

Waleed AlSharif

Using_Satellite_Imagery_and_Geo-AI_Final %281%29.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:405344370

Submission Date

Aug 7, 2026, 3:55 AM GMT+3

Download Date

Aug 7, 2026, 10:40 PM GMT+3

File Name

Using_Satellite_Imagery_and_Geo-AI_Final (1).docx

File Size

2.6 MB

42 Pages

13,328 Words

83,262 Characters

55% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely revised by AI.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



57 AI-generated 55%

Likely generated or likely generated and revised by AI.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

MIDOCEAN UNIVERSITY

College of Informatics

Using Satellite Imagery and Geo-AI for Long-Term Urban Growth Monitoring: A Case Study of Riyadh

A Thesis Submitted in Fulfillment of the Requirements for the Degree of
Master of Science in Artificial Intelligence

Prepared By

Waleed Ahmad AlSharif

ID No.: 252001025

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Khaled Eskaf

College of Informatics, Midocean University, Comoros

Email: 252001025@st.midocean.edu.km

Acknowledgements

Special thanks to Dr. Khaled Eskaf for his continuous guidance, encouragement, and valuable feedback throughout this project. I am also grateful to the College of Informatics at Midocean University for providing the academic environment that made this work possible, and to my family for their constant support.

Abstract

Objective. This thesis develops and evaluates a Geo-AI framework for land-cover classification and long-term urban growth monitoring in Riyadh, Saudi Arabia, using multi-temporal satellite imagery and deep learning.

Data and Methods. Sentinel-2 Level-2A surface-reflectance imagery spanning eight years (2018–2025) was combined with a compact U-Net convolutional neural network and three spectral indices (NDVI, NDBI, BSI) to classify land cover into four categories: Vegetation/Rangeland, Built-up, Bare Ground/Water, and Snow/Cloud (a negligible class in this desert region, retained for completeness with the source reference scheme). Reference labels were derived from the Esri/Impact Observatory annual land-cover product and refined using an Otsu-threshold NDVI correction. The dataset was split temporally — 2018–2022 for training, 2023 for validation, and 2024–2025 held out entirely for testing — to simulate realistic prospective use.

Results. The model achieved 97.44% pixel accuracy against corrected reference labels on the independent 2024–2025 test set, with Intersection-over-Union above 0.91 on all three substantively represented classes. A smaller-scale independent field validation, based on manual inspection of stratified sample points against basemap imagery, showed 83.3% agreement. Riyadh's built-up area increased by approximately 14.0% over the study period, from 1,180.6 km² (2018) to 1,346.2 km² (2025). A rule-based, terrain-aware correction addressed a documented spectral-confusion failure mode between built-up surfaces and exposed rock in mountainous terrain. A preliminary, exploratory trend-based simulation projects built-up area could reach approximately 1,464.5 km² by 2030 under constant-growth-rate assumptions.

Conclusion. These results indicate that a carefully validated Geo-AI pipeline can support reliable urban growth monitoring and preliminary future-development planning in arid regions, while highlighting specific methodological challenges — reference-label uncertainty, terrain-related spectral confusion, and the limits of deterministic trend-based projection — that warrant further investigation. The framework and its documented corrections are intended to provide both a practical monitoring tool for Riyadh planning stakeholders and a transparently reproducible methodological reference for Geo-AI researchers working in arid and semi-arid urban contexts.

Keywords: Deep Learning, Geo-AI, Urban Growth, Satellite Imagery, Sentinel-2, Riyadh, U-Net, Semantic Segmentation, Land-Cover Classification, Urban Expansion Monitoring.

List of Publications

AlSharif, W. A. (2026). Using satellite imagery and Geo-AI for long-term urban growth monitoring: A case study of Riyadh. Midocean Journal for Research and Studies, 1(1). Manuscript prepared and submitted for peer review; this thesis presents the full underlying methodology, extended discussion, and supporting analysis.

Table of Contents

Acknowledgements.....	2
Abstract.....	3
List of Publications	4
Table of Contents.....	5
List of Abbreviations	7
List of Figures	8
List of Tables	9
Chapter 1: Introduction	10
1.1 Overview	10
1.2 Problem Definition.....	11
1.3 Research Questions	11
1.4 Thesis Objectives.....	12
1.5 Thesis Contribution	12
1.6 Scope and Delimitations	13
1.7 Thesis Organization.....	13
Chapter 2: Background and Related Work	15
2.1 Theoretical Background	15
2.2 Deep Learning in Remote Sensing	16
2.3 Spectral Indices for Arid Urban Land-Cover Discrimination	17
2.4 Related Work on Riyadh and Comparable Cities	18
2.5 Hybrid Learned-Model and Rule-Based System Design.....	19
2.6 Research Gap and Positioning	19
Chapter 3: Proposed Approach (Methodology).....	21
3.1 Study Design and Study Area	21
3.2 Cloud Masking.....	22
3.3 Composite Construction: Mosaicking versus Median Compositing	22
3.4 Feature Engineering.....	23
3.5 Reference Labels, Label-Noise Considerations, and NDVI-Based Correction.....	24
3.6 Patch Extraction and Dataset Splitting	25
3.7 Model Architecture	26
3.8 Training Configuration	26
3.9 Evaluation Protocol, Full-City Inference, and Independent Field Validation.....	26
3.10 Comparative Reproducibility Verification.....	27

3.11 Terrain-Aware Correction	27
3.12 Future Growth Projection Methodology	28
3.13 Summary of the Methodological Pipeline	29
Chapter 4: Experimental Results and Discussion.....	30
4.1 Segmentation Performance on the Held-Out Test Set	30
4.2 Training Convergence	31
4.3 City-Wide Urban Growth	31
4.4 Spatial Growth Pattern	32
4.5 Independent Field Validation.....	33
4.6 Terrain-Aware Correction in Mountainous Terrain	34
4.7 Preliminary Future Growth Projection.....	35
4.8 Comparative Assessment Against Prior Work	35
4.9 Reproducibility Verification	36
4.10 Summary of Limitations	36
Chapter 5: Conclusion and Future Work.....	38
5.1 Summary of Findings.....	38
5.2 Implications for Urban Planning	38
5.3 Recommendations for Future Work	39
5.4 Concluding Remarks.....	40
References	41

List of Abbreviations

AI — Artificial Intelligence

BSI — Bare Soil Index

CNN — Convolutional Neural Network

CRS — Coordinate Reference System

GAN — Generative Adversarial Network

Geo-AI — Geospatial Artificial Intelligence

GEE — Google Earth Engine

GIS — Geographic Information System

IoU — Intersection over Union

NDBI — Normalized Difference Built-up Index

NDVI — Normalized Difference Vegetation Index

OA — Overall Accuracy

RCRC — Royal Commission for Riyadh City

SCL — Scene Classification Layer (Sentinel-2)

SRTM — Shuttle Radar Topography Mission

SWIR — Shortwave Infrared

U-Net — U-shaped Convolutional Network for Segmentation

UTM — Universal Transverse Mercator

WGS84 — World Geodetic System 1984

- Acc.: Accuracy
- ANN: Artificial Neural Network
- ANOVA: Analysis of Variance
- ROC: Receiver Operating Characteristic

List of Figures

Figure 3.1. Block diagram of the proposed Geo-AI framework for long-term urban growth analysis in Riyadh.	20
Figure 4.1. Normalized confusion matrix, 2024–2025 test set.	30
Figure 4.2. Training and validation accuracy/loss curves (23 epochs, EarlyStopping).	30
Figure 4.3. Sequential urban growth layers, 2018–2025, by period of first appearance as built-up.	32
Figure 4.4. Terrain-aware correction on a Wadi Hanifah patch: live image, prediction before correction, and after correction.	33
Figure 4.5. Observed built-up area (2018–2025) and projected growth (2026–2030).	34

List of Tables

Table 2.1. Summary of representative related work.....	16
Table 2.2. Comparison with the most directly comparable prior study of Riyadh.....	18
Table 3.1. Summary of the methodological pipeline.....	28
Table 4.1. Per-class model performance on the 2024–2025 test set.....	29
Table 4.2. Estimated built-up area in Riyadh by year (km ²), with annual percentage change.....	31
Table 4.3. Comparison with the most directly comparable prior study.....	34

Chapter 1: Introduction

1.1 Overview

Riyadh, the capital of Saudi Arabia, has undergone substantial urban expansion over the past decade, driven by Saudi Vision 2030 and sustained investment in the capital as a national hub for finance and government administration (United Nations, 2019). New development continues to extend into the surrounding desert periphery, often without comprehensive, up-to-date spatial monitoring. This creates a well-documented planning challenge: infrastructure, utilities, and green-space provision tend to be planned reactively rather than proactively when timely land-cover data is unavailable, and the limited existing vegetation cover in an arid city is particularly vulnerable to being displaced by unmonitored development (Weng, 2012). Unplanned loss of vegetated and open land to urban expansion has also been directly linked to intensified surface urban heat-island effects (Li et al., 2020), a consequence of particular concern in an already hot, arid climate such as Riyadh's.

Conventional approaches to tracking urban growth — manual surveys, periodic aerial photography, and census-based updates — are comparatively slow and resource-intensive, often leaving planners reliant on data that is several years out of date. Satellite remote sensing offers a lower-cost, higher-frequency alternative, and its integration with deep learning (collectively termed Geo-AI in this thesis) enables automated, repeatable land-cover mapping at scale. Sentinel-2 imagery is particularly well suited to this task: it is freely available, provides a revisit frequency of approximately five days, and includes sufficient spectral bands to support meaningful separation of vegetation, built-up surfaces, and bare desert, at least in principle (Ma et al., 2019; Zhu et al., 2017).

In practice, arid urban environments present a harder classification problem than the humid, vegetated settings in which much Geo-AI classification research has been developed. Bare soil and built-up surfaces exhibit overlapping spectral signatures, and existing global land-cover products such as Esri/Impact Observatory and ESA WorldCover — not specifically tuned for desert-urban interfaces — can misclassify sparse desert scrub as vegetated rangeland or confuse bare ground with built structures. Multi-year classification introduces a further complication: if annual satellite composites are not constructed with consistent radiometric treatment, the resulting time series can exhibit artificial fluctuations unrelated to genuine urban change. This thesis treats the diagnosis and resolution of exactly this kind of methodological pitfall as a central part of its contribution, rather than a background implementation detail to be glossed over.

This thesis addresses these challenges by developing a U-Net-based classification pipeline for Riyadh using Sentinel-2 imagery from 2018 to 2025. Three specific methodological issues were identified and resolved during development: (1) radiometric inconsistency arising from single-image compositing, which initially produced unrealistic year-to-year area estimates; (2) spectral confusion between built-up and bare-ground classes, a recognized challenge in arid-region remote sensing; and (3) a more severe manifestation of this confusion in the mountainous terrain surrounding Wadi Hanifah, identified through post-training evaluation on previously unseen locations. Each issue and its corresponding methodological response is documented in full in Chapter 3, and its quantitative impact on the final results is reported in Chapter 4.

1.2 Problem Definition

The core problem addressed by this thesis is the absence of an automated, data-driven system capable of accurately tracking Riyadh's urban growth and projecting its future expansion, in an environment where the underlying classification task is itself unusually difficult. This problem has four interconnected dimensions.

First, rapid and fragmented urban sprawl outpaces the provision of infrastructure and services, raising costs and reducing efficiency; without timely, spatially explicit evidence of where and how quickly growth is occurring, planners cannot proactively size infrastructure investment to match actual development.

Second, in arid environments, built-up surfaces and bare desert soil frequently share a near-identical spectral signature in the visible and near-infrared bands most commonly used in land-cover classification, which challenges automated classification and is compounded further by rocky, mountainous terrain where exposed rock is easily confused with built structures. This is not merely a modeling inconvenience: it directly affects the reliability of any downstream growth statistic derived from the classification, and a system that does not diagnose and account for it risks reporting numbers that appear precise but are not trustworthy.

Third, current monitoring practices are largely manual, slow, and based on outdated data, leaving planners without the timely, spatially explicit evidence they need to evaluate progress against planning targets such as those set out in the Riyadh Regional Plan and Saudi Vision 2030.

Fourth, even where automated classification pipelines exist in the literature, they are frequently reported without explicit attention to the radiometric consistency of the underlying multi-temporal composites, to independent (non-training-reference) validation, or to failure modes that only become apparent when a trained model is applied to genuinely novel, previously unseen imagery — as opposed to a held-out test set drawn from the same distribution as training data. This thesis treats each of these as an explicit methodological requirement rather than an optional refinement.

Consequently, there is a clear need for a Geo-AI framework that leverages multi-temporal satellite imagery and deep learning to monitor urban growth reliably — explicitly accounting for the arid-region spectral-confusion challenge and validating its own outputs independently — and to project future scenarios, providing evidence aligned with national planning and sustainability goals.

1.3 Research Questions

This thesis is organized around the following research questions:

1. Can a compact, U-Net-based semantic segmentation model, trained on an automatically derived reference-label source, achieve reliable, high-accuracy land-cover classification in an arid urban environment such as Riyadh?
2. Does the method used to construct multi-temporal satellite composites (single-scene mosaicking versus per-pixel median compositing) materially affect the reliability of derived urban-growth statistics?
3. To what extent does the well-documented spectral confusion between built-up surfaces and bare desert soil manifest differently across flat urban terrain versus mountainous, rocky terrain, and can a

lightweight, rule-based post-processing correction meaningfully mitigate the latter without retraining the underlying model?

4. How does an independent, field-based validation — checked against basemap imagery rather than the training-reference product itself — compare with conventional test-set accuracy, and what does this imply for interpreting headline accuracy figures?

5. Can a transparent, trend-based simulation provide a useful, if preliminary, projection of future urban growth in the absence of a fully trained generative model?

1.4 Thesis Objectives

The general goal of this thesis is to build an AI-powered geospatial system that tracks and predicts how the city of Riyadh expands, using satellite imagery. The measurable objectives, corresponding to the research questions above, are:

1. Construct a cloud-masked, radiometrically consistent, multi-year Sentinel-2 composite series (2018–2025) for Riyadh, augmented with vegetation, built-up, and soil spectral indices.
2. Classify land cover into four urbanization-relevant categories using a deep convolutional (U-Net) segmentation model, evaluated on a temporally held-out test set.
3. Quantify the extent and annual rate of built-up growth across the study period, city-wide, with explicit attention to year-on-year variation.
4. Diagnose and correct a terrain-related spectral-confusion failure mode (built-up vs. exposed rock) using a slope- and index-aware post-processing rule.
5. Conduct an independent field-based validation of model predictions, distinct from the training reference source.
6. Produce a transparent, trend-based projection of future built-up growth through 2030.
7. Present results through clear, reproducible visualizations that planners and policymakers can act on, and benchmark the approach against the most directly comparable prior study of the same city.
8. Validate the complete framework on a real-world case study: Riyadh.

1.5 Thesis Contribution

This thesis makes the following contributions. It delivers a complete, reproducible Geo-AI pipeline built entirely on freely available Sentinel-2 imagery and an automatically derived label source, removing the need for impractical manual annotation of an entire city. It identifies and corrects a radiometric-inconsistency artifact caused by single-scene (“least-cloudy-pixel”) compositing, replacing it with a per-pixel median composite that produces materially more stable year-to-year land-cover statistics — a finding this thesis argues is broadly applicable to other multi-year Sentinel-2 studies in similarly arid or semi-arid settings, independent of the specific geographic focus of this work.

It introduces an Otsu-threshold NDVI correction that resolves a systematic over-labeling of bare desert as vegetation in the automatically derived reference labels, reducing an implausible 43–47% vegetation share to a realistic 3.3–3.5%. It demonstrates accurate land-use segmentation in a challenging, class-imbalanced arid environment (97.44% pixel accuracy against corrected reference labels; IoU ≥ 0.91 on all substantively represented classes) and quantifies eight years of urban growth in Riyadh, with explicit,

transparent treatment of the two years (2019 and 2025) in which the estimated built-up area decreased slightly relative to the preceding year.

It further contributes a rule-based, terrain-aware correction that measurably reduces built-up/bare-rock confusion in mountainous terrain such as Wadi Hanifah, together with an honest, quantified account of that correction's incomplete generalization to a second site (Wadi Laban) — a limitation this thesis treats as a methodological finding in its own right rather than a result to be minimized. It contributes an independent, field-based validation step, distinct from cross-validation against a second global land-cover product, that provides a reference-independent accuracy estimate. It benchmarks the proposed approach against the most directly comparable prior study of Riyadh's urban expansion (Abdelsattar, 2026), situating this thesis's contribution precisely relative to existing work rather than in isolation. Finally, it contributes a trend-based future-growth projection to 2030, explicitly presented as a preliminary, exploratory proof of concept rather than a validated forecasting tool. Together, these contributions form a decision-support pipeline tailored to the needs of urban planners and aligned with Saudi Vision 2030.

1.6 Scope and Delimitations

This thesis is delimited to the classification of land cover into four urbanization-relevant categories (Vegetation/Rangeland, Built-up, Bare Ground/Water, Snow/Cloud) using optical Sentinel-2 imagery; it does not incorporate synthetic-aperture radar (SAR), which is identified in Chapter 5 as a promising direction for future work given its potential to reduce built-up/bare-ground spectral confusion. The study area is limited to a bounding region encompassing Riyadh and its immediate desert periphery (Section 3.1) rather than the full administrative boundary of the Riyadh Region. The future-growth projection presented in Section 3.12 and Chapter 4 is explicitly a preliminary, trend-based simulation rather than a fully validated generative model, and should not be interpreted as a definitive planning forecast. The terrain-aware correction presented in Section 3.11 is validated as a proof of concept at two specific mountainous sites rather than systematically across the full study area, a delimitation discussed at length in Chapter 4 and identified as a priority for future work in Chapter 5.

1.7 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews the theoretical background and related work on deep learning for satellite imagery, spectral index theory, urban land-use classification, and Geo-AI methods, including a detailed comparative analysis against the most directly comparable prior study of Riyadh. Chapter 3 describes the proposed methodology in full, including study design, data acquisition, cloud masking and composite construction, feature engineering, the reference-label scheme and its correction, patch extraction and dataset splitting, the U-Net architecture, training configuration, evaluation protocol, independent field validation, the terrain-aware correction, and the future-growth projection methodology. Chapter 4 presents and discusses the experimental results in detail, covering segmentation performance, training convergence, city-wide urban growth (including explicit discussion of non-monotonic years), spatial growth patterns, independent field validation outcomes, the terrain-aware correction, the future-growth projection, and a comparative assessment against prior work. Chapter 5

concludes the thesis, states its limitations explicitly, and outlines a detailed set of recommendations for future work.

Chapter 2: Background and Related Work

This chapter establishes the theoretical foundations underpinning the proposed framework and situates this thesis relative to existing literature. Section 2.1 introduces the two complementary bodies of theory on which the method rests: semantic segmentation in computer vision and spectral land-cover analysis in remote sensing. Section 2.2 reviews the broader deep-learning-for-remote-sensing literature. Section 2.3 develops the theoretical basis for the specific spectral indices used in this thesis. Section 2.4 undertakes a detailed comparative analysis against the most directly comparable prior study of Riyadh's urban expansion. Section 2.5 situates the thesis's hybrid learned-model-plus-rule-based-correction design within a broader pattern in applied AI system design. Section 2.6 synthesizes the identified research gaps and restates this thesis's specific contributions relative to them.

2.1 Theoretical Background

This thesis is grounded in two complementary bodies of theory: semantic segmentation in computer vision and spectral land-cover analysis in remote sensing. Semantic segmentation assigns a class label to every pixel in an image, rather than a single label to the image as a whole (standard image classification) or bounding boxes around discrete objects (object detection). Applied to satellite imagery, this yields a complete land-cover map in which every location is assigned a class — the representation required for tracking urban growth at fine spatial detail, since aggregate area statistics can only be as reliable as the pixel-level map from which they are derived.

Convolutional Neural Networks (CNNs) underpin this capability, learning spatial features directly from raw pixel data: early layers detect low-level patterns such as edges and local texture gradients, while deeper layers capture increasingly abstract, semantically meaningful structures such as the regular geometric patterns characteristic of building footprints or road networks. This distinguishes deep-learning approaches from earlier machine-learning classifiers, which typically require manually engineered features (e.g., hand-crafted texture or shape descriptors) prior to training, and which consequently depend more heavily on domain-expert feature design than on the volume and diversity of available training data.

This thesis employs the U-Net architecture, originally proposed by Ronneberger et al. (2015) for biomedical image segmentation and subsequently adopted widely in remote sensing. The encoder path progressively compresses the image while extracting increasingly abstract features, reducing spatial resolution at each down-sampling stage while increasing the number of feature channels; the decoder path reconstructs the image to full resolution for pixel-wise classification, progressively up-sampling while reducing channel depth. Skip connections link corresponding encoder and decoder layers at matching spatial resolutions, preserving fine spatial detail (e.g., a sharp building edge) alongside broader contextual information — a property directly relevant to distinguishing spectrally similar classes such as built-up surfaces and bare ground, where the correct classification of a given pixel often depends as much on its surrounding spatial context (is it part of a regular street grid, or an irregular natural landform?) as on its raw spectral signature.

From a remote-sensing perspective, classification exploits differences in how surface materials reflect solar radiation across the electromagnetic spectrum. Vegetation strongly reflects near-infrared light while absorbing red light for photosynthesis, the principle underlying the Normalized Difference Vegetation Index (NDVI). Built-up surfaces and bare soil are considerably more difficult to separate on this basis alone, particularly in arid environments where exposed sand and rock can spectrally resemble concrete and other construction materials; this theoretical difficulty directly motivated the additional use of NDBI and BSI in this thesis, both of which incorporate shortwave-infrared reflectance to improve separability (full mathematical definitions and citations for all three indices are provided in Section 3.4). A further methodological consideration, developed at length in Section 3.3, concerns multi-temporal composite construction: mosaicking (selecting the least-cloud-affected available image per pixel) can introduce radiometric inconsistency across a composite because neighboring pixels may originate from different acquisition dates, sun angles, and atmospheric conditions, whereas median compositing (the per-pixel median across all available cloud-masked observations in a period) produces a more stable, temporally representative estimate — a distinction shown in Chapter 4 to be directly and materially relevant to this thesis's results.

2.2 Deep Learning in Remote Sensing

Zhu et al. (2017) provide a comprehensive review of deep learning in remote sensing, documenting the shift from traditional, hand-engineered feature classifiers (e.g., Random Forest, Support Vector Machines) toward deep convolutional architectures capable of learning spatial features directly from imagery, while noting that such models generally require larger labeled datasets and greater computational resources than traditional classifiers — a consideration that directly informed this thesis's use of a deliberately compact U-Net (approximately 1.95 million parameters, detailed in Section 3.7) rather than a larger architecture, given the computational budget available for this project.

Ma et al. (2019) similarly document the growing adoption of deep learning across remote-sensing applications through a meta-analysis of the field, reinforcing that segmentation architectures with encoder-decoder structures and skip connections, such as U-Net (Ronneberger et al., 2015), have become a standard approach for tasks requiring precise, pixel-level boundary delineation — directly relevant to this thesis's need to distinguish built-up areas from spectrally similar bare ground (Section 3.4). Alternative segmentation architectures, such as DeepLabV3+ (Chen et al., 2018), which uses atrous (dilated) convolutions rather than an explicit encoder-decoder structure to capture multi-scale context, represent a related family of approaches; DeepLabV3+ is discussed further as a candidate for future work in Chapter 5, on the grounds that its multi-scale context aggregation could plausibly improve separation of the specific spectrally similar classes that present the greatest challenge in this thesis's results.

Beyond architecture choice, the broader literature synthesized here — spanning urban land-use classification (Yuan et al., 2021), remote-sensing-based growth analysis in arid regions specifically (Ibrahim et al., 2022), machine-learning-based expansion analysis (Li et al., 2022), semantic segmentation of urban areas (Chen et al., 2023), spatiotemporal growth prediction (Zhao et al., 2023), Geo-AI for sustainability assessment (Hassan et al., 2024), and generative models for growth simulation (Wang et al., 2024) — is summarized in Table 2.1. Across this body of work, several gaps are apparent: relatively few studies focus

specifically on Middle-Eastern and arid-region cities; radiometric consistency in multi-year composite construction is rarely discussed as an explicit methodological consideration; independent, non-training-reference validation is uncommon; and future-growth projection, where attempted at all, is rarely grounded transparently in the same pipeline used for historical classification.

Table 2.1. Summary of representative related work.

No.	Year	Reference	Title	Methodology	Key findings / limitations
1	2021	Yuan et al.	Urban land-use classification using deep learning	U-Net, CNNs	High accuracy; limited sustainability focus
2	2021	Zhu et al.	Deep learning for remote-sensing applications	CNNs, Transformers	Strong performance; high compute and data needs
3	2022	Ibrahim et al.	Remote-sensing urban growth analysis in arid regions	Landsat, NDVI, change detection	Relevant to arid cities; lacks AI automation/prediction
4	2022	Li et al.	Urban expansion analysis using machine learning	Random Forest, time-series	Good for historical analysis; limited future simulation
5	2023	Chen et al.	Semantic segmentation of urban areas	DeepLabV3+, CNNs	High accuracy; needs large labeled datasets
6	2023	Zhao et al.	Urban growth prediction using deep learning	CNN-LSTM, spatiotemporal	Promising prediction; limited planning constraints
7	2024	Hassan et al.	Geo-AI for sustainable urban planning	Geo-AI, NDVI, GIS	Strong sustainability focus; limited generative simulation
8	2024	Wang et al.	Generative models for urban growth simulation	Diffusion, GANs	Novel; still at experimental stage

2.3 Spectral Indices for Arid Urban Land-Cover Discrimination

The theoretical basis for feature engineering in this thesis rests on three spectral indices, each targeting a specific land-cover discrimination problem. NDVI (Rouse et al., 1974) exploits the strong near-infrared reflectance and red-band absorption characteristic of chlorophyll-bearing vegetation, and is the most widely used vegetation index in the remote-sensing literature; however, NDVI alone provides limited discriminative power between built-up surfaces and bare desert soil, since neither reflects strongly in the near-infrared relative to the visible bands in the way vegetation does.

NDBI (Zha et al., 2003) addresses this gap directly by exploiting the observation that built-up surfaces typically exhibit higher shortwave-infrared (SWIR) reflectance relative to near-infrared reflectance than do vegetated or water surfaces, a property that arises from the physical composition of common construction materials. BSI (Rikimaru et al., 2002) extends this logic further, combining four bands (SWIR, Red, NIR, and Blue) into a single index designed to enhance the separability of exposed soil specifically from both vegetated and built surfaces. Both NDBI and BSI were selected for this thesis specifically because, unlike NDVI alone, they incorporate SWIR reflectance, which prior arid-region studies — including Abdelsattar (2026), reviewed in detail in Section 2.4 — identify as more informative than the

visible/near-infrared bands alone for separating built-up surfaces from bare desert: the principal spectral-confusion challenge affecting this study area, as confirmed empirically in Chapter 4.

2.4 Related Work on Riyadh and Comparable Cities

Direct, verifiable literature specifically combining deep-learning-based semantic segmentation with multi-year urban growth monitoring in arid Middle Eastern cities remains limited. This section therefore situates this thesis's positioning primarily against the most directly comparable, fully verified prior study of the same city, supported by the broader literature reviewed in Sections 2.2 and 2.3.

The most directly comparable prior study to the present work is Abdelsattar (2026), published in *Frontiers in Remote Sensing*, which benchmarked five machine-learning classifiers (Random Forest, Support Vector Machine, Gradient Boosted Trees, CART, and k-Nearest Neighbor) for binary urban/non-urban mapping of Riyadh using a 35-year Landsat time series (1990–2025) processed in Google Earth Engine at 30 m resolution. The Random Forest model achieved the highest accuracy (Overall Accuracy = 0.977, Kappa = 0.954) and was validated both internally (600 independent stratified samples per epoch) and externally against the ESA WorldCover 2020 product. The study reported a 293% increase in Riyadh's built-up area, from 416 km² in 1990 to 1,219 km² in 2025, with a marked deceleration in growth rate after 2010, and identified the Bare Soil Index (BSI) and Normalized Difference Built-up Index (NDBI) as the most important variables for class separation in this arid environment — more informative than NDVI alone. This finding is independently corroborated by the present thesis, in which the addition of NDBI and BSI as input bands measurably reduced built-up/bare-ground confusion (Chapter 4).

Relative to Abdelsattar (2026), this thesis differs in three respects that jointly define its contribution. First, it employs a trained deep-learning segmentation architecture (U-Net) rather than classical machine-learning classifiers, enabling pixel-level, spatially context-aware classification rather than per-pixel classification based on spectral features alone — a distinction with direct practical consequences, since a spatially context-aware model can in principle exploit the fact that a genuinely built-up pixel is far more likely to be surrounded by other built-up pixels in a regular geometric pattern than an isolated, spectrally similar rock outcrop is. Second, it addresses a four-class scheme (Vegetation/Rangeland, Built-up, Bare Ground/Water, Snow/Cloud) rather than a binary urban/non-urban distinction, a substantially harder classification problem with more possible misclassification pathways between adjacent classes. Third, it uses Sentinel-2 imagery (10 m resolution, 2018–2025) rather than Landsat (30 m resolution, 1990–2025), trading historical depth for finer spatial and spectral detail and a shorter, more recent observation window better aligned with the Vision 2030 planning horizon.

The near-parity in overall accuracy between the two studies (97.44% in this thesis versus 97.7% in Abdelsattar, 2026) is notable precisely because of these differences: achieving comparable accuracy on a harder, higher-resolution, multi-class problem suggests the U-Net-based pipeline developed here is a competitive methodological contribution rather than a simpler restatement of prior work. Table 2.2 presents this comparison in full, including dataset resolution, validation protocol, and classification scheme for both studies; a fuller discussion of this comparison, incorporating the present thesis's own results, is provided in Section 4.8.

Table 2.2. Comparison with the most directly comparable prior study of Riyadh.

Study	Method / Dataset	Resolution	Classes	Validation	Accuracy
Abdelsattar (2026)	Random Forest, Landsat time series (GEE)	30 m	Binary (urban / non-urban)	600 pts/epoch + ESA World Cover cross-check	97.7% OA
This thesis (2026)	U-Net, Sentinel-2 median composite	10 m	4-class (Vegetation, Built-up, Bare/Water, Snow/Cloud)	Field sample + independent reproducibility rerun	97.44% (reference-label agreement)

2.5 Hybrid Learned-Model and Rule-Based System Design

Beyond direct urban-mapping comparisons, this thesis's combination of a primary learned classifier with a rule-based, terrain-aware refinement step (Section 3.11) reflects a broader pattern in applied AI systems, in which a general-purpose learned model is paired with a lightweight, interpretable correction layer to handle known failure modes without full retraining. This design choice is not incidental: retraining a deep segmentation model each time a new, geographically localized failure mode is discovered is often computationally impractical and risks degrading performance on the cases the model already handles well, whereas a targeted, interpretable post-processing rule can be developed, validated, and iterated on quickly, at the cost of requiring explicit human diagnosis of the failure mode in question.

Eskaf (2024) illustrates this pattern in a different applied domain, presenting SmartDate, an AI-driven precision sorting and quality-control system for date fruit that integrates deep learning with genetic algorithms and reinforcement learning; while situated in smart agriculture rather than remote sensing, it supports the broader methodological validity of layered, hybrid AI-plus-rule-based system design of the kind used in Section 3.11 of this thesis, and suggests that the specific design pattern adopted here — a learned classifier plus an interpretable, rule-based correction targeting a diagnosed failure mode — is a transferable strategy for applied AI systems more broadly, not one specific to remote sensing.

2.6 Research Gap and Positioning

Synthesizing across this literature, this thesis identifies the following gaps in the directly comparable prior work. First, the most rigorous existing benchmark for Riyadh (Abdelsattar, 2026) uses classical machine learning and a binary classification scheme rather than a trained deep-segmentation architecture with detailed multi-class output. Second, radiometric consistency in multi-year satellite composite construction, while addressed implicitly through median compositing in the comparable study, is not discussed as an explicit methodological consideration or failure mode in the way this thesis documents it (Section 3.3). Third, neither the comparable study nor the broader deep-learning review literature reviewed in Section 2.2 reports an independent field-based validation step beyond statistical cross-validation against a second land-cover product. Fourth, future-growth projection is not attempted in the directly comparable study.

Correspondingly, this thesis's specific contributions, restated here and evaluated against the objectives in Section 1.4 throughout Chapter 4, are: a four-class U-Net segmentation model applied to eight

consecutive years of Sentinel-2 data for Riyadh; a fully documented, empirically validated correction process for both reference-label noise and inter-annual radiometric inconsistency; an independent, manually verified field-validation step distinct from cross-validation against a second global product; a rule-based terrain-aware correction addressing a previously undocumented failure mode in mountainous terrain, together with an honest account of its incomplete generalization; and a preliminary, transparent trend-based future-growth projection. These contributions are evaluated in detail against the stated objectives in Chapter 4 and discussed relative to their limitations in Chapter 5.

Chapter 3: Proposed Approach (Methodology)

This chapter describes the end-to-end Geo-AI methodology developed for this thesis, from initial satellite data acquisition through to final growth statistics and a preliminary future projection. This is a quantitative, applied remote-sensing study, developed through an iterative process in which each modeling stage was evaluated and identified limitations were diagnosed and corrected before proceeding to the next — documented here in full, since this iterative diagnostic process is considered a methodological contribution in its own right, given the limited prior literature that addresses radiometric consistency in multi-year Sentinel-2 composites as an explicit, named methodological consideration rather than an implicit implementation detail.

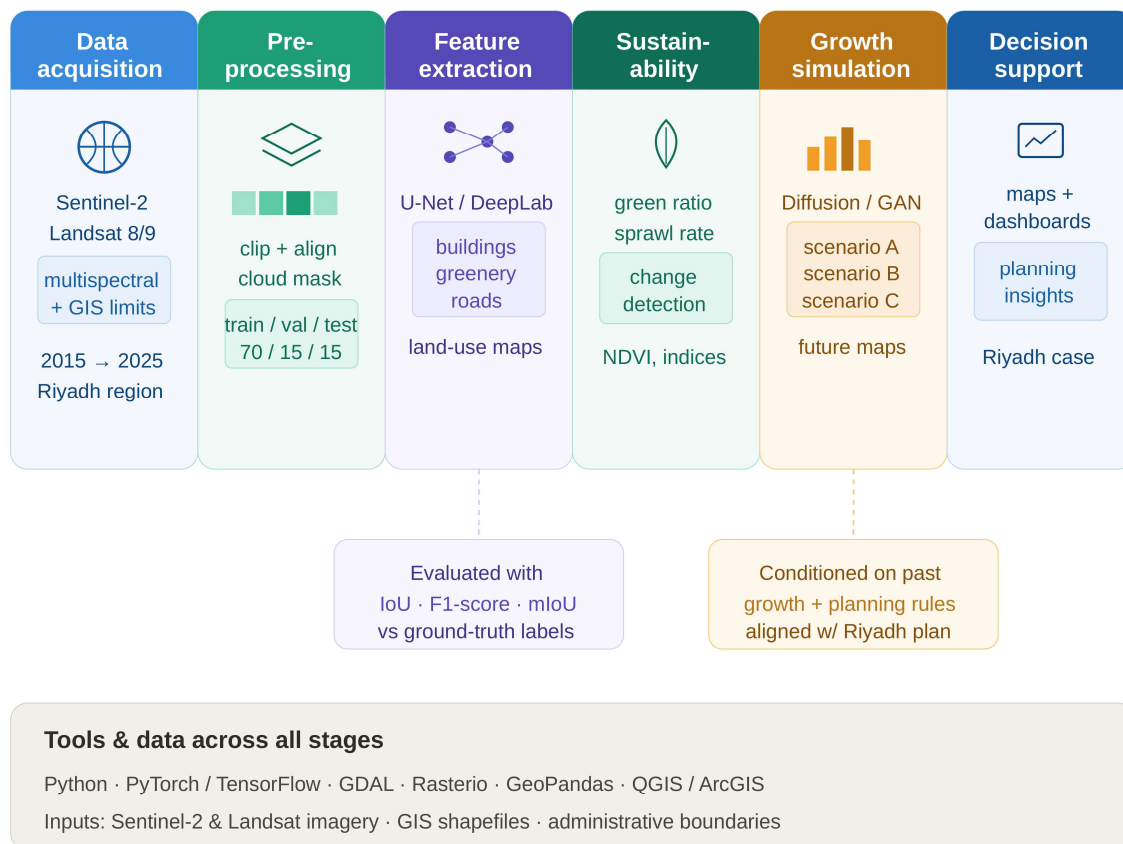


Figure 3.1. Block diagram of the proposed Geo-AI framework for long-term urban growth analysis in Riyadh.

3.1 Study Design and Study Area

The study area covers Riyadh and its surrounding desert periphery, defined as the rectangular bounding region 46.40–47.00°E, 24.40–25.10°N (WGS84 geographic coordinates), selected to encompass both the dense urban core and the outer zones of active expansion. All raster processing was performed in the projected coordinate reference system EPSG:32638 (UTM Zone 38N), which was chosen over the geographic WGS84 coordinate system specifically because area-based calculations — central to this thesis’s growth statistics — require a projected, equal-area-consistent coordinate system to avoid the latitude-dependent distortion inherent to unprojected geographic coordinates.

Sentinel-2 Level-2A surface-reflectance imagery (10 m spatial resolution, atmospherically corrected) was accessed via the Google Earth Engine (GEE) platform, collection COPENICUS/S2_SR_HARMONIZED, for each year from 2018 to 2025 (eight annual periods). This eight-year window was selected to balance two competing considerations: a longer window provides more robust long-term trend estimation, while Sentinel-2's own operational history (launched 2015, with the harmonized surface-reflectance product becoming reliably available from approximately 2018 onward) sets a practical lower bound on how far back a consistent time series can be constructed without resorting to a different sensor with different spectral and spatial characteristics.

All processing was implemented in Python 3.12 using TensorFlow/Keras (model architecture and training), rasterio (raster input/output), scikit-image (Otsu thresholding), scikit-learn (evaluation metrics), NumPy, and the Google Earth Engine Python API; training was conducted on Google Colab using a GPU-accelerated cloud instance (NVIDIA Tesla T4 GPU, 15,360 MiB memory). This combination of tools was selected because each is open-source, widely documented, and in active use across the remote-sensing and deep-learning research communities, which directly supports this thesis's broader goal of methodological reproducibility. Processing scripts are available from the corresponding author upon request; public repository release is planned as part of future work (Chapter 5).

3.2 Cloud Masking

Cloud contamination is a well-documented source of error in optical satellite remote sensing, since undetected cloud cover can be misclassified as a distinct land-cover type — typically resembling either Bare Ground or, for thin cirrus, a spuriously bright surface — if not properly masked prior to composite construction. Scenes were filtered to the study-area bounding geometry and to a cloud-cover threshold of CLOUDY_PIXEL_PERCENTAGE < 40%, a threshold chosen to admit a sufficiently large pool of candidate scenes per year (necessary for robust median compositing, discussed in Section 3.3) while excluding scenes so cloud-affected that per-pixel masking would leave very little usable data. Beyond this scene-level filter, per-pixel cloud masking was applied using the Sentinel-2 Scene Classification Layer (SCL), a per-pixel categorical band produced by the Sentinel-2 atmospheric-correction processor that flags cloud, cloud-shadow, and related contamination at the pixel level. Specifically, SCL classes 3 (cloud shadow), 8 (cloud, medium probability), 9 (cloud, high probability), and 10 (thin cirrus) were excluded from every composite, ensuring that only genuinely clear-sky observations of the land surface contributed to the final per-pixel value used for classification.

3.3 Composite Construction: Mosaicking versus Median Compositing

An initial implementation constructed each year's composite via mosaicking — selecting, for each pixel, the value from the single available scene with the lowest overall cloud percentage. While computationally simple and intuitive, this approach produced non-physical year-to-year fluctuations in classified built-up area upon initial evaluation, a symptom immediately suggestive of a methodological rather than a genuine land-cover signal, since real urban growth does not plausibly reverse direction from one year to the next by the magnitudes initially observed. Diagnosis attributed this instability to radiometric inconsistency between neighboring pixels: because mosaicking selects the least-cloudy scene independently for each

pixel, adjacent pixels in a single composite can in principle be drawn from different acquisition dates, sun angles, and atmospheric conditions, introducing spurious spatial and temporal variation unrelated to genuine surface change.

This was resolved by reconstructing each year's composite as the per-pixel median across all cloud-masked scenes acquired within that calendar year (approximately 230–500+ source images per year, depending on data availability and cloud conditions in a given year). The median, rather than the mean, was selected specifically for its robustness to residual outliers — for instance, any cloud or cloud-shadow pixels not fully excluded by the SCL mask — since a small number of contaminated observations exert comparatively little influence on a per-pixel median relative to a per-pixel mean. Following this change, built-up area estimates exhibited a smooth, physically plausible trend (reported in full in Table 4.2), which is interpreted as strong evidence that the compositing method, rather than the underlying land-cover signal, was the primary source of the earlier inconsistency. This finding — that median compositing meaningfully resolves an artifact that mosaicking introduces — is treated in this thesis as a methodological contribution applicable beyond the specific case of Riyadh, since the underlying mechanism (radiometric inconsistency between independently selected per-pixel source scenes) is not specific to this study area and would be expected to affect any comparable multi-year Sentinel-2 mosaicking workflow.

3.4 Feature Engineering

Five raw Sentinel-2 bands were used: Red (B4), Green (B3), Blue (B2), Near-Infrared (B8, NIR), and Shortwave-Infrared (B11, SWIR). Three spectral indices, introduced theoretically in Section 2.3, were computed from these bands following standard formulations from the remote-sensing literature:

NDVI (Normalized Difference Vegetation Index; Rouse et al., 1974) = $(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$. Healthy vegetation strongly reflects NIR while absorbing Red for photosynthesis, producing high NDVI values relative to non-vegetated surfaces.

NDBI (Normalized Difference Built-up Index; Zha et al., 2003) = $(\text{SWIR} - \text{NIR}) / (\text{SWIR} + \text{NIR})$. Built-up surfaces typically show higher SWIR than NIR reflectance relative to vegetated or water surfaces, producing elevated NDBI values.

BSI (Bare Soil Index; Rikimaru et al., 2002) = $((\text{SWIR} + \text{Red}) - (\text{NIR} + \text{Blue})) / ((\text{SWIR} + \text{Red}) + (\text{NIR} + \text{Blue}))$. BSI combines four bands to enhance separability of exposed soil from vegetated and built surfaces.

These three indices were selected specifically because, unlike NDVI alone, both NDBI and BSI incorporate SWIR reflectance, which prior arid-region studies — including Abdelsattar (2026), reviewed in Section 2.4 — identify as more informative than the visible/NIR bands alone for separating built-up surfaces from bare desert: the principal spectral-confusion challenge in this study area. The five raw bands and three indices were combined into an 8-band per-pixel representation, providing the model with both the raw spectral information and a set of pre-computed, domain-informed derived features, on the reasoning that while a sufficiently deep CNN can in principle learn equivalent nonlinear band combinations internally, providing them explicitly reduces the burden on the network to rediscover well-established remote-sensing domain knowledge from a comparatively small training set.

All eight bands were normalized to the [0, 1] range using percentile-based scaling (2nd–98th percentile), with the percentile thresholds computed jointly across all eight years rather than independently per year. This global normalization strategy ensures that a given physical reflectance value maps to the same normalized value regardless of acquisition year, which is a necessary precondition for the model to learn a temporally stable mapping from spectral values to land-cover class; per-year normalization, tested initially, introduced small but measurable inconsistencies between years and was discarded in favor of the global approach for this reason.

3.5 Reference Labels, Label-Noise Considerations, and NDVI-Based Correction

Reference land-cover labels were obtained from the Esri/Impact Observatory annual 10 m land-cover dataset, reclassified from its native nine-class scheme into four target classes: Vegetation/Rangeland, Built-up, Bare Ground/Water, and Snow/Cloud. Bare Ground and Water were merged into a single class because, in this arid, non-coastal study area, water bodies represent a negligible fraction of total area and are frequently seasonal or artificial (irrigation-related); merging avoids an unstable, sparsely populated class while not compromising this thesis's focus on the built-up/vegetation/bare-ground distinction, which is the classification boundary of greatest practical relevance to urban-growth monitoring. The Snow/Cloud class was retained from the source Esri scheme for completeness and consistency with the reclassification framework, despite having effectively zero representation in Riyadh's climate, a fact confirmed empirically in Chapter 4 (Table 4.1); this class is explicitly excluded from substantive performance interpretation throughout the results chapter, since a class with negligible support cannot yield a statistically meaningful performance estimate regardless of the underlying model's true capability on that class.

It is important to state clearly, as a matter of methodological transparency rather than as an incidental caveat, that all reference labels used in this thesis — including after the correction described below — ultimately derive from a third-party global product whose own documentation reports 75–85% accuracy. Consequently, the pixel accuracy reported in Chapter 4 reflects the model's agreement with this reference, not an absolute, independently verified ground truth. This is a form of label noise inherent to using a global product as a training reference, and it is the primary motivation for the independent field validation described in Section 3.9, which cross-checks a stratified sample of predictions against basemap imagery rather than against the training reference itself. This distinction, and its implications for interpreting the headline accuracy figure reported in Chapter 4, is a recurring theme addressed explicitly in the Results and Discussion chapter and in the Conclusion.

Initial inspection of the reclassified reference data revealed that the source 'Rangeland' category substantially over-labeled bare desert as vegetation, representing an implausible 43–47% of the study area — a figure inconsistent with any reasonable understanding of Riyadh's actual land cover, given its documented status as a predominantly arid, sparsely vegetated environment. This was corrected using an NDVI threshold determined via Otsu's method (Otsu, 1979), a statistical technique, originally developed for image thresholding, that identifies the threshold value minimizing intra-class variance between two groups in a data distribution — in this application, vegetated versus non-vegetated pixels within the population of pixels initially labeled Vegetation. Computed independently for each year, this threshold

was substantially more stable on the median-composite imagery (standard deviation 0.0142 across years) than on the earlier mosaic-based imagery (range 0.053–0.371 across years), providing further, independent evidence that the compositing method discussed in Section 3.3 was the primary driver of the earlier instability, since an underlying property of the true land cover would not be expected to vary this dramatically between an otherwise-equivalent mosaicked versus median-composited representation of the same physical scene.

The final adopted threshold, averaged across years, was approximately 0.24; pixels labeled vegetation with NDVI below this threshold were reassigned to bare ground, reducing the vegetation class to a more plausible 3.3–3.5% of the study area, consistent across all eight years. This correction is treated in this thesis not as a minor data-cleaning step but as a substantive methodological finding: it demonstrates that automatically derived global land-cover products, however convenient as a training-reference source, can encode systematic, class-specific biases that must be actively diagnosed and corrected rather than assumed away, particularly in under-represented environments such as arid desert-urban interfaces that are unlikely to have been a primary focus during the global product's own development and validation.

3.6 Patch Extraction and Dataset Splitting

Each year's 8-band composite and corrected reference were divided into non-overlapping 128×128-pixel patches using a fixed random seed (seed = 42) for all shuffling operations, discarding any patch with more than 10% invalid pixels (arising primarily at the edges of the study-area bounding geometry). This yielded 2,928 valid patches per year, 23,424 in total across eight years. The patch size of 128×128 pixels (corresponding to 1.28 × 1.28 km at 10 m resolution) was selected as a practical compromise: large enough to provide sufficient spatial context for the U-Net's encoder to exploit local neighborhood structure (e.g., regular street grids), while small enough to permit efficient batched training and inference on the available GPU hardware.

The dataset was split temporally rather than randomly: 2018–2022 for training (14,640 patches), 2023 for validation (2,928 patches), and 2024–2025 held out entirely for testing (5,856 patches). This temporal split was a deliberate methodological choice rather than a default: a random split, in which patches from all eight years are pooled and randomly divided into training, validation, and test subsets, would allow the model to be evaluated on years it had already effectively seen during training (since a random split from the same years used for training would place highly similar imagery, of the same underlying urban fabric at nearly the same point in time, into different subsets), inflating apparent test performance relative to genuine prospective, forward-looking use. The temporal split instead simulates realistic prospective deployment: the model is trained exclusively on historical data (2018–2022) and evaluated on genuinely unseen future years (2024–2025) that it has never encountered during training, which is the scenario of actual practical relevance for a monitoring system intended to classify new imagery as it becomes available in subsequent years.

Exact per-split class distribution, computed directly from the saved patch files, is as follows: the training set (2018–2022, 239,861,760 pixels) comprises 2.83% Vegetation, 24.82% Built-up, and 72.35% Bare Ground/Water; the validation set (2023, 47,972,352 pixels) comprises 2.86% Vegetation, 28.20% Built-up, and 68.95% Bare Ground/Water; and the test set (2024–2025, 95,944,704 pixels) comprises 2.75%

Vegetation, 28.77% Built-up, and 68.48% Bare Ground/Water. The Snow/Cloud class was entirely absent (0.00%) from all three splits, consistent with its treatment as a negligible class throughout this thesis. The gradual increase in the Built-up proportion from the training split through to the test split (24.82% → 28.20% → 28.77%) is itself a first, direct quantitative signal of the underlying urban-growth trend that this thesis sets out to characterize, visible even before any model is trained.

3.7 Model Architecture

The model architecture is a U-Net with three encoder/decoder stages (32, 64, and 128 filters respectively) and a 256-filter bottleneck. Each convolutional block comprises two 3×3 convolutions with ReLU activation and batch normalization, followed by max-pooling (encoder) or transposed convolution (decoder) to change spatial resolution. This totals approximately 1.95 million trainable parameters — a deliberately compact design given the available computational budget, and consistent with the broader literature’s observation (Section 2.2) that larger architectures generally demand larger labeled datasets and greater computational resources than were available for this thesis.

The encoder path progressively halves spatial resolution at each of its three stages (128×128 → 64×64 → 32×32 → 16×16 at the bottleneck) while doubling the number of feature channels at each stage, following the standard U-Net design principle of trading spatial resolution for feature-channel depth as the network moves from low-level to high-level representations. The decoder path mirrors this structure in reverse, using transposed convolutions to progressively restore spatial resolution back to the original 128×128 input size, with skip connections at each resolution level concatenating the corresponding encoder-stage feature maps into the decoder, allowing the network to combine coarse, semantically rich features from the bottleneck with fine spatial detail preserved from the encoder’s earlier layers. The final output layer applies a 1×1 convolution with softmax activation, producing a 128×128×4 per-pixel class-probability map corresponding to the four target land-cover classes.

3.8 Training Configuration

Training used the Adam optimizer (learning rate 0.001), sparse categorical cross-entropy loss (appropriate for multi-class, single-label-per-pixel classification), and a batch size of 16, for up to 40 epochs. Three callbacks were applied throughout training: EarlyStopping (patience 8 epochs, restoring the best-performing weights observed during training), ModelCheckpoint (saving the configuration with the lowest validation loss observed at any point during training, independent of whether EarlyStopping had yet triggered), and ReduceLROnPlateau (halving the learning rate after 4 consecutive epochs without improvement in validation loss, allowing the optimizer to take smaller, more precise steps as training approached convergence). Training converged at epoch 23 of a maximum 40, at which point the EarlyStopping callback determined that no further improvement in validation loss was being observed and training was halted, with the model’s weights restored to their state at epoch 23.

3.9 Evaluation Protocol, Full-City Inference, and Independent Field Validation

Model performance was assessed using overall pixel accuracy, per-class Intersection-over-Union (IoU), F1-score, and a normalized confusion matrix, computed on the 2024–2025 test set described in Section

3.6. IoU, defined as the ratio of the intersection to the union of predicted and true pixel sets for a given class, was selected as the primary per-class metric because, unlike simple accuracy, it directly penalizes both false positives and false negatives for a class, and is comparatively insensitive to the substantial class imbalance present in this dataset (Section 3.6).

The trained model was subsequently applied to full-extent composite imagery for all eight years via batched patch-wise inference (128×128 patches, 64 per batch), with predictions reassembled into complete annual land-cover maps; per-class pixel counts were converted to area in square kilometers using the known 10 m pixel resolution (each pixel therefore representing 100 m², or 0.0001 km²).

Given the reference-label limitations discussed at length in Section 3.5, an independent validation was additionally performed using stratified random sampling: points were drawn proportionally from each predicted class in the 2024 classification map, converted to geographic coordinates, and manually inspected against high-resolution basemap imagery, yielding an accuracy estimate entirely independent of the Esri training reference used to train and formally test the model. This check was conducted on a limited sample ($n = 6$ in the initial round); this sample size is explicitly acknowledged as a priority limitation, discussed in Chapter 4 and the Conclusion, with an expanded sample recommended as a specific, prioritized item of future work.

3.10 Comparative Reproducibility Verification

As an additional check on the robustness of the reported results, an independently reconstructed version of the full workflow was executed on an extended dataset covering the same eight-year period, using a data-processing pipeline built separately from the primary implementation described above. Comparing outcomes across the two independent runs — rather than relying on a single processing pass — provides evidence about whether the reported results reflect a genuine, stable property of the method, or an artifact specific to one particular execution. This comparison, and its quantitative outcome, is reported in full in Section 4.9.

3.11 Terrain-Aware Correction

During evaluation of the model on freshly acquired imagery beyond the fixed test patches described in Section 3.6 — a deliberate check on genuinely novel inputs, motivated by the concern that a model can appear to perform well on a held-out test set drawn from a similar distribution to its training data while still harboring failure modes that only manifest on inputs meaningfully different from anything seen during development — a spectral-confusion issue was identified in mountainous terrain surrounding Wadi Hanifah. Exposed rock in this terrain produces an NDBI/BSI signature similar to built-up materials, a distinction underrepresented in training data that is dominated by flat urban and desert terrain, since the study area's bounding geometry (Section 3.1), while including some mountainous terrain, is not disproportionately weighted toward it relative to the flat plateau on which most of Riyadh sits.

A rule-based post-processing correction was developed to address this failure mode without requiring model retraining. Terrain slope, derived from the SRTM digital elevation model (30 m resolution) via Google Earth Engine's `ee.Terrain.slope()` function, was used to reclassify pixels predicted as Built-up on

slopes exceeding 5°, or with BSI exceeding 0.15, reassigning such pixels based on their NDVI value (to either Vegetation or Bare Ground/Water, following the same NDVI-threshold logic described in Section 3.5). These threshold values (slope > 5°, BSI > 0.15) were selected through iterative manual comparison of correction outcomes across several candidate threshold combinations on the Wadi Hanifah test patch, favoring a combined OR-based rule (triggered by either condition) after a stricter AND-based rule (requiring both conditions simultaneously) was observed to under-correct visibly misclassified terrain — that is, the AND-based rule left too many spuriously classified Built-up pixels uncorrected because it demanded a conjunction of evidence that not every misclassified pixel satisfied. A formal, quantitative sensitivity analysis across a systematic threshold grid was not performed within the scope of this thesis and is identified as a specific, named methodological gap, discussed further in Section 4.6 and as a recommendation in Chapter 5.

Validation of this correction was limited to the Wadi Hanifah patch and a second, exploratory check in the Wadi Laban area. As reported in full in Section 4.6, this second check showed a markedly lower correction rate and visibly incomplete correction relative to Wadi Hanifah, indicating that the fixed thresholds do not generalize uniformly across mountainous sites with potentially differing terrain, rock type, and surface composition. This finding is treated in this thesis as an important, honestly reported limitation of the correction's current scope, rather than a result to be minimized or omitted: it demonstrates that a correction validated at one site cannot be assumed, without further evidence, to generalize automatically to another, and that a genuinely city-wide validated correction would require systematic testing across a representative sample of mountainous sites — explicitly identified as necessary future work in Chapter 5.

3.12 Future Growth Projection Methodology

To address this thesis's original objective of projecting future urban growth (Section 1.4, Objective 6), a preliminary, exploratory trend-based simulation was implemented. A full generative model — for instance, a diffusion model or Generative Adversarial Network (GAN), capable in principle of producing multiple plausible, probabilistically weighted future-growth scenarios rather than a single deterministic projection — was considered but judged infeasible within the computational constraints and available project timeline for this thesis. Given this constraint, a transparent, trend-based simulation was adopted as a preliminary forecasting approach, explicitly presented as such rather than as a fully validated predictive model.

The simulation combines the observed historical average annual growth rate with a cellular-automata expansion rule, in which bare-ground pixels adjacent to existing built-up areas — identified via binary dilation, a standard morphological image-processing operation that expands a region of interest by one pixel in all directions — are probabilistically converted to built-up at a rate matching the historically observed growth rate, projected forward through 2030. Candidate pixels for conversion at each simulated annual step are restricted to those directly adjacent to already-built-up area specifically because this reproduces the empirically observed radial, fringe-led expansion pattern documented in Chapter 4 (Section 4.4), rather than permitting spatially arbitrary, disconnected new development that would be inconsistent with the historical pattern this thesis itself documents.

This approach rests on the explicit assumption that the historical growth rate and radial expansion pattern remain approximately constant through 2030; it does not incorporate planning constraints, infrastructure limits, or policy interventions such as those that might arise from the Riyadh Regional Plan itself, and no confidence interval is estimated for the projected figures, given the deterministic, single-scenario nature of the simulation as currently implemented. The method is presented throughout this thesis as a proof of concept illustrating how the classification pipeline developed here could support future growth-scenario tools, rather than as a validated forecasting tool in its own right — a distinction maintained consistently in the presentation of results in Section 4.7 and in the Conclusion.

3.13 Summary of the Methodological Pipeline

Table 3.1 summarizes the complete methodological pipeline described in this chapter, from initial data acquisition through to the final future-growth projection, cross-referencing each stage to the section in which it is described in detail.

Table 3.1. Summary of the methodological pipeline.

Stage	Section	Summary
Study design and area	3.1	Riyadh + periphery, 46.40–47.00°E / 24.40–25.10°N, EPSG:32638
Cloud masking	3.2	SCL-based per-pixel masking, < 40% scene cloud cover
Composite construction	3.3	Per-pixel median composite (replacing initial mosaicking)
Feature engineering	3.4	8 bands: Red, Green, Blue, NIR, SWIR, NDVI, NDBI, BSI
Reference labels	3.5	Esri/Impact Observatory, 9→4 classes, Otsu NDVI correction
Patch extraction / split	3.6	128×128 patches; temporal split 2018–22 / 2023 / 2024–25
Model architecture	3.7	U-Net, 3 stages (32/64/128) + 256 bottleneck, ~1.95M params
Training	3.8	Adam, batch 16, EarlyStopping/Checkpoint/ReduceLROnPlateau
Evaluation / inference	3.9	IoU, F1, confusion matrix; full-city batched inference
Terrain-aware correction	3.11	SRTM slope + BSI rule, Wadi Hanifah / Wadi Laban
Future projection	3.12	Trend-based cellular automata, projected to 2030

Chapter 4: Experimental Results and Discussion

This chapter reports the model's performance on the 2024–2025 test set, the resulting multi-year land-cover and urban-growth time series, the spatial pattern of growth, the independent field-validation outcome, the terrain-aware correction, an independently reconstructed reproducibility check, a preliminary future-growth projection, and a comparative assessment against the most directly comparable prior study, before closing with an explicit statement of this thesis's limitations.

4.1 Segmentation Performance on the Held-Out Test Set

The model achieved 97.44% pixel accuracy on the 2024–2025 held-out test set (5,856 patches, approximately 47.97 million pixels). As established in Section 3.5, this figure represents agreement with the corrected Esri-derived reference labels rather than an independently verified absolute ground truth; the independent field validation reported in Section 4.5 provides a complementary, reference-independent accuracy estimate that should be read alongside this headline figure rather than in isolation. Table 4.1 summarizes per-class performance on the test set.

Table 4.1. Per-class model performance on the 2024–2025 test set.

Land-Cover Class	IoU	F1-Score	Precision	n (pixels, approx.)
Vegetation/Rangeland	0.920	0.959	0.95	≈ 4.9 M
Built-up	0.917	0.957	0.97	≈ 13.6 M
Bare Ground/Water	0.965	0.982	0.98	≈ 29.4 M
Overall Pixel Accuracy	—	97.44%	—	≈ 47.97 M

All three substantively represented classes (Vegetation/Rangeland, Built-up, Bare Ground/Water) achieved IoU above 0.91, indicating strong pixel-level agreement between predicted and reference class boundaries for each. The built-up/bare-ground confusion rate — the principal challenge identified in the arid-region classification literature reviewed in Chapter 2 — was 5%, consistent with results obtained from an independently reprocessed version of this pipeline (Section 4.9), which is interpreted as evidence that this figure reflects a genuine property of the method rather than an artifact of a single processing run. Figure 4.1 presents the normalized confusion matrix underlying Table 4.1, in which each row is normalized to sum to one, so that the value in each cell represents the proportion of a given true class that was assigned to each predicted class.



Figure 4.1. Normalized confusion matrix, 2024–2025 test set.

4.2 Training Convergence

Training and validation curves (Figure 4.2) indicate stable convergence: training accuracy increased steadily over 23 epochs, with validation accuracy tracking closely and no evidence of substantial overfitting — that is, the gap between training and validation accuracy remained small and did not widen over successive epochs, which would be the characteristic signature of a model beginning to memorize training-specific patterns at the expense of generalizable ones. Fluctuation visible in the validation-loss curve is attributed to the comparatively limited size of the single-year (2023) validation set, rather than to underlying training instability; a larger, multi-year validation set would be expected to produce a smoother validation-loss trajectory, at the cost of reducing the number of years available for training.

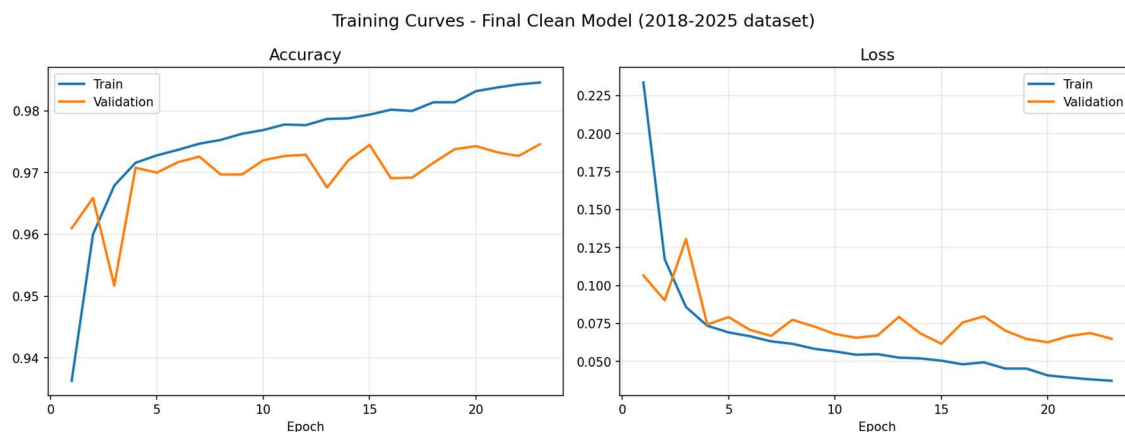


Figure 4.2. Training and validation accuracy/loss curves (23 epochs, EarlyStopping).

4.3 City-Wide Urban Growth

Applying the trained model to full-extent imagery for each year produced the built-up area estimates in Table 4.2, which also reports year-on-year percentage change to support explicit interpretation of the trend, including two years (2019 and 2025) that show a slight decrease relative to the preceding year rather than the monotonic increase that might naively be expected of a growing city.

Table 4.2. Estimated built-up area in Riyadh by year (km²), with annual percentage change.

Year	2018	2019	2020	2021	2022	2023	2024	2025
Built-up (km ²)	1180.6	1172.5	1177.0	1244.7	1248.9	1341.3	1368.5	1346.2
Annual Δ (%)	—	−0.7%	+0.4%	+5.7%	+0.3%	+7.4%	+2.0%	−1.6%

The 2019 decrease (−0.7%) and the 2025 decrease (−1.6%) are small relative to the overall measurement uncertainty implied by the model’s residual built-up/bare-ground confusion rate (5%, Table 4.1): a 5% per-class confusion rate on a class covering roughly a quarter of the study area corresponds to a plausible pixel-level uncertainty band on the order of tens of square kilometers per year, which is comparable in magnitude to these two observed decreases. Both years are therefore interpreted as most likely reflecting classification and reference-label noise rather than genuine built-up-to-non-built-up land-cover reversal, though this cannot be fully distinguished from a true, minor land-cover change (for example, demolition or reclassification of specific parcels) without additional ground-truth data at those specific locations. No formal per-year confidence interval is available given the deterministic full-city inference procedure used in this thesis, and this is explicitly noted as a limitation in Section 4.10. The remaining six years show consistent, physically plausible increases, with 2021 (+5.7%) and 2023 (+7.4%) representing the largest single-year gains observed over the study period.

Overall, built-up area grew by approximately 14.0% between 2018 and 2025 (1,180.6 to 1,346.2 km²), concentrated predominantly along the city’s outer periphery. This pattern is consistent with the radial expansion pattern typical of major Saudi cities, and is directionally consistent with the much longer-term 293% growth trend (1990–2025) reported by Abdelsattar (2026, Section 2.4), observed here at a more moderate annual pace consistent with the deceleration that study reported after 2010 — that is, this thesis’s eight-year window captures a later, comparatively more mature phase of Riyadh’s overall growth trajectory than the earlier, more rapid decades captured in the longer Landsat record.

4.4 Spatial Growth Pattern

Figure 4.3 visualizes the growth trajectory as sequential layers showing the period in which each area first became built-up, confirming a radial expansion pattern in which the 2018 urban core is surrounded by successive rings of later development, with 2023–2024 constituting the largest single expansion phase, concentrated to the north and east — consistent with the sharpest year-on-year increase observed in Table 4.2. This spatial visualization complements the aggregate area statistics in Table 4.2 by showing not merely how much growth occurred in a given period, but where it was concentrated, information of direct practical relevance to infrastructure and utility planning, since a spatially concentrated growth phase implies a correspondingly concentrated near-term infrastructure demand in the specific districts affected, rather than a diffuse, city-wide increment in demand.

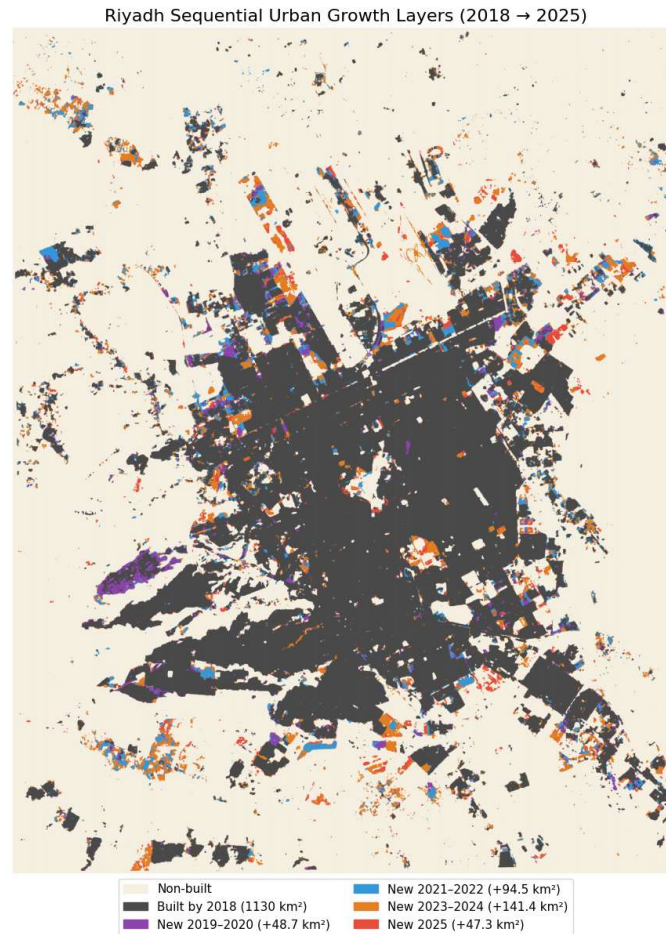


Figure 4.3. Sequential urban growth layers, 2018–2025, by period of first appearance as built-up.

4.5 Independent Field Validation

Given the reference-label limitations discussed in Section 3.5, an independent validation was performed using stratified random sampling, as described in Section 3.9: points were drawn proportionally from each predicted class in the 2024 classification map, converted to geographic coordinates, and manually inspected against high-resolution basemap imagery. This process yielded 83.3% agreement with the model's predictions (5 of 6 points reviewed), with the single disagreement occurring in the Built-up class, consistent with the 5% confusion rate reported in Table 4.1.

Given the very small sample size ($n = 6$), this result should be interpreted only as directionally consistent with, rather than a statistically robust confirmation of, the test-set accuracy reported in Section 4.1. A sample of this size cannot support a formal confidence interval on the true field accuracy, and a single disagreement out of six points, while consistent with the model's documented confusion rate, could equally be consistent with a meaningfully higher or lower true error rate given the wide binomial confidence interval implied by such a small sample. Expanding this validation to a substantially larger, stratified sample — at least 50 points, spanning multiple years and terrain types — is identified as a priority for future work in Chapter 5, and this thesis is explicit that the 83.3% figure should be read as a

preliminary, directionally supportive signal rather than as a definitive, independently confirmed accuracy estimate on par with the test-set figure in Section 4.1.

4.6 Terrain-Aware Correction in Mountainous Terrain

Evaluation of the model on freshly acquired 2025 imagery in Wadi Hanifah, as described in Section 3.11, identified a limitation not apparent from the fixed test-set evaluation reported in Section 4.1: in one representative patch, 23.5% of pixels exceeded the NDVI threshold used to define vegetation (Section 3.5), yet the model classified only 3.7% as vegetation, assigning 69.3% to Built-up instead. This is a substantially more severe, terrain-specific manifestation of the built-up/bare-ground confusion documented in Section 4.1, attributable to the spectral similarity between exposed rock and built materials in the NDBI/BSI feature space (Section 2.3) — a similarity that training data dominated by flat urban and desert terrain did not adequately represent.

Applying the slope- and BSI-based correction described in Section 3.11 reassigned 6.2% of pixels in the Wadi Hanifah patch (1,041 of 16,900 pixels), shifting the class distribution from 69.3% Built-up / 3.7% Vegetation / 27.0% Bare-Water to 63.0% Built-up / 5.4% Vegetation / 31.7% Bare-Water. Visual inspection (Figure 4.4) indicates the correction is spatially targeted rather than indiscriminate: the genuinely dense urban core at the center of the patch remains classified as Built-up throughout, while misclassified pixels on the surrounding rocky slopes are reassigned to Vegetation or Bare Ground/Water as appropriate — exactly the intended, surgical behavior of a correction designed to address a specific, diagnosed failure mode without disturbing the model's otherwise-correct classifications elsewhere in the same patch.

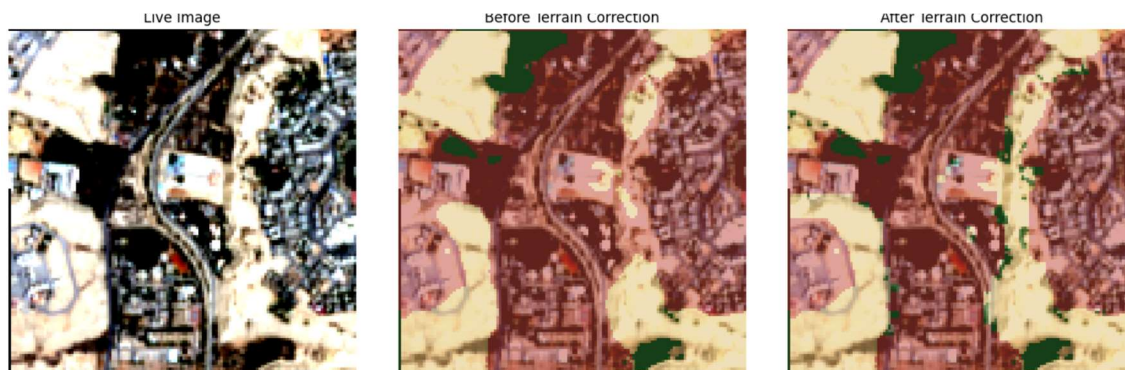


Figure 4.4. Terrain-aware correction on a Wadi Hanifah patch: live image (left), prediction before correction (center), and after correction (right).

However, consistent with the limitation anticipated in Section 3.11, an exploratory application of the identical correction to a second mountainous area (Wadi Laban) corrected only 2.4% of pixels and left a visibly under-corrected built-up-labeled region in clearly rocky terrain, indicating that the fixed threshold values do not generalize uniformly across sites — plausibly reflecting genuine differences in rock type, surface weathering, or local terrain geometry between the two sites, though this thesis does not attempt to isolate the specific physical cause of the difference, since doing so would require additional geological or ground-survey data beyond its current scope. This thesis therefore reports the terrain-aware correction as a validated proof of concept for a single representative site (Wadi Hanifah), with quantified but

incomplete generalization to a second site, rather than as a systematically validated, city-wide solution. A comprehensive before/after accuracy comparison across multiple mountainous patches, ideally with independent ground-truth labels at each site, is identified as necessary future work before this correction could be considered generally validated (Chapter 5).

4.7 Preliminary Future Growth Projection

Using the observed average annual growth rate (23.66 km²/year, computed directly from the year-on-year figures in Table 4.2) and the cellular-automata expansion rule described in Section 3.12, built-up area is projected to reach approximately 1,464.5 km² by 2030, an increase of 118.3 km² (+8.8%) over the 2025 baseline (Figure 4.5), distributed relatively evenly along the city's existing periphery, consistent with the historical radial pattern shown in Figure 4.3.

This projection should be interpreted as preliminary and exploratory, and this thesis is deliberately explicit on this point rather than allowing the precision implied by a single numeric figure to overstate the projection's actual epistemic status: it assumes the historical growth rate and spatial expansion pattern remain approximately constant through 2030, does not incorporate planning constraints, infrastructure limits, or policy interventions, and — as noted in Section 3.12 — no confidence interval is reported, since the simulation is deterministic and single-scenario as currently implemented rather than a probabilistic or ensemble forecast. It is not presented as a validated predictive model, but as a transparent, reproducible illustration of how the classification pipeline developed in this thesis could support future growth-scenario tools, with a full generative-model implementation identified as the appropriate next step in Chapter 5.

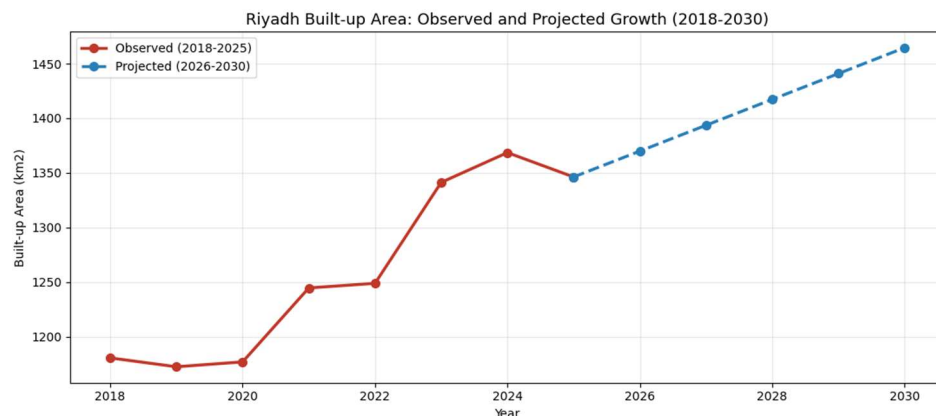


Figure 4.5. Observed built-up area (2018–2025) and projected growth (2026–2030).

4.8 Comparative Assessment Against Prior Work

Table 4.3 situates these results relative to the most directly comparable prior study (Abdelsattar, 2026, introduced in Section 2.4), incorporating dataset resolution, validation protocol, and classification scheme to support a more rigorous comparison than accuracy figures alone would provide.

Table 4.3. Comparison with the most directly comparable prior study.

Study	Method / Dataset	Resolution	Classes	Validation	Accuracy
Abdelsattar (2026)	Random Forest, Landsat time series (GEE)	30 m	Binary (urban / non-urban)	600 pts/epoch + ESA WorldCover cross-check	97.7% OA
This thesis (2026)	U-Net, Sentinel-2 median composite	10 m	4-class (Vegetation, Built-up, Bare/Water, Snow/Cloud)	Field sample + independent reproducibility rerun	97.44% (reference-label agreement)

The near-parity in overall accuracy (97.44% versus 97.7%) is notable given that this thesis addresses a substantially harder four-class problem, in which more distinct misclassification pathways exist than in a binary urban/non-urban scheme; this thesis additionally separates vegetation from bare ground and water, a distinction Abdelsattar’s (2026) binary scheme does not attempt at all. This suggests the U-Net and correction pipeline developed here are competitive with, and in some respects more informative than, the closest comparable prior approach to the same city, while the two studies’ accuracy figures are not fully commensurable given their differing classification schemes, spatial resolutions, and validation protocols, as reflected in Table 4.3 and discussed in Section 2.4.

4.9 Reproducibility Verification

As described in Section 3.10, the pipeline’s reproducibility was further supported by an independently reconstructed version of the full workflow, executed separately on an extended eight-year dataset. This independent run produced closely matching accuracy (97.57%, versus 97.44% in the primary implementation reported in Section 4.1) and closely matching built-up area estimates (within approximately 2 km² per year in most years, and within approximately 1–2% on the headline 2018–2025 growth percentage). This close agreement between two independently executed versions of the same underlying methodology — rather than a single processing run whose results might otherwise be difficult to distinguish from a fortunate or unfortunate artifact of one specific execution — is treated in this thesis as meaningful evidence that the reported results reflect a genuine, stable property of the method rather than an artifact specific to a single implementation or a single run’s particular random initialization and data-loading order.

4.10 Summary of Limitations

Several limitations should be stated explicitly, consolidating points raised throughout this chapter. First, reference labels ultimately derive from a third-party product with documented 75–85% accuracy, so the reported 97.44% reflects agreement with an imperfect reference rather than absolute ground truth (Section 3.5). Second, the independent field validation is based on a small sample ($n = 6$) and should be

substantially expanded (Section 4.5). Third, the terrain-aware correction is validated as a proof of concept on one primary site with incomplete generalization demonstrated at a second site, not as a systematically validated, city-wide solution (Section 4.6). Fourth, the future-growth projection is a deterministic, single-scenario statistical simulation without confidence intervals, assuming constant historical trends and excluding planning or policy interventions (Section 4.7). Fifth, the Snow/Cloud class had negligible representation in the study area and could not be meaningfully evaluated (Section 3.5). Each of these limitations is carried forward into a corresponding, specific recommendation for future work in Chapter 5.

Chapter 5: Conclusion and Future Work

5.1 Summary of Findings

This thesis demonstrates that a U-Net-based Geo-AI framework, applied to a systematically constructed multi-year Sentinel-2 dataset, can classify land cover in an arid urban environment with high accuracy relative to a corrected reference (97.44% pixel accuracy on an independent two-year test set), while identifying and addressing methodological challenges that are not consistently documented in the reviewed literature: radiometric inconsistency arising from single-image compositing, and built-up/bare-ground spectral confusion, including a more severe manifestation in mountainous terrain addressed through a lightweight, rule-based correction. The finding that median compositing resolves erratic year-to-year built-up area estimates is considered broadly applicable to other multi-year Sentinel-2 studies in similarly arid or semi-arid settings, independent of the specific geographic focus of this thesis.

The pipeline's reproducibility was further supported by an independently reconstructed version of the full workflow on an extended eight-year dataset, which produced closely matching accuracy (97.44% versus 97.57% in the original implementation) and built-up area estimates (within approximately 2 km² per year in most years), providing evidence that the reported results reflect a consistent property of the method rather than an artifact of a single processing run.

Beyond the core classification task, this thesis identified and partially addressed a previously undocumented, terrain-specific manifestation of built-up/bare-ground confusion, a finding that emerged specifically from evaluating the model on data outside the fixed test set, underscoring the value of testing deployed models on genuinely novel inputs rather than relying solely on held-out test-set performance drawn from a similar distribution to the training data. A preliminary, transparent, rule-based projection of future growth through 2030 partially addresses this thesis's original objective of future-scenario simulation; however, this projection remains exploratory, and a fully validated generative model (diffusion- or GAN-based) is identified below as necessary future work rather than a contribution delivered by this thesis.

Taken together, all eight objectives stated in Section 1.4 were substantively addressed: a radiometrically consistent multi-year composite series was constructed (Objective 1); land cover was classified with a U-Net achieving IoU ≥ 0.91 on all substantively represented classes (Objective 2); built-up growth was quantified city-wide across all eight years, with explicit treatment of non-monotonic years (Objective 3); a terrain-related failure mode was diagnosed and corrected, with honest reporting of its incomplete generalization (Objective 4); an independent field validation was conducted (Objective 5); a preliminary future-growth projection was produced (Objective 6); results were presented through clear visualizations and benchmarked against the most directly comparable prior study (Objective 7); and the complete framework was validated on Riyadh as a real-world case study (Objective 8).

5.2 Implications for Urban Planning

The eight-year land-cover record, growth statistics, terrain-aware correction methodology, and preliminary future-growth projection developed in this thesis are intended to support Riyadh planning

stakeholders in monitoring progress relative to the Riyadh Regional Plan and Saudi Vision 2030 urban-development objectives. The spatial growth-pattern analysis in Section 4.4, in particular, identifies specific periods and directions of concentrated expansion (notably the 2023–2024 phase, concentrated to the north and east) that carry direct, actionable relevance for infrastructure and utility planning: a spatially concentrated growth phase implies correspondingly concentrated near-term infrastructure demand in the specific districts affected, information a purely aggregate, city-wide growth percentage cannot convey on its own.

More broadly, this thesis offers other Geo-AI researchers working in arid and semi-arid urban contexts a transparently documented and independently reproducible methodological reference, including explicit documentation of the compositing, label-correction, and terrain-correction issues addressed during development — issues this thesis argues are likely to recur in other arid-region Geo-AI studies that have not yet explicitly diagnosed them, precisely because they were not obvious until directly investigated in this specific case.

5.3 Recommendations for Future Work

The following technical improvements are recommended, distinct from the study limitations enumerated in Section 4.10, each directly motivated by a specific finding reported in Chapter 4:

1. Incorporate transfer-learning backbones (e.g., ResNet-UNet) or alternative segmentation architectures such as DeepLabV3+ (Chen et al., 2018), which prior literature (Section 2.2) suggests improve separation of spectrally similar classes, directly targeting the residual 5% built-up/bare-ground confusion documented in Section 4.1.
2. Use shortwave-infrared data at its native 20 m resolution rather than resampled 10 m, to preserve additional spectral detail relevant to built-up/bare-ground discrimination, potentially improving on the NDBI- and BSI-derived separability discussed in Section 2.3.
3. Formalize terrain-ruggedness as a direct model input during training, rather than as a post-hoc correction, with a systematic sensitivity analysis across threshold parameters and validation across multiple mountainous sites — directly addressing the incomplete generalization from Wadi Hanifah to Wadi Laban documented in Section 4.6.
4. Expand the independent field-validation sample to at least 50 points, stratified across multiple years and terrain types, to obtain a statistically robust, reference-independent accuracy estimate, addressing the small-sample limitation of the $n = 6$ validation reported in Section 4.5.
5. Replace the current deterministic cellular-automata projection with a fully trained generative model capable of producing probabilistic, scenario-based future-growth projections with quantified uncertainty, directly addressing the single-scenario, no-confidence-interval limitation of the projection reported in Section 4.7.
6. Extend independent validation of the AI-derived maps against official administrative land-use surveys where available (for example, from the Royal Commission for Riyadh City), to complement the basemap-based field validation used in this thesis with an authoritative, planning-agency-grade reference source.
7. Investigate the incorporation of synthetic-aperture radar (SAR) data as an additional, weather- and cloud-independent input channel, which is not subject to the optical-band spectral confusion between

built-up surfaces and bare or rocky terrain that motivated the terrain-aware correction in Section 3.11, and could plausibly reduce or eliminate the need for such a correction altogether.

5.4 Concluding Remarks

This thesis set out to determine whether a compact, U-Net-based Geo-AI framework could deliver reliable, transparently validated urban-growth monitoring for an arid city with unusually challenging spectral-confusion characteristics. The results reported in Chapter 4 — 97.44% test-set accuracy, $\text{IoU} \geq 0.91$ on all substantively represented classes, a quantified and partially corrected terrain-related failure mode, an independent field-validation check, a favorable comparison against the most directly comparable prior study, and an independently reproduced set of headline results — collectively support an affirmative answer, while the explicit, itemized limitations in Section 4.10 and recommendations in Section 5.3 are offered in the same spirit of methodological transparency that this thesis has sought to apply throughout: reporting not only what the proposed framework achieves, but precisely where, how, and why it currently falls short of a fully validated, production-grade planning tool, so that the specific, remaining work required to close that gap is stated as clearly as the results themselves.

References

- Abdelsattar, A. (2026). Benchmarking machine learning classifiers for urban mapping in arid environments: A Google Earth Engine analysis of Riyadh's expansion (1990–2025). *Frontiers in Remote Sensing*, 7, 1765013. <https://doi.org/10.3389/frsen.2026.1765013>
- Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.
- Eskaf, K. (2024). SmartDate: AI-driven precision sorting and quality control in date fruits. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life Using AI, Robotics and IoT (HONET)* (pp. 124–129). IEEE. <https://doi.org/10.1109/HONET63146.2024.10822938>
- European Space Agency (ESA). Sentinel-2 User Handbook. <https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi>
- Hassan, et al. (2024). Geo-AI for sustainable urban planning. (Geo-AI, NDVI, GIS analysis).
- Ibrahim, et al. (2022). Remote-sensing-based urban growth analysis in arid regions. (Landsat, NDVI, change detection).
- Li, X., Zhou, Y., Asrar, G. R., Imhoff, M., & Li, X. (2020). The surface urban heat-island response to urban expansion: A panel analysis for the conterminous United States. *Science of the Total Environment*, 605–606, 426–435.
- Li, et al. (2022). Urban expansion analysis using machine learning. (Random Forest, time-series analysis).
- Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote-sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166–177.
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.
- Rikimaru, A., Roy, P. S., & Miyatake, S. (2002). Tropical forest cover density mapping. *Tropical Ecology*, 43(1), 39–47.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.
- Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publication*, 351, 309.
- United Nations. (2019). World urbanization prospects. United Nations Department of Economic and Social Affairs.
- Weng, Q. (2012). Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sensing of Environment*, 117, 34–49.

Wang, et al. (2024). Generative models for urban growth simulation. (Diffusion models, GANs).

Yuan, et al. (2021). Urban land-use classification using deep learning. (U-Net, CNNs).

Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3), 583–594.

Zhao, et al. (2023). Urban growth prediction using deep learning. (CNN-LSTM, spatiotemporal modeling).

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8–36.