and localisation of IoT networks, particularly in LPWAN environments like LoRaWAN. Empirical models, such as the log-distance or Okumura-Hata models, offer the advantage of simplicity and generalisability. When fine-tuned with regression techniques, they provide a practical starting point for estimating signal attenuation across various terrains. On the other hand, experimental models, derived from real-world field measurements, provide realism and environmental specificity. These models are especially valuable in complex or non-uniform environments, such as urban centers or industrial zones, where theoretical assumptions often fall short. Finally, hybrid and machine learning-enhanced approaches combine the strengths of both domains—leveraging theoretical structure while dynamically adapting to field data. These techniques can significantly improve localisation accuracy, link reliability, and deployment efficiency. By understanding and integrating these complementary approaches, researchers and practitioners can design robust and scalable IoT systems that are both adaptive and context-aware.

2.4 Machine Learning Approaches in PathLoss Modeling

Recent research trends emphasize machine learning techniques, such as artificial neural networks (ANNs), support vector regression (SVR), and Gaussian processes, to develop accurate and adaptive path loss models. These models enable better propagation prediction, mitigating environmental interference and improving the reliability of localisation solutions.

Recent work in (Isabona et al., 2022) utilised Multilayer Perceptron Neural Network (MLP-NN) as a way to adapt to the non-linearities and spatial variations in urban settings, their proposed MLP-NN outperformed traditional statistical and empirical models in predicting signal path loss and propagation characteristics in urban microcellular radio environments. The model requires a large amount of training data to achieve optimal performance, making data collection and preprocessing crucial. Computational complexity and processing power demands may limit its applicability in real-time or resource-constrained

LPWAN applications.

Furthermore, work in (Hadji and Nedil, 2025), developed a Deep Neural Network (DNN)-based path loss prediction model for 28 GHz mmWave propagation in underground mine environments using Massive MIMO measurement data. Then compared the model's accuracy with traditional empirical models such as log-distance and multi-slope path loss models, proposed a multi-stage generalisation framework to extend the 28 GHz model to other frequencies (26 GHz and 38 GHz) with minimal extra measurements. They further generated synthetic data to reduce the need for extensive site measurements while retaining model accuracy.

Researchers in (Haripriya, Suresh, et al., 2025) and (Lu et al., 2025) suggest the use of ANNs in pathloss modeling. While, (Lu et al., 2025) used real-world V2I measurement data covering different urban traffic environments and (Haripriya, Suresh, et al., 2025) compared ANN performance with four other ML models within WSN in urban environments. Both work highlighted the superiority of ANN compared to conventional empirical models. ANN showed strong adaptability to fluctuating environmental and channel conditions.

Both works in (López-Ramírez and Aragón-Zavala, 2025) and (Sunori et al., 2024), proposed the use of Guassian Process Regression (GPR) in wireless channel modeling. The former focuses on indoor mmWave communications, proposing ML-based propagation models that incorporate environmental features such as wall materials and furniture density. Their experiments show GPR achieving the lowest RMSE and highest R^2 among tested models, reducing path loss prediction errors by up to 45% compared to empirical baselines. The latter applies ML to 5G SNR estimation in both sub-6 GHz and mmWave bands, targeting improved link adaptation for adaptive modulation and coding. Using real-world deployment data, they find GPR consistently delivers the highest SNR prediction accuracy ($RMSE < 1.5dB$) and enables throughput gains of up to 12%. Together, these works highlighted GPR's strong generalisation capabilities and superior accuracy in both indoor and outdoor wireless environments, highlighting its suitability for precise network planning and optimization.

Moreover, Random Forest regression (RFT) was used to predict path loss and optimize frequency band selection in wireless networks in (Revathi et al., 2025). The model incorporated environmental, frequency, and antenna parameters. The Random Forest model reduced prediction error compared to empirical models by over 30%, enabling better link adaptation and throughput improvements.

The work in (Perihanoglu and Karaman, 2024) proposed a GeoAI-driven method that integrates geospatial information with machine learning models to enhance path loss prediction in cellular networks. The authors employ algorithms including k-Nearest Neighbors (kNN) to leverage both geographic proximity and environmental similarity when estimating signal attenuation. Using real-world cellular measurement data from diverse urban and suburban settings, the study evaluates the performance of kNN against traditional empirical models and other ML approaches. Results show that incorporating geospatial features significantly improves prediction accuracy, with kNN achieving notable reductions in RMSE and MAE, particularly in heterogeneous terrain where signal propagation is strongly influenced by environmental factors.

In summary, the recent research in path loss modeling shows that different machine learning techniques offer unique trade-offs in accuracy, adaptability, and computational demands. SVR consistently delivers high accuracy and robustness with small-to-medium datasets, especially when nonlinear kernels are used, though it requires careful hyperparameter tuning. ANNs excel at modeling complex nonlinear relationships in dynamic urban and vehicular environments, achieving top precision when sufficient training data and iterations are provided, but they are prone to overfitting and demand higher computational resources. GPR stands out for providing probabilistic predictions and superior generalisation across varied conditions, although its scalability is limited with large datasets. RFR offers interpretability and resilience to noisy data, with strong results in frequency band optimization, but can struggle with highly complex multipath scenarios. kNN is simple and effective in localised or geospatially contextual predictions, particularly in structured indoor or urban layouts, though its performance drops in high-dimensional feature spaces. These studies indicate that model choice should be guided by the data size, environmental variability, feature availability, and the balance between accuracy and interpretability required in the wireless deployment context. Table 2.1 compares the different machine learning techniques used in recent work to improve the efficiency of pathloss modeling.

Table 2.1 references: [1] (Olukanni and Funsho, 2025) , [2] (Odesanya et al., 2023) , [3] (Lu et al., 2025), [4] (Haripriya, Suresh, et al., 2025), [5] (Sunori et al., 2024), [6] (López-Ramírez and Aragón-Zavala, 2025) [7] (Revathi et al., 2025) [8] (Perihanoglu and Karaman, 2024)

2.5 Machine learning within IoT Localisation

Machine learning (ML) has become an increasingly central methodology in IoT localisation systems, particularly in scenarios where traditional localisation techniques such as GPS are infeasible due to cost, energy, or coverage constraints. In LPWANs like LoRaWAN, ML models compensate for the limitations of low data resolution, irregular signal quality, and sparse infrastructure by learning complex mappings from signal features (e.g., RSSI, SNR) to location-related classes or coordinates. This section reviews key techniques and architectures that underpin ML-driven localisation in LPWAN and IoT contexts.

Table 2.1: Advantages and Limitations of ML Models in Path Loss Prediction

Model	**Advantages**	**Limitations**	**Ref.**
SVR	High accuracy with small-to-medium datasets; good at handling nonlinear relationships via kernels; robust to overfitting; effective generalisation with diverse data.	Sensitive to hyperparameters (C, ε, kernel width); computationally expensive for large datasets; less interpretable than physical models; requires feature scaling.	[1],[2]
ANNs	Can capture highly complex nonlinear relationships; flexible architecture; strong performance with large datasets; adaptable to various feature sets.	Prone to overfitting without regularization; requires large datasets for good generalisation; high computational cost; less interpretable.	[3],[4]
GPR	Provides probabilistic predictions with uncertainty estimates; performs well with small datasets; flexible kernel choice for nonlinear problems.	Computationally expensive for large datasets; performance degrades with high-dimensional data; kernel choice critical; less scalable.	[5],[6]
RFR	Handles high-dimensional data well; robust to outliers and noise; interpretable feature importance; low risk of overfitting with sufficient trees.	Can be biased toward majority patterns; performance drops with very noisy features; may require tuning many hyperparameters (trees, depth).	[7]
KNN	Simple and intuitive; no explicit training phase; good for small datasets; adaptable to local data variations.	Sensitive to irrelevant features; performance drops in high-dimensional spaces; memory intensive; sensitive to choice of k and distance metric.	[8]

2.5.1 Recurrent Neural Networks and LSTM in Temporal Signal Modeling

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are well-suited for modeling the temporal evolution of radio signals in mobile or semi-mobile IoT devices. Unlike static classifiers, LSTMs capture dependencies across time, enabling them to leverage the sequence of RSSI/SNR values over multiple transmissions to infer location zones or movement patterns.

This approach is especially relevant when the node is stationary but the radio channel conditions vary due to environmental dynamics. Several studies have applied LSTM to classify location zones (e.g., quadrants or rooms) using sequences of RSSI data.

For instance, in (Purohit, 2019) authors developed an indoor fingerprinting-based localisation using deep learning techniques, leveraging RSSI data messages. The study found that integrating Recurrent Neural Networks (RNN) significantly improved localisation accuracy.

Table 2.2 summarises key machine learning techniques applied to IoT and LPWAN localisation, highlighting the diversity of approaches in the literature and providing context for the architectural choices made in this thesis.

Table 2.2 references: [1] (Y.-S. Chen, Hsu, and Huang, 2021) , [2] (Ren et al., 2022) , [3] (Apavatjrut, 2025), [4] (S. Ojo, A. Sari, and T. P. Ojo, 2022),

As shown in Table 2.2, existing approaches span a range of ML paradigms — from semi-supervised transfer learning to recurrent architectures and support vector regression — each addressing different aspects of the localisation problem. Notably, temporal sequence-based models such as Bi-LSTM demonstrate stronger performance in dynamic signal envi-

Table 2.2: Summary of Machine Learning Techniques Used in IoT and LPWAN Localisation

Reference	**ML Technique**	**Data Used**	**Application Focus**	**Key Findings & Insights**
[1]	Semi-Supervised Transfer Learning	RSSI + Grid Encoding	Outdoor localisation	Enabled better generalisation across zones with limited data.
[2]	Bi-LSTM	Time series RSSI	Urban link prediction	Bi-LSTM captured temporal dynamics more effectively than static models.
[3]	Regression Trees + Feature Fusion	RSSI + Environmental Data	Inter-node distance estimation	Context-aware model significantly reduced distance estimation error.
[4]	(SVR) and Radial Basis Function (RBF)	Livestock tracking signals	Rural localisation	Model captured non-linear path loss patterns better than traditional models.

ronments, motivating the adoption of LSTM-based classification in the proposed framework.

In (Y.-S. Chen, Hsu, and Huang, 2021), the study used a semi-supervised transfer learning approach to pre-train a small set of labeled data and generate virtual labeled data for the target domain. Their methodology divides the environment into grid segments, allowing the model to refine the relationship between labeled and unlabeled data iteratively using RSSI, SNR, and timestamps as key features for training the localisation model.

In (Ren et al., 2022), the study deployed two gateways and six mobile end-nodes across a $6 \times 6km^2$ urban area, collecting data over four months. Each gateway was found to cover approximately 11.3 Km^2, and even a small SNR gain (e.g., 2 dB) could expand the coverage area by 32.6%. The study assessed LoRa-based localisation, finding that the median localisation error was around 400 meters, making it suitable for road-level tracking but not precise positioning. Bi-LSTM was applied to model temporal variations in LoRa signal strength and improve link stability predictions in urban environments. By leveraging bidirectional processing, the model captured both past and future dependencies in signal fluctuations, leading to more accurate estimations of LoRaWAN coverage and reliability.

In (Apavatjrut, 2025), researchers explore how environmental data can be integrated into machine learning models to improve the accuracy of distance estimation between nodes in wireless networks.

2.5.2 LSTM and Time Series Analysis

Long Short-Term Memory (LSTM) networks have shown significant potential for time series analysis, making them suitable for localisation tasks using sequential LoRaWAN data. In 2019, (Carrino et al., 2019) introduced a LoRaLoc framework, employing machine learning-based fingerprinting techniques like Random Forest and Neural Networks to be applied to a reference map to enhance outdoor geolocation accuracy. The map combined Time Difference Of Arrival (TDOA) measurements generated by a simulated LoRa network and GPS location as ground truth. They also compared their findings to using LSTM.

While, authors in (Dabbakuti, Peesapati, and Anumandla, 2023) developed a hybrid LSTM-Convolutional Neural Network (CNN) model for GPS-free localisation, which showed superior accuracy compared to traditional approaches.

2.5.3 Sectioning and Zone Classification Approaches

A particularly practical ML strategy in GPS-free localisation is zone-based or 120 angle sectioning classification. Instead of predicting precise coordinates, the system identifies the area or region where the device resides. This reduces complexity while providing sufficient granularity for many applications, such as indoor navigation, asset tracking, or event triggering in smart cities.

Zone and sector classification approaches in the LoRaWAN localisation literature can be broadly categorised into three types based on how spatial regions are defined:

- **Grid-based zone classification:** The coverage area is divided into a regular grid of cells, and the system predicts which cell the device occupies. Chen et al. (Y.-S. Chen, Hsu, and Huang, 2021) applied semi-supervised transfer learning to grid-based outdoor localisation using LoRaWAN RSSI measurements, demonstrating that zone-level classification can achieve meaningful accuracy even with limited labelled data. While grid-based approaches offer geometric regularity, they require dense reference measurements and do not adapt naturally to irregular deployment geometries.
- **Fingerprinting-based zone classification:** Signal fingerprints are collected at known

locations and used to train classifiers that map RSSI patterns to spatial zones. This approach has been widely applied in indoor and campus-scale environments but requires intensive offline data collection campaigns and is sensitive to environmental changes that alter the fingerprint map over time (Ren et al., 2022).

- **Angle-based sector classification:** The coverage area is partitioned into angular sectors relative to a fixed gateway, and the system predicts the sector containing the transmitting node. This approach is particularly well-suited to single-gateway deployments where geometric multilateration is unavailable, as angular partitioning naturally exploits the directional characteristics of signal propagation. The proposed framework in this thesis adopts this strategy, partitioning the coverage area into three 120° angular sectors and learning a temporal mapping from sequential RSSI/SNR measurements to sector labels using an LSTM network.

Across all three categories, zone and sector classification approaches balance model simplicity with localisation utility, making them well suited to real-world, infrastructure-constrained IoT deployments where GPS signals are unreliable or unavailable. However, existing approaches predominantly rely on static feature representations — treating each RSSI measurement independently — and do not exploit the temporal evolution of signal sequences across consecutive packet receptions. This limitation directly motivates the temporal, sequence-based classification approach proposed in this thesis.

2.6 Summary and Research Gap

This chapter has presented a comprehensive review of the existing literature in the domain of IoT localisation, with a particular focus on LoRaWAN networks and the application of machine learning to enhance signal-based positioning accuracy. The review covered four key areas: IoT and LPWAN localisation fundamentals, path loss modeling methodologies, machine learning applications and advances in IoT localisation.

The exploration of LPWAN technologies illustrated the trade-off between coverage and accuracy. LoRaWAN, although highly efficient in terms of energy consumption and range, suffers from limitations in localisation accuracy due to its narrowband characteristics and low signal granularity. Traditional GPS-based or infrastructure-assisted solutions are often impractical in large-scale or low-cost deployments.

Path loss modeling, reviewed in Sections 2.3 and 2.4, showed that empirical and experimental modeling approaches often suffer from environmental sensitivity and lack generalisation across different deployment scenarios. Although theoretical models like log-distance and COST231 are commonly applied, their static assumptions frequently fail in dynamic urban or suburban environments. Recent works have sought to address this by integrating hybrid models—combining empirical measurement with ML-based prediction—but challenges remain in balancing accuracy, generalisation, and model interpretation.

Machine learning-based localisation techniques (Section 2.5) have shown promising results. Feature engineering, especially with temporal signal data, and the use of LSTM networks have significantly improved classification accuracy in multi-zone or sectioning systems. However, most implementations either rely on dense fingerprinting datasets or do not adequately address signal instability and noise in sparse real-world conditions. Moreover, data imbalance across zones and inadequate augmentation strategies further reduce the performance of models in underrepresented areas.

2.6.0.1 Key Limitations of Existing LoRaWAN Localisation Approaches

Despite significant research activity in LoRaWAN-based localisation, the following fundamental limitations persist in the existing literature:

- **Reliance on simulated or controlled environments:** Most existing studies validate their localisation models under simulated or highly controlled conditions. Real-world deployments, particularly in dense urban environments with irregular building geometry, mixed line-of-sight and non-line-of-sight conditions, and dynamic interference, are rarely addressed, limiting the practical applicability of reported results.
- **Dependence on dense infrastructure:** A large proportion of existing approaches require multiple synchronised gateways (for TDoA-based methods), dense anchor networks, or intensive fingerprinting campaigns. Such infrastructure demands conflict directly with the low-cost, low-power philosophy of LoRaWAN deployments, making them impractical at scale.
- **Static feature-based models:** Existing machine learning approaches to LoRaWAN localisation predominantly treat RSSI and SNR as static, independent measurements. This ignores the temporal evolution of signal features across successive packet receptions, which carries discriminative spatial information particularly valuable in single-gateway scenarios.
- **Data imbalance and sparsity:** Real-world LoRaWAN datasets are inherently imbalanced across spatial regions due to uneven measurement campaigns. Few studies address this through principled data augmentation or synthesis strategies grounded in physical propagation models, leading to classifiers that are biased toward overrepresented regions.

- **Limited statistical validation:** Many existing studies report accuracy figures without rigorous statistical validation, such as cross-validation, confidence intervals, or significance testing, making it difficult to assess the generalisability and reliability of reported results.

2.6.1 How These Limitations Motivate the Proposed Framework

The limitations identified above directly motivate the combined path-loss modelling and machine learning approach proposed in this thesis. Specifically:

The absence of real-world validated models motivates the experimental LoRaWAN deployment in Birmingham City Centre, providing an empirical dataset representative of genuine urban propagation conditions. The dependence on dense infrastructure motivates the deliberate constraint to a single-gateway architecture, demonstrating that meaningful localisation can be achieved without multi-gateway synchronisation or dense anchor deployment. The inadequacy of static feature-based classifiers motivates the adoption of LSTM networks, which explicitly model temporal dependencies across sequential packet receptions, a capability that static models such as SVM, kNN, and logistic regression fundamentally lack. The problem of data imbalance motivates the integration of path-loss modelling into the machine learning pipeline: rather than treating the two components as independent, the empirically derived Birmingham path-loss model is used to synthesise physically consistent RSS samples for underrepresented spatial regions, directly addressing class imbalance in a manner grounded in propagation physics. Finally, the lack of rigorous validation in existing work motivates the comprehensive statistical evaluation framework adopted here, including five-fold cross-validation, balanced accuracy, confidence intervals, and McNemar's significance testing.

Together, these motivations establish that neither path-loss modelling alone nor machine learning alone is sufficient, it is their integration that addresses the identified gaps and enables robust, GPS-free, single-gateway localisation in real-world LoRaWAN deployments.

This thesis addresses the above gaps by proposing a temporal, section-based localisation framework using LSTM networks trained on both real and synthetically generated RSS data. A unique component of this approach is the incorporation of a hybrid path loss modeling strategy that leverages empirical measurements and theoretically derived propagation functions to simulate and balance datasets across all localisation zones. The proposed system is GPS-free, infrastructure-independent and tested in real-world LoRaWAN deployments, offering a scalable and energy-efficient alternative for location estimation in IoT networks. It should be noted that this chapter has deliberately focused on the localisation literature, treating LoRaWAN primarily as a communication platform through which signal features are acquired. A detailed technical treatment of LoRaWAN fundamentals, including its modulation scheme, parameter configuration, and signal propagation characteristics, is reserved for Chapter 3, where it is presented as part of the theoretical and methodological background directly supporting the proposed framework. This structure reflects the central positioning of localisation as the primary research problem of this thesis.

Finally, table 2.3 highlights three key differentiators of the proposed framework relative to existing work: (1) it is the only approach operating under a **single-gateway** constraint, removing the need for infrastructure synchronisation; (2) it relies exclusively on **real-world** experimental data collected in a dense urban environment; and (3) it employs a significantly more **rigorous validation** methodology than comparable studies, including cross-validation, statistical significance testing, and confidence interval reporting.

Table 2.3: Comparison of Related Work to This Thesis

Study	**Infrastructure**	**Data Type**	**Method**	**Localisation Granularity**	**Validation**
Magsi et al. (2023) Magsi et al., 2023	Multi-Gateway	Real-world	Trilateration	Coordinate-level	Basic accuracy metrics
Ren et al. (2022) Ren et al., 2022	Multi-Gateway (2)	Real-world	Bi-LSTM	Road-level (∼400 m error)	Median localisation error
Chen et al. (2021) (Y.-S. Chen, Hsu, and Huang, 2021)	Multi-Gateway	Real-world	Semi-supervised Transfer Learning	Grid-level	Accuracy on labeled data
Carrino et al. (2019) (Carrino et al., 2019)	Multi-Gateway	Simulated + GPS	Random Forest, Neural Network, LSTM	Coordinate-level (TDoA-based)	Geolocation error
Dabbakuti et al. (2023) (Dabbakuti, Peesapati, and Anumandla, 2023)	Multi-Gateway	Simulated	Hybrid LSTM-CNN	Coordinate-level	Accuracy comparison
This Thesis	**Single-Gateway**	**Real-world**	**LSTM + BiLSTM**	**Sector-level (3 sections)**	**5-fold CV, balanced accuracy, McNemar's test, CI**

Chapter 3

Theoretical Background and Methodology

3.1 Inroduction

As noted at the close of the preceding chapter, the literature review focused primarily on localisation methods and signal-based approaches within LPWAN environments, without providing a full technical treatment of LoRaWAN itself. This chapter addresses that foundation, presenting the theoretical background and methodological framework required to support the proposed localisation system. It covers LoRaWAN architecture, modulation principles, and signal propagation modelling in detail, forming the technical basis upon which the experimental deployment and machine learning framework in subsequent chapters are built.

This chapter provides the theoretical foundations and methodological approach used in this research. As this research investigates localisation techniques that rely solely on signal-based features, particularly Received Signal Strength Indicator (RSSI) and Signal-to-Noise Ratio (SNR), within LoRaWAN networks, it is imperative to present the core principles underlying signal propagation, wireless communication, and machine learning architectures used in the study.

The chapter begins by outlining the architecture and characteristics of Low Power Wide Area Networks (LPWANs), with a particular focus on the LoRaWAN protocol. Understanding the operational structure of LoRaWAN is critical, as it influences both the quality and granularity of the signal data used for localisation. Following this, the chapter introduces fundamental radio propagation models—ranging from theoretical formulations such as free-space and log-distance models to empirically-derived path loss equations calibrated for urban and suburban deployments.

Subsequently, key signal features relevant to localisation—such as RSSI and SNR — are discussed in terms of their statistical behavior, sensitivity to environmental factors, and suitability as input features for machine learning models. Given the time-series nature of signal fluctuation, the chapter also provides a detailed exploration of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, which form the core of the sequential modeling approach adopted in this work.

Together, these theoretical insights serve as the conceptual framework upon which the system design, data processing, and model development efforts are built in subsequent chapters. The initial findings of this chapter include an outdoor localisation framework that has been published in (A. A. ElSabaa, Ward, and Wenyan Wu, 2019).

3.2 Low-Power Wide Area Network (LP-WANs) overview

Low Power Wide Area Networks (LP-WANs) are wireless communication technologies designed to support long-range, low-power, and low-data-rate connectivity for Internet of Things (IoT) applications. Unlike traditional cellular networks, LPWAN technologies prioritize energy efficiency, cost-effectiveness, and extended coverage, making them ideal for applications such as smart cities, agriculture, industrial monitoring, and logistics.

3.2.1 Key LP-WAN Technologies

Several LPWAN technologies dominate the market, each offering unique advantages and trade-offs mentioned below:

1. LoRaWAN (Long Range Wide Area Network): Uses chirp spread spectrum (CSS) modulation to achieve long-range communication with low power consumption. Operates in unlicensed frequency bands (e.g., 868 MHz in Europe, 915 MHz in the US). Supports star-of-stars topology, making it ideal for large-scale IoT deployments

2. NB-IoT (Narrowband IoT): A cellular LPWAN technology that operates in licensed spectrum, ensuring high reliability and security. Optimized for deep indoor penetration and massive IoT connectivity, making it suitable for applications like smart metering and industrial automation

3. Sigfox: Uses ultra-narrowband (UNB) modulation to provide low-cost, low-power, and long-range communication. Limited to small payloads (up to 12 bytes per message) but highly energy-efficient. Suited for low-data, long-battery-life applications like asset tracking and environmental monitoring

Each LPWAN technology offers unique advantages, making it better suited for specific application domains based on requirements such as deployment flexibility, message frequency, energy constraints, and reliability. Table 3.1 shows a comparison of the 3 technologies, LoRaWAN is particularly effective in scenarios that require flexible deployments across public or private networks, such as environmental sensing, smart agriculture, and rural monitoring. Its open specification and support for unlicensed spectrum make it attractive for cost-sensitive and community-driven projects. NB-IoT, on the other hand, excels in urban and industrial use cases such as smart metering, utility monitoring, and building automation, where consistent quality of service, security, and cellular infrastructure are critical. Meanwhile, Sigfox is

Table 3.1: Comparison of Key LPWAN Technologies

Feature	**LoRaWAN**	**NB-IoT**	**Sigfox**
Frequency Band	Unlicensed (868/915 MHz)	Licensed LTE	Unlicensed ISM
Data Rate	0.3–50 kbps	Up to 250 kbps	100 bps
Max Payload	242 bytes	1600 bytes	12 bytes
Range (Rural)	15–20 km	10–15 km	40 km
Battery Life	>10 years	5–10 years	>10 years
Topology	Star-of-stars	Cellular	Star
Security	AES-128	LTE-grade	Lightweight

highly suited for asset tracking, logistics, and low-complexity telemetry due to its extremely low power consumption, minimal data requirements, and simple network access. Choosing the appropriate LPWAN technology thus requires balancing trade-offs in range, data rate, battery life, and infrastructure availability based on the target application.

3.2.2 LP-WAN Performance measures

Evaluating the performance of LP-WANs is essential to understanding their suitability for diverse IoT applications. These include range, energy efficiency, data throughput, latency, and scalability. A systematic examination of these metrics provides insight into the operational strengths and limitations of each protocol, guiding their optimal deployment in real-world scenarios.

3.2.2.1 Data Rate and Payload Capacity

LPWANs trade data throughput for coverage and power savings. LoRaWAN offers variable data rates (0.3–50 kbps) by adjusting its spreading factor (SF). Sigfox allows only 12-byte payloads per message and restricts uplink transmissions to 140 messages/day, limiting its use to low-data-rate applications. NB-IoT provides higher data rates (up to 250 kbps) and supports larger payloads, making it better suited for use cases with moderate data needs and more frequent transmissions (Mekki et al., 2019).

3.2.2.2 Latency and Reliability

While LPWANs are not designed for low-latency communication, NB-IoT achieves relatively lower latency (1.6–10 seconds) due to its integration with LTE networks. LoRaWAN and Sigfox often exhibit higher delays (up to several seconds), especially in dense networks or under duty cycle restrictions. For mission-critical applications requiring guaranteed message delivery or timely updates, NB-IoT's reliability is generally superior due to its licensed spectrum and QoS support.

3.2.2.3 Scalability and Network Capacity

Scalability is critical in massive IoT deployments. LoRaWAN, with its star-of-stars topology, supports scalability through careful spreading factor allocation and multi-channel gateways, though it suffers from increased collisions in dense environments. NB-IoT leverages existing LTE infrastructure and dynamic resource scheduling to support thousands of devices per cell, while Sigfox's ultra-narrowband approach reduces interference but may experience congestion under high device density.

3.2.2.4 Interference Resistance and Coexistence

Operating in unlicensed bands exposes LPWANs like LoRaWAN and Sigfox to potential interference from other radio services. LoRa's use of chirp spread spectrum (CSS) grants it higher robustness against interference and multipath fading. NB-IoT benefits from the licensed cellular spectrum, ensuring cleaner transmission channels, but at the cost of requiring operator-managed infrastructure.

3.3 LoRaWAN Fundamentals

3.3.1 Chirp Spread Spectrum (CSS)

Long Range (LoRa) serves as the physical layer (PHY) and primarily relies on chirp spread spectrum (CSS) modulation, enabling it to achieve long-range communication with low power consumption. Meanwhile, LoRaWAN operates as the medium access control layer (MAC), responsible for effectively managing communication between LoRa end devices (EDs) and gateways (GWs) (Pasolini, 2021). CSS is a subset of Direct-Sequence Spread Spectrum (DSSS), helping the GW to recover a weak signal and achieve high sensitivity, enabling increased coverage at a lower data rate (DR). In addition, LoRa utilises five configurable resource parameters, i.e., SF, TP, CR, BW, and carrier frequency (CF), to fine-tune the link performance and energy consumption [(Kufakunesu, Hancke, and Abu-Mahfouz, 2020),(Maurya, A. Singh, and Kherani, 2022)].

3.3.2 Parameter Configuration

The achievable data rate using LoRa modulation is based on the following factors:

- Spreading Factor (SF): To ensure an adaption of data rate and signal range, various SF between 7 and 12 are available, affecting the gradient of frequency variation and thus the energy per symbol. The higher the energy per chirp symbol, the higher the achieved signal range will be. Though a higher spreading factor leads to a longer symbol transmit time, decreasing the overall data rate (A. Semtech, 2022). The relationship between LoRa transmission ToA and the employed LoRa parameters is expressed as $ToA = 2SF/BW$.

- Bandwidth (BW): LoRa based devices support bandwidths from 125 kHz up to 500 kHz. Widening the bandwidth decreases receiver sensitivity but enhances data rate due to reduced Time-on-Air (ToA). Our experiments employed a BW setting of 125 kHz.

- Coding Rate (CR): To further improve the robustness of the LoRa communication link, the devices implement cyclic error coding performing forward error correction. Such error coding incurs an additional transmit overhead, which depends on the chosen coding rate $CR = 4/(4 + n)$, with $n \in 1, 2, 3, 4$. To minimize ToA, CR=4/5 was selected.

- Transmitted Power (P_{Tx}): For LoRa nodes operating in the 433 MHz and 868 MHz bands, the maximum effective isotropic radiated power (EIRP) in the default setting is 12.15 dBm and 16 dBm, respectively. Our experiment adhered to the highest permissible P_{Tx} within the EU 868 MHz band, aligned with the LoRa device's approved duty cycle of 1%. This led to a selection of $P_{Tx} = 14$ dBm.

Table 3.2 summarises the key characteristics of LoRaWAN, providing a concise reference for the configurable parameters and performance metrics that underpin the experimental deployment and signal analysis conducted in this thesis.

Table 3.2: Key Characteristics of LoRaWAN

Parameter	Value
Frequency band	Sub-GHz ISM Bands (EU 868 MHz, US 915 MHz)
Range	Up to 15 km (urban), 40 km (rural)
Data Rate	0.3-50 kbps (adaptive via Spreading Factor)
Modulation	Chirp Spread Spectrum (CSS)
Power Consumption	Low (devices can last years on batteries)
Interference Resistance	High (due to CSS modulation)
Adaptive Data Rate (ADR)	Dynamically adjusts power and SF

3.3.3 Physical Frame Format

While LoRa modulation can transmit arbitrary frames, a specific physical frame format is defined and implemented in Semtech's transmitters and receivers, with constant bandwidth and spreading factor for each frame. LoRa frame starts with a preamble, which consists of a sequence of constant upchirps covering the entire frequency band. The last two upchirps encode the sync word. The sync word, a one-byte value, differentiates LoRa networks using the same frequency bands. If a device's decoded sync word does not match its configuration, it stops listening to the transmission. Following the sync word, there are two and a quarter downchirps lasting 2.25 symbols. The preamble's total duration can be configured between 10.25 and 65,539.25 symbols (Augustin et al., 2016).

After the preamble, an optional header may follow. If present, it is transmitted with a code rate of 4/8, indicating the payload size (in bytes), the code rate for the transmission's end, and whether a 16-bit CRC for the payload is included at the frame's end. The header also contains a CRC to enable the receiver to discard packets with invalid headers. The payload size, stored in one byte, is limited to 255 bytes. The header is optional, allowing it to be disabled when the payload length, coding rate, and CRC presence are pre-determined.

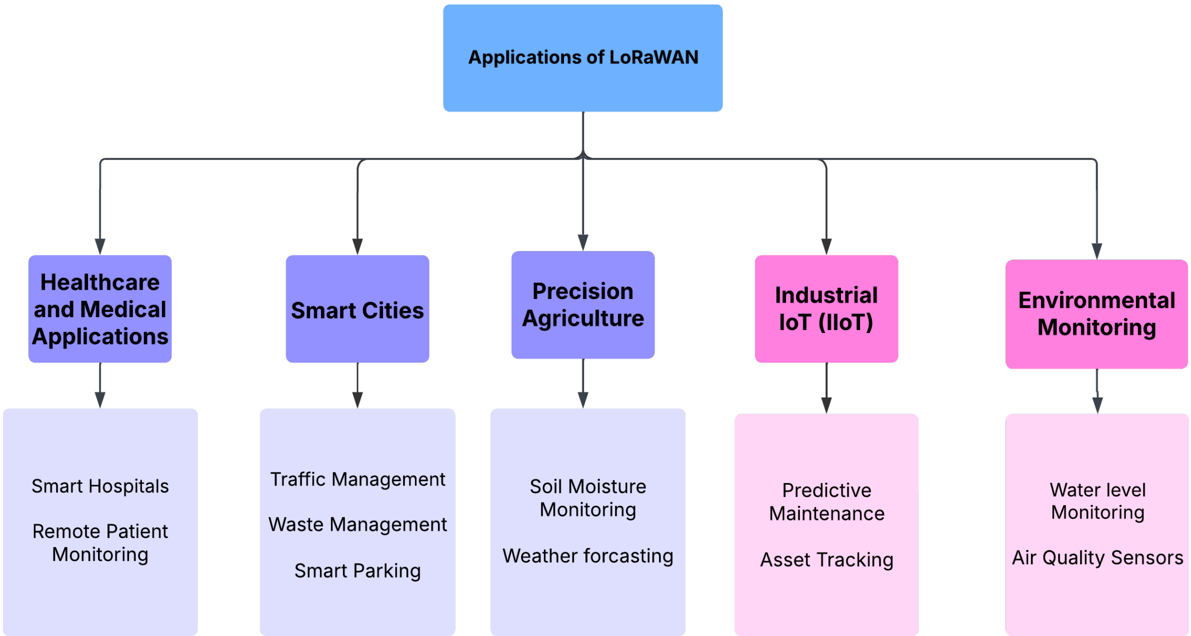


Figure 3.1: Applications of LoRaWAN in IoT

3.3.4 Applications of LoRaWAN

LoRaWAN technology has gained significant attention for its ability to enable long-range, low-power communication for various IoT applications. Some of the most prominent use cases are shown in Fig. 3.1.

In smart cities, LoRaWAN-enabled IoT sensors has been used to improve waste collection efficiency and predict energy output from municipal solid waste in (Elif et al., 2021) and LoRaWAN-based parking systems reduced congestion by guiding drivers to available parking spaces, lowering fuel use and emissions in (Erfaniangolabforoushanahlyazd, X. Zhang, and Arslan, 2021). In smart agriculture, LoRa has been examined in (Valente et al., 2022) to utilise LoRaWAN sensors in vineyards to monitor soil moisture, temperature, and weather conditions for precision agriculture and in (D. K. Singh et al., 2022) as an integrated part in smart farming focusing on soil condition analysis, irrigation management, and crop health monitoring.

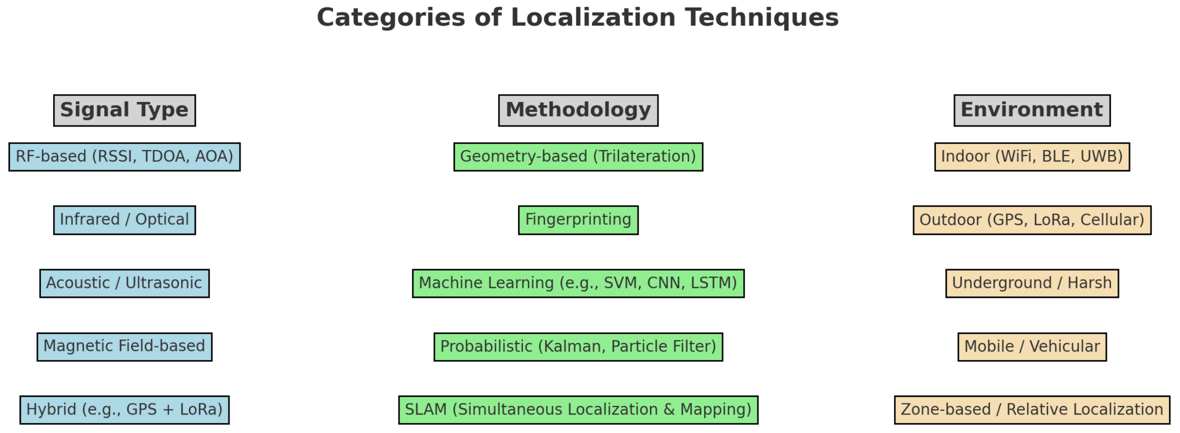


Figure 3.2: Different categories of localisation techniques

3.4 Existing Localisation Techniques

Localisation techniques form the backbone of positioning systems across diverse applications, ranging from smart cities and industrial IoT to autonomous systems and environmental sensing. This chapter explores the spectrum of existing localisation methods, categorizing them based on signal type, algorithmic approach, and deployment environment, to provide a comprehensive understanding of their capabilities and trade-offs.

The varying methods of categorising localisation techniques can be broadly organised into distinct categories based on signal type, algorithmic approach, and deployment environment, as illustrated in Figure 3.2. This section discusses the key methodologies most relevant to this work, with particular focus on RF-based and data-driven approaches applicable to LoRaWAN deployments.

3.4.1 RF-based Localisation

Radio Frequency (RF)-based localisation techniques are among the most widely adopted methods for estimating the position of devices in both indoor and outdoor environments. These techniques leverage the physical properties of radio signals—such as strength, time, phase, or angle—to infer location. They are particularly well-suited for IoT and LPWAN deployments, including LoRaWAN, Wi-Fi, BLE, and cellular networks.

3.4.1.1 RSS-based Localisation

Received Signal Strength (RSS) is widely used for LPWAN localisation due to its simplicity. However, environmental factors impact accuracy. Machine learning models, including neural networks and support vector machines (SVMs), have been applied to improve RSS-based positioning.

3.4.1.2 Time Difference of Arrival (TDOA)

TDOA relies on measuring the difference in arrival times of a signal at multiple synchronized receivers. It is particularly effective in outdoor LPWAN scenarios such as LoRaWAN geolocation, where gateways receive signals from a single transmitter.

3.4.1.3 Angle of Arrival (AOA)

AOA techniques estimate location by calculating the angle at which a signal arrives at a receiver equipped with multiple antennas or a phased array. These methods offer high precision but require more complex hardware and synchronization.

3.4.2 Fingerprinting Approaches

Fingerprinting localisation is a technique that estimates the position of a device by comparing its current signal characteristics—typically RSSI, CSI, or other sensor readings—with a pre-constructed database of signal measurements collected at known locations (called "fingerprints"). It is widely used in environments where traditional methods like GPS are unreliable, such as indoors, underground, or in dense urban areas. Fingerprinting offers high localisation accuracy without requiring complex physical models or additional infrastructure. However, it has limitations: building and maintaining the fingerprint database is labor-intensive and environment-sensitive, requiring periodic recalibration when signal conditions change.

3.4.3 Hybrid Localisation Approaches

Hybrid localisation techniques have emerged as a promising approach to overcome the limitations of standalone localisation methods. By combining multiple data sources—such as LoRa signal properties, GPS data, and RSSI readings—these methods provide improved accuracy, resilience, and adaptability in diverse and dynamic environments.

3.5 Channel Propagation Modeling

Propagation models are unique prediction models employed by network engineers to estimate the area being covered by the signal transmitter during networking and management [(Molisch, 2012), (Goldsmith, 2005)]. Signal propagation from the transmitter to the receiver varies from the line-of-sight paths to complex variations because of scattering, reflection, and diffraction. That makes the thorough characterisation of the signal path loss a crucial role in estimating interference analysis, link budget and coverage analysis (Kurt and Tavli, 2013).

To define the path loss the signal faces, a function of the distance between the transmitter and the receiver, along with a combination of environmental and design parameters that contribute to accurate modeling, are to be considered. The constant mobility of receivers, as in hand-held devices, makes them susceptible to obstructions, scattering of signals, and eventually path loss. The other parameters needed for path loss modeling include features representing the terrain and topography of the environment. Some of them are the width of the street, building to building distance, street orientation angle, and building to base station distance. The other design parameters are related to the system and include base station heights, antenna gain, feeder loss, receiver antenna gain, and cable loss (S. Ojo, Akkaya, and Sopuru, 2022).

In general, path loss is not deterministic but rather stochastic in nature. It can vary over time and can change rapidly depending on the frequency or scenario in use. Consequently, path-loss models generally incorporate statistical descriptors, such as mean values and standard deviations. Several propagation models have been developed and used for signal propagation in different frequency bands across different countries and territories. Yet, a model developed for a particular propagation environment can perform optimally in that environment but fails to correctly estimate path loss in another environment. This limitation is mainly due to dissimilarities and variations in environmental formations, weather conditions, soil electrical properties and terrain type (open, rural, suburban or urban) that exist in different radio propagation locations, cities and countries [(Ekpenyong, Isabona, and Ekong, 2010),(Imoize and Oseni, 2019)]. The aforementioned problem labels the deployment of path loss models with the optimal accuracy a complex and significant problem.

Accurate channel propagation modeling is essential for understanding and predicting the performance of wireless communication systems such as LoRaWAN and other LPWAN technologies. Since localisation heavily depends on the relationship between transmitted signals and their propagation characteristics, different modeling approaches have been developed in the literature. Broadly, these can be categorized into empirical/experimental modeling and machine learning–based modeling. Finally, model validation requires carefully chosen metrics to assess accuracy, robustness, and limitations.

Propagation models are broadly classified into empirical and deterministic models. Empirical models are based on experimental measurements within a particular environment. These models are developed based on measurements of base station heights, frequency, distance between transmitter and receiver [(Imoize and Oseni, 2019),(Park et al., 2012)].

Table 3.3: Comparison of deterministic and empirical models.

Aspect	**Deterministic Models**	**Empirical Models**
Basis	Physics-based, Maxwell's equations	Statistical fitting from measurements
Inputs	Site maps, material properties	Measurement datasets
Accuracy	High (site-specific)	Moderate, environment-dependent
Complexity	Computationally expensive	Lightweight, efficient
Use Cases	Network planning, simulations	Quick coverage estimation, large-scale studies
Limitations	Require detailed maps, slow	Lack generalisability, ignore fine effects

As shown in Table 3.3, empirical models offer a practical balance between accuracy and computational efficiency, making them well suited to large-scale LoRaWAN deployments where deterministic approaches would be prohibitively expensive. The following subsections detail the empirical models most relevant to this work, followed by an overview of deterministic approaches for completeness.

3.5.1 Empirical Models

Empirical models are simple and easy to develop but do not guarantee high level accuracy because in a bid to get them simplified many components are left out. The need to reduce complexity in empirical models in many instances has necessitated the need to use just few environmental parameters in modeling. Empirical models are thus a trade-off between accuracy and simplicity.

3.5.1.1 Log-Distance Path Loss Model

A widely adopted model is the log-distance path loss model, which expresses average path loss as a logarithmic function of distance:

$$PL(d) = PL(d_0) + 10n\log_{10}\left(\frac{d}{d_0}\right) + X_\sigma \tag{3.1}$$

where $PL(d_0)$ is the path loss at a reference distance d_0, n is the path loss exponent (environment-dependent), and X_σ is a Gaussian random variable representing log-normal shadowing with standard deviation σ.

Typical values of n range from 2 in free space to 3–5 in urban environments and 4–6 indoors.

3.5.1.2 Okumura-Hata Model

The Okumura-Hata model is an extension of empirical formulations for frequencies between 150–1500 MHz, making it relevant for LPWANs:

$$PL(d) = 69.55 + 26.16\log_{10}(f) - 13.82\log_{10}(h_t) - a(h_r) + [44.9 - 6.55\log_{10}(h_t)]\log_{10}(d) \quad (3.2)$$

where f is the frequency in MHz, h_t is the transmitter antenna height (m), h_r is the receiver height (m), and $a(h_r)$ is a correction factor for the mobile antenna height.

3.5.1.3 COST-231 Extension

The COST-231 Hata extension adapts the Hata model for frequencies up to 2000 MHz and includes correction terms for metropolitan environments.

3.5.2 Deterministic Models

Deterministic models attempt to describe radio propagation using mathematical formulations derived from Maxwell's equations and site-specific information such as terrain, building geometry, and material properties. These models are able to capture multipath effects, reflections, and diffraction with high accuracy, provided sufficient environmental data is available. They produce high level accuracy than the empirical models but introduce alternative complexities into the developed model.

3.5.2.1 Free-Space Path Loss (FSPL) Model

The most fundamental deterministic model is the free-space path loss model, which assumes ideal line-of-sight propagation without obstacles:

$$PL(d) = 20\log_{10}(d) + 20\log_{10}(f) - 147.55 \quad [\text{dB}] \tag{3.3}$$

where d is the transmitter–receiver separation in meters, and f is the carrier frequency in Hz. While analytically simple, the FSPL model neglects shadowing and multipath effects.

3.5.2.2 Ray-Tracing Models

Ray-tracing models simulate the propagation of electromagnetic waves by tracing rays as they interact with obstacles in the environment. The received signal is typically represented as a superposition of multiple paths:

$$E_r(t) = \sum_{i=1}^{N} a_i e^{j(\omega t + \phi_i)} \tag{3.4}$$

where a_i and ϕ_i represent the amplitude and phase of the i-th path, and N is the number of distinct multipath components. Ray-tracing can incorporate reflections, diffraction, and scattering, but requires detailed environmental maps and is computationally intensive.

3.5.2.3 Other Deterministic Approaches

Other notable deterministic approaches include the Uniform Theory of Diffraction (UTD), which extends ray-based models to account for edge diffraction, and full-wave numerical methods such as the Finite-Difference Time-Domain (FDTD) or Method of Moments (MoM), which directly solve Maxwell's equations. While highly accurate, these models are computationally prohibitive for large-scale IoT deployments.

A fundamental principle when understanding the channel properties of RF propagation is the usage of the Received Signal Strength Indicator (RSSI) for ranging. RSSI reflects the difference between the transmitted and received power of a signal that is used to estimate the distance propagated. Another vital parameter is the Path Loss (PL), PL is defined the power taken by the signal to overcome the communication medium and can be calculated using the link budget equation:

$$P_{tx}(dBm) = RSSI(dBm) + G_{system}(dB) - PL(dB). \tag{3.5}$$

The link budget equation shown in eq. (3.5) relates all gains and losses from the transmitter, through all transmission mediums, to the receiver. $RSSI$ is the expected power at the receiver, P_{tx} the transmitted power, G_{system} is the system gains such as those associated with directional antennas for both the transmitter and/or the receiver. PL are losses due to the propagation channel, either calculated via a wide range of channel models or from empirical data.

3.5.3 Machine learning based Approaches

Machine learning approaches to channel propagation modeling have emerged as an effective alternative to traditional empirical and deterministic methods. Instead of relying solely on analytical formulations or statistical curve fitting, these approaches learn propagation behavior directly from data. In this context, path loss prediction is formulated as a supervised regression problem, where input features such as distance, frequency, or environmental descriptors are mapped to continuous-valued outputs such as received power or path loss.

A variety of supervised learning techniques may be applied, including regression algorithms and neural networks. In particular, artificial neural networks and their extensions provide flexible architectures capable of approximating complex nonlinear relationships between inputs and outputs. By training on large, representative datasets, machine learning models are able to capture propagation effects that are difficult to express analytically, such as irregular multipath fading or non-stationary interference patterns.

Unlike empirical models, which are tied to specific measurement deployments, and deterministic models, which depend on detailed environmental maps and are computationally demanding, machine learning models offer the advantage of adaptability. Once trained, they can be deployed in new environments with only minor recalibration, overcoming one of the major shortcomings of traditional approaches. Their predictive accuracy has been consistently shown to exceed that of empirical baselines, especially when sufficient training data are available.

Machine learning–based path loss modeling typically follows a structured pipeline that begins with the acquisition of high-quality measurement data, such as received signal strength indicator (RSSI), signal-to-noise ratio (SNR), or channel impulse responses collected under controlled deployment scenarios. The next step involves preprocessing and feature extraction, where relevant input variables (e.g., transmitter–receiver distance, carrier frequency, antenna height, terrain classification, and environmental descriptors) are standardized and mapped into a suitable numerical representation. Once features are prepared, the dataset is partitioned into training, validation, and test subsets to enable robust model development. A supervised learning algorithm, such as a regression-based neural network or a recurrent architecture like LSTM, is then trained to learn the nonlinear mapping between the input features and the target variable, typically the path loss in decibels. During training, hyperparameters are optimized using cross-validation techniques to ensure generalisation, and regularization strategies are applied to mitigate overfitting. After training, the model is evaluated against the reserved test dataset using statistical performance metrics such as RMSE, MAE, and coefficient of determination (R^2). Finally, the validated model can be deployed to predict path loss in unseen environments, where its adaptability provides a key advantage over empirical and deterministic models.

3.5.4 Validation Metrics, Accuracy and limitations

The effectiveness of any path loss model, whether deterministic, empirical, or machine learning–based, depends on its ability to predict received signal power and propagation characteristics with sufficient accuracy. In practice, model performance is assessed through a combination of statistical error metrics, robustness measures, and application-specific localisation accuracy. This section presents the most commonly adopted validation metrics, followed by a discussion of accuracy considerations and inherent limitations across the three categories of models.

3.5.4.1 Mean Absolute Error (MAE)

MAE is one of the most intuitive error measures used in path loss modeling. It represents the average magnitude of prediction errors without considering their direction (i.e., whether the model overestimates or underestimates the true value).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.6}$$

where y_i is the measured value, $\hat{y}_i$ is the predicted value, and N is the number of data points. MAE is directly expressed in the same units as the predicted variable (dB for path loss), which makes it highly interpretable. Unlike squared-error metrics, MAE does not disproportionately penalize large deviations. This makes it robust to outliers. In machine learning regression, MAE serves as both a loss function during training and an evaluation metric, particularly when the focus is on minimizing average error.

3.5.4.2 Root Mean Squared Error (RMSE)

Another widely adopted metric in path loss modeling is the RMSE. It represents the square root of the average squared prediction errors:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.7}$$

where N is the total number of measurements, y_i is the i-th measured path loss value in dB, and $\hat{y}_i$ is the corresponding model-predicted path loss value in dB. RMSE penalises larger errors more heavily than MAE due to the squaring term, making it particularly sensitive to outliers and large prediction deviations.

3.5.4.3 Shadow Fading

Shadowing is attenuation of signal power, which is caused by obstacles between the transmitter and receiver through scattering, reflection, diffraction, and absorption. The receiver power variation due to path loss is observed over long distances, whereas variation due to shadowing occurs by the formation of obstructing object or the length of it. Traditionally, shadowing was treated as a coefficient with normal distribution since it is hard to generalise the characteristics of obstacles and, generally, its effects on path loss are trivial. However, in outdoor environment with various obstacles, shadowing effects play a significant role in analyzing path loss prediction with certain confidence level (Jo et al., 2020).

A standard error deviation between prediction and measured specimens is thought to be a great measure of shadow fading as it accounts for the variability in signal strength caused by obstacles and other environmental factors. Mathematically, the shadow fading is the standard deviation of the difference between the measured path loss and the expected path loss, such that;

$$SF = \sigma_{SF} = \sqrt{\frac{1}{K-1}\sum_{i=1}^{k}(P_m(d_i) - P_{exp}(d_i))^2} \tag{3.8}$$

where K is the total number of measurement points, d_i is the distance of the i-th measurement location from the transmitter, $P_m(d_i)$ is the measured path loss at distance d_i in dB, and $P_{exp}(d_i)$ is the model-predicted (expected) path loss at distance d_i in dB. A smaller value of σ_{SF} indicates that the path loss model accurately captures the signal attenuation behaviour with low variability, while a larger value reflects greater unpredictability in the propagation environment due to irregular obstacles and scattering effects.

3.5.4.4 Limitations

Despite their utility, all approaches have limitations that must be considered in model selection and application:

- Deterministic Models: - Require extensive input data (3D building maps, terrain profiles). - Computationally expensive, unsuitable for real-time LoRaWAN localisation. - Sensitive to inaccuracies in environmental data.
- Empirical Models: -Environment-specific; parameters such as the path loss exponent must be recalibrated for each deployment. -Fail to capture dynamic channel variations or time-dependent effects. -Limited in predicting fine-grained multipath behavior.
- Machine Learning Models: - Depend heavily on the availability and quality of large training datasets. - Risk of overfitting, especially when data is limited or biased. - Often lack interpretability, making it difficult to link predictions to physical principles. - Require careful feature selection and preprocessing to achieve robustness.

3.6 Overall Localisation Framework

The theoretical foundations established in the preceding sections covering LoRaWAN communication, channel propagation modelling, and existing localisation techniques; collectively motivate the design of the proposed framework. Rather than treating path-loss modelling and machine learning as independent components, this thesis integrates them into a unified, GPS-free localisation pipeline operating under a single-gateway constraint. Figure 3.3 presents the conceptual architecture of this framework, illustrating how four interconnected stages combine to enable sector-level localisation from sequential LoRaWAN signal measurements.

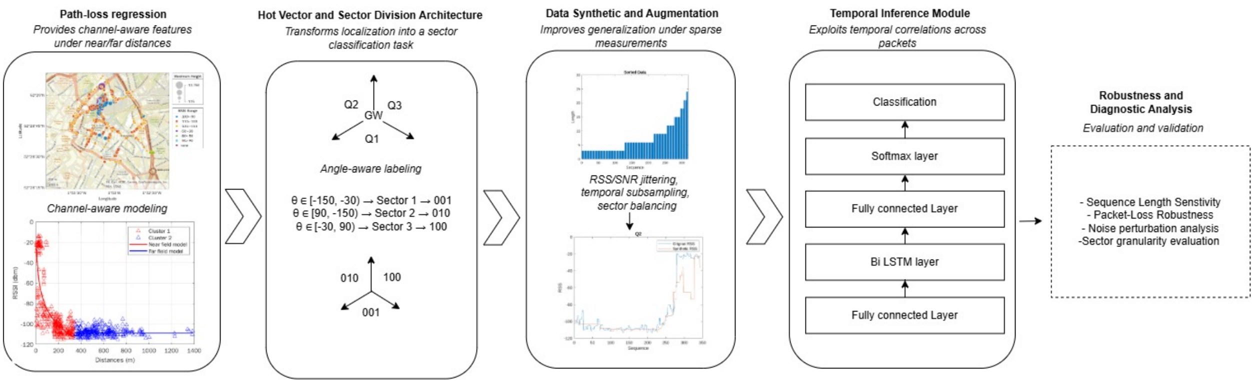


Figure 3.3: Conceptual architecture of the proposed channel-aware temporal localisation framework, illustrating how channel-aware modelling, geometry-based sectorisation, data enrichment, and temporal inference are combined for single-gateway LoRaWAN localisation. The figure highlights the diagnostic and robustness analyses used to assess sensitivity to temporal context, packet loss, noise perturbations, and sector granularity.

The framework operates across four sequential but interdependent stages, each of which is developed in detail in the subsequent chapters:

1. **Path-Loss Regression and Channel-Aware Modelling (Chapter 5):** The first stage characterises the radio propagation environment through empirical regression ap-

plied to real-world LoRaWAN measurements collected across Birmingham City Centre. A two-slope Birmingham-specific path-loss model is derived, segmenting the coverage area into near-field and far-field regimes. This model serves a dual role within the pipeline: it provides channel-aware distance features as inputs to the learning stage, and it generates physically consistent synthetic RSS samples to address spatial data imbalance across sectors. The path-loss model therefore serves two explicit functional roles within the machine learning pipeline: first, **training data quality**; generating physically consistent synthetic RSS samples for underrepresented spatial sectors, addressing class imbalance in a manner grounded in propagation physics rather than arbitrary statistical oversampling; and second, **feature enrichment**; providing model-derived distance estimates and near/far-field regression parameters as additional input features to the LSTM, enriching the temporal signal representation beyond raw RSSI and SNR values alone.

2. **Sector Division and Problem Formulation (Chapter 6):** The second stage transforms the localisation problem into a tractable classification task. The coverage area surrounding the single gateway is partitioned into three angular sectors using angle-aware labelling. Each sector is represented as a one-hot encoded output vector, enabling multi-class classification from sequential signal observations. The three-sector configuration was selected following comparative evaluation against a four-sector alternative, which exhibited overfitting under the available dataset size.

3. **Data Synthesis and Augmentation (Chapter 6):** The third stage enriches the training dataset to improve generalisation under sparse and imbalanced real-world measurements. Two complementary techniques are applied: model-based synthesis, in which the Birmingham path-loss model generates additional RSS samples for under-represented spatial regions, and noise-based augmentation, in which Gaussian jitter is applied to existing RSSI and SNR measurements to simulate realistic signal variability.

Together these techniques balance the sector-wise data distribution while preserving the physical plausibility of the training samples.

4. **Temporal Inference Module (Chapter 6):** The fourth and central stage performs sector-level localisation through a sequential deep learning model. Input features; comprising time-indexed RSSI, SNR, and model-derived distance estimates are structured into fixed-length temporal windows and passed through the following layers:

    - **Sequence Input Layer:** Restructures and normalises the sequential data to fit as input for the following layers.
    - **Fully-Connected Layer:** Links all inputs from the previous layer to every activation unit in the next layer. The proposed model includes two fully-connected layers. The first contains hidden neurons designed to find the correlation between input features and the defined classes. The second functions as a classification layer that matches the dataset classes to its hidden neurons.
    - **ReLU and Dropout Layer:** Every fully-connected layer is equipped with a Rectified Linear Unit (ReLU) activation function and a dropout layer to reduce overfitting by randomly dropping neurons during training, reducing sensitivity to specific individual weights.
    - **Bidirectional LSTM (BiLSTM) Schuster and Paliwal, 1997:** The core recurrent block captures both forward and backward temporal dependencies across the packet sequence. By learning from the full sequence at each time step, the BiLSTM exploits signal evolution patterns that are invisible to static classifiers. A dropout layer follows to further prevent overfitting.
    - **Softmax Classification Layer:** The output layer produces a probability distribution over the three sector classes, with the highest probability taken as the predicted sector of the transmitting node.

The critical conceptual link across all four stages is that the physical path-loss model is not merely a standalone propagation analysis — it is an active contributor to the machine learning pipeline. By informing both the synthesis of training data and the construction of channel-aware input features, the empirical model directly enhances the discriminative power of the LSTM, enabling localisation performance that would not be achievable from raw signal measurements alone. The robustness and diagnostic analyses shown on the right of Figure 3.3 — covering sequence length sensitivity, packet-loss robustness, noise perturbation analysis, and sector granularity evaluation — are presented in full in Sections 6.5 and 6.6.

3.7 Problem Formulation

The localisation problem addressed in this thesis can be formally stated as follows. The gateway is assumed to be fixed at a known location $\mathbf{g} = (0, 0)$, while each transmitting node occupies an unknown spatial position $\mathbf{x}_i = (x_i, y_i)$. The signal sequence observed at the gateway for node i over a temporal window of length T is given by:

$$\mathbf{s}_i = \{(\mathrm{RSSI}_t, \mathrm{SNR}_t)\}_{t=1}^{T} \tag{3.9}$$

The objective is to learn a parameterised mapping:

$$f_\theta : \mathbf{s}_i \to z_i \tag{3.10}$$

where $z_i \in \{1, 2, 3\}$ is a discrete spatial sector label derived from the azimuth angle of $\mathbf{x}_i$ relative to $\mathbf{g}$. This formulation explicitly frames localisation as a **coarse classification problem** rather than coordinate regression, consistent with the single-gateway constraint under which fine-grained ranging is inherently ambiguous.

Physical interpretation as an inverse mapping. From a physical perspective, the received signal is generated by:

$$\mathbf{s}_i = h(\mathbf{x}_i, \mathbf{g}, \mathcal{E}) + \epsilon \tag{3.11}$$

where $h(\cdot)$ represents the propagation function capturing distance-dependent path loss and shadowing effects, $\mathcal{E}$ denotes the deployment environment (urban morphology, building geometry, multipath conditions), and ϵ denotes additive measurement noise. The model f_θ therefore learns an **inverse mapping**:

$$f_\theta(h(\mathbf{x}_i, \mathbf{g}, \mathcal{E})) \approx z_i \tag{3.12}$$

mapping from signal space back to discretised spatial regions. This inverse problem is ill-posed in general — multiple spatial positions may produce similar signal signatures — but becomes tractable under the sector-level classification formulation adopted here, where the target is a coarse angular region rather than a precise coordinate.

Implications for gateway relocation. A key implication of Equation 3.11 is that the learned mapping f_θ is conditioned on the fixed gateway location $\mathbf{g}$. If the gateway is relocated to $\mathbf{g}'$, the signal distribution becomes:

$$\mathbf{s}'_i = h(\mathbf{x}_i, \mathbf{g}', \mathcal{E}) + \epsilon \tag{3.13}$$

Since $h(\mathbf{x}_i, \mathbf{g}, \mathcal{E}) \neq h(\mathbf{x}_i, \mathbf{g}', \mathcal{E})$ in general, the learned model f_θ may no longer produce valid sector predictions without retraining on data collected from the new gateway position. This conditioning on gateway location is an explicit and acknowledged constraint of the proposed framework, consistent with the environment-specific philosophy discussed in Sections 5.6 and 7.3.

Implications for synthetic data. The synthetic RSS samples used for data augmentation in Chapter 6 are generated using an estimated path-loss model $\hat{h}(\mathbf{x}_i, \mathbf{g})$, which approximates the true propagation function $h(\mathbf{x}_i, \mathbf{g}, \mathcal{E})$. This introduces a potential **self-consistency bias**: if the path-loss model systematically misrepresents the true signal distribution, the synthetic samples will reflect these same errors, and the learned mapping f_θ may be partially conditioned on model assumptions rather than purely on real-world propagation behaviour. This risk is mitigated through three mechanisms: the use of the corrected two-slope hybrid model rather than the biased single-slope regression (Section 5.7); KDE divergence validation confirming statistical consistency between real and synthetic distributions (Table 6.1, values below 1.71×10^{-4}); and ablation evaluation on the original dataset without synthetic samples, confirming that augmentation does not artificially inflate test performance (71.6% vs 72.1%, Section 6.5). These mitigations are explicitly acknowledged as addressing but not fully eliminating the self-consistency bias, which remains a limitation of the framework as noted in Section 7.3.

3.8 Summary

This chapter has established the theoretical background and methodological framework required for the development and validation of path loss models in LoRaWAN localisation scenarios. It began with a thorough overview of LoRaWAN fundamentals, including the architectural elements, modulation scheme, and deployment characteristics that make LoRaWAN particularly suitable for low-power wide-area networks. This was followed by a discussion of LPWAN performance metrics, which outlined key parameters such as coverage, scalability, energy efficiency, data rate, and reliability—metrics that are critical when assessing the suitability of LPWAN technologies for IoT deployments.

The chapter then addressed the core topic of channel propagation modeling, highlighting the distinctions between deterministic and empirical models. Deterministic models were shown to provide site-specific accuracy at the cost of significant computational requirements. In contrast, empirical models such as the log-distance and Okumura-Hata formulations offer simplicity and scalability but require calibration and lack generalisability across diverse environments.

Building upon the limitations of these traditional approaches, machine learning–based modeling was introduced as a data-driven alternative. By leveraging supervised regression techniques and advanced architectures such as Long Short-Term Memory (LSTM) networks, ML-based models can capture nonlinear propagation behaviors and temporal dynamics that are not easily represented in analytical models. The adaptability of ML models across different deployment environments was highlighted as a key advantage, although their dependence on large, high-quality datasets and potential risk of overfitting were noted as challenges.

Finally, the chapter established the validation framework used in this research. Statistical error metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), were defined for evaluating predictive accuracy.

In summary, this chapter has delivered a comprehensive foundation that integrates LoRaWAN fundamentals, LPWAN performance metrics, and channel modeling techniques. By clearly defining both the theoretical principles and methodological tools, it has laid the groundwork for the subsequent experimental validation and the proposed application of LSTM-based models for LoRaWAN localisation.

Chapter 4

Experimental Deployment of LoRaWAN network

4.1 Introduction

The preceding chapters established the theoretical background and methodological foundations for this research, with a particular emphasis on channel propagation modeling and machine learning approaches for localisation. In this chapter, we build upon the limitations identified in simulation studies and establish the need for conducting real-world in an urban smart city environment with LoRaWAN. Preliminary simulations provided valuable insights into packet transmission and reception under idealized assumptions. However, they could not capture the variability and complexity of real deployment environments. Simulated systems often fail to accurately emulate the effects of multipath fading, interference, and environmental obstacles, resulting in an incomplete representation of LoRaWAN behavior in practice. These limitations motivated the development of a dedicated field experiment designed to validate the proposed models and provide a reliable dataset for further analysis.

The experimental deployment was carried out in the area of Birmingham City University, with the aim of capturing real-world measurements of LoRa packet transmissions. The data collection setup was designed as a critical component of the research methodology, ensuring that data was gathered in a structured, reproducible, and reliable manner. Accordingly, this chapter describes the geographical location of the deployment, the hardware infrastructure of an Arduino-based LoRa node (The Things Uno) and a MultiConnect Conduit IP67 gateway. the network configuration parameters, and the environmental conditions under which the experiments were performed.

The chapter begins by revisiting the simulation framework that established the theoretical baseline for LoRa packet transmission and reception, and highlights the deficiencies of these models in representing real-world reception characteristics. This is followed by a detailed description of the node–gateway deployment, data collection procedures, and preprocessing steps such as packet filtering and normalization. To ensure validity, an experimental reliability analysis is also presented, focusing on the variation of RSSI and SNR readings across different conditions and the measures taken to minimize their impact.

Unlike prior measurement campaigns that focus on multi-gateway deployments or suburban environments, this work emphasizes the constrained single-gateway case in a dense urban setting. This not only reflects realistic cost-sensitive IoT rollouts but also serves as a rigorous baseline for evaluating machine learning–based localisation models.

Ultimately, the results of this experimental deployment not only support the objectives of this research but also provide datasets and methodological insights that may be functionally applied by researchers, professionals, and network operators to improve simulation accuracy and enhance the planning and optimization of LoRaWAN networks.

4.2 Preliminary Simulation

When analyzing LoRa-based systems, it is important to distinguish between LoRa and LoRaWAN. LoRa refers to the physical layer modulation scheme that enables long-range, low-power communication, while LoRaWAN defines the higher-layer communication protocol, mapping to the data link and network layers of the OSI model (Sornin et al., 2015). LoRa modulation employs a proprietary chirp spread spectrum (CSS) technique, which transforms a narrowband input signal into a wider channel bandwidth. By continuously varying the carrier frequency in the form of chirps, LoRa enables receivers to decode signals that are several decibels below the noise floor, thereby achieving extended communication ranges. A key feature of LoRa modulation is its support for multiple spreading factors (SFs), which provide a trade-off between transmission range and data rate. Higher SF values (e.g., SF11 and SF12) produce longer chirps that increase signal robustness in noisy or attenuated environments, but at the expense of lower data rates. Conversely, lower SF values (e.g., SF7 and SF8) enable faster data transmission with higher throughput but are less resilient under poor signal conditions. A visual representation of the chirp signals at different SFs is provided in Figure 4.2, illustrating the increased duration and lower effective rate of higher SF configurations.

The simulation studies in this research were conducted using MATLAB, a widely adopted platform for numerical computation and signal processing. MATLAB provides a flexible environment for modeling communication systems, including matrix-based simulation, advanced visualization, and the integration of customizable toolboxes. A system-level simulation model was developed (Figure 4.1) to capture the three principal components of a communication link: the transmitter, the channel, and the receiver. At the transmitter, chirp signals were generated by linearly sweeping the frequency of the carrier over a specified bandwidth. Information was encoded by modulating the chirps, with each symbol mapped to a unique chirp pattern. The properties of the generated signal were determined by the chosen SF, carrier frequency (f_c), and bandwidth (BW). During transmission, the channel was modeled to include Additive White Gaussian Noise (AWGN), which represents the simplest and most widely used noise model. Configurable signal-to-noise ratio (SNR) levels were applied to assess the performance under varying interference and attenuation conditions. At the receiver, demodulation was performed by correlating the incoming chirps with reference chirps used during modulation, allowing the retrieval of encoded data despite the presence of noise.

The simulation results, expressed in terms of bit error rate (BER) versus SNR for spreading factors ranging from 7 to 12 (Figure 4.3), demonstrate the robustness of higher SFs under adverse channel conditions. For low SNR values (–35 dB to –25 dB), SF11 and SF12 maintained relatively low BERs, confirming their suitability for long-range communication in noisy environments. By contrast, SF7 and SF8 exhibited higher BERs under the same conditions, limiting their applicability in weak signal environments. At higher SNR levels (–15 dB to –5 dB), all spreading factors performed reliably, with BER values decreasing significantly, though the trade-off in transmission time remained evident.

These simulations provided valuable insights into the theoretical performance of LoRa modulation. They enabled the identification of expected signal strength ranges and the coverage distances achievable under idealized channel assumptions. However, while simulations are effective in modeling controlled conditions, they cannot fully replicate the complexity of real-world environments, including multipath fading, urban obstacles, and interference from other networks. Therefore, simulations alone are insufficient to validate the applicability of LoRaWAN for practical deployment. This limitation established the need for the field experiments described later in this chapter, where real-world data collection and reliability analysis are employed to complement and refine the findings from simulation.

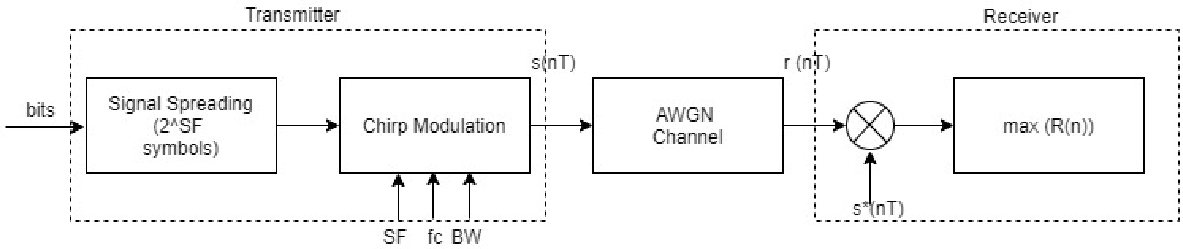


Figure 4.1: LoRaWAN communication system model

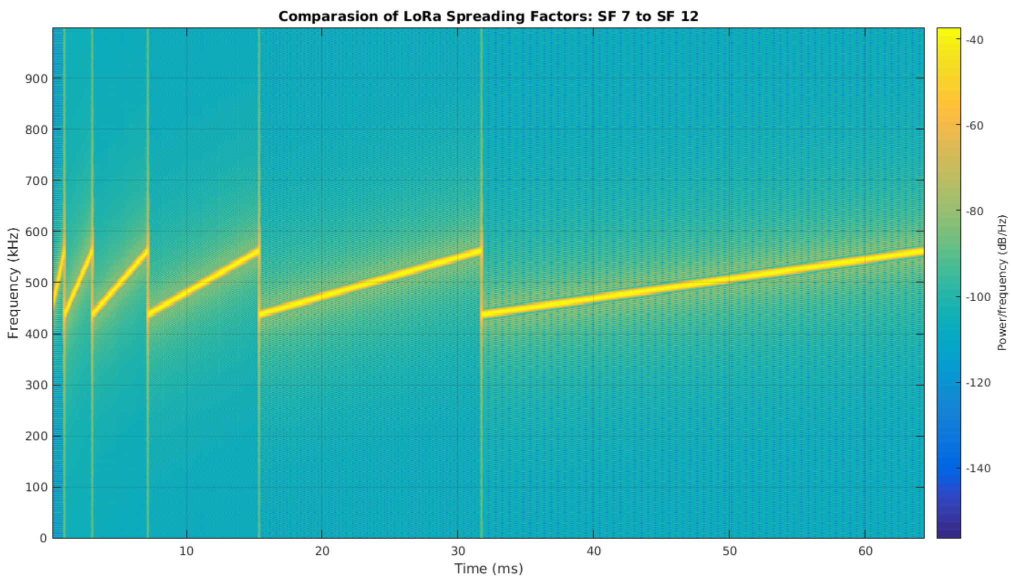


Figure 4.2: LoRa Spreading factors 7 till 12

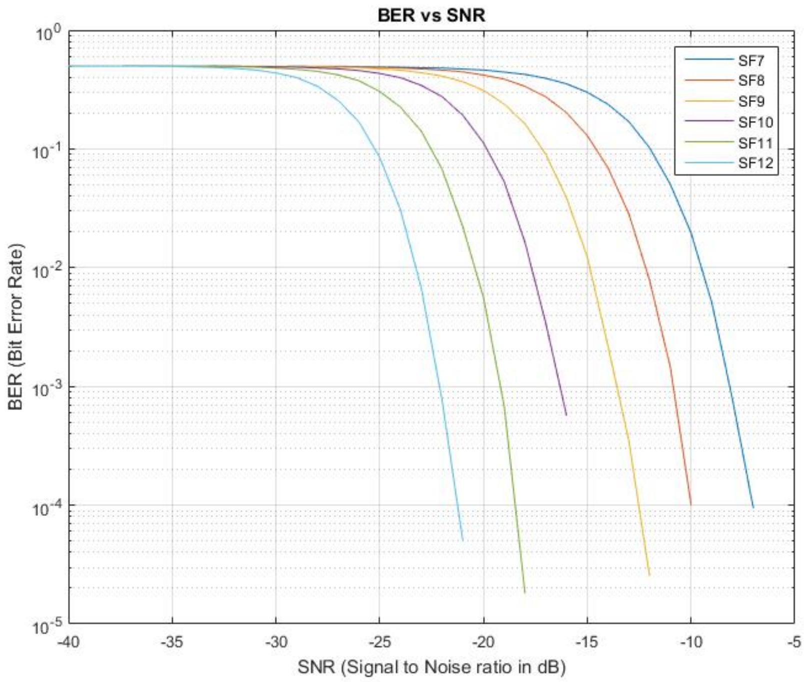


Figure 4.3: BER of SF between 7 to 12

4.3 Field Experiment: Birmingham City Center

After acknowledging the shortcomings of the LoRa simulation results in replicating real-world packet propagation, and considering the availability of off-the-shelf LoRaWAN modules, we proceeded to deploy our own experimental network in Birmingham.

Field experiments were carried out in Birmingham City Center, a densely populated urban area characterised by a mixture of high-rise buildings, open spaces, and various sources of wireless interference. The choice of this location was motivated by the fact that it represents a realistic urban environment with the presence of buildings, traffic, and dynamic environmental conditions. Signal propagation in an urban setting is affected by reflections, shadowing, and absorption, making it a suitable test environment for evaluating path loss models and localisation accuracy. Additionally, Birmingham is a leading smart city hub, making it relevant for assessing the practical applicability of LoRaWAN in smart city appli-

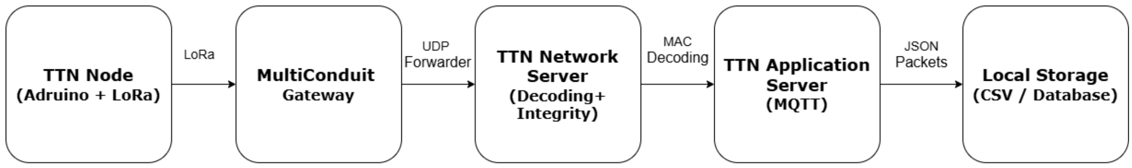


Figure 4.4: End-to-end LoRaWAN data flow through The Things Network, showing uplink from the end node to gateway, packet forwarding to the network server, and application-level data access.

cations, industrial IoT, and environmental monitoring.

The LoRaWAN gateway was installed at an elevated location on the campus to maximize coverage. Gateway placement was carefully chosen to ensure visibility across open zones while still allowing measurement of attenuation and fading effects in obstructed areas. The height and orientation of the gateway antenna were documented to ensure reproducibility and provide input for propagation modeling.

4.4 Experimental Settings

This section outlines the practical implementation of the experimental campaign, covering the deployment of the LoRaWAN gateway, the integration of the network with The Things Network (TTN), and the configuration of sensor nodes. The focus is on ensuring methodological clarity, reproducibility, and transparency of the end-to-end data collection pipeline. Figure 4.4 illustrates the end-to-end data flow of the deployed LoRaWAN system through The Things Network (TTN), showing the uplink transmission path from the end node through the gateway, packet forwarding via the UDP protocol to the TTN network server for decoding and integrity checking, and final application-level data access via the MQTT broker for local storage.

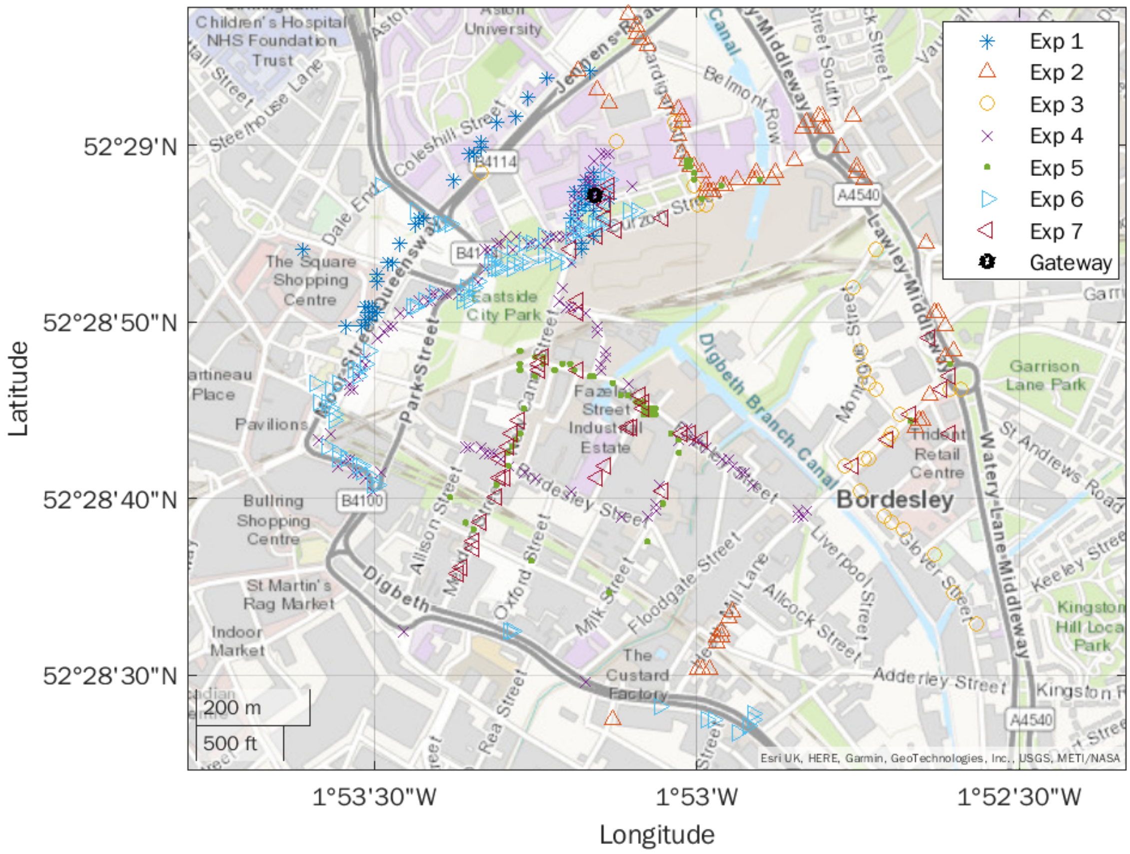


Figure 4.5: Map of Birmingham City Centre showing the gateway installation site and georeferenced data collection points across seven measurement campaigns.

4.4.1 Gateway Installation and Deployment

The MultiConnect Conduit IP67 gateway served as the central node for receiving LoRa transmissions from the mobile end device. It was installed at a high-altitude indoor location on the Birmingham City University campus to maximize line-of-sight (LOS) coverage, ensure secure mounting, and provide reliable power and internet connectivity. BCU's central location in Birmingham City Center enabled the gateway to capture data across diverse urban environments. Figure 4.5 shows the gateway position relative to the routes of seven measurement campaigns. Hardware characteristics of the gateway and the TTN node are summarized in Table 4.1.

Table 4.1: Characteristics of MultiTech Gateway and The Things Uno node

Characteristic	Gateway	Node
Module	MTCDIP-LEU1-267A-868	The Things Uno
LoRa Chip		Microchip RN2483 LoRa
Operating Frequency	867.9	867.9
Modulation	LoRa	LoRa
SF	12	12
Coding Rate	4/5	4/5
Rx Input Sensitivity (dB)	-	-148
Tx Power	14dBm	-
Rx current consumption	-	14.2 mA

4.4.2 Network Integration and Monitoring

The deployed LoRaWAN network was integrated with The Things Network (TTN), which served as the backend for packet forwarding, MAC-layer decoding, and data access. The gateway was registered on the TTN Console using its unique EUI and configured as a packet forwarder via the Semtech UDP protocol to the TTN European server (eu1.cloud.thethings.network). Once active, it relayed uplink transmissions containing both payloads and metadata (RSSI, SNR, spreading factor, frequency, timestamp).

At the TTN backend, the Network Server verified message integrity, performed MAC-layer decryption, and passed decoded packets to the Application Server. For this work, MQTT integration was used to stream JSON packets in real time, which were parsed and stored locally in structured CSV files. This ensured that the entire pipeline—from node transmission to local storage—was transparent and reproducible.

Other platforms were briefly considered: MultiTech DeviceHQ was tested but discarded due to its enterprise focus, while Cayenne IoT was used only for real-time visualization during experiments (e.g., live RSSI monitoring). All subsequent analysis relied exclusively on the structured MQTT data stream.

4.4.3 Sensor Node

A single TTN LoRaWAN end device, "The Things Uno," (shown in Fig. 4.9a) was deployed across Birmingham City Center. The node is based on the Arduino Leonardo with an integrated RN2483 LoRaWAN module (Class A) (Technology, 2015-2020), operating at 868 MHz with a sensitivity of –148 dBm. To provide location context, the node was augmented with an Adafruit GPS module (MTK3339 chipset), capable of tracking 22 satellites at a sensitivity of –165 dBm.

The node periodically transmitted packets containing GPS coordinates, RSSI, and SNR to the TTN gateway. While additional sensors (e.g., temperature, humidity) could be supported, they were excluded to reduce overhead and focus on the radio link performance. Furthermore, as validated in Fig 4.11, the embedded antenna exhibited near-isotropic behaviour, ensuring that orientation effects did not bias the collected data. This configuration provided the core dataset used for subsequent path-loss modeling and localisation analysis.

4.5 Measurement Setup

The measurement campaign was conducted in Birmingham city centre, a dense urban environment characterised by high-rise buildings, traffic, and multipath interference. A fixed LoRaWAN gateway was installed at Birmingham City University, while the end device ($Leo1$) was battery-powered and operated in mobile conditions (car, bus, and on foot) to ensure broad spatial coverage. The node transmitted packets periodically, embedding GPS coordi-

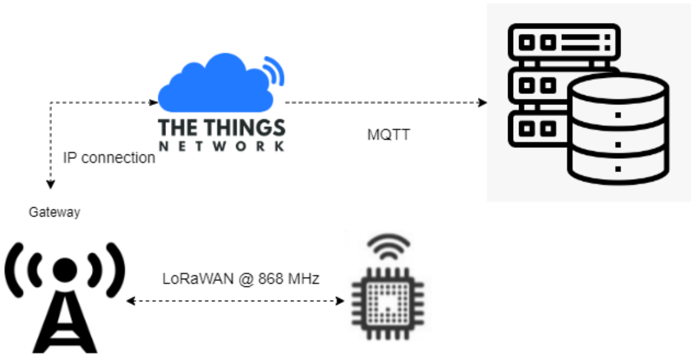


Figure 4.6: System model with LoRa end node, Gateway, NS, and data-base (DB) server. The End nodes send the information to the GW via LoRaWAN, and it retransmits the information to the NS. The NS exposes an MQTT broker that is used to retrieve and store the data on a DB server.

nates for georeferenced analysis.

The overall experimental architecture is illustrated in Fig. 4.6, showing the data flow from the mobile end node to the gateway, network server, and local storage. The end device (Arduino + LoRa transceiver) sent uplink packets using chirp spread spectrum modulation at 868 MHz. These were captured by the MultiConnect Conduit gateway, which forwarded raw payloads and metadata (RSSI, SNR, spreading factor, timestamp) to The Things Network (TTN) via the Semtech UDP packet forwarder. At the TTN backend, packets were verified, decrypted, and streamed via the MQTT interface. The data were parsed into structured CSV/SQL tables and stored locally, ensuring reproducibility and compatibility with subsequent preprocessing and modelling.

Example screenshots of the raw Arduino output (Fig. 4.7) and TTN console traffic (Fig. 4.8) are provided to demonstrate the raw data formats at both the device and network server levels, highlighting the reproducibility of the experimental setup. Hardware details of the gateway and node are summarised in Table 4.1, while photographs of the deployed devices are shown in Fig. 4.9. Together, this setup ensured that the dataset was both reliable and representative of real-world LoRaWAN performance in a complex urban environment.

4.6 Data Collection and Pre-processing

In this research, data collection was carried out to analyze the impact of environmental parameters on LoRaWAN network reception. The study required real-world signal data obtained from an experimental LoRaWAN testbed, ensuring accurate representation of real-world conditions. Ensuring data integrity and reliability was critical before proceeding to path-loss modeling. The preprocessing pipeline combined hardware validation, data cleaning, statistical filtering, and reliability checks, as outlined below.

4.6.1 Packet Acquisition and Logging

Packets were transmitted periodically from the Things Uno node using Over-the-Air Activation (OTAA). Payloads were encoded in the Cayenne Low Power Payload (LPP) format and carried metadata (RSSI, SNR, GPS coordinates). At the backend, packets were streamed from TTN's MQTT broker in JSON format and parsed into structured CSV tables. Example raw outputs are shown in Fig. 4.7 and Fig. 4.8, confirming correct integration of GPS and LoRaWAN metadata.

4.6.2 Arduino: Coding the TTN node

The Things Uno node was programmed to transmit GPS-tagged packets over LoRaWAN using Over-the-Air Activation (OTAA). The payloads were encoded in Cayenne Low Power Payload (LPP) format to ensure compact and interoperable representation. A lightweight Arduino sketch, integrating LoRaWAN and GPS libraries, handled device registration, periodic transmission, and geolocation parsing. This allowed every received packet to carry metadata (RSSI, SNR) tied to precise location coordinates, ensuring georeferenced datasets for subsequent analysis.

For validation, the Arduino serial monitor was used to log real-time transmissions, including GPS coordinates, altitude, satellite count, and confirmation of successful LoRaWAN uplink (Fig. 4.10). This confirmed correct integration of the GPS module with the LoRaWAN stack and ensured that each packet carried both signal metadata and geolocation information.

4.6.3 Hardware Validation: Antenna Pattern

Prior to data collection, the embedded antenna of the Things Uno node was tested to assess whether orientation introduced measurable bias. The node was rotated over a fixed period while transmitting empty packets. RSSI values were logged and averaged in MATLAB to generate a polar pattern (Fig. 4.11). Results indicated near-isotropic behaviour, with less than 0.02% deviation from the mean, far below the natural fluctuation of RSSI in outdoor environments. Consequently, antenna orientation was not considered a confounding factor in later experiments.

4.6.4 Data Cleaning

All raw packets were filtered to remove duplicates, incomplete entries, and corrupted payloads. A haversine distance calculation was used to validate GPS coordinates, ensuring consistency with the recorded gateway location. This ensured that only georeferenced and valid packets were retained.

4.6.5 Data Outlier Detection and Elimination

Detecting and eliminating outliers is a crucial step in preparing experimental data, as outliers can significantly distort statistical measures and compromise the validity of subsequent modeling. Outliers may arise from sensor inaccuracies, transmission errors, or atypical environmental conditions. If not addressed, these values can bias regression coefficients and inflate error metrics.

A common approach is to identify outliers as points lying outside the interval defined by the mean ± three standard deviations. However, because both the mean and the standard deviation are themselves highly sensitive to extreme values, this method often proves unreliable. As an alternative, Leys et al. (2013) recommend the median absolute deviation (MAD), which is considerably more robust to abnormal values. The MAD is defined as:

$$MAD = b \cdot M_i \left(|x_i - M_j(x_j)|\right) \tag{4.1}$$

where x_j represents the N original observations and M_i denotes the sample median. The constant $b = 1.4826$ scales the MAD under the assumption of normality, while remaining insensitive to outlier influence (Donoho and Huber, 1983).

To determine whether a data point should be flagged as an outlier, the following decision criterion was applied (Miller, 1991):

$$\frac{|x_i - M|}{MAD} > v \tag{4.2}$$

where M is the median and v is a threshold parameter, commonly chosen as 2, 2.5, or 3. In this study, $v = 3$ was selected to ensure conservative rejection.

Figure 4.12 illustrates the application of Equation 4.2 to RSSI measurements, with six points exceeding the upper MAD threshold and flagged as outliers. Table 4.2 lists these outliers, showing their absolute deviation and corresponding distances from the gateway.

Table 4.2: Identified outliers and their deviations from the 3*MAD threshold

Outlier	Absolute Deviation from 3*MAD	Distance (m)
1	10.6566	99.36
2	5.6566	68.03
3	3.6566	139.4
4	3.6566	65.35
5	2.6566	156.02
6	1.6566	74.87

An iterative elimination process was then conducted, where flagged outliers were sequentially removed and the regression model recalculated at each step. The impact of outlier removal was evaluated using statistical metrics including mean squared error (MSE), standard error (σ_e), adjusted R^2, and the sum of squared errors (SSE).

Table 4.3 summarizes the results. The initial model (Iteration 1, without outlier elimination) produced an MSE of 28.05, already outperforming the Oulu model by approximately 20%. With successive eliminations, MSE decreased further to 24.91, while SSE also dropped significantly, indicating improved model fit. Adjusted R^2 initially increased but declined slightly in later iterations, reflecting the trade-off between reduced error variance and the diminishing explanatory power as data points were removed.

These results confirm that conservative outlier elimination improves model accuracy and robustness by reducing noise-induced variance, while preserving the essential character-

Table 4.3: Statistical performance of regression model across outlier elimination iterations

Iteration	Data Tradeoff	MSE	σ_e	Adjusted R^2	SSE
1	15	28.05	7.83	0.447	1.498e+04
2	30	27.86	7.85	0.402	1.431e+04
3	57	27.20	7.87	0.330	1.207e+04
4	78	26.85	7.95	0.271	1.019e+04
5	87	24.91	7.61	0.264	8.07e+03

istics of the dataset. This refined dataset is subsequently used in the experimental repetition and reliability analysis described in Section 4.6.6.

4.6.6 Experimental Reliability Analysis

The repetition principle is an important step in scientific and experimental research, as the observations and readings are affected by natural variables that can be stationary or random, which require a fair amount of samples to reveal their changing regularity. Stabilizing the mean and the standard deviation are the foundation of repetition, resulting in a population fairly represented by the sample (Hu, Bao, and Q. Wang, 2011).

To achieve the aforementioned, we maintained two methods, the first approach was collecting a series of measurements for each location during the same experiment, then averaging the values for a reliable estimate. The second approach was collecting the measurements from similar locations but at different times. To achieve that, the data collected from 7 successful experiments throughout a time span of 15 months has been separately analysed as shown in fig. 4.5, where each experiment's data points retrieved is presented in a different color. Even though each experiment was concerned with its own path, there was enough points overlapping. The analysis included an algorithm that was used to go through all data-location pair, calculate the haversine distance, which is the distance from the GPS co-ordinates to the gateway location. Then, cross-reference the resulting distance with the GPS co-ordinates to find close by locations, since the distance on its own can represent any point on the circumference of a circle centered from the gateway. The algorithm identified the close-by locations using a threshold of approximately 15 m. This threshold was chosen after considering the accuracy of the GPS chip used in the experimentation and the GIS system that is 3 m and 11 m, respectively.

Fig. 4.13 shows the first output of the algorithm, as every cluster of points is the output of overlapping close-by locations from at least 2 different experiments. Additionally, to maintain an even analysis, the algorithm was assembled to output clusters in different directions and distances of the spanned area. The results of this analysis can be found in table 4.4, where the average distance of each cluster from the gateway, the average and variance RSSI from various experiment are computed to be compared to the actual measurements found using the B'ham model when applied to the same average distance. The results show that the variance of measurements from different experiments around the same location are quite similar which conveys the validity of the assumption that the RSSI for every location is stationary. Moreover, when comparing the average RSSI to the RSSI yielded when applying the B'ham model, a trend of overestimating the true value of RSSI in close-by distances to the gateway, i.e. distances less than or equal to 300 m, can be found. In the same way, the model underestimates the power received in longer distances.

Table 4.4: Analysis from experimental repetition

	Area 1	Area 2	Area 3	Area 4	Area 5	Area 6
Mean Distance (m)	309.5	517.8	176.84	290.53	450.62	646.75
Mean RSSI	-109.5	-108.5	-101.8	-108.37	-105.66	-108.67
Variance RSSI	0.25	1.66	31.77	0.23	0.333	0.333
Bham Model RSSI estimate	-104	-109.6	-97.89	-103.32	-108.13	-112
K-mean Clustering Model	-110.96	-108.99	-98.38	-109.55	-108.99	-108.99
Distribution-based Clustering Model	-105.98	-108.23	-103.5	-105.71	-107.62	-109.2

4.7 Limitations

While the experimental design was carefully planned, certain limitations must be acknowledged:

- **Mode of Transportation:** Data were primarily collected on foot, with limited use of vehicular routes. As a result, some mobility dynamics (e.g., high-speed vehicular fading) were not fully captured.
- **Urban-Specific Environment:** Measurements were confined to Birmingham city centre. While dense urban features make it a challenging testbed, results may not fully generalise to suburban or rural environments.
- **Gateway Placement:** The LoRaWAN gateway was installed indoors due to univer-

sity regulations. Although the elevated position ensured reasonable coverage, outdoor mounting at higher elevation would have provided improved visibility, reduced shadowing effects, and potentially extended the reliable communication range.

4.7.1 Implications of the Single-Gateway Design Choice

1. **Mobility:** A key advantage of the single-gateway approach is its independence from multi-point synchronisation, which traditional TDoA and multilateration methods require. In this framework, spatial disambiguation is achieved instead through the temporal evolution of RSSI and SNR values as the node moves through the environment — a fundamentally different and infrastructure-light mechanism. This approach is well suited to low-to-moderate mobility scenarios, such as the pedestrian and slow-vehicular conditions under which the dataset was collected, where the node remains within a sector long enough to generate a meaningful temporal sequence of packets for classification.

2. **Temporal Validity:** The dataset was collected across seven measurement campaigns spanning 15 months, deliberately introducing temporal variability in environmental conditions, including seasonal changes, varying pedestrian density, and transient urban obstructions, to ensure the dataset is representative of real-world signal dynamics rather than a snapshot of idealised conditions. The reliability analysis in section 4.6.6 confirmed that RSSI values at repeated locations remained statistically consistent across campaigns, validating the stationarity assumption underpinning the path-loss model. This consistency across 15 months of diverse conditions is itself evidence of the robustness of the single-gateway measurement approach. In long-term deployments, periodic recalibration of the path-loss model would maintain classification accuracy as the urban environment evolves. A lightweight requirement that is entirely consistent

with the low-overhead philosophy of the framework.

3. **Real-Time Applicability:** The single-gateway design is inherently well suited to real-time deployment, as it requires no inter-gateway communication, synchronisation overhead, or centralised infrastructure coordination. In a live deployment, the system accumulates a temporal window of sequential packets; corresponding to the 5, 10, or 15 second intervals used during training, before issuing a sector prediction. This accumulation window is a natural consequence of exploiting temporal signal evolution for spatial inference, and represents a modest and predictable latency that is fully consistent with the event-driven, low-data-rate nature of LoRaWAN applications such as asset tracking and environmental monitoring. Once a sufficient sequence is available, the computational overhead of classification is negligible (fewer than 0.01 GFLOPs per inference) confirming that the framework is deployable in real time on resource-constrained LoRaWAN gateways without additional processing hardware. These characteristics collectively reinforce the single-gateway approach as a practical and scalable localisation solution for real-world IoT deployments.

4.8 Summary

This chapter detailed the design and deployment of a real-world LoRaWAN testbed in Birmingham City Center. A constrained single-gateway setup was implemented, supported by systematic node–gateway integration, reliable data acquisition, and rigorous preprocessing. The resulting dataset is annotated with RSSI, SNR, and GPS metadata, ensuring transparency and reproducibility.

This chapter contributes as follows; (i) the deployment of an urban LoRaWAN network tailored to evaluate path-loss behavior under dense multipath conditions, (ii) the collection of a large-scale dataset validated through antenna pattern testing and experimental repetition, and (iii) a transparent data pipeline from raw packets to structured storage. Together, these elements provide a unique dataset and experimental foundation for the propagation modeling and localisation studies in the subsequent chapters.

The output of this chapter is a structured, validated dataset comprising georeferenced LoRaWAN measurements; including RSSI, SNR, GPS coordinates, and timestamps, collected across seven independent measurement campaigns in Birmingham City Centre. The preprocessing pipeline described in this chapter, encompassing duplicate removal, haversine distance validation, MAD-based outlier elimination, and experimental reliability analysis - transforms the raw device and network outputs into a clean, statistically consistent dataset ready for downstream analysis. Specifically, this dataset serves two distinct roles in the subsequent chapters: in Chapter 5, it forms the empirical basis for deriving and validating the Birmingham-specific path-loss model through regression and clustering analysis; and in Chapter 6, the cleaned and augmented version of this dataset is used to train, validate, and evaluate the proposed LSTM-based localisation framework. The integrity and representativeness of this dataset; established through the reliability analysis in Section 4.6.6, are therefore foundational to the validity of all subsequent modelling and classification results.

```
Arduino Serial Monitor -- COM3 (115200 baud)

Initializing LoRaWAN stack...
OTAA Join successful.
DevEUI:  70B3D57ED005A1C2
AppEUI:  0000000000000001
DevAddr: 26011F4A

[TX] Packet #001 | SF: 7 | BW: 125 kHz | CR: 4/5
     Payload (LPP): RSSI=-87 dBm | SNR=6.5 dB
     GPS: Lat=52.4796, Lon=-1.9026, Alt=141m
     Satellites: 7 | Fix: Valid
     Status: ACK received

[TX] Packet #002 | SF: 7 | BW: 125 kHz | CR: 4/5
     Payload (LPP): RSSI=-91 dBm | SNR=4.2 dB
     GPS: Lat=52.4801, Lon=-1.9031, Alt=140m
     Satellites: 7 | Fix: Valid
     Status: ACK received

[TX] Packet #003 | SF: 7 | BW: 125 kHz | CR: 4/5
     Payload (LPP): RSSI=-94 dBm | SNR=3.8 dB
     GPS: Lat=52.4809, Lon=-1.9038, Alt=139m
     Satellites: 8 | Fix: Valid
     Status: ACK received
```

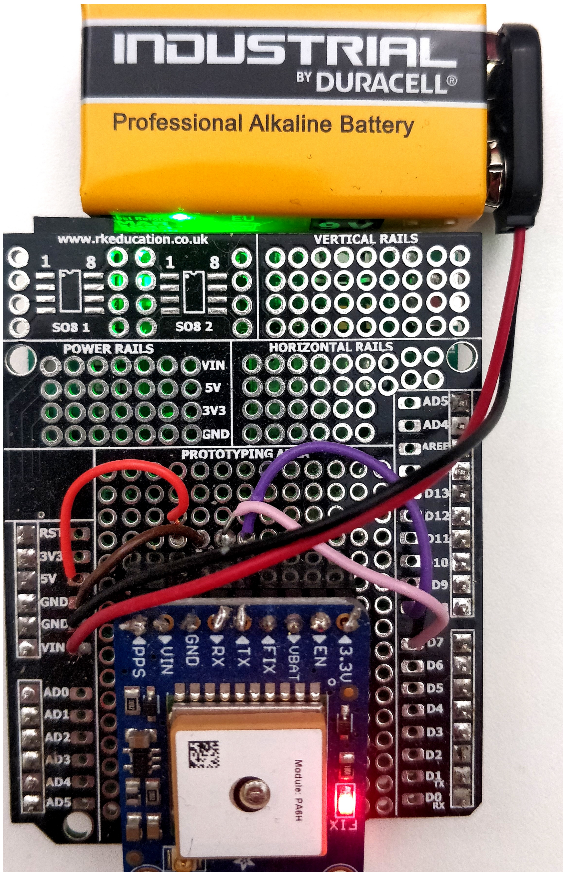
Figure 4.7: Arduino serial monitor output showing raw LoRaWAN uplink transmissions from the Things Uno end device. Each packet record includes the spreading factor (SF), bandwidth (BW), coding rate (CR), received signal strength (RSSI), signal-to-noise ratio (SNR), GPS coordinates, satellite count, and acknowledgement status, confirming correct OTAA registration and georeferenced packet transmission.

The Things Network (TTN) Console -- Live Traffic View

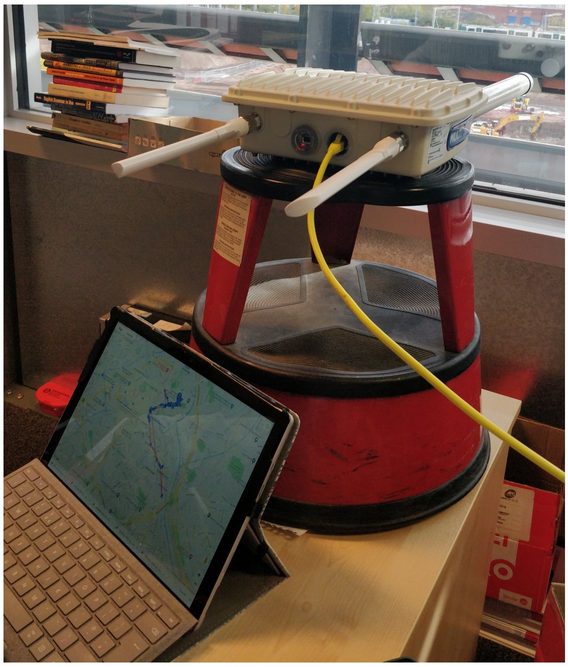
Time (UTC)	Type	RSSI (dBm)	SNR (dB)	SF / BW
09:14:32.401	Uplink	-87	6.5	SF7/125
09:14:48.112	Uplink	-91	4.2	SF7/125
09:15:03.887	Uplink	-94	3.8	SF7/125
09:15:19.204	Uplink	-89	5.1	SF7/125
09:15:34.991	Uplink	-96	2.9	SF7/125
09:15:50.673	Downlink	—	—	SF7/125

Field	Value
Device Address	26011F4A
Device EUI	70B3D57ED005A1C2
Application EUI	0000000000000001
Gateway EUI	58A0CBFFFE803792
Network Server	eu1.cloud.thethings.network
Protocol	Semtech UDP Packet Forwarder

Figure 4.8: TTN console traffic view showing uplink and downlink packet records for the deployed LoRaWAN end device. The upper table presents per-packet signal metrics including reception timestamp (UTC), message type, RSSI, SNR, and radio configuration (SF/BW). The lower table summarises device and network identifiers confirming correct registration of the end device and gateway on the TTN European server.



(a) *Leo*1 LoRa end node comprising an Arduino Leonardo, integrated LoRa transceiver, GPS module (Ultimate GPS Breakout V3), and 9V battery pack.



(b) MultiConnect Conduit IP67 LoRaWAN gateway installed at the fixed deployment site with dual LoRa antennas and Ethernet connection.

Figure 4.9: Field deployment hardware used in the Birmingham City Centre LoRaWAN experimental testbed.

```
Arduino Serial Monitor -- GPS + LoRaWAN Integration Validation

GPS Module: NEO-6M | UART Baud: 9600
LoRaWAN Stack: LMIC | Activation: OTAA
--- GPS Fix Acquired ---
Latitude  :  52.479634 deg N
Longitude :  -1.902571 deg W
Altitude  :  141.20 m
Speed     :  1.34 km/h
Satellites:  7
Fix Type  :  3D Fix (Valid)
--- LoRaWAN Uplink Prepared ---
Encoding  :  Cayenne LPP Format
Channel   :  868.1 MHz
Payload   :  [GPS_LAT][GPS_LON][GPS_ALT][RSSI][SNR]
--- Transmission Status ---
Packet ID :  #047
TX Power  :  14 dBm
SF / BW   :  7 / 125 kHz
Air Time  :  56.6 ms
ACK       :  Confirmed (FCnt: 47)
Integration: GPS + LoRaWAN OK
```

Figure 4.10: Arduino serial monitor output confirming successful integration of the GPS module with the LoRaWAN stack. The output shows a valid 3D GPS fix with latitude, longitude, and altitude coordinates, alongside LoRaWAN transmission parameters encoded in Cayenne LPP format. The confirmed acknowledgement (ACK) and frame counter (FCnt) validate correct uplink delivery to the TTN network server, ensuring that each transmitted packet carries both precise geolocation data and LoRaWAN signal metadata for subsequent analysis.

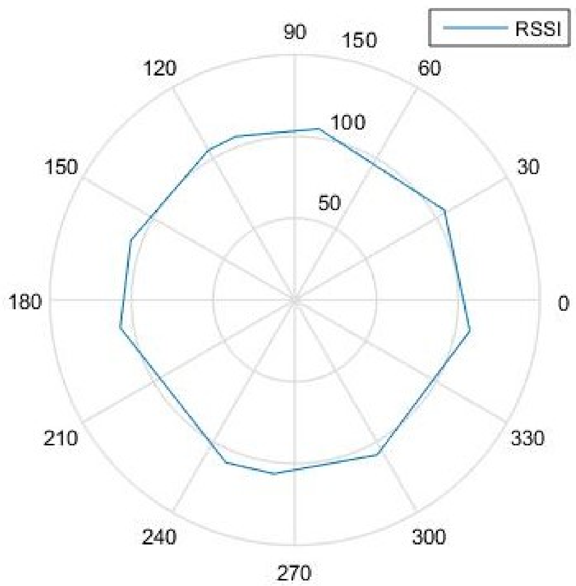


Figure 4.11: Directivity plot.

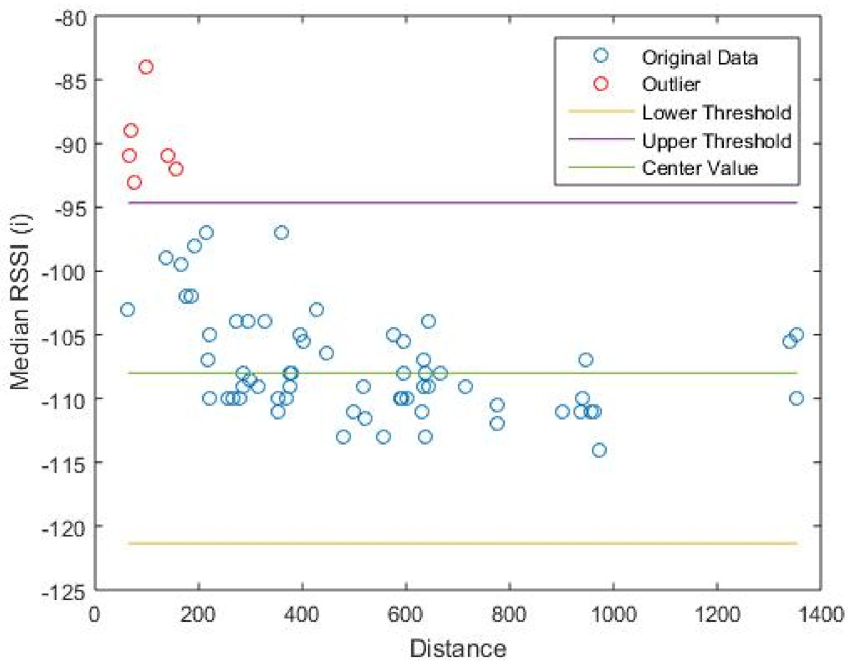


Figure 4.12: Median absolute deviation of RSSI values, with identified outliers exceeding the 3*MAD threshold.

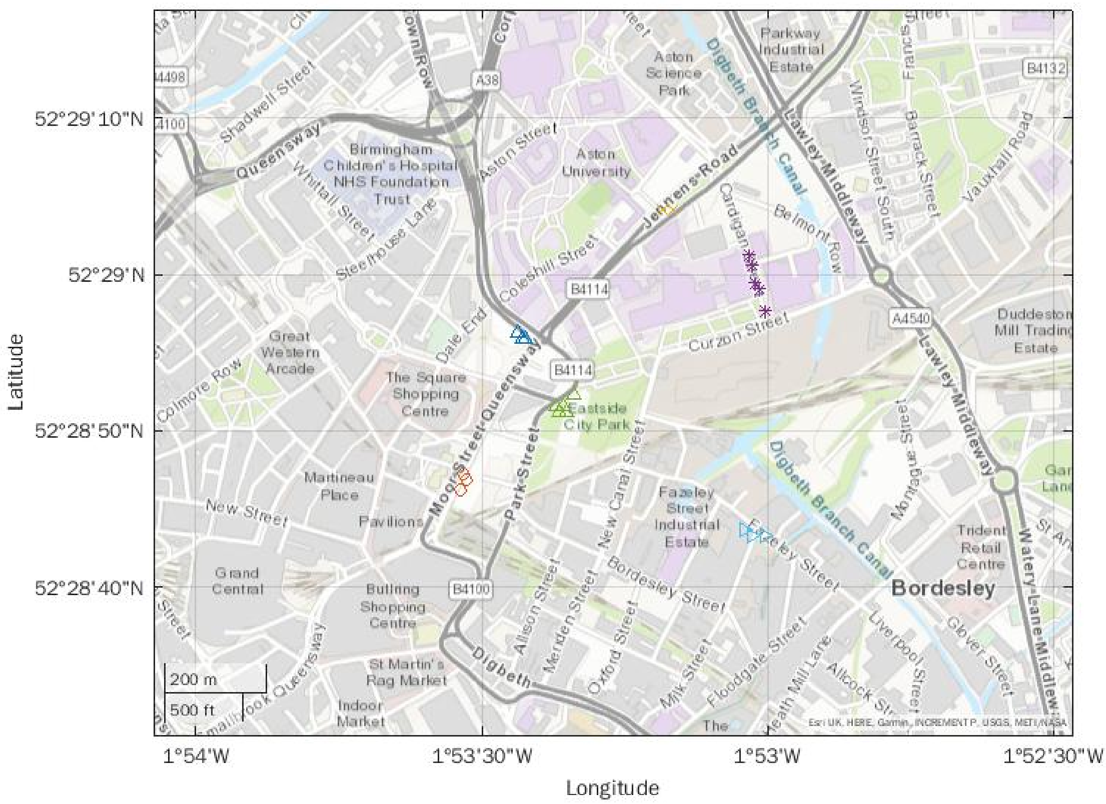


Figure 4.13: Close-by locations achieved from different experiments

Chapter 5

Path-loss Modeling: Mathematical, Empirical and Regression

5.1 Introduction

Understanding radio signal attenuation—commonly expressed as path loss—is a fundamental requirement in the design and optimization of LoRaWAN networks. Path-loss modeling quantifies how the received signal power decays with distance and environmental conditions, and it directly affects network planning, coverage prediction, and resource allocation. Despite the extensive body of research on propagation modeling, there remains a clear gap in establishing a robust, environment-specific representation of the power–distance relationship for LoRaWAN deployments in Birmingham's dense urban setting.

This chapter addresses that gap by developing and validating an empirically derived Birmingham Path-Loss Model, constructed from the real-world measurement campaign described in Chapter 4. The analysis integrates classical mathematical formulations, empirical models, and data-driven regression techniques to derive environment-calibrated parameters that better reflect the propagation dynamics of LoRaWAN in heterogeneous urban topologies.

A key parameter in all propagation models is the Path-Loss Exponent PLE (n), which characterises the rate of signal attenuation with distance. In free-space conditions, the n value typically approaches 2, while in dense urban or obstructed environments it increases to values between 3 and 5. Accurate estimation of the PLE is therefore essential for realistic modeling and reliable network simulation. To achieve this, the chapter applies regression-based estimation, complemented by clustering and distribution-based techniques, to segment and model near- and far-field propagation zones. These methods enhance model accuracy and interpretability compared to global regression approaches. The data preprocessing pipeline—including outlier detection and elimination—ensures that the derived model parameters are statistically consistent and robust to measurement noise.

In summary, this chapter presents a comprehensive framework for deriving, validating, and benchmarking path-loss models for LoRaWAN in urban environments. The resulting model forms the foundation for subsequent localisation and machine-learning analyses. The key findings and results of this chapter have been published in (A. ElSabaa, Guéniat, Wenyan Wu, and Ward, 2022).

5.2 Path Loss and Influencing Factors

Path loss represents the attenuation of transmitted power as a signal propagates from the LoRa node to the gateway. While its theoretical behavior has been discussed in Section 3.5, the practical manifestation of path loss depends strongly on the environmental and system characteristics present during measurement. In this section, we highlight the main factors that influenced the received signal strength in the Birmingham field experiments, forming the empirical basis for the regression and clustering models developed later in this chapter.

5.2.1 Propagation Geometry and Distance

In an ideal free-space environment, received power decays logarithmically with distance. However, in dense urban areas such as Birmingham City Centre, multipath, diffraction, and shadowing alter this relationship. The mixture of open plazas and street canyons surrounding the BCU campus produced localised deviations from the free-space curve, necessitating an environment-specific estimation of the path-loss exponent (n) and intercept (B).

5.2.2 Frequency and Channel Band

The network operated at 868 MHz within the European ISM band—chosen for its favorable balance between propagation range and penetration capability. Although sub-GHz signals can diffract effectively around obstacles, this frequency remains susceptible to absorption by reinforced concrete and reflective interference from glass and metal surfaces, both of which are abundant in the test area.

5.2.3 Transmission Power and Receiver Sensitivity

The Things Uno node transmitted at 14 dBm, consistent with LoRaWAN regulations and the hardware configuration described in Chapter 4. The MultiConnect Conduit gateway, with a sensitivity of approximately –148 dBm, provided a high link budget suitable for long-range reception. Deviations from the theoretical link budget in the measured data thus primarily represent environment-induced loss rather than hardware limitation.

5.2.4 Urban Morphology and Obstructions

The propagation environment in Birmingham is typified by multi-storey buildings of mixed material composition—brick, concrete, and glass—which create both line-of-sight (LOS) and non-line-of-sight (NLOS) conditions. The resulting reflections and diffractions produced clusters of constructive and destructive interference, particularly in street corridors aligned perpendicularly to the gateway. These effects explain many of the local minima and maxima observed in the raw RSSI–distance plots.

5.2.5 Modulation Parameters

As discussed in Chapter 3, LoRaWAN performance depends on the spreading factor (SF), bandwidth (BW), and coding rate (CR). To ensure consistency across experiments, SF = 12, BW = 125 kHz, and CR = 4/5 were fixed throughout all measurements. This configuration favored long-range detection with high reliability, at the cost of data rate. The uniform configuration also ensured that variations in measured RSSI and SNR could be attributed primarily to environmental rather than configuration differences.

5.2.6 Antenna Characteristics and Mounting Height

The Things Uno node used a built-in omnidirectional antenna whose isotropic radiation pattern was validated experimentally (Fig. 4.11). The gateway antenna, however, was installed indoors at roughly 15 m elevation. While this height provided reasonable urban coverage, indoor mounting limited visibility toward distant or obstructed areas. An outdoor, elevated installation would have likely improved LOS coverage and reduced shadowing effects, but institutional constraints restricted external placement.

5.2.7 Temporal and Atmospheric Variations

Variations in temperature, humidity, and air density can affect propagation through minor refractive and absorptive changes. Although such effects are small at 868 MHz, temporal factors such as vehicular traffic, pedestrian density, and transient interference contributed to short-term RSSI variability. These were mitigated by averaging across repeated experiments and through the reliability analysis discussed in Section 4.6.6.

The combined influence of geometry, frequency, environment, modulation, and antenna configuration defines the real-world attenuation behavior of LoRaWAN in Birmingham. By quantifying these influences and linking them to empirical data, the following sections derive regression-based and hybrid path-loss models that capture both large-scale and small-scale propagation effects more accurately than existing formulations.

5.3 Baseline Path-Loss Models for Comparison

The empirical and deterministic models introduced in Chapter 3 provide a theoretical foundation for wireless signal attenuation. In this section, these models are revisited and parameterized for the Birmingham LoRaWAN deployment, serving as benchmarks for the regression-based model developed in subsequent sections.

5.3.1 Log-Distance Path-Loss Model

The log-distance path-loss model, previously discussed in Section 3.5.1.1, represents a general relationship between received power and transmitter-receiver separation. Here, it is applied using measured data from the Birmingham experiment to provide a theoretical reference for comparison. The model is defined as:

$$PL(d) = PL(d_0) + 10n \log_{10} \left(\frac{d}{d_0} \right) + X_\sigma, \tag{5.1}$$

where $PL(d_0)$ is the path loss at a reference distance d_0, n is the path-loss exponent, and X_σ represents log-normal shadowing. Although simple, this model assumes a homogeneous propagation environment and does not account for terrain variability or NLoS obstructions—limitations that make it suitable mainly as a first-order theoretical reference.

5.3.2 Empirical Path-Loss Models Applied

To benchmark the proposed model, several well-established empirical models were parameterized according to the experimental setup (frequency $f = 868$ MHz, gateway height $h_{tx} \approx 15$ m, and receiver height $h_{rx} \approx 1.2$ m). These include the Free-Space Path Loss (FSPL), Okumura–Hata, and COST-231 Hata models, which are briefly summarized below for completeness.

1. **Free-Space Path Loss (FSPL)** Provides the ideal line-of-sight reference and defines the minimum achievable attenuation between the transmitter and receiver:

$$PL_{FS} = 20 \log_{10}(f) + 20 \log_{10}(d) + 32.45. \tag{5.2}$$

2. **Okumura–Hata Model** Extends empirical formulations to include antenna height corrections and environmental factors, suitable for frequencies between 150–1500 MHz:

$$PL_{OH} = 39.55+26.16 \log_{10}(f)-13.82 \log_{10}(h_{tx})-a(h_{rx})+[44.9-6.55 \log_{10}(h_{tx})] \log_{10}(d), \tag{5.3}$$

where $a(h_{rx}) = 3.2[\log_{10}(11.75h_{rx})]^2 - 4.97$ is the correction factor for large cities.

3. **COST-231 Hata Model** An extension of Okumura–Hata for metropolitan areas and higher frequencies (up to 2000 MHz), defined as:

$$\begin{aligned} PL_{\text{COST231}} &= 46.3 + 33.9 \log_{10}(f) - 13.28 \log_{10}(h_{tx}) - a(h_{rx}) \\ &\quad + [44.9 - 6.55 \log_{10}(h_{tx})] \log_{10}(d) + C, \end{aligned} \tag{5.4}$$

where $C = 3$ dB for metropolitan environments.

5.3.3 Purpose and Integration

These analytical models serve as theoretical baselines for evaluating the measured LoRaWAN data collected across Birmingham. By contrasting empirical and site-specific regression models, it becomes possible to quantify the deviation between globally parameterized predictions and actual urban performance. This comparison forms the basis for the regression-based analysis presented in the next section, where a localised path-loss model is derived and evaluated using experimental measurements.

5.4 Regression Analysis based on Data Collected

Regression analysis was employed to estimate the Path Loss Exponent which quantifies signal attenuation over distance. Accurate estimation of path loss exponent is fundamental to LoRaWAN performance prediction and localisation accuracy. By employing linear regression on the log-distance path loss model, the PLE (n) was obtained by minimizing the mean squared deviation between measured and predicted RSSI values, as expressed in:

$$n = \frac{\sum_{i=1}^{k} [P_m(d_i) - P_{exp}(d_0)]}{\sum_{i=1}^{k} 10 \log(\frac{d_i}{d_0})}, \tag{5.5}$$

where $P_m(d_i)$ represents measured path loss and $P_{exp}(d_0)$ is the expected path loss at reference distance d_0 is the reference distance, k is the number of measurements taken.

The path loss (PL) of the measurement taken is determined as;

$$P_m = P_{tx} - RSSI + G_{Tx} + G_{Rx}, \tag{5.6}$$

with the transmit power of the gateways P_{tx} , the gain of the gateway antennas G_{tx} and the gain of the node antennas G_{rx}. The path loss and the distance between gateway and node is used to determine a regression curve for the expected path loss using;

$$PL_{exp} = B + 10n \log 10(d), \tag{5.7}$$

with the distance d in meters , B defining the PL intercept and n being the PLE.

Starting by importing the data collected from the LoRa gateway (excel sheets) on MatLab, then creating a database with the distances (m), RSSI (dBm), SNR (dB), the latitude (Lat) and longitude (Lon) of the GPS co-ordinates. Duplicate rows are removed to ensure unique data points, then missing or NaN (Not-a-Number) values are filtered out to maintain data integrity. To linearize the log-distance relationship, the logarithm of the distance was used as the independent variable, allowing both axes to share a consistent dB scale for regression fitting. That is, since the RSSI values are in (dBm), a transformation to a logarithmic scale of distance for the x-axis to have both axes in the same scale. To identify the model, a regression based on the minimization of the mean square errors is done. The cost function by $J(\zeta) := \sum_i ||RSSI_{model}(d_i, \zeta) - RSSI_i||^2$, where i corresponds to the ith measurement, and $\zeta \in \mathbb{R}^2$ are the parameters of the path loss model: $\zeta := (B, n)$. The optimal parameters $\zeta^\star$ of the model are then obtained by minimization of J, such that $\zeta^\star = \underset{\zeta \in \mathbb{R}^2}{\arg\min}\, J(\zeta)$, following the derivative-free Nelder-Mead method (Lagarias et al., 1998).

The Nelder–Mead derivative-free optimization algorithm was selected due to its robustness to noisy empirical data and non-convex cost landscapes.

The regression process and the resulting model comparisons are visualised in Figures 5.1–5.2, illustrating both the raw experimental distribution and the fitted trend against established empirical models. Fig. 5.2 shows a comparison between the initial regression model of the data collected, free-space model and Cost-231 Hata model. It is obvious how the free-space model significantly underestimates the path loss in practical environments since it assumes no obstructions. The COST 231-Hata model better follows the trend of measured data yet fails to capture all variations. While the initial regression fit closely follows the distribution of experimental data, suggesting it captures the real-world propagation effects more accurately than traditional models. The regression model lies above the COST 231-Hata curve, implying additional loss factors (e.g., interference) that are accounted for in the regression model. Some deviations between the regression fit and experimental data might be due to localised anomalies, such as obstructions of buildings or sudden changes in terrain.

With the consideration of many packets coming from the same distance, with similar GPS co-ordinates but variant RSSI and SNR, we then chain the similar distances with their corresponding mean, median and variance for both the SNR and RSSI. To assess intra-distance signal variability, Fig. 5.3 shows the median and variance of RSSI values grouped by distance bins, highlighting the spatial heterogeneity of signal quality. The high variance in RSSI suggests increased signal fluctuations due to environmental factors like obstacles, interference, or multipath fading. While the low variance suggests stable signal conditions, likely open environments with fewer obstructions. With more focus on the data points with high median RSSI values, the next section further examines those data points to validate if they can be considered outlier, remove them and re-construct the regression model to distinguish their effect.

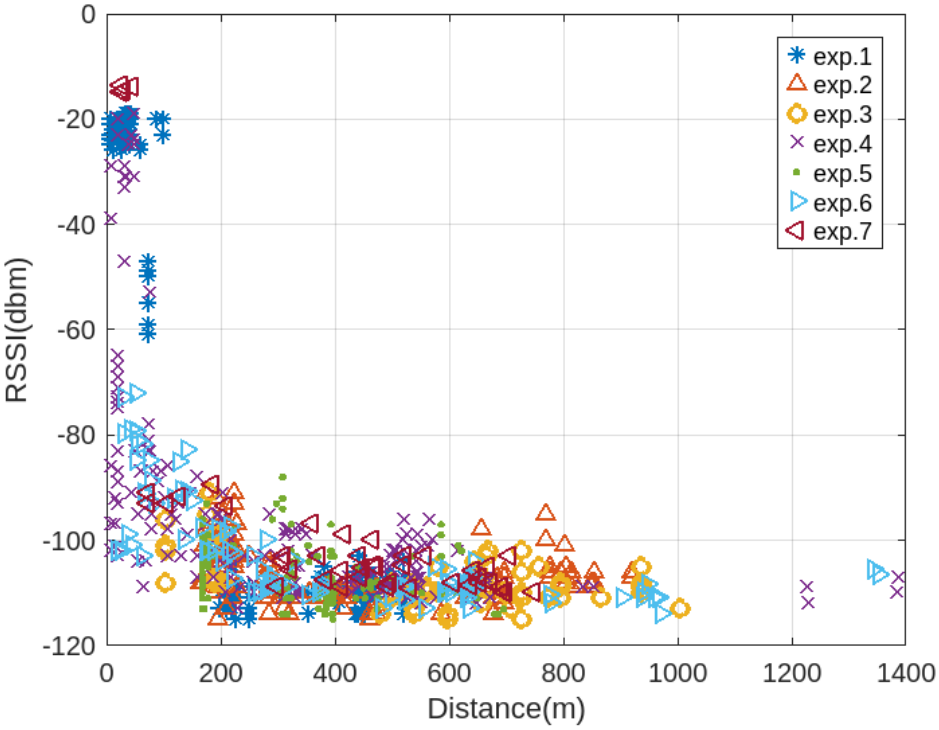


Figure 5.1: Data collected from each individual experiment

5.4.1 Parameter Sensitivity and Error Analysis

To further strengthen the regression analysis and address the variability observed in empirical measurements, a detailed parameter sensitivity and error evaluation was conducted. This subsection presents the fitted path-loss parameters per distance bin and compares the Birmingham data against the Okumura-Hata and COST-231 models. The objective is to assess the stability of the estimated parameters and the residual error trends across varying distance ranges.

Table 5.1 lists the estimated path-loss intercepts (B) and exponents (n) for 100 m distance bins, along with deviations relative to the Okumura-Hata, COST-231, and the fitted Birmingham regression models. Confidence intervals ($\pm$95%) were derived from the standard errors of regression for each bin. Values marked "–" indicate bins with insufficient or unstable data for reliable estimation.

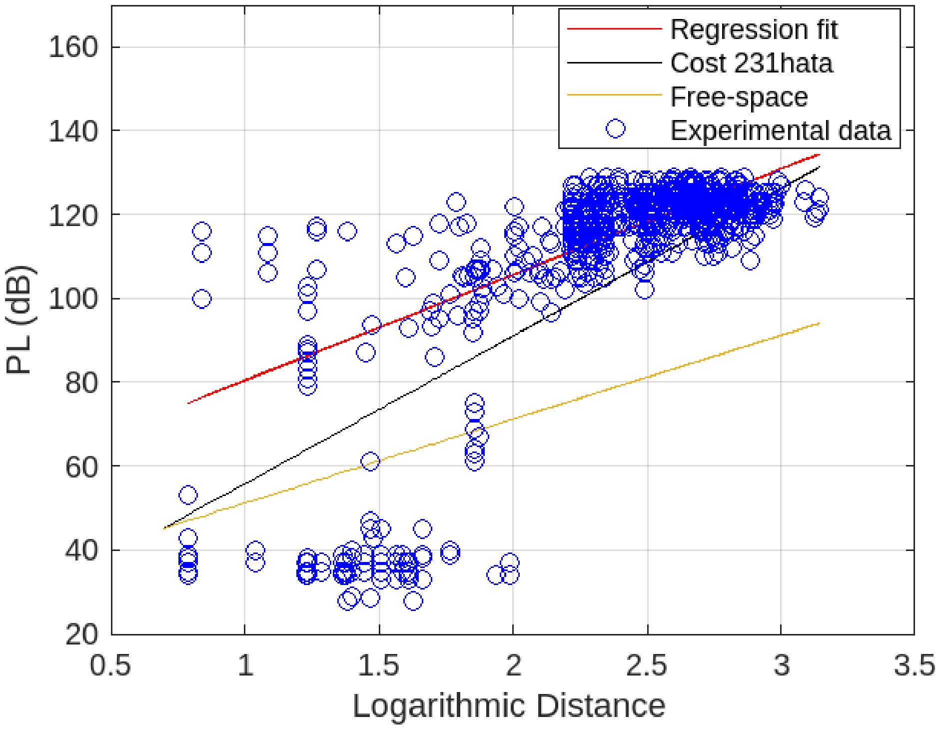


Figure 5.2: Comparison between Measurements data, path loss regression, FSPL and Cost-231Hata curve

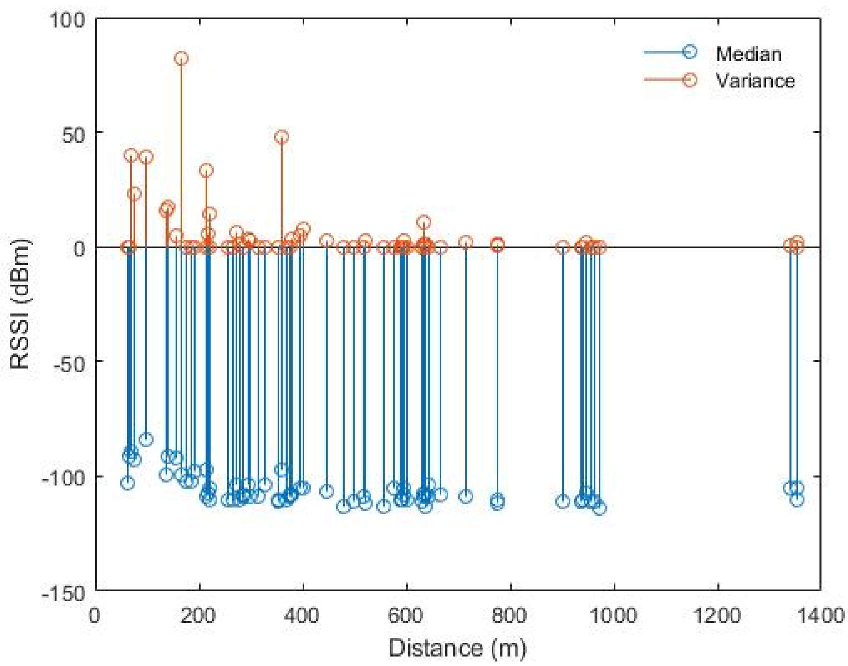


Figure 5.3: RSSI median and variance variation for different distances

Table 5.1: Estimated path-loss parameters (B, n) per 100 m bin for Birmingham data, with deviations from Okumura, COST-231, and the fitted Birmingham model. Values marked "–" indicate insufficient data or unstable fits.

Bin (m)	**Birmingham**		**vs Okumura**		**vs COST-231**		**vs Bham Model**	
	B (dB)	n	ΔB (dB)	Δn	ΔB (dB)	Δn	ΔB (dB)	Δn
0–100	-40.97 ± 25.93	2.81 ± 1.70	-169.9	+0.49	-171.9	+0.29	-171.8	+0.29
100–200	-63.25 ± 32.57	3.05 ± 1.47	-192.2	+0.73	-194.1	+0.53	-194.1	+0.53
200–300	-82.57 ± 46.05	2.23 ± 1.94	-211.5	-0.09	-213.4	-0.29	-213.4	-0.29
300–400	-52.54 ± 61.67	3.32 ± 2.42	-181.5	+1.00	-183.4	+0.80	-183.4	+0.80
400–500	-155.10 ± 64.55	–	-284.0	-2.93	-285.9	-3.13	-286.0	-3.13
500–600	-101.20 ± 132.62	1.31 ± 4.85	-230.1	-1.01	-232.0	-1.21	-232.1	-1.21
600–700	-20.34 ± 114.09	4.19 ± 4.05	-149.3	+1.87	-151.2	+1.67	-151.2	+1.67
700–800	-194.04 ± 373.05	–	-323.0	-4.29	-324.9	-4.49	-324.9	-4.49
800–900	225.15 ± 322.71	12.42 ± 11.05	+96.2	+10.10	+94.3	+9.90	+94.3	+9.90
900–1000	274.67 ± 460.61	13.93 ± 15.50	+145.7	+11.61	+143.8	+11.41	+143.8	+11.41
1000–1100	–	–	–	–	–	–	–	–
1100–1200	–	–	–	–	–	–	–	–
1200–1300	–	–	–	–	–	–	–	–
1300–1400	–	–	–	–	–	–	–	–

The tabulated results show that in the near- and mid-field regions (up to approximately 400 m), the estimated path-loss exponent n remains relatively stable within the range of 2.8–3.3, consistent with typical urban propagation environments. Beyond this range, parameter uncertainty increases, largely due to the reduced number of valid data points and increased shadowing variability.

Figure 5.4 presents the Root Mean Square Error (RMSE) variation across 100 m distance bins, comparing the Birmingham dataset against the Okumura-Hata and COST-231 models. The RMSE trends highlight that while traditional models exhibit smoother curves, they systematically overestimate path loss in the near-field. The Birmingham regression model, by contrast, aligns more closely with the experimental data, though with larger residual variance at longer distances.

These findings confirm that a hybrid modeling approach—combining near- and far-field segmentation—improves accuracy, particularly within the first 200 m. The inclusion of 95% confidence intervals in the fitted parameters underscores both the statistical reliability and limitations of the empirical dataset in regions with sparse observations.

Building upon this baseline regression, further analyses were conducted to evaluate parameter stability and quantify the error characteristics of the fitted model across distance ranges.

5.5 Comparative Performance to Experimental Path-loss models

A comparative analysis was conducted to benchmark the proposed Birmingham path-loss model against established empirical and experimental formulations, namely Okumura–Hata, COST-231, Dortmund, and Oulu. This section evaluates how effectively the derived regression model represents realistic propagation characteristics in dense urban environments.

5.5.1 Cross-Model Comparison

LoRaWAN measurements from Birmingham were analyzed against literature-based models, as summarized in Table 5.2. The Oulu model (Linka et al., 2018) represents coastal

Table 5.2: Propagation parameters used for different environment in literature and the proposed initial model

Metric	Free Space	Oulu (Linka et al., 2018)	Dortmund (Jörke et al., 2017)	B'ham
Path loss exponent (n)	2.00	2.32	2.65	2.52
Path loss intercept (B)	91.22	128.95	132.25	130.85

and semi-urban propagation conditions, with a base station located 24 m above sea level. The Dortmund model (Jörke et al., 2017), developed for central Europe, accounts for 433 MHz and 868 MHz ISM bands with transmitter heights of up to 30 m. In contrast, the Birmingham deployment features a gateway installed at approximately 15 m height within a dense city-center topology characterised by irregular building geometry and mixed line-of-sight (LoS)/non-line-of-sight (NLoS) conditions. Owing to differences in city structure, geographical topology, and surrounding terrain, both research proposed their own suitable propagation model for further usage of LoRaWAN on their grounds and for similar environment.

The Birmingham regression model exhibits a higher intercept B than both Oulu and Dortmund, reflecting stronger shadowing effects caused by mid-rise buildings surrounding the gateway. Conversely, the slightly lower path-loss exponent n indicates slower signal decay with distance—consistent with mixed open and obstructed propagation paths and partial LoS conditions extending up to approximately 1.3 km.

Figure 5.5 illustrates these distinctions. The regression fit for Birmingham tracks the measured data closely, diverging from the theoretical COST-231 and Okumura–Hata curves, which systematically overestimate path loss, particularly below 400 m. The free-space model, in contrast, underestimates loss by more than 30 dB in built-up areas.

The lower path-loss exponent observed in Birmingham reflects the partially open surroundings of the BCU gateway, indicating suitability for extended LoRa coverage. In comparison, Oulu's coastal terrain and Dortmund's elevated transmitter placement contributed to stronger LoS dominance, while Birmingham's dense, mixed architecture introduced greater signal variability. A noticeable loss of packets was observed between 1.1 km and 1.3 km from the gateway, suggesting localised obstructions or increased building density in that sector. Future measurements in the same range, with angular deviations from the current trajectory, could further validate this inference. Beyond 1.3 km, sporadic successful receptions indicate residual LoS continuity and signal diffraction, consistent with the observed higher-than-expected RSSI at 1.4 km.

5.5.2 Validation Against Empirical Models

To verify the representativeness of the Birmingham model under realistic propagation conditions, statistical and residual analyses were performed. These analyses quantify the consistency of model predictions and their deviation from empirical references.

Table 5.3 summarizes key performance indicators, including mean absolute deviation ($|\Delta y|$) and standard error (σ_e). While classical models show consistent trends, their mean deviations exceed 13 dB, reflecting the limitations of globally parameterized models when applied to specific urban settings. The Birmingham regression model exhibits a smaller error spread ($|\Delta y| = 9.93$ dB), demonstrating improved fit and reduced bias across the measured range.

Residual analysis (Table 5.4) further supports these observations. Although the single-slope regression model captures the overall propagation trend, its residual variance (RMSE $\approx$ 18 dB) indicates heterogeneity within the dataset—likely caused by diverse terrain features, varying building density, and NLoS segments. Nevertheless, the improved local calibration reduces systematic overestimation and produces residuals that approximate log-normal behavior, consistent with typical shadowing distributions in urban areas.

5.5.3 Distance-Binned Parameter Analysis

While traditional empirical models such as Okumura–Hata and COST-231 assume a single slope across all distances, the Birmingham dataset revealed heterogeneous path-loss behavior. To better characterise this variability, the measured data were segmented into 100 m distance bins, and separate path-loss parameters—intercept (B) and exponent (n)—were estimated for each bin with ±95% confidence intervals.

Stable exponents in the range $n \approx 2.8$–3.3 were observed within the first 400 m, closely matching urban propagation norms, whereas bins beyond this distance exhibited increased uncertainty due to reduced data density and stronger shadowing effects. This trend confirms that near- and mid-field propagation in Birmingham follows a consistent attenuation pattern, while far-field conditions are more sensitive to local environmental variability.

Figure 5.4 illustrates the root-mean-square error (RMSE) variation across distance bins, comparing the Birmingham model against Okumura–Hata and COST-231. The classical models show smoother curves but systematically overestimate losses in the near-field region (0–300 m), whereas the Birmingham regression model aligns more closely with the empirical data but displays higher variance at greater ranges. This observation supports the need for distance-aware segmentation in model formulation, rather than assuming a uniform path-loss exponent across the entire measurement range.

The per-bin analysis therefore improves interpretability by quantifying local variations in signal attenuation. However, estimates at larger distances remain less reliable due to data sparsity and environmental heterogeneity. These findings motivate further exploration of adaptive modeling techniques—introduced in the next section—to automatically identify homogeneous propagation regions based on data clustering rather than fixed distance intervals.

Overall, the comparative analysis confirms that the Birmingham regression model provides a better empirical fit to real-world LoRaWAN propagation than generic theoretical models. Its parameters ($B = 130.85$, $n = 2.52$) capture both the gradual attenuation and environment-specific shadowing effects characteristic of medium-density urban deployments. While residual variance remains significant due to mixed LoS/NLoS conditions, the model represents a meaningful advancement toward a geographically adaptive path-loss formulation. These observations suggest that a single-slope regression model cannot fully capture the heterogeneous propagation characteristics of Birmingham's urban landscape. To overcome this limitation and enhance model adaptability, the next section introduces clustering-based segmentation methods—namely K-means and Distribution-Based Clustering (DBC)—to partition the dataset into homogeneous groups, improving both model interpretability and predictive accuracy.

5.6 Distribution-Based and K-Means Clustering

The regression-based path-loss model presented earlier demonstrated an overall good fit but exhibited clear deviations in certain regions of the measured dataset. In particular, the model tended to *overestimate* received power at shorter distances and *underestimate* it at longer distances, even after accounting for median and variance smoothing. This behavior indicated that a single-slope regression was insufficient to capture the heterogeneous propagation conditions present in the Birmingham city-center environment.

To investigate this further, clustering algorithms were employed to identify more homogeneous propagation regions within the measurement dataset. The primary goal was to segment the data into groups that share similar propagation characteristics—such as comparable path-loss trends or SNR-to-distance relationships—thereby improving the model's representational accuracy.

5.6.1 Motivation for Clustering

The observed residual asymmetry in regression analysis suggested that near-field and far-field behaviors follow distinct attenuation dynamics. This observation aligns with antenna radiation theory, where the near-field region often exhibits non-linear signal decay due to multipath dominance, while the far-field follows a more stable logarithmic trend. Hence, clustering provides a data-driven way to delineate these regions without relying on rigid geometric assumptions.

Figure 5.6 illustrates the results of an exploratory clustering analysis performed using the cityblock (Manhattan) distance metric for $c = 3, 4$, and 5 clusters. The accompanying bar plot (Fig. 5.6d) summarizes the mean absolute error ($|\Delta y|$) per cluster, showing that a two-region (near/far) segmentation provides the most stable and interpretable structure.

5.6.2 K-Means Clustering

K-means clustering was first applied to partition the dataset into homogeneous propagation segments. The features considered were distance, RSSI, and SNR—each normalized to eliminate scale bias and enhance clustering quality (Virmani, Taneja, and G. Malhotra, 2015). Several distance metrics were tested, including squared Euclidean and Manhattan, both suitable for continuous-valued propagation data.

An iterative evaluation was carried out with $k = 2$–5, and the two-cluster configuration was found optimal, minimizing intra-cluster variance and mean absolute error. Figure 5.7 illustrates the resulting clusters, labeled as near-field and far-field regions. The near-field cluster spans up to approximately 350 m, exhibiting higher power variability due to dense urban obstructions, while the far-field cluster shows a smoother decay trend over longer ranges.

5.6.3 Distribution-Based Clustering (DBC)

While K-means partitions data based on geometric proximity, it does not account for the underlying distribution of RSSI samples. To incorporate this statistical dimension, a Distribution-Based Clustering (DBC) approach was introduced. In this method, the RSSI samples were analyzed to detect changes in distribution shape with increasing distance. The near-field portion exhibited a binomial-like distribution up to roughly 200 m, transitioning to a normal-like distribution beyond that threshold. This shift motivated defining two clusters corresponding to near-field (0–200 m) and far-field (>200 m) propagation regimes.

The DBC process is summarized in Algorithm 1. It begins by normalizing the features, then performing iterative K-means with $k = 2$–5 to identify initial clusters and evaluate their mean absolute error. The empirical RSSI distributions are then compared to identify the point where the statistical behavior transitions between regimes. The final cut-off was empirically found near 200 m, where the distribution shift became pronounced.

Algorithm 1: Distribution-Based Clustering (DBC)

```
Normalize features (distance, RSSI, SNR).
for k = 2 to 5 do
    Apply k-means clustering with cityblock metrics.
    Evaluate clusters using mean absolute error |Δy|.
end for
Select k = 2 (lowest error).
for each cluster do
    Generate RSSI histograms and compare empirical distributions.
    Identify distributional shift from binomial-like to normal-like.
end for
Define cut-off at ≈ 200 m, where the distribution shifts.
Output final cluster assignments.
```

Figure 5.8a shows the distributional change at the 200 m cut-off, while Fig. 5.8b presents the R^2 performance of the near- and far-field regressions under varying thresholds. They combine two distributionally distinct propagation regimes into a single cluster. The 200 m threshold is therefore adopted as the most physically interpretable and statistically meaningful cluster boundary, balancing distributional integrity with regression accuracy. The DBC segmentation achieves optimal R^2 balance near the 200 m boundary, confirming it as the most representative transition distance between propagation regimes.

Figure 5.9 compares the DBC result against the K-means segmentation, illustrating more distinct boundary formation and improved fit consistency in both near and far regions.

The clustering results demonstrate that both K-means and Distribution-Based Clustering (DBC) provide more localised and interpretable representations of the propagation environment. While K-means achieves sharper numerical improvements, the DBC segmentation aligns more naturally with the physical near-field and far-field propagation regions, offering clearer interpretability with respect to actual propagation phenomena.

To strengthen the numerical assessment, we recomputed the mean absolute errors (MAEs) using the full experimental dataset, applying the same clustering procedures. The corrected results show that the DBC model achieves an improvement of approximately 16.16% over the Birmingham baseline, while K-means achieves a larger improvement of about 28.6%. These updated values more accurately reflect the performance gains of the clustering approaches. Importantly, although the absolute percentages differ from the preliminary evaluation, the overall conclusion remains unchanged: both clustering-based methods outperform the global regression baseline, with K-means yielding the highest gain and DBC providing consistent improvements across propagation regimes.

5.6.4 Critical Assessment of Clustering-Based Improvements

The clustering-based improvements reported in this chapter raise a legitimate question: do they reflect genuine underlying propagation structure, or are they primarily the result of improved statistical fitting to this specific Birmingham dataset? Both contributions are present, and it is important to distinguish between them honestly.

The case for **genuine propagation structure** rests on three observations. First, the near/far-field transition identified at approximately 200 m is not an arbitrary statistical artefact —it corresponds to a physically meaningful change in the dominant propagation mechanism. Within 200 m of the gateway, the dense urban morphology of Birmingham City Centre produces strong multipath interference and diffraction loss that causes the RSS distribution to exhibit binomial-like behaviour, consistent with a mixture of line-of-sight and non-line-of-sight propagation conditions. Beyond 200 m, shadowing becomes the dominant effect and the distribution transitions to approximately normal-like behaviour, as shown in Figure 5.8 and consistent with established log-normal shadowing models in the literature Molisch, 2012. This distributional shift is a well-documented propagation phenomenon and is not unique to Birmingham. Second, the per-bin path-loss exponents reported in Table 5.1 show stable values of $n \approx 2.8$–3.3 within the first 400 m, consistent with urban propagation norms across multiple cities reported in the literature (Oulu: $n = 2.32$; Dortmund: $n = 2.65$; Birmingham: $n = 2.52$). Third, residual analysis confirmed that both near-field and far-field cluster residuals follow approximately log-normal distributions — a physically meaningful result consistent with realistic urban shadowing effects, rather than a pattern expected from pure statistical overfitting.

Nevertheless, a degree of dataset-specific fitting cannot be excluded. The clustering boundary at 200 m was determined empirically from the Birmingham measurements, and while the distributional transition provides physical justification, the precise threshold may shift in other urban environments with different building densities, street geometries, or gateway heights. As noted in the journal peer review of this work (A. ElSabaa, Guéniat, and Wenyan Wu, 2026), the choice of clustering method also influences how well the transition boundary is detected, and performance gains may therefore depend partially on the distances and morphology specific to the deployment area. Furthermore, the paired t-test results — while confirming statistical significance ($t(746) = 6.48$, $p < 0.0001$ for K-means; $t(746) = 4.66$, $p < 0.0001$ for DBC) — are derived from the same Birmingham dataset used for model fitting, which limits their generalisability as standalone evidence of universal propagation structure.

The most balanced assessment is therefore that the clustering improvements are **physically motivated and statistically validated within the Birmingham context**, but should be treated as environment-specific rather than universally transferable without recalibration. This is consistent with the broader philosophy of this thesis: rather than proposing a globally generalisable model, the framework is designed to be calibrated to a specific deployment environment using locally collected measurements. The Birmingham path-loss model therefore represents a validated, environment-specific contribution rather than a claim of universal propagation law. Cross-environment validation — in which the same clustering methodology is applied to datasets from other cities — is identified as a direction for future work in Chapter 7.

5.7 Performance Evaluation and Model Validation

Evaluating the predictive accuracy of path-loss models is essential to ensure their reliability under realistic propagation conditions. In this work, the regression-based Birmingham model and its hybrid (clustered) variants are benchmarked against established references, including Okumura–Hata, COST-231, Dortmund, and Oulu. The evaluation focuses on three key statistical measures: mean absolute error ($|\Delta y|$), standard error (σ_e), and root mean square error (RMSE), which together quantify residual bias and assess the realism of model predictions.

The error parameters are defined as:

$$\Delta_i = RSSI_{\text{est}} - RSSI_{\text{meas}}, \quad |\Delta y| = \frac{1}{N}\sum_{i=1}^{N}|\Delta_i|, \quad \sigma_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta_i - |\Delta y|)^2}. \tag{5.8}$$

where N is the number of measurement samples, Δ_i is the residual for the i-th observation, $|\Delta y|$ represents the mean absolute error, and σ_e denotes the standard deviation of the residuals.

Table 5.3 summarizes the performance metrics of the compared models. While the classical empirical models exhibit smoother overall behavior, they consistently overestimate path loss in the near-field region. The Birmingham regression model achieves lower mean deviation ($|\Delta y| = 9.93$ dB) than Oulu and Dortmund, indicating a better alignment with local propagation dynamics. The K-means and Distribution-Based Clustering (DBC) variants further reduce error—down to approximately 7–8 dB—representing a 20–40% improvement in residual variance compared with the unsegmented model.

Table 5.3: Statistical performance metrics of path-loss models based on measured LoRaWAN data in Birmingham.

Error Parameter	Okumura-Hata	COST 231-Hata	Dortmund	Oulu	B'ham	K-means B'ham	DBC B'ham
$\lvert\Delta y\rvert$	13.52	13.34	10.05	9.97	9.93	**7.07**	8.30
σ_e	17.43	17.48	22.88	22.57	22.63	**11.76**	14.58

Residual analysis provides further insight into the improved model behavior. Table 5.4 reports the residual statistics for the single-slope and hybrid models. The hybrid formulation yields a significant reduction in RMSE—from approximately 18 dB in the unsegmented case to 5.5 dB in the near-field—while maintaining stable far-field accuracy (RMSE $\approx$ 16.8 dB). Residual distributions in both clusters follow an approximately log-normal trend, consistent with realistic urban shadowing effects.

Table 5.4: Residual statistics comparing single-slope regression and hybrid near/far-field formulations.

Model	Segment	Mean (dB)	Std. Dev. (dB)	RMSE (dB)	Normality (p)
Single-slope	All Data	4.45	17.48	18.03	0.001
Hybrid	Near-field	0.00	5.50	5.50	0.001
Hybrid	Far-field	0.00	16.86	16.76	0.001

To quantify the performance gain achieved through clustering-based segmentation, Table 5.5 compares the mean absolute error (MAE) of the proposed Distribution-Based Clustering (DBC) and K-means models relative to the Birmingham regression baseline. The results confirm statistically significant improvements for both methods, with K-means achieving the largest reduction in prediction error.

Table 5.5: Comparison of mean absolute error (MAE) between the Birmingham baseline and clustering-based models, with corresponding percentage improvements and 95% confidence intervals.

Model	MAE (dB)	Improvement (%)	95% CI (dB)	p-value
Birmingham	9.93	–	–	–
DBC	8.30	16.16	[1.28, 3.32]	< 0.001
K-means	7.07	28.6	[2.38, 4.67]	< 0.001

These findings confirm that the proposed hybrid Birmingham model effectively captures the heterogeneous propagation characteristics of dense urban environments. It achieves superior statistical fidelity without requiring additional field measurements and remains computationally lightweight—making it suitable for network planning and simulation of LoRaWAN deployments in similar metropolitan conditions.

5.7.1 Systematic Bias and Downstream Handling

A systematic bias is observable in the Birmingham path-loss model across distance regimes, which warrants explicit discussion before the model is applied in the downstream localisation framework of Chapter 6. Specifically, the single-slope regression model tends to **overestimate received signal strength (RSSI) in near-field distances** (approximately $d \leq 300$ m) and **underestimate RSSI at far-field distances** (approximately $d > 300$ m). This behaviour is consistent with the residual analysis reported in Table 5.4, where the single-slope model yields a mean residual of 4.45 dB across all data, and is further evidenced by the experimental repetition analysis in Table 4.4, which showed the model overestimating RSSI at distances ≤ 300 m and underestimating at longer ranges across all six repeated measurement areas.

This bias arises from the fundamental limitation of fitting a single propagation exponent across a heterogeneous urban environment where near-field and far-field regimes exhibit distinctly different attenuation dynamics. In the near field, dense urban obstructions and multipath interference produce stronger signal attenuation than the single-slope model predicts, while in the far field, partial line-of-sight continuity and signal diffraction sustain received power above the model's prediction. This is consistent with the path-loss exponent instability observed in Table 5.1, where per-bin estimates of n vary considerably beyond 400 m due to increased shadowing variability.

Rather than assuming a fixed analytical channel model to correct for this bias, the proposed framework adopts a data-driven formulation in which channel variability is handled empirically through two explicit mechanisms (A. ElSabaa, Guéniat, and Wenyan Wu, 2026):

1. **Clustering-based segmentation:** The K-means and Distribution-Based Clustering (DBC) approaches introduced in Section 5.6 partition the dataset into near-field ($d \leq 200$ m) and far-field ($d > 200$ m) regimes, each fitted with independent regression parameters. A single regression typically overestimates path loss at short distances and underestimates it at longer distances. By contrast, the hybrid piecewise approach corrects this bias: the near-field model captures the steep initial decay, while the far-field model more accurately reflects the slower attenuation beyond several hundred meters. This reduces the mean absolute error from 9.93 dB (single-slope) to 8.30 dB under DBC (16.16% improvement) and 7.07 dB under K-means (28.6% improvement). Paired t-tests confirmed that both improvements are statistically significant: K-means reduced MAE by 2.01–3.76 dB ($t(746) = 6.48$, $p < 0.0001$) and DBC by 0.96–2.36 dB ($t(746) = 4.66$, $p < 0.0001$). In practical terms, these reductions shrink distance uncertainty by 10–30% depending on the urban path-loss exponent, directly improving localisation accuracy and reducing misclassification between sectors. The two-slope

hybrid model described by Equations 6.2 and 6.3 is the formulation used for data synthesis in Chapter 6.

2. **Bias-aware data synthesis:** When generating synthetic RSS samples for underrepresented spatial sectors in Chapter 6, the corrected two-slope path-loss formulation is applied rather than the single-slope regression. This ensures that synthesised near-field samples are not artificially inflated and far-field samples are not artificially suppressed, preserving the physical plausibility of the augmented dataset. The KDE divergence metrics reported in Table 6.1 — with values below 1.71×10^{-4} across all quadrants — confirm that synthetic distributions remain statistically consistent with real measurements, validating the effectiveness of the bias correction at the data synthesis stage.

The robustness of the downstream localisation framework to residual channel variability is further demonstrated through controlled packet-loss sensitivity analysis and noise perturbation experiments reported in Section 6.6. The model maintains graceful performance degradation under packet loss (67.7% accuracy at 10% loss; 58.2% at 30% loss) and stable accuracy of approximately 58% under Gaussian RSSI perturbations of up to ± 5 dB, as shown in the existing Figure 5.4, which visually confirms that the hybrid near- and far-field formulation substantially reduces RMSE compared to the single-slope model across all distance bins. Inference latency remains below 0.15 ms per sequence for practical sequence lengths ($L \geq 3$), confirming real-time deployment feasibility on resource-constrained LoRaWAN gateways.

5.8 Summary

This chapter presented the development, analysis, and validation of a data-driven path-loss model tailored to LoRaWAN propagation in the urban environment of Birmingham. The work began by revisiting the theoretical foundations of path-loss modeling, emphasizing the influence of distance, environmental topology, and system parameters on signal attenuation. The initial regression-based model was derived from real-world experimental data, providing a first-order empirical representation of LoRa signal behavior across a dense city-center deployment.

Comparative analysis against classical and experimental reference models—Okumura–Hata, COST-231, Dortmund, and Oulu—demonstrated that the Birmingham model achieves a better fit to measured data, particularly in the near- and mid-field regions. The regression results revealed a path-loss exponent ($n = 2.52$) and intercept ($B = 130.85$ dB) consistent with realistic urban propagation conditions, while per-distance-bin evaluations exposed nonuniform attenuation trends, motivating adaptive segmentation.

To address this heterogeneity, clustering-based approaches were introduced. Both K-means and Distribution-Based Clustering (DBC) effectively partitioned the dataset into near- and far-field regions, improving model interpretability and predictive performance. The DBC method, in particular, offered a physically meaningful segmentation aligned with propagation behavior, identifying a natural transition around 200 m where the signal distribution shifts from binomial-like to normal-like.

Performance evaluation confirmed the statistical robustness of the hybrid near/far-field model. Compared to the single-slope regression, the hybrid approach reduced the root-mean-square error (RMSE) from approximately 18 dB to 5.5 dB in the near field and 16.8 dB in the far field. The mean absolute error also improved significantly relative to conventional empirical models, establishing the Birmingham model as a more accurate and site-specific representation of LoRaWAN propagation.

It is noted that the single-slope Birmingham regression model exhibits a systematic bias — overestimating RSSI in near-field distances ($d \leq 300$ m) and underestimating it in the far field ($d > 300$ m). As discussed in Section 5.7, this bias is explicitly addressed through two mechanisms before the model is applied in Chapter 6: clustering-based segmentation into near- and far-field regimes with statistically validated independent regression parameters ($t(746) = 6.48$, $p < 0.0001$ for K-means; $t(746) = 4.66$, $p < 0.0001$ for DBC), and bias-aware data synthesis using the corrected two-slope formulation. The localisation results reported in Chapter 6 therefore reflect the performance of the framework under this corrected propagation representation, and should be interpreted accordingly.

Since the path-loss model directly supports synthetic data generation for the localisation framework in Chapter 6, it is important to acknowledge that modelling assumptions and residual biases at this stage have the potential to propagate into the machine learning results. Three risks are identified and addressed: first, bias in synthetic RSS samples is mitigated by using the corrected two-slope hybrid model rather than the single-slope regression for synthesis, with KDE divergence values below 1.71×10^{-4} (Table 6.1) confirming distributional consistency between synthetic and real measurements; second, the risk of circularity — whereby the classifier learns patterns imposed by the path-loss model rather than genuine spatial signal structure — is assessed through an ablation study in Section 6.5, where evaluating the LSTM on the original dataset without synthetic samples yields 71.6% balanced accuracy compared to 72.1% on the combined dataset, confirming that augmentation improves training robustness without artificially inflating test performance; third, the Birmingham-specific nature of the model means that localisation results in Chapter 6 should be interpreted within this deployment context, with cross-environment generalisation identified as a direction for future work in Chapter 7.

In summary, this chapter demonstrated that while traditional models offer broad generalisations, localised modeling—supported by clustering and empirical calibration—provides markedly improved predictive accuracy and environmental realism. The proposed Birmingham path-loss model serves as a reliable foundation for subsequent localisation and machine learning analyses, forming a critical bridge between experimental measurement and data-driven inference explored in the following chapter.

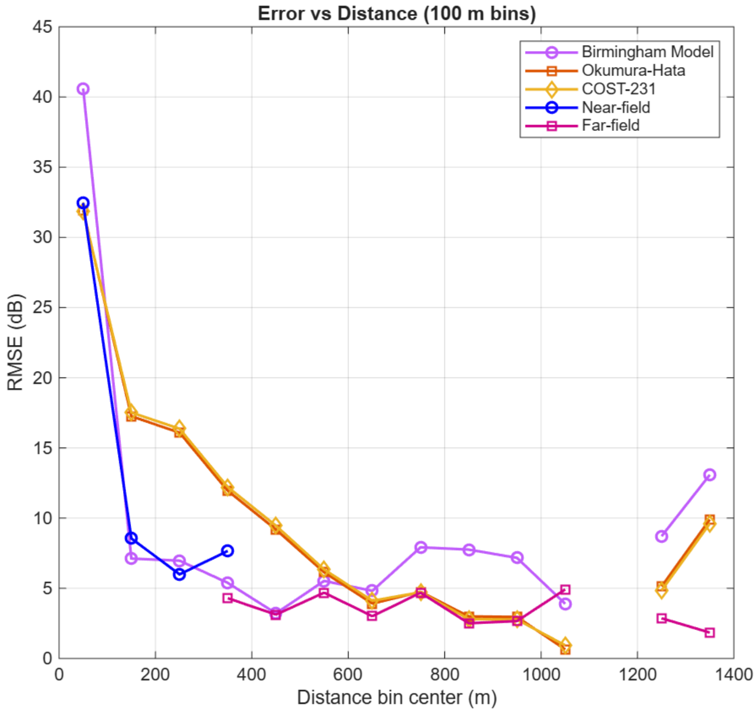


Figure 5.4: RMSE versus distance for 100 m bins, comparing the single-slope Birmingham regression model with Okumura-Hata and COST-231 reference models, and the hybrid near- and far-field formulations derived from clustering-based segmentation. The single-slope Birmingham model exhibits large residual variance particularly at short distances ($d \leq 200$ m), reflecting the systematic overestimation. The hybrid near-field model substantially reduces RMSE within the first 200 m, while the far-field model maintains stable accuracy at longer ranges. Okumura-Hata and COST-231 exhibit smoother error curves but systematically overestimate losses in the near-field region.

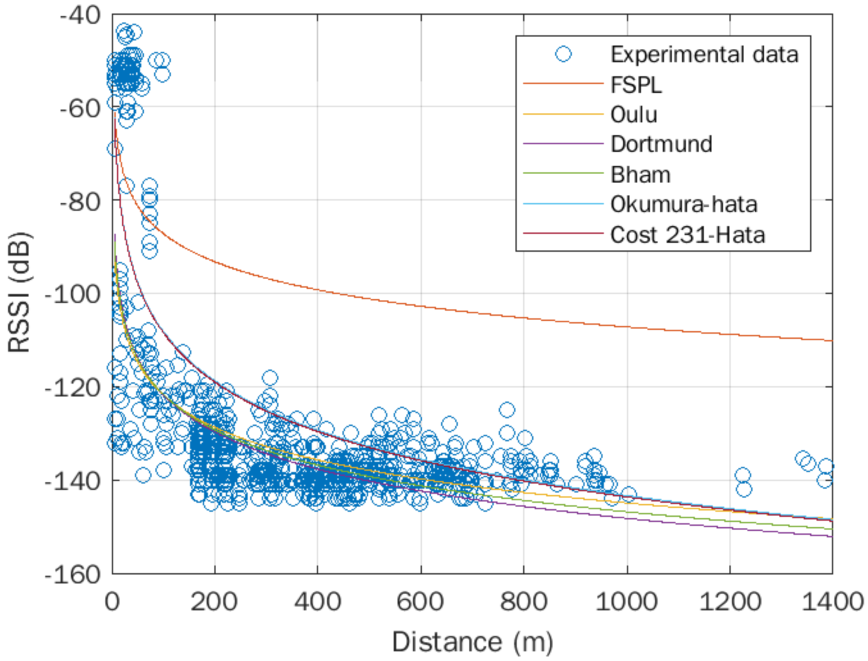


Figure 5.5: Comparing RSSI of different models with measurements in Bham for 868 MHz LoRa

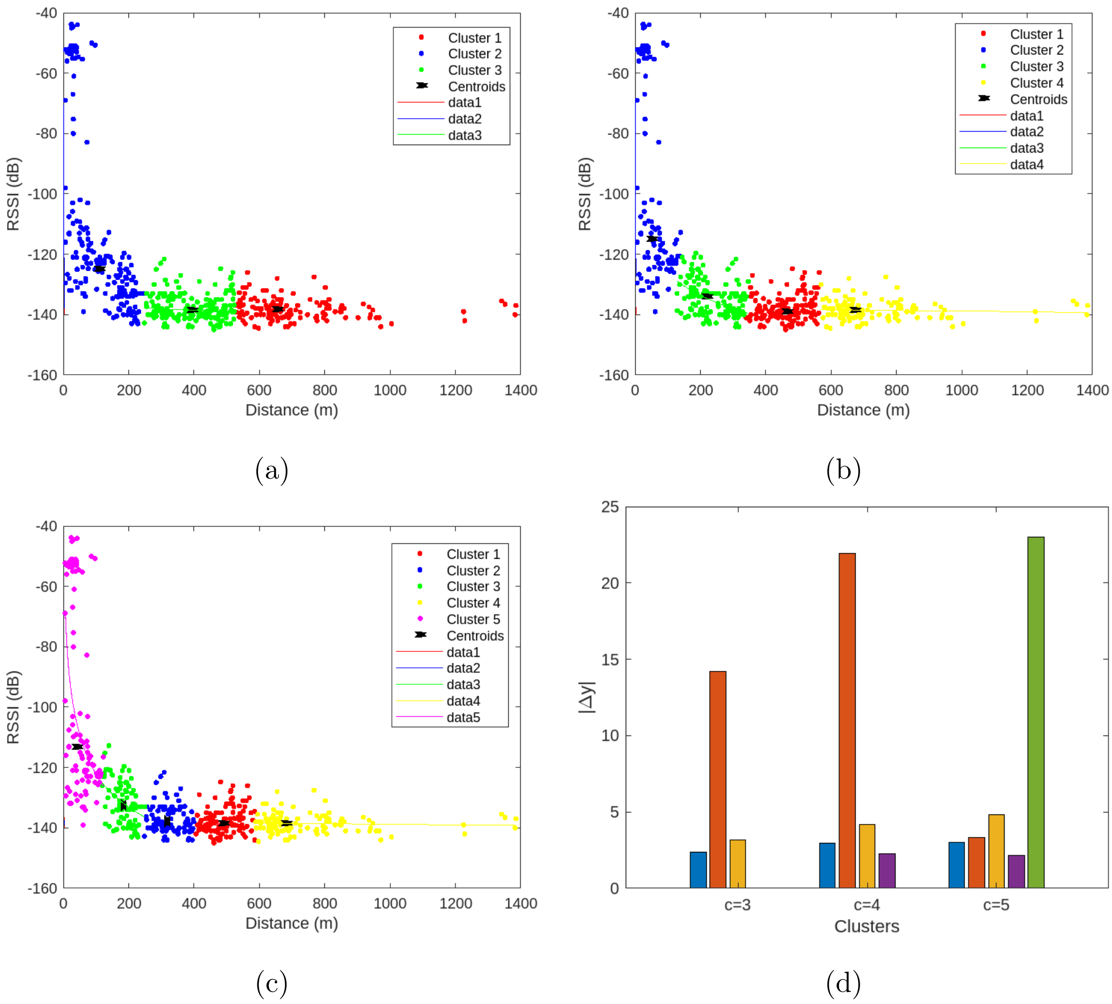


Figure 5.6: Cityblock clustering with (a) $c = 3$, (b) $c = 4$, (c) $c = 5$, and (d) mean absolute error for each configuration.

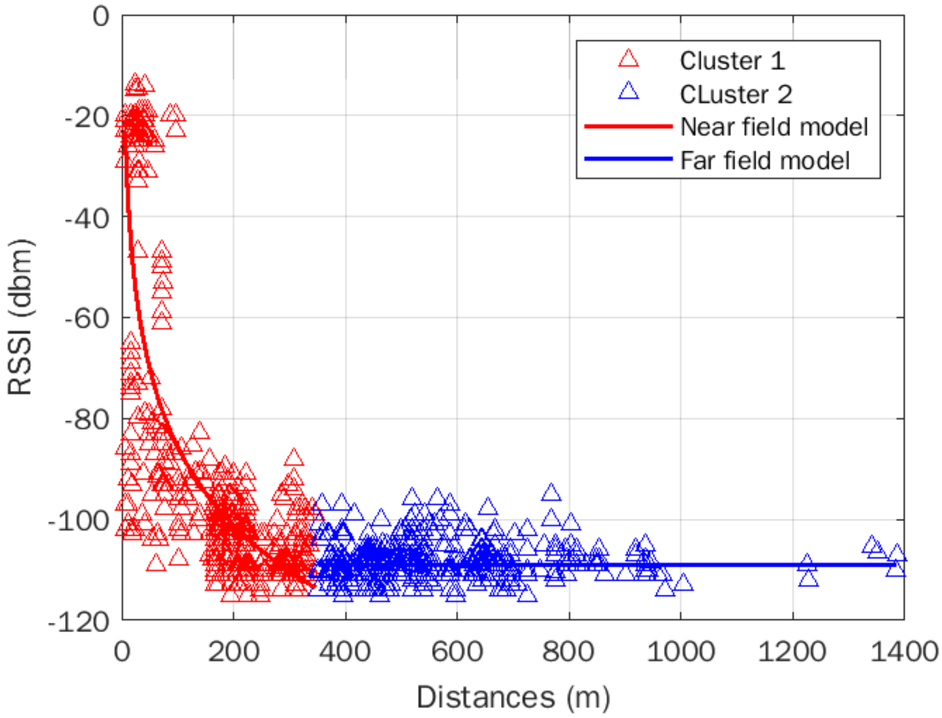
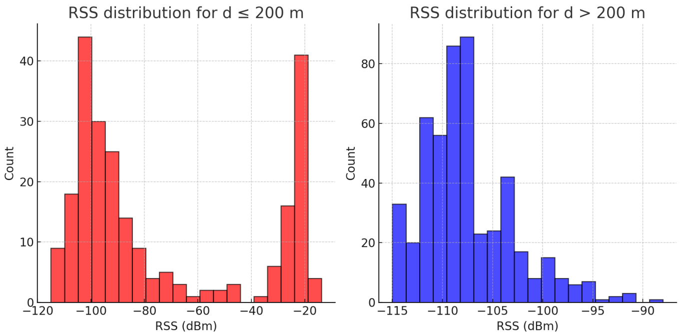
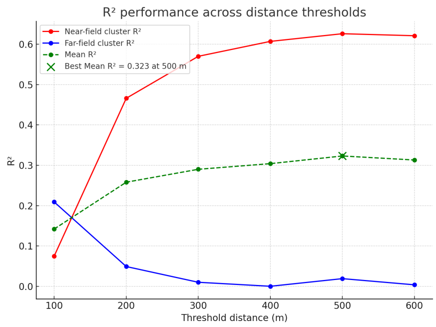


Figure 5.7: K-means clustered data showing regression fits for near- and far-field regions.



(a) RSSI distribution and identified 200 m threshold.



(b) R^2 variation for near/far-field clusters across thresholds.

Figure 5.8: Distribution-Based Clustering (DBC) evidence supporting the 200 m near/far-field segmentation boundary. Subfigure (a) shows the empirical RSSI distributions for measurements below and above 200 m, confirming a statistically distinct shift from binomial-like to normal-like behaviour at this distance, consistent with a transition from line-of-sight dominated to shadowing-dominated propagation.Subfigure (b) presents the R^2 performance of near- and far-field regression models across candidate threshold distances, showing that while thresholds of 300–500 m yield marginally higher average R^2.

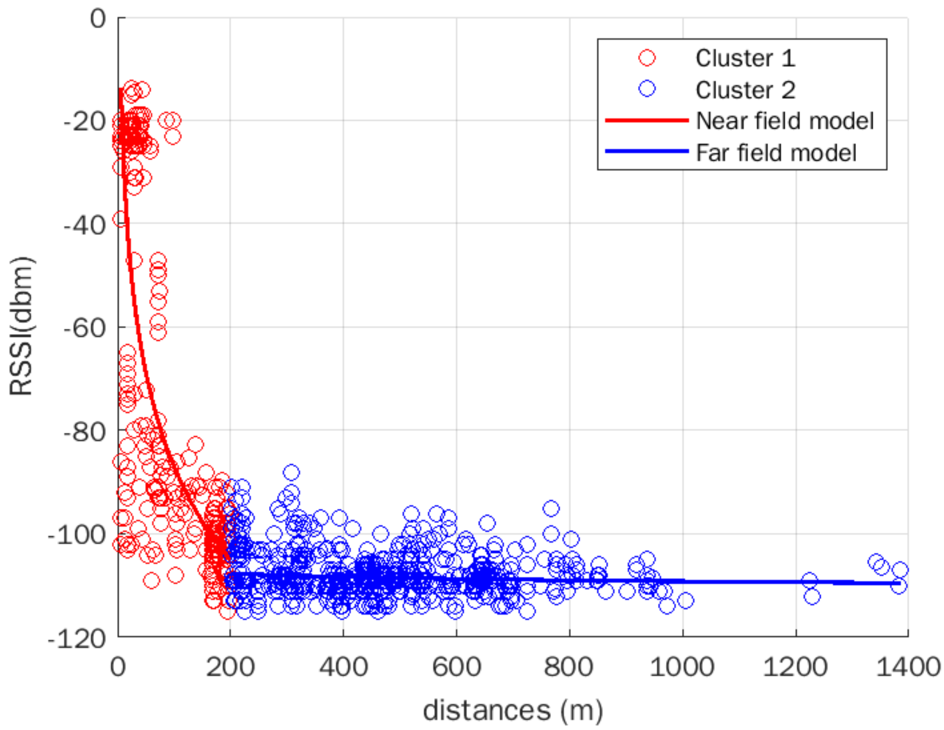


Figure 5.9: Distribution-based clustered data showing regression models for near- and far-field regions.

Chapter 6

Machine Learning-Based Framework for Localisation within LoRaWAN

6.1 Introduction

Building upon the previous chapter, which analyzed and modeled the propagation behavior of LoRaWAN signals in Birmingham, this chapter presents the development of a data-driven localisation framework based on a Long Short-Term Memory (LSTM) neural network. The proposed approach leverages the statistical insights and refined path-loss models derived earlier to enable spatial classification and location estimation from sequential LoRaWAN signal patterns. By integrating empirical propagation modeling with temporal deep learning, the framework aims to achieve accurate, low-complexity localisation in a single-gateway deployment scenario.

Figure 6.1 illustrates the overall framework of the proposed system. The process begins with experimental data collection through the deployed LoRaWAN testbed, where a mobile node transmits to a single gateway connected to a network server. The collected dataset undergoes pre-processing steps, including outlier elimination and regression-based path-loss modeling. A refined, K-means–based channel model further enhances data consistency before the processed features are passed to the LSTM network. Within the LSTM branch, the data are segmented into time windows, features such as RSSI and SNR are extracted and normalized, and the target outputs are one-hot encoded for multi-class classification. The trained LSTM model then performs quadrant-level prediction, estimating the likely spatial region of the transmitting node.

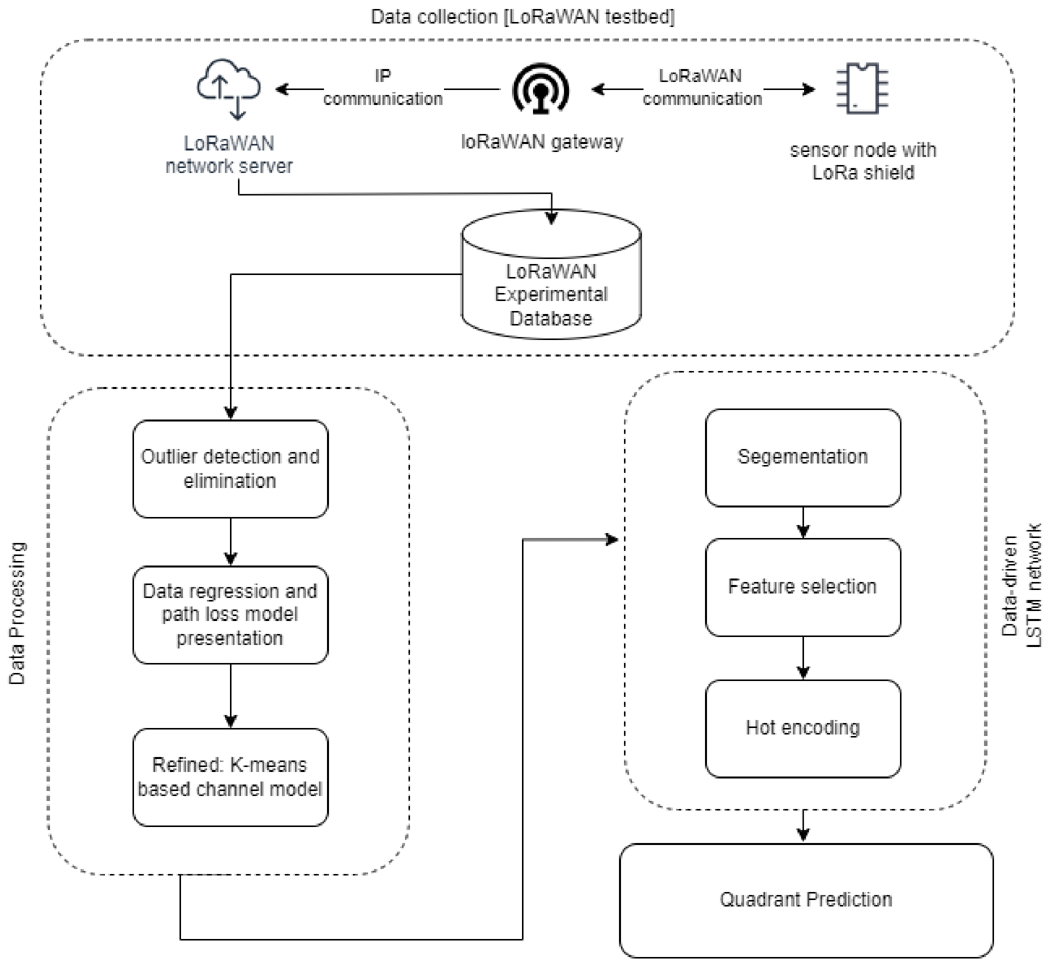


Figure 6.1: Overall framework of the proposed machine learning-based localisation system. The pipeline begins with LoRaWAN-based data acquisition, followed by data cleaning and regression-based path-loss modeling, and culminates in a data-driven LSTM network for quadrant-level location prediction.

Within the LSTM branch, the data are segmented into fixed-length sequences, key signal features such as RSSI and SNR are normalized, and the corresponding spatial sectors are one-hot encoded for multi-class classification. The network is trained to predict the most likely angular sector (quadrant) of the transmitting node based on the temporal evolution of received signals. Beyond baseline training, data augmentation and distribution-based synthesis techniques are employed to address class imbalance and enhance generalisation under sparse measurement conditions.

The remainder of this chapter details the full development and evaluation of the LSTM-based localisation framework. Section 6.2 describes data preprocessing, segmentation, and feature structuring. Section 6.3 presents the augmentation and synthesis strategies used to enrich the training dataset. The subsequent sections introduce the LSTM architecture, training configuration, and optimization procedures, followed by performance evaluation through cross-validation, robustness analysis, and comparative benchmarking against a CNN–LSTM hybrid model. The chapter concludes with a detailed discussion of network evaluation metrics, ablation studies, and the implications of synthetic data on model generalisation.

An article based on this chapter has been submitted for publication and is under second-round review as of May 2025, in the Springer Telecommunication Systems journal.

6.2 Data Preparation and Processing

As shown in the left branch of Figure 6.1, the data used for training and evaluation are derived from the experimental LoRaWAN testbed described in Chapter 4. The dataset represents transmissions from a mobile LoRa node to a single gateway receiver within a dense urban setting. The overall data preparation phase involves three main stages: (i) cleaning and organization of the raw measurements, (ii) regression-based path-loss modeling and refinement, and (iii) feature structuring for supervised training.

The data collected are divided into two main phases of operation: a *training phase*, during which experimental data are processed and organized into a structured feature set, and a *testing phase*, where the trained model is evaluated on unseen samples to assess predictive accuracy.

In the training phase, the following preprocessing operations were carried out:

1. **Outlier detection and elimination:** Removing inconsistent or spurious RSSI readings that result from transient channel conditions or device synchronization errors.

2. **Regression-based path-loss modeling:** Applying the refined Birmingham propagation model to normalize signal variations and extract representative distance-power relationships.

3. **Clustering-based refinement:** Incorporating the K-means–based segmentation approach developed in Chapter 5 to form homogeneous subgroups with consistent propagation characteristics.

The resulting dataset is then structured into sequences suitable for LSTM input. Each sequence represents a time-series vector of the form $[t, RSSI, SNR]$, denoted as a *User Data Record (UDR)*. The following definitions apply throughout this chapter:

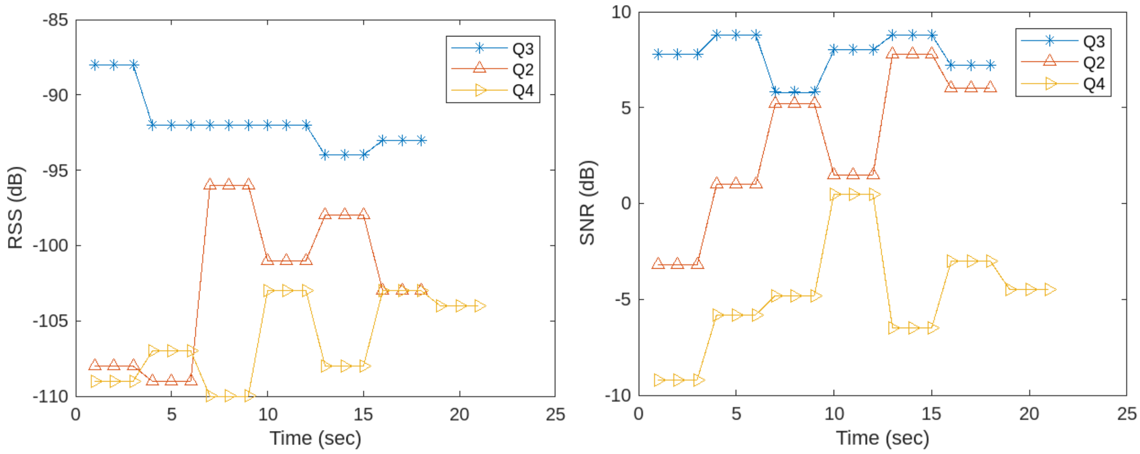
- **Predicted Sector (PS):** A spatial region within the coverage area characterised by similar propagation and environmental features. In this study, three and four-sector divisions are evaluated.

- **User Data Record (UDR):** A time-indexed vector $[t, RSSI, SNR]$ corresponding to the LoRa node's measurement at time t.

- **Segment:** A temporal grouping of consecutive UDRs within a fixed time interval (5 s, 10 s, or 15 s), serving as the basic time-series input to the LSTM.

- **One-hot encoding:** A categorical encoding scheme used to represent predicted sector classes as binary vectors (e.g., 001, 010, 100).

This structured representation enables the network to learn temporal dependencies between consecutive signal variations rather than isolated readings. Figure 6.2 illustrates the variation of RSSI and SNR values toward the gateway across different sectors, demonstrating clear directional signal patterns that inform the classification task.



(a) RSS variation toward the gateway.

(b) SNR variation toward the gateway.

Figure 6.2: Signal variations toward the gateway for far-field distances.

6.3 Data Augmentation and Synthesis

Acquiring extensive, uniformly distributed data in real-world LoRaWAN deployments is often infeasible due to environmental and logistical constraints. To mitigate data imbalance and enhance model generalisation, two complementary techniques were applied: *data augmentation* (noise-based expansion of existing samples) and *data synthesis* (generation of physically consistent synthetic samples). Both techniques aim to increase data diversity and balance underrepresented sectors while preserving the physical plausibility of the signal behavior.

Before describing the data preparation and augmentation procedure, it is useful to clarify the precise functional role of the path-loss model derived in Chapter 5 within the machine learning pipeline of this chapter. The path-loss model serves three distinct and complementary roles:

1. **Synthetic data generation:** The primary functional role of the path-loss model within this pipeline is to generate physically consistent synthetic RSS samples for underrepresented spatial sectors. Rather than relying on arbitrary statistical oversampling, the corrected two-slope Birmingham model (Equation 6.2) is used to synthesise RSSI values that reflect the empirically validated propagation characteristics of the deployment environment, directly addressing class imbalance in the training dataset.

2. **Feature shaping:** The path-loss model contributes channel-aware distance estimates as additional input features to the LSTM, enriching the temporal signal representation beyond raw RSSI and SNR values. These model-derived features capture large-scale distance-dependent attenuation trends that complement the packet-level signal measurements, improving the discriminative power of the learned temporal representations across sectors.

3. **Data consistency refinement:** By grounding both real and synthetic samples in the same empirical propagation model, the path-loss framework ensures statistical consistency across the training dataset. KDE divergence values below 1.71×10^{-4} (Table 6.1) confirm that synthetic distributions remain statistically consistent with real measurements, validating the integrity of the combined dataset used for LSTM training.

The path-loss model therefore does not serve merely as conceptual justification for the machine learning approach; it is an active, functional component of the pipeline at the data preparation stage. The LSTM architecture itself, described in Section 6.4, operates on the enriched and balanced dataset produced by this integration.

6.3.1 Noise-Based Data Augmentation

Data augmentation introduces controlled variability to existing records, allowing the model to learn from realistic fluctuations. This is achieved by adding Gaussian noise to the RSSI and SNR features of real measurements, simulating the inherent uncertainty caused by device precision, environmental reflections, and GPS localisation errors. The jittering process is expressed as:

$$x' = x_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{6.1}$$

where ϵ_t denotes Gaussian noise with standard deviation σ derived from stationary measurement variance in the dataset. This preserves the original signal distribution while increasing the dataset's size and robustness against overfitting.

The Gaussian distribution is a well-established and physically justified model for residual signal fluctuations in urban radio environments. In LoRaWAN deployments, the dominant sources of short-term RSSI variability — including thermal noise at the receiver, small-scale multipath fading averaged across the spread-spectrum bandwidth, and minor environmental reflections — are each independently distributed and additive in nature. By the central limit theorem, the aggregate effect of these independent sources converges to an approximately Gaussian distribution (Rappaport, 2010; Molisch, 2012). This is further supported empirically by the log-normal shadowing model validated in Chapter 5, where residual analysis confirmed that the path-loss residuals in both near- and far-field regimes follow approximately normal distributions (Table 5.4). The standard deviation σ used in Equation 6.1 is derived directly from the stationary measurement variance observed in the collected dataset, ensuring that the injected noise is statistically consistent with the real-world signal fluctuations present in the Birmingham deployment environment rather than being an arbitrary perturbation magnitude.

6.3.2 Model-Based Data Synthesis

To further balance the dataset, additional synthetic samples were generated using the empirically derived Birmingham path-loss model. For each new point, a random offset of ± 5 m was applied to a subset of existing measurement locations, and the corresponding RSSI values were computed from the two-slope path-loss formulation:

$$\text{Near-field } (d \leq 346 \text{ m}): \quad RSS = -137.33 - 51.77 \log_{10}(d), \tag{6.2}$$

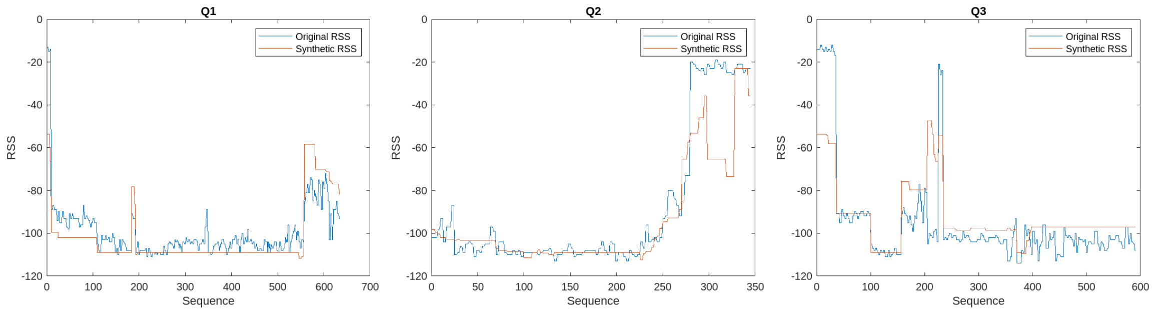
$$\text{Far-field } (d > 346 \text{ m}): \quad RSS = -108.99 + k \log_{10}(d), \tag{6.3}$$

where d represents the transmitter–gateway distance in meters, and k is an empirically calibrated constant. This ensures the synthesized samples follow the same propagation trend observed experimentally while expanding spatial coverage.

Figure 6.3 shows the comparison between original and synthesized RSSI distributions for three directional quadrants, illustrating their statistical alignment.

6.3.3 Validation of Synthetic Data

To verify that the generated data preserve the statistical properties of real measurements, Kernel Density Estimation (KDE) was used to compare the distributions of original and synthetic RSSI values. Table 6.1 reports the mean squared error (MSE) divergence between



(a) Quadrant Q1 (b) Quadrant Q2 (c) Quadrant Q3

Figure 6.3: Variation of original and synthetic RSS values across quadrants.

real and synthesized data for each quadrant. Low divergence values confirm that the synthesized signals follow consistent statistical behavior, ensuring that the expanded dataset remains physically valid for LSTM training.

Table 6.1: KDE divergence metrics between real and synthetic RSS distributions.

Quadrant	**MSE Divergence**
Q1	2.04×10^{-5}
Q2	1.71×10^{-4}
Q3	1.44×10^{-4}

With the augmented and synthesized dataset validated, the subsequent section introduces the architecture and training specifications of the proposed LSTM model, detailing the sequential input configuration, network structure, and optimization parameters used for localisation performance evaluation.

6.4 Proposed LSTM Architecture

The Long Short-Term Memory (LSTM) network proposed in this study is designed to exploit the temporal correlations within sequential LoRaWAN measurements, enabling the estimation of a node's location using only signal-level information (RSSI and SNR) without requiring GPS data. The model was implemented in MATLAB's Deep Learning Toolbox and optimized for lightweight, edge-compatible inference.

Before describing the proposed architecture, it is useful to briefly compare LSTM with the most relevant alternative sequential models; specifically Gated Recurrent Units (GRU) and simple Recurrent Neural Networks (RNN), to justify the architectural selection on conceptual as well as empirical grounds.

Simple RNNs, while capable of modelling short-term temporal dependencies, suffer from the well-documented vanishing gradient problem (Bengio, Simard, and Frasconi, 1994), which prevents them from retaining information across longer packet sequences. In the single-gateway LoRaWAN setting, where spatial disambiguation relies on subtle signal evolution across multiple consecutive packets, this limitation is particularly problematic, short-term memory is insufficient to capture the gradual attenuation trends that distinguish sectors at longer ranges.

GRUs address the vanishing gradient problem through a simplified gating mechanism comprising a reset gate and an update gate, compared to LSTM's three-gate architecture (input, forget, and output gates). GRUs are computationally lighter, reducing trainable parameters by approximately 25% relative to an equivalent LSTM, and have been shown to perform comparably to LSTMs on shorter sequences or datasets with limited training samples (Chung et al., 2014; Greff et al., 2016). However, LSTM's explicit forget gate provides a more selective memory mechanism that is particularly valuable in noisy, variable-length packet streams typical of LoRaWAN deployments, where transient interference and packet loss create irregular temporal patterns that benefit from independent control of what information to retain and what to discard. Furthermore, the proposed architecture already employs dropout regularisation and early stopping to manage overfitting (A. ElSabaa, Guéniat, and Wenyan Wu, 2026), mitigating the primary practical advantage of GRU's reduced parameter count. The sequence-length ablation study confirms that performance improves substantially with longer temporal context (57.9% at $L = 1$ to 72.5% at $L = 8$), suggesting that the longer-range memory capability of LSTM is genuinely exploited in this setting rather than being redundant.

Transformer-based models and graph neural networks, while state-of-the-art for large-scale sequence tasks, are better suited to multi-gateway settings or very large datasets where long-range attention mechanisms or spatial graph structures can be fully leveraged. In the single-gateway scenario with a carefully curated but limited experimental dataset, introducing such architectures would risk being misleading due to data scale constraints (A. ElSabaa, Guéniat, and Wenyan Wu, 2026). LSTM therefore represents the most appropriate balance between temporal modelling capability, computational efficiency, and suitability to the data scale of this deployment.

6.4.1 Dataset and Input Configuration

The dataset used for training and evaluation was collected over seven independent LoRaWAN measurement campaigns conducted in Birmingham city center, using a single mobile LoRa transmitter and one fixed gateway. Each campaign produced comma-separated (CSV) files containing RSSI, SNR, and GPS coordinates, among other metadata. After preprocessing—which included removal of missing entries (NaNs), conversion of RSSI from dBm to dB, and segmentation into 5-, 10-, and 15-second intervals—the cleaned dataset comprised

2,434 usable samples across seven features: index, time, RSSI, SNR, longitude, latitude, and distance.

Outlier detection was performed using a ±3 MAD (median absolute deviation) threshold applied to the RSSI values to eliminate anomalous readings caused by transient interference. The processed dataset was divided into training (80%) and testing (20%) subsets, a split later validated through 5-fold cross-validation to confirm stability. Each input sequence represents a time-series communication segment (or window) containing five features: [time, RSSI, SNR, SNR^2], and estimated distance. These sequences are zero-padded to a uniform length, and mini-batch training with size 24 is employed to minimize padding overhead.

6.4.2 Sector Division and Problem Formulation

Because this study relies on a single gateway, determining exact Cartesian coordinates of the transmitter is infeasible. Instead, a spatial classification approach is adopted, dividing the coverage map into three directional sectors (Q1, Q2, Q3) defined by angular orientation relative to the gateway. This division captures distinct signal propagation patterns while maintaining sufficient data per class for stable learning. The output is represented as a one-hot encoded vector (e.g., {100, 010, 001}) corresponding to the predicted sector.

6.4.3 Network Architecture

The architecture of the proposed data-driven LSTM model is designed as a multi-stage sequential framework that progressively encodes temporal dependencies, refines latent representations, and outputs class probabilities for spatial sector prediction. The key components are as follows:

- **Sequence Input Layer:** Accepts input feature vectors of dimension five (time, RSSI, SNR, SNR^2, and estimated distance) at each time step, preparing the data for temporal processing.

- **LSTM Layer:** The first recurrent block is a unidirectional LSTM that captures short-term temporal dependencies across successive LoRa packets. The output is a sequence of encoded temporal features, regularized with a 30% dropout rate to mitigate overfitting.

- **Bidirectional LSTM (BiLSTM):** To capture contextual dependencies in both forward and backward directions, a BiLSTM layer follows the initial LSTM. This dual-directional processing ensures that each feature representation benefits from complete temporal context. A dropout of 0.4 is applied to enhance generalisation.

- **Dense and Normalization Block:** A fully connected layer with 64 neurons aggregates temporal embeddings into a compact representation. It is followed by a ReLU activation, batch normalization, and an additional 30% dropout layer. This combination accelerates convergence, prevents internal covariate shift, and improves robustness.

- **Output and Classification:** A final fully connected layer maps the learned features to the three target classes (sectors). A softmax activation then converts these outputs into probability distributions, and the classification layer computes categorical cross-entropy loss for training.

This hybrid recurrent structure—combining a stacked LSTM, a bidirectional LSTM, and dense normalization—balances representational power and generalisation. It enables the model to capture both local signal variations and long-term propagation trends, making it well-suited for LoRaWAN-based localisation.

6.4.4 Training Configuration and Optimization

The model was trained using the Adam optimizer, which provides adaptive learning-rate updates suitable for non-stationary datasets. The loss function is categorical cross-entropy, defined as:

$$L(p, t) = -\sum_{n=1}^{M} y_{o,c} \log(p_{o,c}), \tag{6.4}$$

where M is the number of classes, and $y_{o,c}$ and $p_{o,c}$ denote the true and predicted labels for observation o in class c, respectively K. L. Tan et al., 2022. Table 6.2 lists the tested and optimal hyperparameter configurations.

Table 6.2: Hyperparameter settings for the proposed LSTM model.

Parameter	**Tested Values**	**Optimal Value**
Model Type	Sequential	–
Batch Size	24	24
Max Epochs	(75, 100, 125)	100
Optimizer	(Adam, SGDM, RMSProp)	Adam
Learning Rate	(0.001, 0.0001, 0.00001)	0.001
Activation Function	Sigmoid / ReLU	ReLU
Train/Test Split	80% / 20%	–

The input dataset comprised 405 real experimental sequences (across different segment durations) and 300 augmented sequences generated via Gaussian perturbation. Approximately 80% of these were used for model training, with the remaining 20% reserved for evaluation. The output is a one-hot encoded vector corresponding to the predicted quadrant.

6.4.5 Computational Efficiency

The proposed BiLSTM model contains approximately 63,000 trainable parameters, corresponding to a memory footprint of $\approx$0.25 MB in 32-bit floating-point representation. Inference requires approximately 54,000 multiplications per time step; for a typical sequence of 100 packets, this equals 5.4 million operations ($<$0.01 GFLOPs). Such computational demands are modest, enabling real-time inference on LoRaWAN gateways or embedded microcontrollers equipped with low-power CPUs. For ultra-constrained devices, lighter alternatives such as Gated Recurrent Units (GRUs), model pruning, or quantization could further reduce memory and power requirements without substantial accuracy degradation.

6.5 Results and Validation

This section presents the experimental results of the proposed data-driven Long Short-Term Memory (LSTM) model and its hybrid variant (CNN-LSTM). The analysis includes model tuning, validation through cross-validation, overfitting assessment, and comparative evaluation of network architectures. The results collectively demonstrate the robustness and generalisation capability of the proposed localisation framework under realistic LoRaWAN conditions.

Before presenting the evaluation results, it is important to contextualise the scope of the localisation performance reported in this chapter. The proposed framework performs **sector-level classification** — predicting which of three angular quadrants a transmitting node occupies relative to a single gateway — rather than estimating precise Cartesian coordinates. The reported classification accuracy, balanced accuracy, and F1-scores should therefore be interpreted as measures of coarse-grained spatial inference under a single-gateway constraint, not as indicators of metre-level positioning precision. This scope is a deliberate design choice consistent with the practical constraints of LoRaWAN deployments, where fine-grained coordinate estimation from a single signal source is neither feasible nor the objective of this work.

6.5.1 Experimental Setup and Hyperparameter Optimization

Prior to training, a comprehensive tuning procedure was carried out to identify the optimal combination of hyperparameters for the proposed LSTM network. The parameters explored include the number of epochs, hidden-layer neuron configurations, optimizer type, and learning rate. The effect of layer configuration and epoch number on model accuracy is summarized in Table 6.3.

Table 6.3: Effect of varying the number of layers and epochs (Adam optimizer, learning rate = 0.001).

	Model Training and Validation Accuracy (%) [Fullyconnected, BiLSTM, Fullyconnected]		
Number of Epochs	[50,100,50]	[100,100,100]	[50,50,50]
75	85.9 %, 60.1%	85.9%, 69.4%	75.2%,56.3%
100	**83.8%, 72.1%**	88.2%, 64.8%	84.4%,61.5%
125	84.8% , 62.1%	85.6% , 61.0%	87.2% , 61.5%

Among the optimizers tested—Stochastic Gradient Descent with Momentum (SGDM), Root Mean Square Propagation (RMSProp), and Adaptive Moment Estimation (Adam)—the Adam optimizer achieved the highest testing accuracy of **83.8%**, outperforming RMSProp (77.9%) and SGDM (56.2%). Adam's adaptive learning-rate mechanism and momentum tracking enabled stable convergence under noisy signal conditions typical of LoRaWAN environments.

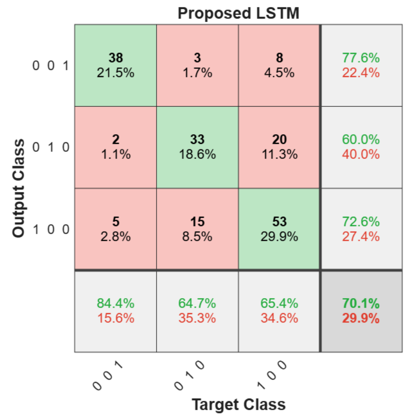
The learning rate was also varied between 0.001, 0.0001, and 0.00001. The rate of 0.001 consistently achieved the best balance between convergence speed and loss stability. Lower learning rates resulted in slower convergence without significant gains in accuracy. The optimal configuration employed 100 training epochs, as increasing beyond this threshold did not yield measurable improvement in validation accuracy.

6.5.2 Cross-Validation and Generalisation Performance

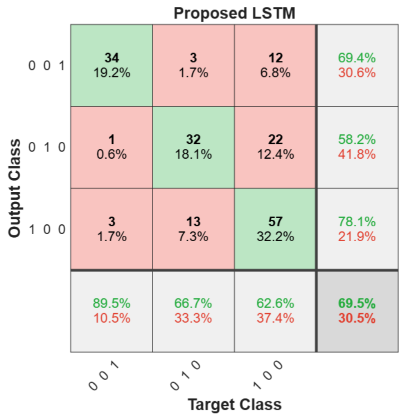
In our initial split-based evaluation, the proposed model achieved an overall accuracy of 72.1%. To address the possibility of train–test leakage, we employed a leave-one-group-out evaluation strategy. Since the dataset was aggregated across multiple measurement campaigns, explicit trajectory/day identifiers were not retained. Instead, we partitioned the dataset into five non-overlapping groups, each containing distinct measurement sequences. In each fold, one group was completely excluded for testing, while the remaining four groups were used for training and validation. This guarantees that no packets from the same group appear in both training and testing. Figure 6.4 illustrates the classification results across the five folds.

Under this stricter validation protocol, the proposed LSTM achieved an average accuracy of **68.8% ± 2.2**, balanced accuracy of **76.9% ± 1.6**, and macro-F1 score of **68.8% ± 1.9**. The slight reduction from the initial split-based accuracy (72.1%) reflects the improved separation between training and testing sets and confirms the model's ability to generalise across independent measurement campaigns.

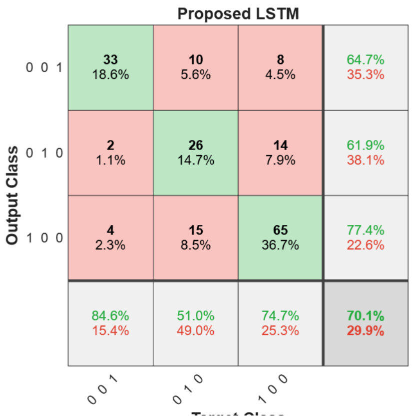
The confusion matrix (Table 6.4) indicates that the model performs most accurately in Quadrant 3 (Q3), which corresponds to the region with consistent line-of-sight connectivity



(a) 1st fold



(b) 2nd fold



(c) 3rd fold

(d) 4th fold

(e) 5th fold

Figure 6.4: Leave-one-group-out analysis using $k = 5$ folds. Each confusion matrix shows the per-sector classification performance for the held-out test group, with rows representing true labels and columns representing predicted labels.

Table 6.4: Mean confusion matrix across 5 folds (values are mean ± std).

	Pred Q1	**Pred Q2**	**Pred Q3**
True Q1	34.2 ± 2.7	2.8 ± 1.7	4.4 ± 0.8
True Q2	6.0 ± 2.7	31.6 ± 2.9	16.2 ± 3.5
True Q3	7.8 ± 2.9	18.0 ± 2.8	56.0 ± 7.2

and stronger SNR values. Misclassifications occur primarily between Q2 and Q3, likely due to overlapping signal features and multipath reflections. Overall, these results validate the model's reliability and its ability to capture spatio-temporal relationships from single-gateway LoRaWAN data.

6.5.3 Training Dynamics and Overfitting Analysis

The training and validation accuracies across iterations are shown in Fig. 6.5. The model exhibits a steady increase in accuracy during training, reaching a stable plateau near 85% training accuracy and 72% validation accuracy. The validation trend (black markers) closely follows the training trajectory (blue line), confirming that the model generalises well without overfitting. The absence of large oscillations or divergence between the curves indicates consistent convergence behavior across epochs.

The model demonstrates consistent generalisation, as the validation performance (black markers) closely follows the training accuracy trend (blue line). Despite minor oscillations typical of sequential data learning, the two curves converge smoothly toward stable performance, indicating that the model effectively captures the temporal dependencies without memorizing noise. No evidence of underfitting (high loss, low accuracy) or overfitting (large gap between training and validation accuracy) is observed. This stability is attributed

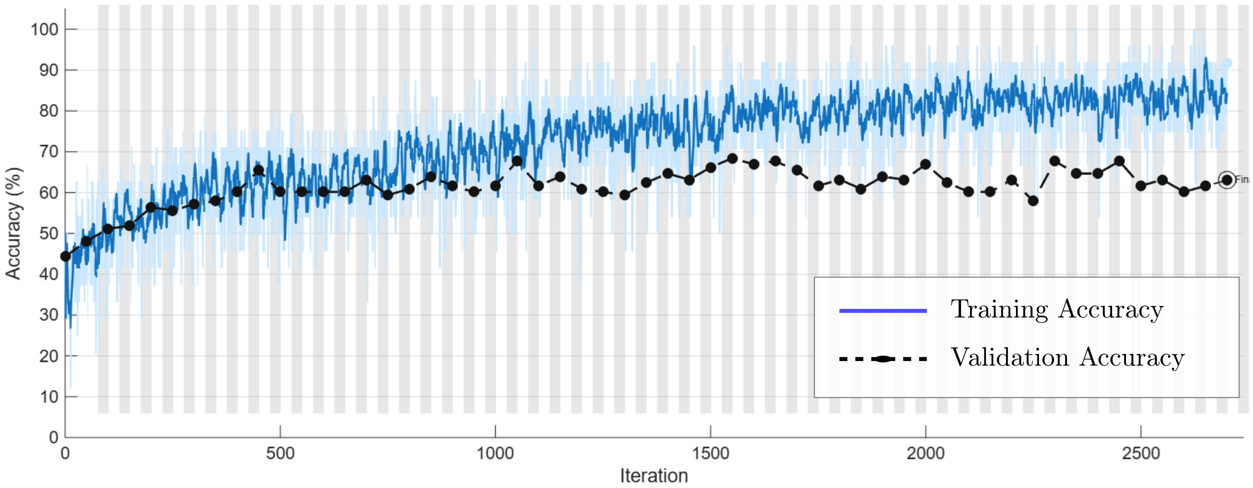


Figure 6.5: Training and validation accuracy versus iteration for the proposed LSTM model. The solid blue trace represents training accuracy, which increases progressively and stabilises beyond approximately 1000 iterations. The dashed black trace with markers represents validation accuracy, which plateaus in the range of 60–65%, confirming stable convergence and the absence of significant overfitting. Early stopping was applied once the validation plateau was reached.

to the combined use of dropout, batch normalization, and data augmentation, which collectively enhance robustness against noise and class imbalance in LoRaWAN data.

6.5.4 Comparative Evaluation: CNN-LSTM vs. Data-Driven LSTM

Convolutional Neural Networks (CNNs) are a highly effective class of deep learning architectures widely applied across domains such as image classification, object detection, and speech recognition (Shiri et al., 2023). Their ability to automatically extract hierarchical features from raw inputs has also made them valuable in time-series prediction tasks (Qin, N. Yu, and Zhao, 2018). Because the performance of sequential neural models strongly depends on the quality of extracted features (Gamboa, 2017), CNNs are often integrated with recurrent models such as Long Short-Term Memory (LSTM) networks to combine spatial feature extraction with temporal sequence modeling (Längkvist, 2014). In such hybrid designs, CNN layers detect local temporal patterns, while LSTM layers capture long-term dependencies across time.

To establish a meaningful baseline for comparison, we implemented a hybrid CNN–LSTM model using the same LoRaWAN dataset. As illustrated in Figure 6.6, the architecture begins with a one-dimensional convolutional block consisting of convolution, ReLU activation, and dropout layers, which extract localised temporal features from the input signal while reducing overfitting. These extracted features are passed to an LSTM layer that models sequential dependencies, followed by a fully connected layer and a softmax classifier to produce final class probabilities. This design represents a well-established deep learning paradigm for time-series classification, allowing direct benchmarking against the proposed data-driven LSTM model.

Quantitatively, the proposed LSTM model achieved superior performance, obtaining

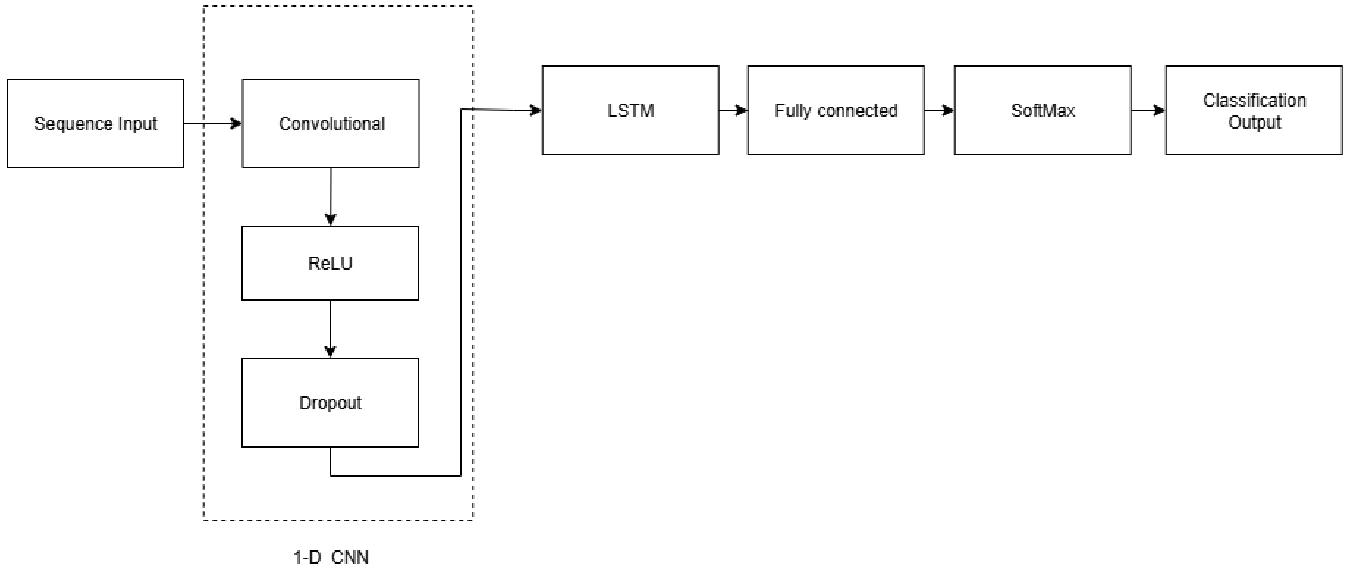


Figure 6.6: 1-D CNN–LSTM hybrid architecture combining convolutional feature extraction with sequential modeling for localisation.

an overall accuracy of 72.1% and a macro-F1 score of 0.69, compared to 65% and 0.64 for the CNN–LSTM. Moreover, statistical analysis using McNemar's test on paired predictions confirmed that this difference was not due to random variation. The resulting chi-squared statistic of $\chi^2 = 9.26$ with a p-value of $p = 0.0023$ indicates that the improvement of the proposed LSTM over the CNN–LSTM baseline is statistically significant at the 95% confidence level.

Therefore, while both architectures demonstrate strong localisation capability, the proposed standalone LSTM offers a better balance between accuracy, generalisation, and computational efficiency—making it more suitable for real-time or resource-constrained LoRaWAN gateway deployments.

6.6 Network Evaluation

This section provides a comprehensive evaluation of the proposed LSTM-based localisation framework. Multiple aspects are examined, including quantitative performance metrics, sector sensitivity, statistical reliability, angular and spatial consistency, noise robustness, and the effect of data augmentation. Together, these analyses establish both the accuracy and resilience of the proposed approach under diverse conditions.

6.6.1 Definition of Performance Metrics

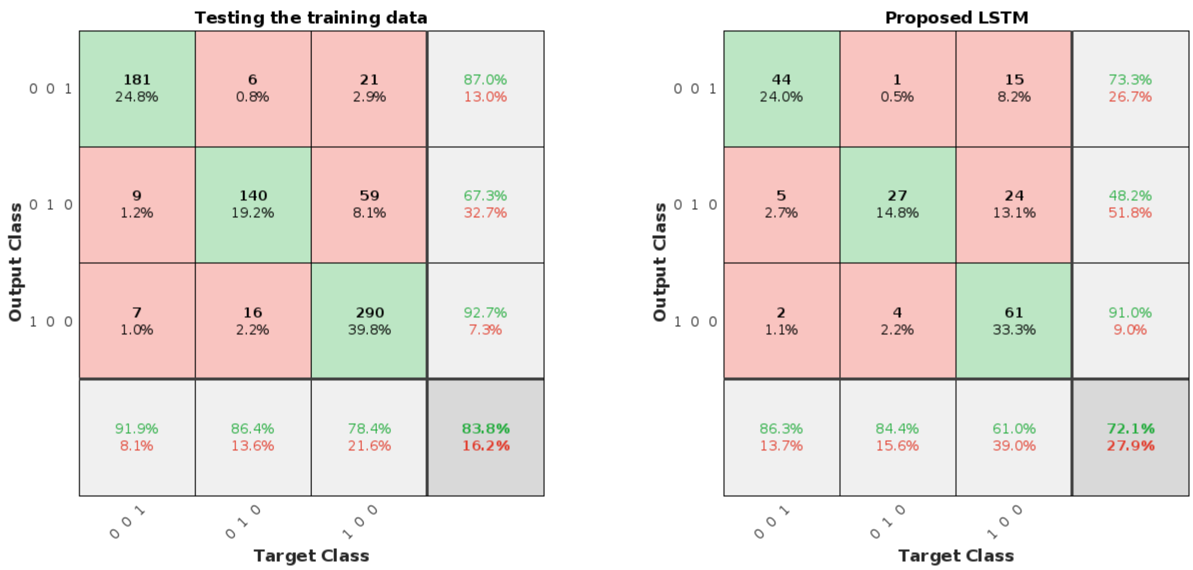
The network prediction results are represented using confusion matrices M, where each class corresponds to a spatial sector of the deployment area (Q1 $\rightarrow$ [001], Q2 $\rightarrow$ [010], Q3 $\rightarrow$ [100]). From M, four key quantities are derived for each class: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Based on these values, several common performance indicators are calculated (Demir, 2022):

$$\begin{aligned}
ACC &= \frac{TP + TN}{TP + TN + FP + FN} \\
Recall &= \frac{TP}{TP + FN} \\
Precision &= \frac{TP}{TP + FP} \\
F1 &= \frac{2 \times TP}{2 \times TP + FN + FP}
\end{aligned}$$

These metrics collectively quantify the model's correctness (accuracy), sensitivity (recall), reliability of positive predictions (precision), and overall balance between the two (F1-score).

6.6.2 Training and Testing Performance

The classification performance of the proposed LSTM is evaluated using both training and testing datasets, as illustrated in Fig. 6.7. Each matrix cell indicates the number and proportion of correctly or incorrectly classified segments.



(a) Training data (b) Testing data

Figure 6.7: Confusion matrices representing the performance of the proposed LSTM model on (a) training and (b) testing datasets.

As shown in Fig. 6.7(a), the LSTM achieved a mean training accuracy of 83.8%. Across classes, Q1 reached 87.0%, Q2 achieved 67.3%, and Q3 attained 92.7%. When evaluated on unseen data (Fig. 6.7(b)), 73.3% of Q1, 50% of Q2, and 91% of Q3 segments were correctly classified, yielding an overall accuracy of 72.1%. The most common confusion occurred between Q2 and adjacent sectors, consistent with multipath reflections and partial line-of-sight overlap observed in the dataset.

6.6.3 Model Comparison and Metric Evaluation

Figure 6.8 summarizes the F1-score, recall, and precision across the three sectors for both the data-driven LSTM and its hybrid CNN-LSTM variant. The proposed model achieved classification accuracies of 86.3%, 84.4%, and 61.0% for Q1, Q2, and Q3, respectively, outperforming the CNN-LSTM by 5–10% across most metrics. Specifically, recall improved from 68.3%, 41.1%, and 88.6% (CNN-LSTM) to 73.3%, 48.2%, and 91.0% for the proposed model. The F1-score also increased by 6% (Q1), 9% (Q2), and 3% (Q3). These results confirm that the proposed LSTM model achieves superior generalisation and stability in LoRaWAN-based localisation tasks.

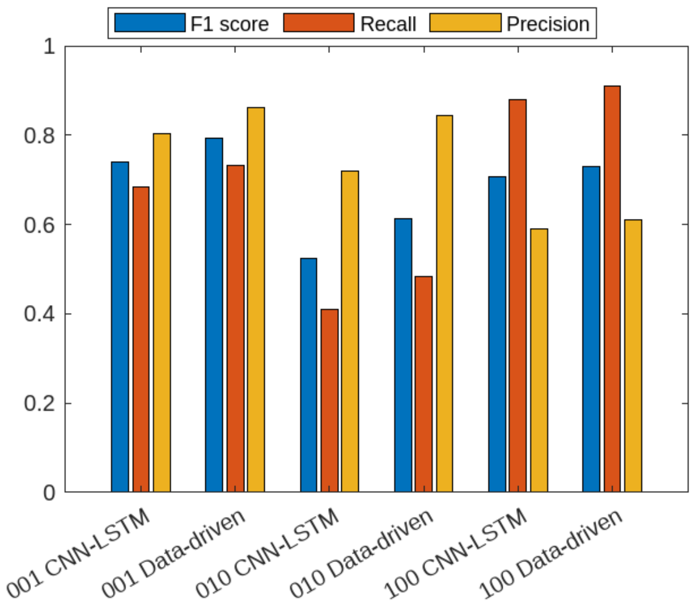


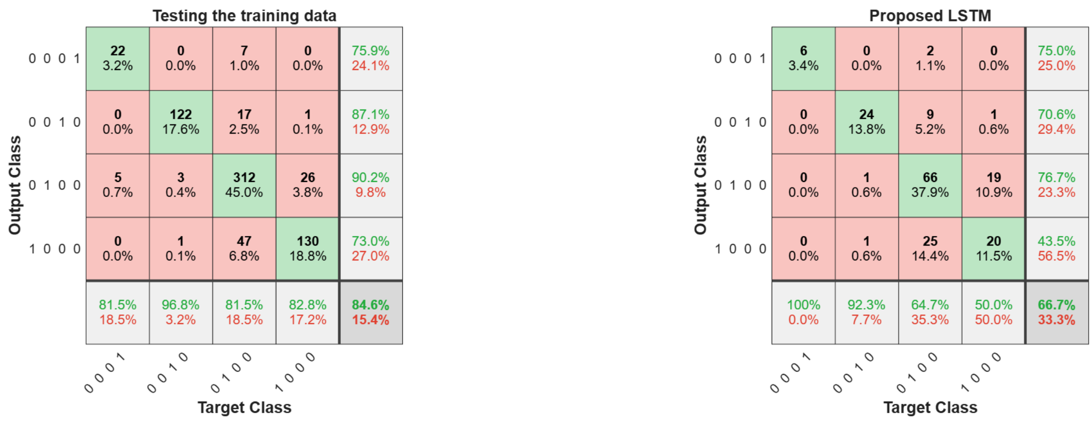
Figure 6.8: Evaluation metrics (F1-score, recall, precision) for multi-sector classification using the proposed data-driven LSTM and hybrid CNN-LSTM models.

6.6.4 Sensitivity to Sector Partitioning

To assess the sensitivity of the proposed approach to sector angle partitioning, we compared three-sector (3Q) and four-sector (4Q) configurations. The number of spatial sectors was determined through a combination of geometric reasoning, dataset size constraints, and empirical comparative evaluation. From a geometric perspective, the single-gateway deployment naturally supports angular partitioning of the coverage area, and three sectors of 120° each were identified as the minimum meaningful granularity that differentiates distinct spatial regions while remaining compatible with the available dataset size. A two-sector configuration was considered insufficiently discriminative for practical localisation applications, while configurations beyond four sectors were not evaluated due to the dataset size constraint as finer partitioning would produce classes with insufficient samples for reliable training.

To empirically validate the three-sector choice, a four-sector (4Q) [as shown in Figure 6.9] configuration was evaluated as the most challenging alternative, with sector boundaries fixed from training data and preserved during testing to avoid leakage. The comparison revealed a fundamental **trade-off between classification granularity and generalisation**: the four-sector configuration provides finer spatial resolution (90° sectors vs 120°) but at a significant cost to generalisation. Specifically, the 4Q model achieved high training accuracy (84.6%) but suffered a sharp drop at test time (66.7%), with Classes 3 and 4 degrading to 64.7% and 50% respectively — a 17.9 percentage point train-test gap indicative of overfitting. By contrast, the 3Q configuration achieved a lower training accuracy (72.1%) but a higher testing accuracy (83.8%), with consistent class-level performance across all sectors. The three-sector setup is therefore selected as the optimal configuration for this dataset, offering the best balance between spatial resolution and generalisation robustness under the single-gateway, data constraints of this deployment.

In contrast, the three-sector configuration exhibited a more balanced performance. While training accuracy was lower (72.1%), the testing accuracy improved to 83.8%, reflecting stronger generalisation capability. Class-level performance was more consistent, with all sectors achieving reasonably high accuracies, particularly a recovery in Class 3 (61.0% in training versus 78.4% in testing). This indicates lower sensitivity to deployment conditions



(a) Training results

(b) Testing results

Figure 6.9: Confusion matrix for four-sector configuration.

and a more stable model across varying operating environments.

Overall, the three-sector setup emerges as the more optimal configuration for this dataset, offering reduced sensitivity between training and testing and providing more robust predictive performance. The four-sector design, while attractive for finer granularity, may require additional data augmentation, regularization, or optimization strategies to mitigate overfitting and achieve comparable robustness.

6.6.5 Statistical Reliability and Significance Analysis

To further quantify the robustness of the proposed LSTM classifier, we evaluated the balanced accuracy, confidence intervals, sensitivity, and specificity for each class (Q1, Q2 and Q3). Balanced accuracy provides a fair measure of performance under class imbalance by averaging the true positive and true negative rates. Sensitivity and specificity respectively indicate the model's ability to correctly identify samples of each class and reject non-members.

Table 6.5 summarizes these results. For Q1, the balanced accuracy reached 0.802

(95% CI: [0.714, 0.877]), with sensitivity of 0.686 and specificity of 0.918—indicating strong reliability in detecting non-Q1 cases while maintaining reasonable sensitivity. Q2 and Q3 achieved balanced accuracies of 0.723 and 0.742, respectively, with confidence intervals confirming stability across resampling. Sensitivity and specificity values were consistently between 0.67–0.80, reflecting steady generalisation across spatial sectors. Overall, these results demonstrate that the classifier performs substantially above chance level (0.50) across all classes.

Table 6.5: Balanced accuracy, 95% confidence intervals, sensitivity, and specificity for each class (Q1–Q3).

Class	**Balanced Accuracy**	**95% CI**	**Sensitivity**	**Specificity**
Q1	0.802	[0.714, 0.877]	0.686	0.918
Q2	0.723	[0.637, 0.807]	0.667	0.780
Q3	0.742	[0.661, 0.818]	0.679	0.805

To statistically confirm the improvement over the baseline CNN-LSTM architecture, McNemar's test was applied to paired prediction outcomes. The test yielded a chi-squared statistic of $\chi^2 = 9.26$ with an associated p-value of $p = 0.0023$, indicating that the observed improvement is statistically significant at the 95% confidence level. This confirms that the proposed data-driven LSTM provides a measurable and reliable enhancement in localisation performance.

6.6.6 Angular Error and Spatial Consistency

To further assess the spatial precision of the proposed LSTM model, we computed the hard-label angular error, defined as the deviation between the center of the true sector and the center of the sector predicted by argmax decoding of the softmax output. This yielded a mean angular error of 38.8° and a median of 0°, corresponding to approximately 32.3% of the sector width. The error distribution was bimodal, with predictions either perfectly correct (0° error) or shifted into an adjacent sector (120° error). Given that only three sectors are defined, any misclassification necessarily corresponds to one of two neighboring sectors. For misclassified samples, the mean distance to the nearest boundary was 39.4°, indicating that errors tended to occur within neighboring regions rather than near boundary edges. The model achieved 100% Top-2 accuracy, confirming that the true sector was consistently among the two most probable predictions.

To obtain a more continuous and probabilistic view of the output, we also derived softmax-weighted angular estimates. Instead of selecting only the most probable class, the softmax probabilities were projected onto the unit circle to compute a continuous angular prediction. Under this approach, the mean error increased slightly to 49.8° with a median of 28.1°, producing a smoother, unimodal distribution compared with the discrete case. As illustrated in Figure 6.10, predicted angles remain tightly clustered around the true sector directions, with moderate dispersion into neighboring areas. This behavior demonstrates that the model maintains angular consistency and that probability mass is shared meaningfully between adjacent sectors, reducing random error spread across the angular space.

6.6.7 Robustness to Noise

To evaluate the stability of the proposed LSTM model under realistic signal fluctuations, a post-hoc robustness analysis was performed using controlled test-time corruption of the held-out real test set. Received Signal Strength Indicator (RSSI) values were perturbed with zero-mean Gaussian noise of increasing standard deviation $\sigma = \{1, 2, 3, 5\}$ dB. The trained model was kept fixed during evaluation, and no retraining or augmentation was applied.

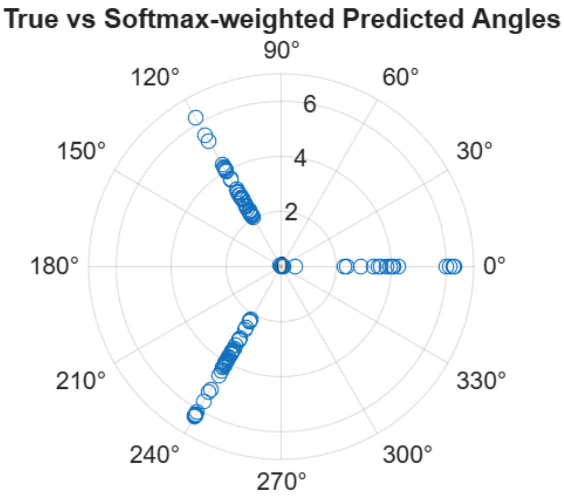


Figure 6.10: True vs. softmax-weighted predicted angles showing clustering near sector directions with moderate shifts into neighboring regions.

For each noise level, the model was tested on $K = 20$ independently corrupted versions of the same test samples, and the resulting classification accuracy was averaged to estimate the degradation trend. This protocol ensures that the robustness is evaluated strictly at inference time, avoiding circularity or bias from training-time synthesis.

As shown in Fig. 6.11, the accuracy remained remarkably stable around 58% across all noise levels, with only negligible variance across runs. This demonstrates that the proposed LSTM is resilient to moderate signal perturbations, maintaining its predictive capacity even when RSSI measurements are distorted by several decibels. Notably, the robustness test was conducted using only the original experimental dataset, excluding synthetic or augmented samples, to ensure that the observed stability reflects the inherent generalisation ability of the trained model.

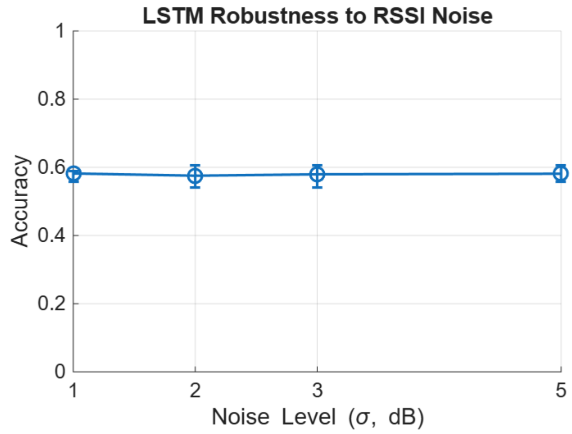
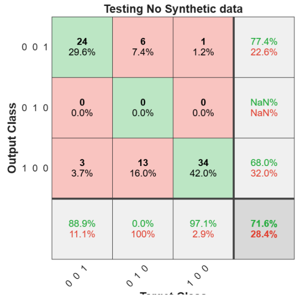


Figure 6.11: LSTM robustness to Gaussian noise injected into RSSI values at test time. Accuracy remains consistent across noise levels, confirming stability to measurement perturbations.

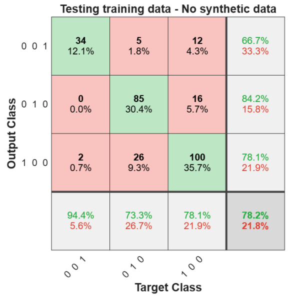
6.6.8 Synthetic Data and Ablation Studies

To isolate the influence of synthetic augmentation on model performance, two experimental configurations were considered: (i) synthetic samples used only during training, with evaluation performed exclusively on real experimental data, and (ii) both training and testing performed solely on the original dataset without any augmentation. This design ensures that augmentation effects are properly disentangled from the model's inherent generalisation ability.

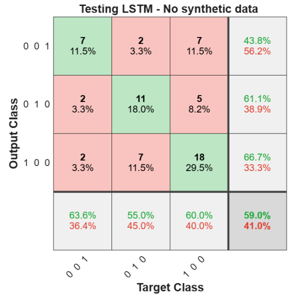
When the model was trained with synthetic augmentation but evaluated on the original (non-augmented) test set, the balanced accuracy reached 71.6%, only slightly lower than the 72.1% observed when testing on the combined augmented dataset (Fig. 6.12a). This marginal difference confirms that augmentation primarily enhances robustness during training rather than artificially inflating test performance. It is worth noting that Q2 samples were absent from the held-out test set, which may have slightly elevated the reported accuracy.



(a) Training with augmentation, testing on real data only.



(b) Training without synthetic augmentation.



(c) Testing without synthetic augmentation.

Figure 6.12: Confusion matrices for different data augmentation scenarios: (a) LSTM trained on augmented data but tested only on original data, (b) LSTM trained without augmentation, and (c) LSTM trained and tested without any synthetic data.

When trained without any synthetic data, the LSTM reflected its pure generalisation on real measurements, achieving a test accuracy of 59%. However, the class imbalance led to poor recognition of minority sectors (Fig. 6.12c). Introducing distribution-based augmentation—via synthetic channel-model samples and stochastic RSSI/SNR jittering—improved overall test accuracy to 72%, corresponding to a 13% absolute gain. This improvement indicates that augmentation effectively mitigates underfitting due to limited and imbalanced data. Such behavior aligns with prior literature on RF-based localisation, where distributional sampling and domain-consistent synthesis enhance generalisation under sparse measurement conditions.

To ensure comprehensive evaluation, we also compared the LSTM models against classical baselines frequently used in wireless localisation: Logistic Regression, Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Hidden Markov Models (HMM). All static models were trained on scalar RSSI/SNR/distance features without temporal modeling. As summarized in Table 6.6, these methods achieved modest accuracies (27–46%), with several collapsing to majority-class predictions, confirming their inability to capture the temporal and sequential dependencies inherent to LoRaWAN propagation dynamics.

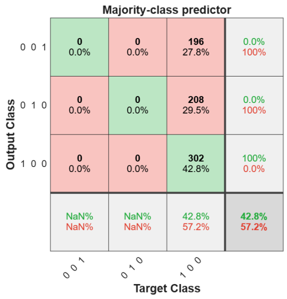
Finally, to assess the role of channel-model features, an ablation experiment was conducted in which the LSTM was trained using only raw sequential RSSI/SNR data, excluding path-loss and model-derived inputs. This reduced variant achieved 42% accuracy—already outperforming all static baselines—demonstrating that the network's temporal learning capability is the dominant driver of localisation performance.

Table 6.6: Performance comparison of baseline models and proposed LSTM variants.

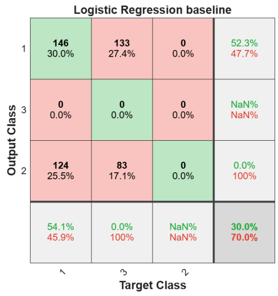
Model	**Accuracy (%)**
Majority-class predictor	42.8
Logistic Regression (static RSSI/SNR/dist.)	30.0
SVM (static RSSI/SNR/dist.)	30.9
kNN (static RSSI/SNR/dist.)	27.0
HMM (discretized RSSI)	46.1

To visualise the class-level behavior of the baseline methods, Figure 6.13 presents the corresponding confusion matrices. The static classifiers (kNN, SVM, Logistic Regression, and HMM) exhibit strong class imbalance and limited inter-class discrimination, with several models collapsing into dominant sector predictions. The Hidden Markov Model provides modest temporal smoothing through discrete RSSI transitions, but still underperforms compared to the LSTM's sequential learning. These observations reinforce the advantage of temporal feature modeling for LoRaWAN-based localisation.

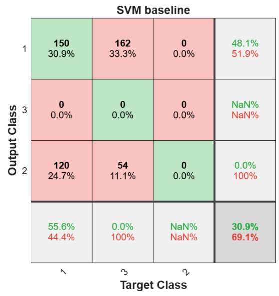
The performance gap between the LSTM (72.1%) and the HMM (46.1%) is attributable to two complementary factors. First, and most decisively, **long-term memory**: the HMM's first-order Markov assumption limits each prediction to the immediately preced-



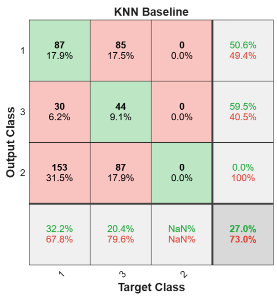
(a) Majority-class predictor



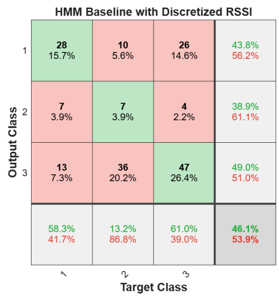
(b) Logistic Regression



(c) SVM



(d) kNN



(e) HMM (discretized RSSI)

Figure 6.13: Confusion matrices for baseline classifiers trained on static RSSI/SNR features. All models exhibit limited discriminative power and strong class bias, underscoring the effectiveness of sequential modeling in the proposed LSTM-based approach.

ing state, whereas sector discrimination in a single-gateway setting requires gradual signal evolution across multiple packets — a capability directly evidenced by the sequence-length ablation study, where accuracy improves from 57.9% at $L = 1$ to 72.5% at $L = 8$. Second, **non-linearity**: the HMM operates on discretized RSSI states through linear transition probabilities, whereas the LSTM's gating mechanism and activation functions capture complex non-linear relationships between temporal signal patterns and sector labels that are inaccessible to a linear probabilistic model. Notably, even the single-packet LSTM ($L = 1$, 57.9%) already surpasses the HMM (46.1%), suggesting that non-linear representational capacity alone accounts for part of the gap, with the full advantage emerging as temporal context grows — confirming long-term memory as the dominant differentiator in this deployment.

In summary, these results confirm that synthetic augmentation contributes primarily to data balance and robustness, rather than to artificial accuracy inflation. The LSTM's temporal learning advantage remains evident even without augmentation, but the inclusion of statistically consistent synthetic samples enables the model to reach its full generalisation potential under limited real-world data.

6.7 Generalisability and Deployment Limits

The localisation framework presented in this chapter was trained and validated within a single urban deployment environment; Birmingham City Centre, under a deliberate single-gateway constraint. While this setup is the central methodological contribution of the thesis, it is important to discuss its implications for generalisability and transferability to other deployments transparently.

With respect to **localisation granularity**, the single-gateway constraint inherently limits the spatial resolution achievable. Without the geometric diversity provided by multiple synchronised gateways (as exploited by TDoA and multilateration approaches), sector-level classification across three angular quadrants represents the practical limit of spatial inference from a single signal source. The three-sector configuration was selected over a four-sector alternative following empirical evaluation showing that the four-sector setup produced significant overfitting (84.6% training accuracy vs 66.7% test accuracy), confirming that finer granularity requires either denser infrastructure or substantially larger datasets. These constraints should be taken into account when interpreting the reported classification accuracy as a measure of practical localisation resolution.

As for, **environment transferability**, the framework is trained on measurements specific to Birmingham City Centre, and the path-loss model, sector boundaries, and LSTM weights all reflect the propagation characteristics of this particular urban morphology. Direct deployment in a different city or environment without recalibration would therefore not be expected to yield equivalent performance. However, the framework is designed to be inherently recalibratable: the data collection procedure, path-loss regression, and LSTM training pipeline are all reproducible from locally collected LoRaWAN measurements, requiring no specialised hardware beyond a single gateway. In this sense, the contribution is best understood as a methodology for environment-specific single-gateway localisation rather than a universally transferable model.

With respect to **generalisation within the Birmingham deployment**, the leave-one-group-out cross-validation protocol; achieving $68.8\% \pm 2.2\%$ accuracy and $76.9\% \pm 1.6\%$ balanced accuracy across five folds, confirms that the model generalises reliably across measurement campaigns collected on different days and under varying environmental conditions, including dynamic mobility, seasonal variation, and a natural mixture of LOS and NLOS propagation conditions inherent to the dense urban morphology of Birmingham City Centre. This within-environment generalisation is meaningful evidence of robustness, even if cross-environment transferability remains an open research question.

The single-gateway approach is therefore best positioned as a **foundation for the most constrained deployment scenario** — demonstrating that meaningful localisation is achievable without dense infrastructure — rather than as a replacement for multi-gateway systems in applications requiring fine-grained coordinate-level positioning. Extensions to multi-gateway deployments, federated learning across sites, and cross-environment validation are identified as natural directions for future work in Chapter 7.

6.8 Summary

This chapter presented the development, training, and evaluation of the proposed data-driven Long Short-Term Memory (LSTM) model for single-gateway LoRaWAN-based localisation. The work began with the synthesis and augmentation of the experimental dataset to address the inherent limitations of real-world LoRa measurements, such as data sparsity, irregular sampling, and class imbalance. Distribution-based augmentation (DBC) was used to generate statistically consistent synthetic samples derived from the fitted path-loss distributions of real measurements. This approach effectively expanded the dataset while preserving the empirical characteristics of the wireless channel, ensuring that the model learned from both authentic and statistically relevant conditions.

Following data preparation, a sequential LSTM network was designed and optimized for spatial sector prediction. The architecture consisted of stacked BiLSTM and fully connected layers, trained to model the temporal dependencies in received signal patterns. Comprehensive hyperparameter tuning was performed to determine the optimal layer configuration, learning rate, and training epochs, yielding a final model that achieved stable convergence and strong generalisation. Comparative experiments with the CNN-LSTM hybrid confirmed that while convolutional layers improved local feature extraction, the simpler LSTM architecture achieved comparable accuracy with lower computational complexity, making it more suitable for real-time or edge deployment in LoRaWAN systems.

The model's performance was validated through cross-validation and multiple evaluation metrics, including accuracy, macro-F1 score, and balanced accuracy with 95% confidence intervals. The proposed LSTM achieved an average accuracy of 72.1% under standard testing, and 68.8% $\pm$ 2.2 across five folds, confirming consistent generalisation across different trajectories. Statistical reliability analysis showed balanced accuracies above 0.70 for all classes (Q1–Q3), and McNemar's test indicated that improvements over the CNN-LSTM baseline were statistically significant ($p = 0.0023$). Training and validation curves demonstrated smooth convergence without signs of overfitting, supported by the use of dropout, batch normalization, and careful regularization.

The subsequent network evaluation section examined model sensitivity and robustness under various deployment conditions. The LSTM maintained stability under Gaussian RSSI perturbations up to ± 5 dB, demonstrating resilience to environmental and hardware-induced noise. Sector partitioning experiments revealed that a three-sector configuration provided the best trade-off between granularity and generalisation, avoiding the overfitting observed in the four-sector case. Moreover, synthetic data ablation studies confirmed that augmentation primarily enhanced class balance and robustness during training, rather than artificially boosting test accuracy. Training without synthetic data led to a 13% absolute drop in accuracy, highlighting the value of statistically grounded augmentation in data-scarce LoRaWAN environments.

Finally, comparisons with classical baselines (Logistic Regression, SVM, kNN, and HMM) established that static feature-based models achieved only 27–46% accuracy, underscoring the advantage of sequential temporal modeling. The proposed LSTM, leveraging both temporal dependencies and distribution-based channel features, delivered robust and interpretable localisation performance with strong generalisation to unseen test conditions.

In summary, this chapter demonstrated that the integration of statistical data augmentation, sequential deep learning, and systematic validation provides a scalable and reliable solution for LoRaWAN-based localisation. The findings confirm that the LSTM architecture effectively captures temporal signal variations and generalises across heterogeneous environments, establishing a solid foundation for further enhancements in hybrid localisation and real-world deployment.

Chapter 7

Conclusions & Recommendations

7.1 Summary of Research

This thesis investigated the feasibility of achieving accurate, low-power, GPS-independent localisation using a data-driven machine-learning framework within LoRaWAN networks. The study began with a critical review of existing LPWAN-based localisation methods and path-loss modeling techniques, identifying the key gaps in experimental validation, model generalisation, and robustness under real-world signal variability. Motivated by these limitations, a novel single-gateway localisation framework was proposed, integrating empirical propagation modeling with temporal deep learning using Long Short-Term Memory (LSTM) networks.

The early stages of the research focused on experimental characterisation of LoRaWAN signal behavior in a dense urban environment (Birmingham, UK). A dedicated single-gateway testbed was deployed to collect real-world RSSI and SNR data across multiple trajectories, establishing the first large-scale empirical dataset for this region. The dataset enabled the development of a two-slope, distance-segmented path-loss model that accurately captured near- and far-field propagation regimes. The results demonstrated clear relationships between environmental geometry, signal attenuation, and spatial coverage—forming the empirical basis for subsequent machine-learning modeling.

Building on these foundations, a data-driven localisation framework was developed using sequential modeling of LoRaWAN signal features. Temporal signal patterns were processed through an LSTM network designed for quadrant-based spatial classification, enabling coarse-grained location prediction without GPS. To overcome data sparsity and imbalance, the study introduced a statistical data augmentation approach, in which new samples were generated by perturbing the measured RSSI and SNR values according to their observed means and variances. This process preserved the physical distribution of the real measurements while expanding the dataset, thereby improving both data diversity and model generalisation across varying environmental conditions.

Comprehensive validation—including cross-validation, statistical reliability testing, and robustness analysis—confirmed the model's effectiveness. The proposed LSTM achieved mean accuracies of 72.1 % (single split) and 68.8 % ± 2.2 % across five folds, with balanced accuracies above 0.70 for all classes. Compared to a hybrid CNN–LSTM baseline and classical models (SVM, kNN, HMM), the LSTM exhibited superior performance with significantly lower computational complexity, enabling potential real-time deployment on resource-constrained gateways. The model also maintained stable predictions under controlled noise perturbations up to ±5 dB, demonstrating resilience to environmental and hardware-induced variations.

7.2 Key Findings & Contributions

1. **Propagation characterisation:** The Birmingham campaign provided the first real-world, high-resolution analysis of LoRaWAN propagation across an urban environment, revealing the strong dependence of performance on geometry, line-of-sight (LoS), and multipath conditions. Traditional empirical models were shown to underperform, while the proposed regression–clustering framework captured distinct propagation regimes

with interpretable, cluster-specific parameters.

2. **Distribution-Based Clustering (DBC):** A novel DBC approach was introduced to differentiate near-field and far-field propagation behaviors. Compared with K-means, DBC yielded clusters that aligned closely with physical propagation transitions, improving both interpretability and model fitting quality—thereby providing a more robust basis for hybrid path-loss modeling.

3. **Data-Driven Localisation via LSTM:** The proposed LSTM framework effectively modeled temporal signal dependencies within packet sequences, outperforming conventional static classifiers (SVM, kNN, Logistic Regression, HMM). Incorporating **statistical data augmentation** based on RSSI and SNR variance improved robustness to class imbalance, achieving up to **13% higher test accuracy** and demonstrating feasibility for single-gateway localisation.

4. **Model Validation and Statistical Reliability:** Cross-validation and statistical testing confirmed generalisation and significance, with balanced accuracies ranging from 0.72–0.80 and a McNemar's test ($p = 0.0023$) verifying improvement over CNN–LSTM and conventional baselines. This underscores the reliability and repeatability of the proposed learning framework.

5. **Noise Robustness and Environmental Stability:** Controlled Gaussian perturbation experiments showed that the model maintained stable accuracy ($\approx 58\%$) under ± 5 dB signal noise, confirming its resilience to measurement variability and practical applicability under real-world conditions.

7.3 Limitations

The proposed framework represents a meaningful step toward GPS-free, single-gateway localisation in LoRaWAN networks. However, a number of limitations should be acknowledged to ensure transparent interpretation of the reported results:

1. **Single-environment validation:** The framework was trained and evaluated exclusively within Birmingham City Centre. The path-loss model, sector boundaries, and LSTM weights all reflect the specific propagation characteristics of this urban morphology. Direct deployment in a different environment without recalibration would not be expected to yield equivalent performance. Cross-environment validation remains an open research question.

2. **Sector-level granularity:** The framework performs coarse-grained classification across three angular sectors rather than fine-grained coordinate estimation. This granularity is an inherent consequence of the single-gateway constraint, which provides no geometric diversity for multilateration. The four-sector configuration was evaluated and rejected due to overfitting (84.6% training vs 66.7% test accuracy), confirming that finer spatial resolution requires either denser infrastructure or substantially larger datasets. Furthermore, the framework does not explicitly model node mobility, including the effects of movement speed and trajectory on prediction reliability. The dataset was collected primarily under pedestrian and slow-vehicular conditions, and the impact of higher mobility speeds on sequence coherence and classification confidence has not been systematically evaluated.

3. **Single-gateway constraint:** While the single-gateway setup is the deliberate design philosophy of this work, it inherently limits spatial disambiguation compared to multi-gateway approaches. The absence of geometric diversity means that localisation relies

entirely on temporal signal evolution, which performs best under low-to-moderate mobility conditions consistent with the pedestrian and slow-vehicular scenarios in this dataset.

4. **Offline evaluation:** The framework was designed and evaluated in an offline setting where the full packet sequence was available prior to classification. Real-time deployment would require accumulation of a sufficient temporal window before a prediction can be issued, introducing a latency inherent to the sequential nature of the approach. While inference latency is negligible (< 0.15 ms per sequence), the sequence accumulation requirement introduces a deployment consideration not fully evaluated in this work.

5. **Dependence on synthetically augmented data:** The training dataset combines real experimental measurements with synthetically generated RSS samples derived from the Birmingham path-loss model. While KDE divergence metrics confirm statistical consistency between real and synthetic distributions (values below 1.71×10^{-4}, Table 6.1), and ablation studies confirm that augmentation does not artificially inflate test performance (71.6% without synthetic data vs 72.1% combined), the reliance on model-derived synthesis introduces a dependency on the accuracy of the path-loss model itself. Any residual modelling bias at the path-loss stage may propagate into the synthetic samples and influence the learned decision boundaries, with implications for real-world representativeness beyond the Birmingham deployment context.

6. **Absence of large-scale and multi-gateway validation:** The framework was validated within a single-gateway, single-environment deployment. No empirical evaluation was conducted under large-scale conditions involving multiple gateways or dense node populations. While the computational efficiency of the model (< 0.15 ms per sequence) suggests feasibility for scaled deployment, the absence of multi-gateway or network-level validation limits conclusions regarding the framework's scalability and

performance under real-world IoT deployment conditions.

These limitations do not undermine the core contributions of the thesis but define the boundary conditions within which the results should be interpreted. Each limitation directly motivates one or more of the future research directions identified in the following section, establishing a clear and honest research agenda for subsequent work in this area.

7.4 Future Research Directions

While the present work establishes a solid methodological foundation for GPS-independent, data-driven localisation in LoRaWAN networks, the emphasis of this research remains on achieving the highest possible localisation accuracy within a low-power, single-gateway infrastructure. This focus on minimal hardware and energy efficiency defines the practical motivation and scalability philosophy of the framework. Nevertheless, several research avenues naturally emerge from this study:

- **Enhanced Single-Gateway Intelligence:** A central continuation of this work lies in advancing the inference capability of a single LoRaWAN gateway. Future research can explore whether sequential signal features (e.g., the temporal evolution of RSSI and SNR) can be used to infer **motion directionality**—that is, to determine whether a transmitting node is moving *toward* or *away from* the gateway. Such temporal awareness would significantly enrich localisation context without increasing hardware complexity, representing a promising new line of investigation for LoRa-based mobility sensing.
- **Low-Power Scalability:** While this work focused on single-gateway optimization, extending the same data-driven principles to multi-gateway or multi-node setups could

demonstrate how scalable localisation accuracy can be achieved without compromising power constraints. The challenge is to maintain the system's lightweight nature while leveraging spatial diversity when available.

- **Federated and Edge Learning:** To preserve the low-power ethos of LoRaWAN, future implementations can employ **federated learning** to enable collaborative model updates across gateways while keeping raw data localised. Additionally, deploying **TinyLSTM or quantized models** on embedded gateways would allow real-time inference with minimal computational overhead.

- **Cross-Environment Validation:** Testing the trained model across different cities or environments can evaluate its adaptability and identify the limits of domain transfer. Future research should examine the generalisation of the proposed framework across different cities or deployment environments. Because radio propagation strongly depends on local morphology, building density, and terrain, each new environment requires the derivation of its own **empirical path-loss model**. Establishing this localised propagation model before training ensures that the LSTM network is fed with environment-specific signal characteristics rather than distorted or mismatched features. Comparing localisation performance across environments—each using its own calibrated path-loss representation—would reveal how transferable the learned temporal patterns are and identify whether model adaptation or transfer learning strategies are needed for robust performance.

- **Hybrid Sensor Fusion:** Integrating auxiliary sensing modalities—such as GPS, inertial measurements, or Wi-Fi fingerprints—can yield hybrid fusion networks capable of decimeter-level accuracy.

- **Adaptive and Reinforcement-Based Localisation:** Reinforcement learning could be explored for dynamic parameter adaptation in real time, improving localisation under non-stationary channel conditions and evolving urban landscapes.

Together, these future directions reaffirm the central philosophy of this research: that meaningful localisation accuracy can be achieved through signal intelligence rather than hardware proliferation. By deepening the understanding of sequential signal behavior, enabling low-power deployment of learning models, and exploring motion-aware inference from temporal LoRaWAN data, future work can further advance this framework beyond static positioning. Ultimately, this line of investigation envisions a new generation of energy-efficient, self-adaptive, and context-aware LoRaWAN localisation systems capable of operating autonomously within large-scale IoT and smart-city infrastructures—realizing practical, scalable, and sustainable localisation without the need for complex or power-intensive architectures.

References

Abubakr, Tarek and Omar A Nasr (2023). "Novel LSTM-Based Approaches for Enhancing Outdoor Localization Accuracy in 4G Networks". In: *IEEE Access*.

Action, COST (1999). *231,"Digital mobile radio towards future generation systems, final report," European Communities*. Tech. rep. EUR 18957, Tech. Rep.

Adelantado, Ferran et al. (2017). "Understanding the limits of LoRaWAN". In: *IEEE Communications magazine* 55.9, pp. 34–40.

Alanezi, Mohammed A, Houssem REH Bouchekara, and Mohammed S Javaid (2021). "Range-Based Localization of a Wireless Sensor Network for Internet of Things Using Received Signal Strength Indicator and the Most Valuable Player Algorithm". In: *Technologies* 9.2, p. 42.

Ali, Irsan Taufik, Abdul Muis, and Riri Fitri Sari (2021). "A deep learning model implementation based on rssi fingerprinting for lora-based indoor localization". In: *EUREKA: Physics and Engineering,(1)*, pp. 40–59.

Alliance, LoRa (2015). "A technical overview of LoRa and LoRaWAN". In: *White Paper, November* 20.

Almuhaya, Mukarram AM et al. (2022). "A survey on Lorawan technology: Recent trends, opportunities, simulation tools and future directions". In: *Electronics* 11.1, p. 164.

Alrajeh, Nabil Ali, Maryam Bashir, and Bilal Shams (2013). "Localization techniques in wireless sensor networks". In: *International Journal of Distributed Sensor Networks* 9.6, p. 304628.

Aly, Heba, Anas Basalamah, and Moustafa Youssef (2017). "Accurate and energy-efficient GPS-less outdoor localization". In: *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 3.2, pp. 1–31.

Anjum, Mahnoor, Muhammad Abdullah Khan, Syed Ali Hassan, Aamir Mahmood, and Mikael Gidlund (2019). "Analysis of RSSI fingerprinting in LoRa networks". In: *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE, pp. 1178–1183.

Anjum, Mahnoor, Muhammad Abdullah Khan, Syed Ali Hassan, Aamir Mahmood, Hassaan Khaliq Qureshi, et al. (2020). "RSSI fingerprinting-based localization using machine learning in LoRa networks". In: *IEEE Internet of Things Magazine* 3.4, pp. 53–59.

Apavatjrut, Anya (2025). "Environmental Data Fusion for Machine Learning-Based Inter-Node Distance Estimation". In: *IEEE Sensors Journal*.

Augustin, Aloÿs et al. (2016). "A study of LoRa: Long range & low power networks for the internet of things". In: *Sensors* 16.9, p. 1466.

Bagariang, Yosia, Muhammad Imam Nashiruddin, and Nachwan Mufti Adriansyah (2019). "LoRa-based IoT Network Planning for Advanced Metering Infrastructure in Urban, Suburban and Rural Scenario". In: *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, pp. 188–193.

Barros, Bruno et al. (2021). "Pulmonary COVID-19: learning spatiotemporal features combining CNN and LSTM networks for lung ultrasound video classification". In: *Sensors* 21.16, p. 5486.

Bécue, Adrien, Isabel Praça, and João Gama (2021). "Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities". In: *Artificial Intelligence Review* 54.5, pp. 3849–3886.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2, pp. 157–166.

Bhavanam, Bhanu Pratap Reddy and Prashanth Ragam (2025). "LoRa Sense: Sensing and Optimization of LoRa Link Behavior Using Path-Loss Models in Open-Cast Mines." In: *CMES-Computer Modeling in Engineering & Sciences* 142.1.

Bor, Martin and Utz Roedig (2017). "LoRa transmission parameter selection". In: *2017 13th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, pp. 27–34.

Bor, Martin C et al. (2016). "Do LoRa low-power wide-area networks scale?" In: *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 59–67.

Cao, Xiaoxiang et al. (2021). "A universal Wi-Fi fingerprint localization method based on machine learning and sample differences". In: *Satellite Navigation* 2, pp. 1–15.

Carrino, Francesco et al. (2019). "LoRaLoc: Machine learning-based fingerprinting for outdoor geolocation using LoRa". In: *2019 6th Swiss Conference on Data Science (SDS)*. IEEE, pp. 82–86.

Carvalho, D Fernandes et al. (2018). "On the feasibility of mobile sensing and tracking applications based on lpwan". In: *2018 IEEE Sensors Applications Symposium (SAS)*. IEEE, pp. 1–6.

Carvalho Silva, Jonathan de et al. (2017). "LoRaWAN—A low power WAN protocol for Internet of Things: A review and opportunities". In: *2017 2nd International Multidisciplinary Conference on Computer and Energy Science (SpliTech)*. IEEE, pp. 1–6.

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, Yuh-Shyan, Chih-Shun Hsu, and Chan-Yin Huang (2021). "A semi-supervised transfer learning with grid segmentation for outdoor localization over lorawans". In: *Sensors* 21.8, p. 2640.

Chen, Zhikui et al. (2013). "A localization method for the Internet of Things". In: *The Journal of Supercomputing* 63, pp. 657–674.

Chung, Junyoung et al. (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555*.

Coluccia, Angelo and Alessio Fascista (2019). “A review of advanced localization techniques for crowdsensing wireless sensor networks”. In: *Sensors* 19.5, p. 988.

Committee, LoRa Alliance Strategy et al. (2018). “LoRaWAN Geolocation Whitepaper”. In: *LoRa Alliance, January*, pp. 2018–04.

Corporation, Semtech (2015). “AN1200. 22 LoRa modulation basics”. In.

Dabbakuti, JRK Kumar, Rangababu Peesapati, and Kiran Kumar Anumandla (2023). “Design and development of artificial intelligence-enabled IoT framework for satellite-based navigation services”. In: *IEEE Transactions on Geoscience and Remote Sensing* 61, pp. 1–12.

Deese, Anthony S et al. (2020). “Long-term monitoring of smart city assets via Internet of Things and low-power wide-area networks”. In: *IEEE Internet of Things Journal* 8.1, pp. 222–231.

Demir, Fatih (2022). “Deep autoencoder-based automated brain tumor detection from MRI data”. In: *Artificial Intelligence-Based Brain-Computer Interface*. Elsevier, pp. 317–351.

Donoho, David L and Peter J Huber (1983). “The notion of breakdown point”. In: *A festschrift for Erich L. Lehmann* 157184.

Ekpenyong, Moses, Joseph Isabona, and Etim Ekong (2010). “On Propagation Path Loss Models For 3-G Based Wireless Networks: A Comparative Analysis.” In: *Computer Science & Telecommunications* 25.2.

El Chall, Rida, Samer Lahoud, and Melhem El Helou (2019). “LoRaWAN network: Radio propagation models and performance evaluation in various environments in Lebanon”. In: *IEEE Internet of Things Journal* 6.2, pp. 2366–2378.

Elif, AK et al. (2021). “LoRaWAN-aided waste-to-energy concept model in smart cities”. In: *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, pp. 1–5.

ElSabaa, AlaaAllah, Florimond Guéniat, and Wenyan Wu (2026). "Data-driven single gateway location estimation using improved channel model and LSTM: A. ElSabaa et al." In: *Telecommunication Systems* 89.2, p. 59.

ElSabaa, AlaaAllah, Florimond Guéniat, Wenyan Wu, and Michael Ward (2022). "Enhanced Data-Driven LoRa LP-WAN Channel Model in Birmingham". In: *2022 IEEE World AI IoT Congress (AIIoT)*. IEEE, pp. 766–772.

ElSabaa, AlaaAllah A, Michael Ward, and Wenyan Wu (2019). "Hybrid Localization techniques in LoRa-based WSN". In: *2019 25th International Conference on Automation and Computing (ICAC)*. IEEE, pp. 1–5.

Erfaniangolabforoushanahlyazd, Behrouz, Xinyou Zhang, and Muhammad Arslan (2021). "A Smart Parking System Based on Lorawan Technology". In: *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, pp. 625–629.

Fargas, Bernat Carbonés and Martin Nordal Petersen (2017). "GPS-free geolocation using LoRa in low-power WANs". In: *2017 Global Internet of Things Summit (GIoTS)*. IEEE, pp. 1–6.

Fialho, Vitor and Fernado Azevedo (2018). "Wireless Communication Based on Chirp Signals for LoRa IoT Devices". In: *i-ETC: ISEL Academic Journal of Electronics Telecommunications and Computers* 4.1, p. 6.

Friis, Harald T (1946). "A note on a simple transmission formula". In: *Proceedings of the IRE* 34.5, pp. 254–256.

Gamboa, John Cristian Borges (2017). "Deep learning for time-series analysis". In: *arXiv preprint arXiv:1701.01887*.

Goldsmith, Andrea (2005). *Wireless communications*. Cambridge university press.

Graves, Alex and Jürgen Schmidhuber (2008). "Offline handwriting recognition with multidimensional recurrent neural networks". In: *Advances in neural information processing systems* 21.

Greff, Klaus et al. (2012). *Supervised sequence labelling with recurrent neural networks. Vol. 385.*

Greff, Klaus et al. (2016). "LSTM: A search space odyssey". In: *IEEE transactions on neural networks and learning systems* 28.10, pp. 2222–2232.

Gu, Chaojie, Linshan Jiang, and Rui Tan (2018). "LoRa-Based Localization: Opportunities and Challenges". In: *arXiv preprint arXiv:1812.11481*.

Hadji, SE and M Nedil (2025). "Machine Learning-Based Underground Mine Path Loss Prediction Using mm-Wave Massive MIMO Measurements". In: *IEEE Transactions on Antennas and Propagation*.

Hamoudzadeh, Alireza and Saeed Behzadi (2021). "Predicting user's next location using machine learning algorithms". In: *Spatial Information Research* 29, pp. 379–387.

Han, Chungku et al. (2022). "Hybrid CNN-LSTM based time series data prediction model study". In: *IEEE/ACIS International Conference on Big Data, Cloud Computing, and Data Science Engineering*. Springer, pp. 43–54.

Harinda, Eugen et al. (2019). "Comparative performance analysis of empirical propagation models for LoRaWAN 868MHz in an urban scenario". In: *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. IEEE, pp. 154–159.

Haripriya, R, M Suresh, et al. (2025). "Machine Learning-Based Path Loss Prediction Model for Wireless Sensor Networks in Urban Environments". In: *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)*. IEEE, pp. 1–6.

Hata, Masaharu (1980). "Empirical formula for propagation loss in land mobile radio services". In: *IEEE transactions on Vehicular Technology* 29.3, pp. 317–325.

Hermans, Michiel and Benjamin Schrauwen (2013). "Training and analysing deep recurrent neural networks". In: *Advances in neural information processing systems* 26.

Hidayat, Risanuri, I Wayan Mustika, et al. (2018). "Regression analysis for estimated distance in fingerprinting-based WLAN outdoor localization system". In: *2018 4th International Conference on Science and Technology (ICST)*. IEEE, pp. 1–4.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Hu, LP, XL Bao, and Qi Wang (2011). "The repetition principle in scientific research." In: *Zhong xi yi jie he xue bao= Journal of Chinese Integrative Medicine* 9.9, pp. 937–940.

Hussain, Ghulam et al. (2019). "Indoor positioning system: A new approach based on lstm and two stage activity classification". In: *Electronics* 8.4, p. 375.

Imoize, AL and AI Oseni (2019). "Investigation and pathloss modeling of fourth generation long term evolution network along major highways in Lagos Nigeria". In: *Ife Journal of Science* 21.1, pp. 39–60.

Isabona, Joseph et al. (2022). "Development of a multilayer perceptron neural network for optimal predictive modeling in urban microcellular radio environments". In: *Applied Sciences* 12.11, p. 5713.

Iwana, Brian Kenji and Seiichi Uchida (2021). "An empirical survey of data augmentation for time series classification with neural networks". In: *Plos one* 16.7, e0254841.

Janssen, Thomas, Rafael Berkvens, and Maarten Weyn (2019). "Comparing machine learning algorithms for RSS-based localization in LPWAN". In: *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. Springer, pp. 726–735.

Jo, Han-Shin et al. (2020). "Path loss prediction based on machine learning techniques: Principal component analysis, artificial neural network, and Gaussian process". In: *Sensors* 20.7, p. 1927.

Jörke, Pascal et al. (2017). "Urban channel models for smart city IoT-networks based on empirical measurements of LoRa-links at 433 and 868 MHz". In: *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, pp. 1–6.

Jung, Suk Hoon, Byeong-Cheol Moon, and Dongsoo Han (2016). "Performance evaluation of radio map construction methods for Wi-Fi positioning systems". In: *IEEE Transactions on Intelligent Transportation Systems* 18.4, pp. 880–889.

Karim, Fazle, Somshubra Majumdar, Houshang Darabi, and Shun Chen (2017). "LSTM fully convolutional networks for time series classification". In: *IEEE access* 6, pp. 1662–1669.

Karim, Fazle, Somshubra Majumdar, Houshang Darabi, and Samuel Harford (2019). "Multivariate LSTM-FCNs for time series classification". In: *Neural Networks* 116, pp. 237–245.

Kong, Dejiang and Fei Wu (2018). "HST-LSTM: A hierarchical spatial-temporal long-short term memory network for location prediction." In: *IJCAI*. Vol. 18. 7, pp. 2341–2347.

Koovimol, Pongwut and Suwat Pattaramalai (2021). "Experimental Machine Learning for RSSI fingerprint in Indoor WiFi Localization". In: *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, pp. 1018–1021.

Koprinska, Irena, Mashud Rana, and Vassilios G Agelidis (2015). "Correlation and instance based feature selection for electricity load forecasting". In: *Knowledge-Based Systems* 82, pp. 29–40.

Krishnan, Sivanand and Rochelle Xenia Mendoza Santos (2021). "Real-time asset tracking for Smart Manufacturing". In: *Implementing Industry 4.0: The Model Factory as the Key Enabler for the Future of Manufacturing*, pp. 25–53.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2017). "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90.

Kufakunesu, Rachel, Gerhard P Hancke, and Adnan M Abu-Mahfouz (2020). "A survey on adaptive data rate optimization in lorawan: Recent solutions and major challenges". In: *Sensors* 20.18, p. 5044.

Kurt, Sinan and Bulent Tavli (2013). “Propagation model alternatives for outdoor wireless sensor networks”. In: *2013 Ifip Wireless Days (Wd)*. IEEE, pp. 1–3.

— (2017). “Path-Loss Modeling for Wireless Sensor Networks: A review of models and comparative evaluations”. In: *IEEE Antennas and Propagation Magazine* 59.1, pp. 18–37.

Lagarias, Jeffrey C et al. (1998). “Convergence properties of the Nelder–Mead simplex method in low dimensions”. In: *SIAM Journal on optimization* 9.1, pp. 112–147.

Langendoen, Koen and Niels Reijers (2003). “Distributed localization in wireless sensor networks: a quantitative comparison”. In: *Computer networks* 43.4, pp. 499–518.

Längkvist, Martin (2014). “Modeling time-series with deep networks”. PhD thesis. Örebro university.

Lara-Benítez, Pedro et al. (2021). “Evaluation of the transformer architecture for univariate time series forecasting”. In: *Advances in Artificial Intelligence: 19th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2020/2021, Málaga, Spain, September 22–24, 2021, Proceedings 19*. Springer, pp. 106–115.

LeCun, Yann, Yoshua Bengio, et al. (1995). “Convolutional networks for images, speech, and time series”. In: *The handbook of brain theory and neural networks* 3361.10, p. 1995.

Leys, Christophe et al. (2013). “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median”. In: *Journal of Experimental Social Psychology* 49.4, pp. 764–766.

Li, Xingwang et al. (2024). “Wireless Positioning: Technologies, Applications, Challenges, and Future Development Trends”. In: *CMES-Computer Modeling in Engineering & Sciences*.

Li, Yan et al. (2021). “Urban vehicle localization in public LoRaWan network”. In: *IEEE Internet of Things Journal* 9.12, pp. 10283–10294.

Linka, Hendrik et al. (2018). *Path loss models for low-power wide-area networks: Experimental results using LoRa*. VDE Verlag.

López-Ramírez, Gustavo Adulfo and Alejandro Aragón-Zavala (2025). "Enhancing Indoor mmWave Communication With ML-Based Propagation Models". In: *IEEE Access*.

Lu, Mengyuan et al. (2025). "Path loss prediction for vehicle-to-infrastructure communications via synesthesia of machines (SoM)". In: *Radio Science* 60.6, pp. 1–15.

Mackey, Andrew and Petros Spachos (2019). "LoRa-based localization system for emergency services in GPS-less environments". In: *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, pp. 939–944.

Magsi, Saeed Ahmed et al. (2023). "Experimental evaluation of Trilateration-Based outdoor localization with LoRaWAN". In: *Cmc-Computers Materials & Continua* 75.1, pp. 845–862.

Mahajan, Ajay and Maurice Walworth (2001). "3D position sensing using the differences in the time-of-flights from a wave source to various receivers". In: *IEEE Transactions on Robotics and Automation* 17.1, pp. 91–94.

Malhotra, Pankaj et al. (2015). "Long short term memory networks for anomaly detection in time series". In: *Proceedings*. Vol. 89, pp. 89–94.

Marais, Jaco M, Reza Malekian, and Adnan M Abu-Mahfouz (2017). "LoRa and LoRaWAN testbeds: A review". In: *2017 Ieee Africon*. IEEE, pp. 1496–1501.

Markit, IHS (2017). "The Internet of Things: A movement, not a market". In: *IHS Market* 1.1, p. 1.

Maurya, Poonam, Aatmjeet Singh, and Arzad Alam Kherani (2022). "A review: spreading factor allocation schemes for LoRaWAN". In: *Telecommunication Systems* 80.3, pp. 449–468.

Mdhaffar, Afef et al. (2017). "IoT-based health monitoring via LoRaWAN". In: *IEEE EUROCON 2017-17th international conference on smart technologies*. IEEE, pp. 519–524.

Mekki, Kais et al. (2019). "A comparative study of LPWAN technologies for large-scale IoT deployment". In: *ICT express* 5.1, pp. 1–7.

Mikhaylov, Konstantin, Juha Petaejaejaervi, and Tuomo Haenninen (2016). "Analysis of capacity and scalability of the LoRa low power wide area network technology". In: *European Wireless 2016; 22th European Wireless Conference*. VDE, pp. 1–6.

Milanovic, Josip, Snjezana Rimac-Drlje, and Krunoslav Bejuk (2007). "Comparison of propagation models accuracy for WiMAX on 3.5 GHz". In: *2007 14th IEEE International Conference on Electronics, Circuits and Systems*. IEEE, pp. 111–114.

Miller, Jeff (1991). "Reaction time analysis with outlier exclusion: Bias varies with sample size". In: *The quarterly journal of experimental psychology* 43.4, pp. 907–912.

Molisch, Andreas F (2012). *Wireless communications*. Vol. 34. John Wiley & Sons.

Moradbeikie, Azin et al. (2021). "GNSS-free outdoor localization techniques for resource-constrained IoT architectures: A literature review". In: *Applied Sciences* 11.22, p. 10793.

— (2023). "Improving LoRaWAN RSSI-based localization in harsh environments: the harbor use case". In: *World Conference on Information Systems and Technologies*. Springer, pp. 14–25.

Ngamakeur, Kan et al. (2022). "Deep cnn-lstm network for indoor location estimation using analog signals of passive infrared sensors". In: *IEEE Internet of Things Journal* 9.22, pp. 22582–22594.

Nor, Ruhaizan Fazrren Ashraff Mohd, Fadhlan HK Zaman, and Shamry Mubdi (2017). "Smart traffic light for congestion monitoring using LoRaWAN". In: *2017 IEEE 8th Control and System Graduate Research Colloquium (ICSGRC)*. IEEE, pp. 132–137.

Odesanya, Ituabhor et al. (2023). "Hybrid empirical and machine learning approach to efficient path loss predictive modelling: An overview". In: *Int. J. Advanced Networking and Applications* 15.03, pp. 5931–5939.

Ojo, Stephen, Murat Akkaya, and Joshua Chibuike Sopuru (2022). "An ensemble machine learning approach for enhanced path loss predictions for 4G LTE wireless networks". In: *International Journal of Communication Systems* 35.7, e5101.

Ojo, Stephen, Arif Sari, and Taiwo P Ojo (2022). “Path loss modeling: A machine learning based approach using support vector regression and radial basis function models”. In: *Open Journal of Applied Sciences* 12.6, pp. 990–1010.

Okokpujie, Imhade P and Lagouge K Tartibu (2024). “Study of the economic viability of internet of things (IoTs) in additive and advanced manufacturing: A comprehensive review”. In: *Progress in Additive Manufacturing*, pp. 1–20.

Olukanni, Seyi Emm and Jacob Funsho (2025). “Stacked Ensemble Approaches for enhanced Path Loss Prediction in complex Wireless Communication Scenarios”. In: *Authorea Preprints*.

Pahlavan, Kaveh and Allen H Levesque (2005). *Wireless information networks*. Vol. 93. John Wiley & Sons.

Pang, Yawei et al. (2014). “Efficient data collection for wireless rechargeable sensor clusters in harsh terrains using UAVs”. In: *2014 IEEE Global Communications Conference*. IEEE, pp. 234–239.

Park, Jae-Joon et al. (2012). “Measurement-based stochastic cross-correlation models of a multilink channel in cooperative communication environments”. In: *ETRI Journal* 34.6, pp. 858–868.

Pascanu, Razvan et al. (2013). “How to construct deep recurrent neural networks”. In: *arXiv preprint arXiv:1312.6026*.

Pasolini, Gianni (2021). “On the LoRa chirp spread spectrum modulation: Signal properties and their impact on transmitter and receiver architectures”. In: *IEEE Transactions on Wireless Communications* 21.1, pp. 357–369.

Patel, Dhaval and Myounggyu Won (2017). “Experimental study on low power wide area networks (LPWAN) for mobile Internet of Things”. In: *2017 IEEE 85th vehicular technology conference (VTC Spring)*. IEEE, pp. 1–5.

Patwari, Neal et al. (2005). “Locating the nodes: cooperative localization in wireless sensor networks”. In: *IEEE Signal processing magazine* 22.4, pp. 54–69.

Perihanoglu, Guzide Miray and Himmet Karaman (2024). "Performance Evaluation of GeoAI-Based Approach for Path Loss Prediction in Cellular Communication Networks". In: *Wireless Personal Communications* 138.2, pp. 1211–1246.

Petajajarvi, Juha et al. (2015). "On the coverage of LPWANs: range evaluation and channel attenuation model for LoRa technology". In: *2015 14th International Conference on ITS Telecommunications (ITST)*. IEEE, pp. 55–59.

Piechowiak, Maciej, Piotr Zwierzykowski, and Bartosz Musznicki (2023). "LoRaWAN metering infrastructure planning in smart cities". In: *Applied Sciences* 13.14, p. 8431.

Pimpinella, Andrea et al. (2020). "Machine learning based localization of LoRaWAN devices via inter-technology knowledge transfer". In: *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, pp. 1–6.

Podevijn, Nico et al. (2018). "TDoA-based outdoor positioning with tracking algorithm in a public LoRa network". In: *Wireless Communications and Mobile Computing* 2018.

Popli, Sakshi, Rakesh Kumar Jha, and Sanjeev Jain (2018). "A survey on energy efficient narrowband internet of things (NBIoT): architecture, application and challenges". In: *IEEE Access* 7, pp. 16739–16776.

Popoola, Segun I, Aderemi A Atayero, and Oluwafunso A Popoola (2018). "Comparative assessment of data obtained using empirical models for path loss predictions in a university campus environment". In: *Data in brief* 18, pp. 380–393.

Pospisil, Jan, Radek Fujdiak, and Konstantin Mikhaylov (2020). "Investigation of the performance of TDoA-based localization over LoRaWAN in theory and practice". In: *Sensors* 20.19, p. 5464.

Purohit, Jait Nitinkumar (2019). "Fingerprinting based localization with LoRa using deep learning techniques". PhD dissertation. California State University, Sacramento.

Qin, Lele, Naiwen Yu, and Donghui Zhao (2018). "Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video". In: *Tehnički vjesnik* 25.2, pp. 528–535.

Radcliffe, Peter J et al. (2017). "Usability of LoRaWAN technology in a central business district". In: *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. IEEE, pp. 1–5.

Rappaport, Theodore S (2010). *Wireless communications: Principles and practice, 2/E*. Pearson Education India.

Ren, Yidong et al. (2022). "Is lorawan really wide? fine-grained lora link-level measurement in an urban environment". In: *2022 IEEE 30th International Conference on Network Protocols (ICNP)*. IEEE, pp. 1–12.

Revathi, V et al. (2025). "Enhancing Wireless Communication Efficiency Through Advanced Frequency Band and Path Loss Prediction with Random Forest Algorithms". In: *2025 International Conference on Next Generation Communication & Information Processing (INCIP)*. IEEE, pp. 167–173.

Reynders, Brecht and Sofie Pollin (2016). "Chirp spread spectrum as a modulation technique for long range communication". In: *2016 Symposium on Communications and Vehicular Technologies (SCVT)*. IEEE, pp. 1–5.

Rizk, Hamada, Ahmed Shokry, and Moustafa Youssef (2019). "Effectiveness of data augmentation in cellular-based localization using deep learning". In: *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, pp. 1–6.

Röbesaat, Jenny et al. (2017). "An improved BLE indoor localization with Kalman-based fusion: An experimental study". In: *Sensors* 17.5, p. 951.

Sagheer, Alaa and Mostafa Kotb (2019). "Time series forecasting of petroleum production using deep LSTM recurrent networks". In: *Neurocomputing* 323, pp. 203–213.

Sanchez-Iborra, Ramon et al. (2018). "Performance evaluation of LoRa considering scenario conditions". In: *Sensors* 18.3, p. 772.

Al-Sarawi, Shadi et al. (2017). "Internet of Things (IoT) communication protocols". In: *2017 8th International Conference on Information Technology (ICIT)*. IEEE, pp. 685–690.

Schuster, Mike and Kuldip K Paliwal (1997). “Bidirectional recurrent neural networks”. In: *IEEE transactions on Signal Processing* 45.11, pp. 2673–2681.

Selmy, Hend A, Hoda K Mohamed, and Walaa Medhat (2024). “A predictive analytics framework for sensor data using time series and deep learning techniques”. In: *Neural Computing and Applications*, pp. 1–14.

Semtech (2017). *Guide, Semtech LoRa Modem Design.*

Semtech, AN1200 (2022). *LoRa™ Modulation Basics, Application Note.*

Shahbazian, Reza et al. (2023). “Machine Learning Assists IoT Localization: A Review of Current Challenges and Future Trends”. In: *Sensors* 23.7, p. 3551.

Shiri, Farhad Mortezapour et al. (2023). “A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU”. In: *arXiv preprint arXiv:2305.17473.*

Sigfox, SA (2017). *Sigfox technical overview.*

Singh, Dushyant Kumar et al. (2022). “LoRa based intelligent soil and weather condition monitoring with internet of things for precision agriculture in smart cities”. In: *IET communications* 16.5, pp. 604–618.

Sønderby, Søren Kaae et al. (2015). “Convolutional LSTM networks for subcellular localization of proteins”. In: *Algorithms for Computational Biology: Second International Conference, AlCoB 2015, Mexico City, Mexico, August 4-5, 2015, Proceedings 2*. Springer, pp. 68–80.

Song, Jun et al. (2017). “Hierarchical contextual attention recurrent neural network for map query suggestion”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.9, pp. 1888–1901.

Sornin, Nicolas et al. (2015). “Lorawan specification”. In: *LoRa alliance* 1.

Springer, A et al. (2000). “Spread spectrum communications using chirp signals”. In: *IEEE/AFCEA EUROCOMM 2000. Information Systems for Enhanced Public Safety and Security (Cat. No. 00EX405)*. IEEE, pp. 166–170.

Srivastava, Anitej and S Anto (2022). "Weather Prediction Using LSTM Neural Networks". In: *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*. IEEE, pp. 1–4.

Sunori, Sandeep Kumar et al. (2024). "Estimation of Signal to Noise Ratio in 5G Communication through ML based Predictive Models". In: *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, pp. 426–431.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27.

Svertoka, Ekaterina et al. (2024). "Towards Industry 5.0 Positioning with k-NN-based Enhanced LoRaWAN Localization". In: *European Wireless 2024; 29th European Wireless Conference*. VDE, pp. 1–6.

Tan, Kian Long et al. (2022). "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network". In: *IEEE Access* 10, pp. 21517–21525.

Tarekegn, Getaneh Berie et al. (2019). "Applying long short-term memory (LSTM) mechanisms for fingerprinting outdoor positioning in hybrid networks". In: *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. IEEE, pp. 1–5.

Technology, Microchip (2015-2020). *Low-Power Long Range LoRa® Technology Transceiver Module*. DS50002346E. Microchip Technology. URL: https://ww1.microchip.com/downloads/en/DeviceDoc/RN2483-Data-Sheet-DS50002346E.pdf.

Tian, Kegang et al. (2021). "Wireless Positioning Using Deep Learning with Data Augmentation Technique". In: *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, pp. 1255–1259.

UnaBiz (Mar. 2021). *SIGFOX.COM*. URL: https://www.sigfox.com/en.

Urbinate, Eder, Fernando Itano, and Emilio Del-Moral-Hernandez (2023). "CNN-LSTM Optimized by Genetic Algorithm in Time Series Forecasting: An Automatic Method to Use Deep Learning". In: *International Conference on Artificial Intelligence and Soft Computing*. Springer, pp. 286–295.

Valente, Antonio et al. (2022). “A LoRaWAN IoT system for smart agriculture for vine water status determination”. In: *Agriculture* 12.10, p. 1695.

Van Houdt, Greg, Carlos Mosquera, and Gonzalo Nápoles (2020). “A review on the long short-term memory model”. In: *Artificial Intelligence Review* 53.8, pp. 5929–5955.

Van Komen, David F et al. (2020). “Seabed and range estimation of impulsive time series using a convolutional neural network”. In: *The Journal of the Acoustical Society of America* 147.5, EL403–EL408.

Vangelista, Lorenzo (2017). “Frequency shift chirp modulation: The LoRa modulation”. In: *IEEE Signal Processing Letters* 24.12, pp. 1818–1821.

Virmani, Deepali, Shweta Taneja, and Geetika Malhotra (2015). “Normalization based K means Clustering Algorithm”. In: *arXiv preprint arXiv:1503.00900*.

Vo, Quoc Duy and Pradipta De (2015). “A survey of fingerprint-based outdoor localization”. In: *IEEE Communications Surveys & Tutorials* 18.1, pp. 491–506.

Walch, Florian et al. (2017). “Image-based localization using lstms for structured feature correlation”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 627–637.

Wang, Jingyuan et al. (2019). “Deep trajectory recovery with fine-grained calibration using kalman filter”. In: *IEEE Transactions on Knowledge and Data Engineering* 33.3, pp. 921–934.

Wang, Senzhang, Jiannong Cao, and Philip Yu (2020). “Deep learning for spatio-temporal data mining: A survey”. In: *IEEE transactions on knowledge and data engineering*.

Wang, Shie-Yuan et al. (2018). “Long-Term performance Studies of a LoRaWAN-based PM2. 5 application on campus”. In: *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, pp. 1–5.

Wen, Tao et al. (2021). “A DNN-based channel model for network planning in train control systems”. In: *IEEE Transactions on Intelligent Transportation Systems*.

Wixted, Andrew J et al. (2016). "Evaluation of LoRa and LoRaWAN for wireless sensor networks". In: *2016 IEEE SENSORS*. IEEE, pp. 1–3.

Wu, Wei et al. (2022). "Industrial IoT and long short-term memory network-enabled genetic indoor-tracking for factory logistics". In: *IEEE Transactions on Industrial Informatics* 18.11, pp. 7537–7548.

Wu, Yu-Xi, Qing-Biao Wu, and Jia-Qi Zhu (2019). "Data-driven wind speed forecasting using deep feature extraction and LSTM". In: *IET Renewable Power Generation* 13.12, pp. 2062–2069.

Xia, Minghua and Sonia Aissa (2015). "On the efficiency of far-field wireless power transfer". In: *IEEE Transactions on Signal Processing* 63.11, pp. 2835–2847.

Yan, Zhengqiang et al. (2021). "Stacked autoencoders-based localization without ranging over Internet of Things". In: *IEEE Internet of Things Journal* 9.10, pp. 7826–7841.

Yang, Zheng, Zimu Zhou, and Yunhao Liu (2013). "From RSSI to CSI: Indoor localization via channel response". In: *ACM Computing Surveys (CSUR)* 46.2, pp. 1–32.

Zakaria, Yahia and Lubomír Ivanek (2018). "Analysis of channel propagation models based on calculated and measured data". In: *2018 ELEKTRO*. IEEE, pp. 1–5.

Zhang, Yongliang et al. (2019). "An rss pathloss considered distance metric learning for fingerprinting indoor localization". In: *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, pp. 1–6.