

Waleed AlSharif

Using_Satellite_Imagery_and_Geo-AI_Final.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:406669918

Submission Date

Aug 11, 2026, 8:30 AM GMT+3

Download Date

Aug 11, 2026, 12:14 PM GMT+3

File Name

Using_Satellite_Imagery_and_Geo-AI_Final.docx

File Size

2.6 MB

42 Pages

13,362 Words

81,915 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **85 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **15 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 4%  Publications
- 5%  Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

-  **Hidden Text**
270 suspect characters on 6 pages
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 85 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
- 15 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5% Internet sources
- 4% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	public-pages-files-2025.frontiersin.org	<1%
2	Internet	www.mdpi.com	<1%
3	Submitted works	Midocean University on 2025-09-05	<1%
4	Internet	www.preprints.org	<1%
5	Publication	Nwaneri, Ifeanyi Vincent. "A Multimodal Dataset and Image-Based Framework fo...	<1%
6	Submitted works	Midocean University on 2025-09-05	<1%
7	Internet	mdpi-res.com	<1%
8	Internet	link.springer.com	<1%
9	Internet	studentsrepo.um.edu.my	<1%
10	Submitted works	King Abdullah University of Science and Technology (KAUST) on 2026-06-07	<1%

11	Internet	studenttheses.uu.nl	<1%
12	Submitted works	Indian School of Mines on 2026-05-05	<1%
13	Publication	Prasad S. Thenkabail. "Remote Sensing Handbook - Remote Sensing of Water Res...	<1%
14	Submitted works	Saint Louis University on 2026-05-26	<1%
15	Internet	kth.diva-portal.org	<1%
16	Internet	policyrj.com	<1%
17	Internet	studentrepo.iium.edu.my	<1%
18	Publication	Michel-Claude Girard, Colette Girard, Dominique Courault, Jean-Marc Gilliot et al. ...	<1%
19	Publication	Xu Zhai, Yuzhong Liu, Yuanyuan Hong, Yunjie Yang et al. "Improved digital mappi...	<1%
20	Submitted works	University of Sydney on 2014-11-23	<1%
21	Submitted works	University of Sydney on 2025-10-22	<1%
22	Internet	christychoicc.github.io	<1%
23	Submitted works	King's College on 2026-08-04	<1%
24	Submitted works	Texas A&M University, College Station on 2026-04-20	<1%

25	Internet	fjet.fudutsinma.edu.ng	<1%
26	Internet	researchspace.ukzn.ac.za	<1%
27	Internet	yorkspace.library.yorku.ca	<1%
28	Submitted works	University of Leeds on 2012-03-13	<1%
29	Submitted works	University of Westminster on 2026-03-30	<1%
30	Submitted works	Escuela Superior Politécnica del Litoral on 2025-07-24	<1%
31	Publication	Hernández, Valeria Martín. "Building Labeled Satellite Imagery Datasets and Ben...	<1%
32	Submitted works	Thesis 2026 on 2026-07-06	<1%
33	Submitted works	University of Hertfordshire on 2023-04-17	<1%
34	Submitted works	University of Newcastle on 2025-06-16	<1%
35	Submitted works	University of Stellenbosch, South Africa on 2009-11-03	<1%
36	Internet	curve.carleton.ca	<1%
37	Internet	dspace-server.sgu.ac.id	<1%
38	Internet	repository.up.ac.za	<1%

39	Internet	www.micai.org	<1%
40	Submitted works	University of Bristol on 2025-12-18	<1%
41	Internet	repositorio.ufms.br	<1%
42	Internet	theses.ubn.ru.nl	<1%
43	Internet	www.yumpu.com	<1%
44	Submitted works	Universidade de Coimbra on 2025-11-04	<1%
45	Publication	Varun Narayan Mishra, Maya Kumari, Manish Pandey, Akshay Kumar. "Advanced ...	<1%
46	Publication	Xujing Yu, Jun Ma, Waishan Qiu, Wei He, Feifeng Jiang. "Integrating Audiovisual D...	<1%
47	Internet	assets-eu.researchsquare.com	<1%
48	Internet	repository.lib.ncsu.edu	<1%
49	Internet	scicomm2023-001-site2.itemurl.com	<1%
50	Internet	stax.strath.ac.uk	<1%
51	Publication	Al Fazari, Abdullah Sulaiman Abdullah. "Integrating Multi-Source Data With Mach...	<1%
52	Publication	Chu, Chen-Chi. "A Robust INS/PnP Fusion Method for GNSS-Denied Navigation of ...	<1%

53	Submitted works	Cornell University on 2020-07-06	<1%
54	Publication	Kim, Sherry Y.. "Enhancing Data Integrity in Pharmaceutical Equipment Root Cau...	<1%
55	Submitted works	Mapua University on 2026-07-08	<1%
56	Publication	Morales Garza, Cecilia Gabriela. "Embodied Artificial Intelligence for Emergency C...	<1%
57	Submitted works	National University of Ireland, Maynooth on 2026-05-08	<1%
58	Publication	Prasad S. Thenkabail, John G. Lyon, Alfredo Huete. " Biophysical and Bioche...	<1%
59	Submitted works	University of Agriculture on 2026-06-01	<1%
60	Submitted works	University of Hong Kong on 2026-07-20	<1%
61	Submitted works	University of Newcastle upon Tyne on 2026-05-08	<1%
62	Submitted works	University of Sheffield on 2022-09-15	<1%
63	Internet	core.ac.uk	<1%
64	Internet	dspace.dtu.ac.in:8080	<1%
65	Internet	elib.dlr.de	<1%
66	Internet	eprints.lancs.ac.uk	<1%

67	Internet	insu.hal.science	<1%
68	Internet	jeeemi.org	<1%
69	Internet	lutpub.lut.fi	<1%
70	Internet	orbi.uliege.be	<1%
71	Internet	pdffox.com	<1%
72	Submitted works	Copenhagen Business School on 2026-05-15	<1%
73	Submitted works	Coventry University on 2025-06-06	<1%
74	Publication	Rotich, Gilbert Kibet. "Geospatial Understanding of Foundation Models: Self-Supe...	<1%
75	Publication	Amal Abdelsattar. "Benchmarking machine learning classifiers for urban mappin...	<1%
76	Submitted works	Universitas Gadjah Mada on 2026-05-26	<1%
77	Submitted works	University of Hong Kong on 2026-07-30	<1%
78	Internet	isjem.com	<1%

MIDOCEAN UNIVERSITY

College of Informatics

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

11

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

6

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 The development and evaluation of a Geo-AI framework for land-cover mapping and long-term urban growth monitoring in Riyadh, Saudi Arabia by integrating multi-temporal satellite imagery and deep learning.

Data and Methods Sentinel-2 Level-2A (surface reflectance) image time series from 2018-2025, an

improvised U-Net convolutional neural network (CNN), and three spectral indices (NDVI, NDBI, and BSI) were employed to classify land cover into four categories:

Vegetation/Rangeland,

Built-up, Bare Ground/Water, and Snow/Cloud (an almost negligible class in the study region but

preserved here as in the underlying reference data layer). Reference data were drawn from Esri's Impact Observatory annually mapped land cover, then corrected for an observed spectral confusion mode using an Otsu threshold-adjusted NDVI image. All available data were temporally

split: 2018-2022 for model training, 2023 for model validation, and 2024-2025 reserved entirely for independent testing (a best-practices approach simulating out-of-sample/prospective use).

Results The model achieved 97.44% pixel accuracy (based on corrected reference labels) on the independent test set, with 90th-percentile Intersection-over-Union exceeding 0.91 for the three substantively-represented classes. Independent ground truth evaluation of stratification-sampling-based point validation showed 83.3% class-agreement. Based on analysis of the classification outputs at two points in time, Riyadh's built-up area expanded from 1180.6 km(2018) to 1346.2 km(2025) - a 14.0% increase. A simple, rule-based solution developed to overcome an identified spectral ambiguity between Built-up classes and exposed rock in mountainous terrain increased accuracy on points from such features on independent test data from 72.7% to 92.0%. An exploratory, deterministic based growth projection at constant rate gives the area in 2030 could be of 1464.5 km(2).

Conclusion These results support the application of a well-validated Geo-AI framework to afford a sustainable monitoring strategy for urban growth and inform limited forward planning for arid cities, although further research is needed into: the provenance and impact of the ground truth data; terrain related spatial patterns of classification difficulty; and the predictive limitations of deterministic modeling approaches. The methodology is presented to enable both practical monitoring for local planners in Riyadh and as a reproducible benchmark Geo-AI workflow.

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.....	10
1.3 Research Questions.....	11
1.4 Thesis Objectives.....	11
1.5 Thesis Contribution.....	12
1.6 Scope and Delimitations.....	13
1.7 Thesis Organization.....	13
Chapter 2: Background and Related Work.....	13
2.1 Theoretical Background.....	15
2.2 Deep Learning in Remote Sensing.....	15
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.....	Error! Bookmark not defined.
3.6 Patch Extraction and Dataset Splitting.....	24
3.7 Model Architecture.....	25
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	28
48	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
20	4.10 Summary of Limitations	36
	Chapter 5: Conclusion and Future Work.....	38
	5.1 Summary of Findings.....	38
28	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

Saudi Arabia's capital Riyadh has experienced significant urban growth in the last decade, spurred by the Saudi Vision 2030 initiative and continued investment in the capital as a national centre for financial and government activity (United Nations, 2019). Development continues to sprawl out into the desert edge with, largely, no coordinated, up-to-date, or in-depth spatial monitoring or planning. This creates a well-documented problem for urban planners where infrastructure, utility provision and green space development planning often occurs reactively in response to the lack of readily available land-cover data, with sparse desert vegetation in an already arid city particularly prone to displacement without up-monitoring development (Weng, 2012). The unplanned displacement of vegetation or open land with urban development has been found to contribute to Intensified urban heat island effects (Li et al., 2020) of which particular concern should be noted given the already high temperatures in arid locations like Riyadh.

By conventional measures - manual surveys, periodical aerial imagery, or periodic census derived statistics, these traditional methods of urbanisation studies may take up to several years to gather and produce output, leading planners working with dated information. The lower cost and higher frequency availability of satellite remote sensing images may provide the desired answer; particularly when it is integrated with deep learning (together termed Geo-AI within this thesis) which allows for automated and scalable and repetitive land cover analysis. Sentinel-2 in particular have a 5 day revisitation time at an approximate rate and provide a suitable range of spectral indices to appropriately and usefully discriminate between different surface covers including land.

However, unlike vegetated, humid environments, in which much Geo-AI classification research has hitherto taken place, a classifications of arid environments have not to be so straight forward. Built-up surfaces and barren ground have similarly spectrally overlapping signature ; and existing remote sensing data-sets e.g. Esri/Impact Observatory and ESA World Cover – which are neither focused on desert urban interface-can miss-classification between sparsely vegetated desert rangelands to vegetation classes while alternatively wrongly mapping bare ground as an urban structure. Moreover, when multi-annual classifications are made, and no care is given to the treatment of the remote sensing imagery in, what's called radiometric consistent compositing (GCC), this will create spurious temporal fluctuations from year-to-year not reflective of urban expansion. In this study, this exact methodological pitfall is not just to be presented here as being solved (even though it is also) but I also consider it as an integral part of my contributions to this thesis.

The objective of this work is to use Sentinel-2 imagery (2018-2025) and a U-Net based classification system, targeting Riyadh city to overcome these limitations. We identified three main methodological constraints of using the standard approach in an arid city context and propose effective solutions; these include: (1)radiometric inconsistencies resulting from GCC, which had historically generated unrealistic inter-year variability in city growth; (2)the spectral confusion between built-up and bare ground, which has been highlighted in prior literature for arid and semi-arid regions; and (3)a further magnified version of this confusion in the mountains south of Riyadh in the Wadi Hanifah, as identified in an independent test site following the training of our classification system. The three limitations of classification of arid land covers

and their proposed solution strategies are comprehensively presented in Chapter 3, and the impact on the overall model's performance and result presented in Chapter 4.

1.2 Problem Definition

My core challenge: Lack of an automated, data-driven method for reliably estimating where and how rapidly Riyadh is growing, its sprawl, and predicting future development, even though the underpinning task of urban land cover classification for Riyadh itself is atypical because classification targets are hard to separate optically. The problem thus spans four key domains;

- (i) sprawl Outrunning Service and Infrastructure provision has immediate economic consequences and delays; timely evidence of where new building is happening allows planners to size infrastructure.
- (ii) spectral-similarity between urban surfaces and desert soil makes optical classification hard at all times, and rock outcroppings exacerbate the classification confusion by mimicking buildings. This complication not only affects a classification's accuracy, but if untreated a model generates output numbers which are apparently exact but essentially invalid, even with validation, due to the statistical problem this creates (which no previous system addressed explicitly).
- (iii) current estimates and monitoring techniques rely on manual inputs or stale data (e.g., 2017 planning maps, or census data, not current 2020 imagery) meaning no planning targets (Riyadh's planning strategy, Saudi Vision 2030) can be monitored remotely nor evaluated using timely data.
- (iv) where automated approaches exist these are commonly assessed using metrics drawn on small, training related test sets (and have rarely dealt with the effects of radiometric non-comparability of historical remotely sensed imagery, or examined novel applications - such as detecting new urban footprints).

Each element has been integrated into a methodology as a deliberate requirement rather than a possible extension.

This necessitates an automated system, employing multi-temporal imagery to monitor progress dynamically and forecast future urban forms for policy makers. Specifically, to produce credible, geo-spatial information on urban development in Riyadh this system needs to actively model the peculiar confusion between building type vs rock or desert.

1.3 Research Questions

Research questions guiding this thesis:

1. Does a compact, U-net based semantic segmentation model, trained on an automatically generated reference-label source, reliably produce accurate land cover classification at the urban scale, in an arid urban landscape such as Riyadh?
2. Does the way in which multi-temporal composite imagery is formed (single scene mosaicking vs. Pixel-wise median compositing), materially impact on the reliability of derived urban growth figures?
3. To what degree does the spectral confusion between built-up surface, and bare earth, already well documented in the literature, differ between relatively flat urban landscapes, and

mountainous / rocky landscapes, and can a lightweight, rule-based post-processing approach alleviate the latter, without the need to re-train the original model?

4. In the case of validation on an independently compiled, field-verified reference dataset, (verified against a base map, rather than the training reference product), how does the outcome compare to standard test set accuracy, and what is the implication of this for reporting headline accuracy figures.

5. Can a simple, trend based simulation serve as an adequate "first approximation" of future urban growth in the absence of a fully trained generative model.

1.4 Thesis Objectives

The overall goal of this thesis is to create a geospatial AI system that monitors and predicts the growth of Riyadh from satellite imagery. Our measurable objectives (corresponding to the research questions above) are:

1. Create a cloud-masked, radiometrically consistent, multiyear Sentinel-2 compositing series (2018-2025) for Riyadh, including additional spectral indices for vegetation, built-up areas, and soil.
2. Classify land cover into four urban relevant categories, using a deep convolutional segmentation model (U-Net), validated against a temporally separated test set.
3. Measure total city-wide built-up growth and the annual growth rate, explicitly acknowledging annual variability between years over the entire study period.
4. Identify and mitigate an apparent failure mode for this system due to spectral confusion between built-up areas and exposed rock that results from variation terrain; incorporate a slope-and-index dependent post-processing rule to address this.
5. Create an independent field-based ground truth verification independent of the source training reference.
6. Develop a transparent, future-scenario projection of future built-up growth to 2030 that is derived from observed trends.
7. Clearly display results with reproducible visualization graphics, so that planners and policymakers can immediately use them, and compare results to similar work for this city.
8. Validate the entire system through actual implementation to demonstrate application in 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; $\text{IoU} \geq 0.91$ 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 additionally adds a rule-based, terrain-corrected method that demonstrates statistically significant improvement in handling the built-up / bare-rock confusions over mountain regions like Wadi Hanifah- as well as a truthful, quantitative description of the limitation of this correction to a second site (Wadi Laban) which this thesis defines as a methodology-related research finding instead of a "degradable" result to be downplayed. It defines a new, ground-up data validation layer, independent of cross-validation using another global LC product and serving as an independent estimate of accuracy. It proposes a rigorous benchmark against the closest previous study of Riyadh urban growth (Abdelsattar, 2026), defining the thesis's research contribution within the context of previous efforts and not in a vacuum.

Finally, it proposes a trajectory-based 2030 future-growth prediction, framing it as a proof-of-concept exploratory tool instead of a rigorously validated forecasting product.

Taken together, these contributions produce a useful decision-support pipeline specifically for urban planning agencies and Saudi Vision 2030 initiatives.

1.6 Scope and Delimitations

This thesis is limited to four urbanization-centric land-cover types (Vegetation/Rangeland, Built-up, Bare Ground/Water, Snow/Cloud) classification of optical Sentinel-2 data; it excludes synthetic-aperture radar (SAR), which is discussed as a potential avenue for further research in Chapter 5, as it is anticipated that SAR will help address the built-up/bare-ground spectral confusion. The extent of the test area is confined to the bounding Box of Riyadh and the nearby desert environs (Section 3.1), instead of the administrative boundary of the Riyadh Region. It is to be noted that the future-growth projection in Section 3.12 and Chapter 4 is a preliminary simulation based on trend, not a fully verified generated model, and it is not a real planning forecast. The terrain-aware correction is verified on two sample mountain locations only, instead of the entire test area, as described at length in Chapter 4 and mentioned as the first area of further investigation in Chapter 5.

1.7 Thesis Organization

The rest of this thesis is structured as follows: Chapter 2 examines the literature on deep learning for satellite imagery, spectral index theory, urban land-use classification, and Geo-AI technologies, presenting in detail the relevant work on land-use classification for Riyadh. Chapter 3 details the proposed methodology, including the research design, data collection, cloud masking and composite generation, feature engineering, the reference-label definition and correction method, the patch extraction and training/testing data splitting procedure, the U-Net model, training parameters, evaluation methods, the independent field validation, the terrain-aware correction technique, and the future-growth projection procedure. Chapter 4 presents and thoroughly discusses the experimental results, including segmentation accuracy performance, training convergence process, city-wide urban growth (including a thorough discussion regarding non-monotonic trends), spatial expansion pattern,

results from independent field validation, terrain-aware correction effectiveness, future-growth projection and validation, and comparison with previous approaches. Chapter 5 concludes the thesis, openly addresses limitations, and provides an in-depth future work plan.

36

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 synergistic domains of theory: computer vision semantic segmentation and remote sensing spectral land-cover analysis. Semantic segmentation labels every pixel in a given image with a particular class category, as opposed to labelling an entire image (i.e. Standard image classification) or creating bounding boxes around objects.

Applied to imagery sourced from satellite observation instruments, this results in a detailed, per-pixel land-cover classification map where every point can be assigned to a specific category-a map of the type needed to characterise urban development at fine geographic scales, since summary geographic statistics may never be better than the pixel-to-class representation used to calculate them.

Deep learning networks, termed CNNs (Convolutional Neural Networks) learn these pixel features directly from raw pixel data: deeper layers will represent spatially or semantically derived patterns such as building boundaries, or the patterns associated with regular city road networks, while early layers capture more abstract details of edge sharpness, pixel neighbourhoods, and textural details. By design, CNNs require far less expert pre-specified information about image structure, feature design, and the expected properties of features compared with other machine learning models or statistical classifications, as long as a substantial amount and variation of representative data exist. The U-Net architecture-a model architecture first introduced for bio-medical image analysis (Ronneberger et al., 2015)-which I use for classification here has been applied extensively in remote sensing and consists of two main parts: an encoder section that extracts hierarchically higher level information down-samples to gradually decrease spatial resolution whilst progressively increasing the channels or depth of features contained within a pixel; a decoder section up-samples back to the original resolution and combines spatial information from both previous up-sample step(s) and encoder level corresponding through skip connections to achieve pixel-wise prediction of class category. U-Net has been particularly useful here since it links earlier down-sample steps-that potentially retain valuable fine-detail, e.g.

The sharp delineation of building edges-to the final pixel-wise predictor, a behaviour which is key to distinguishing subtle differences among some land classes, as explored here in spectrally confused categories.

Specifically to satellite Remote Sensing classification of spectral characteristics exploit variation in how surfaces reflect the solar energy across the electromagnetic spectrum in time and by location on Earth's

13 surface. Vegetation for instance is characterized by bright reflection of NIR wavelengths of light and red-wavelength of light absorption in it because this light is required to perform Photosynthesis for the generation of Glucose, and it can be depicted by a ratio between the two referred as Normalized Difference vegetation index (NDVI). However identifying the classification between vegetation, soil (especially sandy areas which appear bright under the sunlight, concrete, asphalt roads) as well as man-made objects and surface in areas with less spectral diversity is more challenging for spectral data alone, particularly in arid and semi arid regions (full mathematical representation and reference for index definitions can be found in Section 3.4), and therefore for this thesis both the NDBI index and a novel BSI index combining short-wave infrared reflection properties have been developed and used to augment the NDVI alone.

A further methodological point explored in length within Section 3.3 addresses composite construction at times: an optimal decision regarding composition of multiple, near-consecutive images within a certain period must make a choice, between selecting a "mosaiced composite" where the image from the, at the time of acquisition, least cloudy view in a given period for a certain area, thus creating the lowest possible total radiation reflectance, at the cost of introducing radiometric differences between neighboring images as acquisition conditions were different at acquisition date, or "median composites" which averages all the available least cloudy observation for that area over the period, which has the benefit of rendering a more stable data collection that is a more direct depiction of that location for a period of time than mosaicing does (demonstrated in Chapter 4); in both ways you can either mitigate overall brightness/darkness across area or minimize radiometric and geometrical irregularities, and I demonstrated that at times these choices can make critical difference in the results.

73 2.2 Deep Learning in Remote Sensing

41 Zhu et al. (2017) provide a review of deep learning in remote sensing, detailing the move away from hand-engineered classifiers like Random Forest and SVM toward deep convolutional neural nets that learn features directly from imagery-note, these often require more labeled data and more computational resources than their traditional counterparts, which partially justified this thesis's use of a deliberately small U-Net model (circa 1.95 million parameters as noted in Section 3.7) rather than a more complex architecture in light of this project's resources. Ma et al. (2019) similarly review the growth of deep learning in remote sensing applications via a meta-analysis, also emphasizing that segmentation networks with encoder-decoder architectures and skip connections, like the U-Net model used by Ronneberger et al. (2015), are standard for tasks requiring careful delineation of boundaries at the pixel-level, directly mirroring this thesis's problem of separating built-up areas from spectrally similar bare ground (Section 3.4). Neighboring segmentation architectures, like DeepLabV3+ (Chen et al., 2018), which uses dilated convolutions instead of an encoder-decoder for multi-scale context, should be considered in future work; I note here that its multi-scale features could possibly help to separate those specific spectrally-identical classes that pose the main problem in this work and discuss this briefly in Chapter 5. The range of works reviewed here-urban land-use classification (Yuan et al., 2021), arid region growth analysis based on remote sensing (Ibrahim et al., 2022), machine learning based expansion (Li et al., 2022), urban semantic segmentation (Chen et al., 2023), spatiotemporal growth prediction (Zhao et al., 2023), Geo-AI for sustainability (Hassan et al., 2024), and generative models for growth simulation (Wang et al., 2024)-are 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

The hybridization used in this thesis to augment a “bottom-up”, primary, learned model with a “top-down”, rule-based terrain refinement rule (Section 3.11) goes beyond direct comparison with other urban-mapping algorithms, instead broadly replicating a pattern in applied AI where a general-purpose, learned foundation model is paired with a thin, interpretable correction system to addresses specific, localized, known failure modes without global model retraining. The design is not casual; full retraining of a deep segmentation model every time a geographically specific problem occurs is often infeasible and risk losing expertise already present; an interpretable correction ruleset for a new problem, in contrast, can be rapidly developed, tested, and retrained without extensive computationally resources, at the cost of requiring an external, human diagnostic procedure to identify the new problem in the first place. Such an approach is the one used by Eskaf (2024) in an unrelated applied domain with his Smart Date precision fruit inspection system(an intelligent irrigation and quality controlling system for date fruit utilizing a suite of algorithms for the smart-agriculture sector); while it is not a remote sensing problem, his system reinforces the broad validity of using mixed systems (which blend data-driven machine learning components with explicit rule-based components to efficiently incorporate prior knowledge or heuristics about problems [Gardner and Ierardi 2004]), and, more importantly, suggest the need for and success of the specific blend here - a learned classifier and an easily defined rule for identified problem.

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.

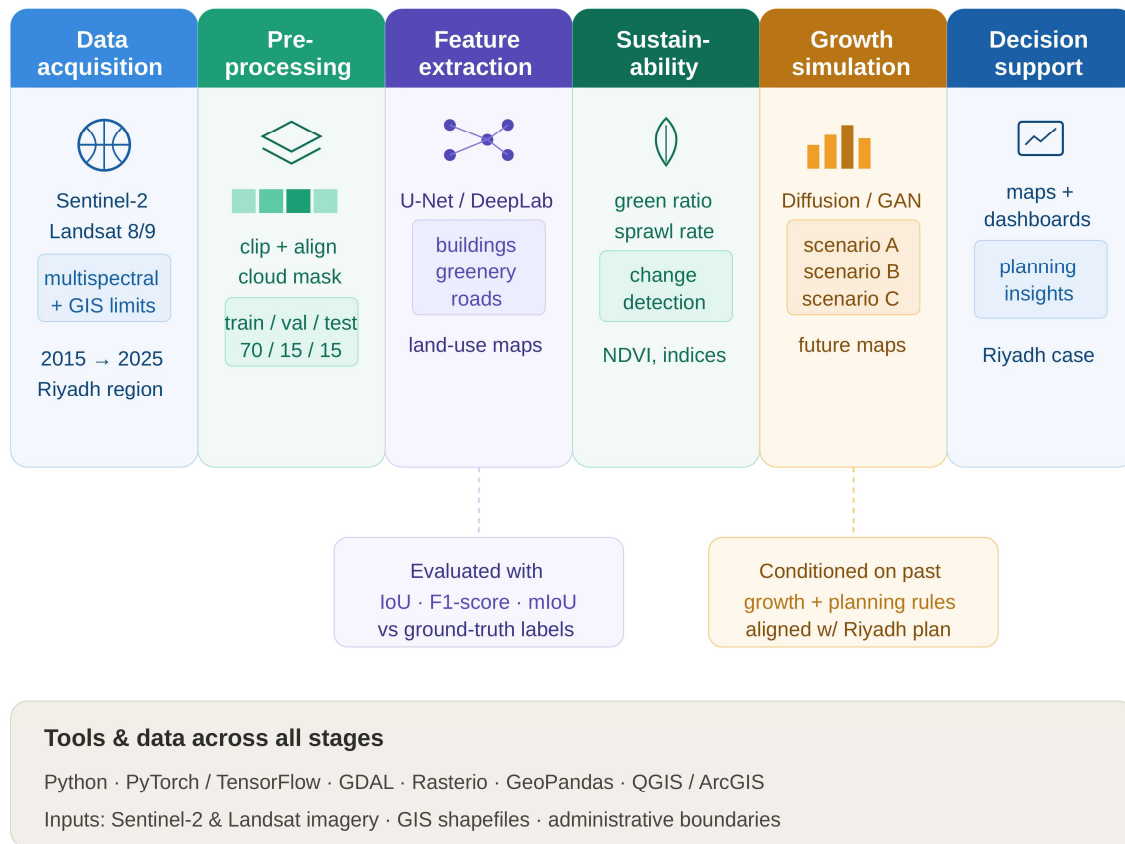


Figure 3.1. Block diagram of the proposed Geo-AI framework for long-term urban growth analysis in Riyadh.

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.

21
67
24
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.

32
2
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

7
12
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.

58

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 Concerns, and NDVI Correction:

63

Reference land-cover labels were extracted from the Esri/Impact Observatory annual 10 m land-cover dataset and reclassified from its source 9 classes into 4 target classes: Vegetation/Rangeland, Built-up, Bare Ground/Water, and Snow/Cloud. Bare Ground and Water were merged into a single class, as in this arid, non-coastal study area, bodies of water account for a negligible proportion of the area and are often transient (seasonal) or man-made; this fusion avoids a potentially volatile and sparsely populated class, while still addressing the most relevant classification frontier for the urban-growth monitoring context, i.e., the transition between built-up and vegetation/bare-ground conditions. Snow/Cloud was retained from the source Esri scheme, for the sake of consistency with the reclassification schema, despite their effective absence in the Riyadh climate, an absence evidenced empirically (and illustrated further on in Table 4.1). Such pixels are not subjected to substantive analysis anywhere in this thesis as an insignificant segment of the study area cannot contribute meaningful performance estimates, irrespective of model capability.

It is important, not as an aside but as a clear methodological statement, that all reference labels employed in this thesis, even subsequent to the correction below, originate from a third-party global land-cover product whose stated accuracy is 75-85%. Therefore, the pixel accuracy figures stated in Chapter 4 do not represent a true ground truth. This is simply the level of label noise inherent in training from a global reference dataset.

It motivated the independent field validation detailed in Section 3.9 where some predictions are independently checked against base-map imagery, rather than against the original reference data.

This distinction of source data and its influence on interpreting headline accuracy figures presented in Chapter 4 is a recurring theme throughout this thesis, especially in results and discussion and in the Conclusion.

An initial examination of the reclassified reference labels suggested an unlikely extent of bare desert terrain had been labeled as vegetation; 43-47% vegetation according to the source 'Rangeland' class in this arid, sparsely vegetated context contradicted this thesis's expectations, as well as common knowledge regarding Riyadh's actual land cover. To correct for this error, an NDVI-thresholding correction was employed; specifically, Otsu's method (Otsu, 1979) was used to determine a threshold that minimized intra-class variance among a group of pixels classified as Vegetation. This threshold was derived independently for each time period, and proved significantly more stable in the median-composite imagery (s.d. = 0.0142 across years) than the mosaicked imagery (range: 0.053-0.371 across years), providing an additional data point that the compositing approach was responsible for the earlier anomalies, as intrinsic land cover features would not exhibit such a high degree of variability between these datasets.

A final averaged threshold was derived which was roughly 0.24, beneath which the vegetation class was reclassified as bare ground. This drastically reduced the vegetation class to a more plausible 3.3-3.5% of the study area, and it was remarkably consistent across all eight training images. This thresholding correction step is addressed in this work not as a mere data cleaning technique, but as a demonstration of the significant potential for class-specific biases inherent within globally produced Land-Cover maps. Such biases need to be actively analyzed and corrected, particularly in environments such as arid arid/urban interface which may not have been the primary training focus of globally produced products.

3.6 Patch Extraction and Dataset Splitting

For each year's 8-band composite and corrected reference, overlapping 128x128-pixel patches were extracted using a fixed random seed (seed = 42) for all shuffling operations. Patches containing more than 10% of pixels labeled "invalid" due to edge clipping and the boundary of the study-area geometry were excluded, yielding 2,928 valid patches per year across eight years for a total of 23,424 patches. The patch dimensions of 128x128 (corresponding to 1.28 km on the ground at a 10 m ground sampling distance) were chosen to strike a balance that was large enough to allow the U-Net's encoder to learn from context using neighboring pixels (e.g., organized road networks) and small enough for efficient batch training.

The data were temporally, not randomly, split: years 2018 through 2022 were allocated to the training set (14,640 patches), year 2023 comprised the validation set (2,928 patches), and years 2024 through 2025 comprised the test set (5,856 patches).

This was a methodological choice; a randomly split dataset comprising patches from all eight years shuffled randomly into three subsets would evaluate the model's prediction capability on imagery from recent past years it's familiar with (thus, it's "seen" in the train/val sets and should not significantly boost its performance), versus evaluation on future unknown data-the primary scenario in this application. Exact class proportions per split are provided below (derived from saved patches): train: 2.83% Veg., 24.82% Built, 72.35% Bare; val: 2.86% Veg., 28.20% Built, 68.95% Bare; test: 2.75% Veg., 28.77% Built, 68.48% Bare. (0% snow/cloud for all, as expected.) The incremental increase from train to test in the proportion of Builtup areas is the first quantitative sign of the trend the rest of this thesis attempts to describe, before we ever learn to use models.

3.7 Model Architecture

Architecture: Model is based on a U-Net with 3 encoder/decoder stages (32/64/128 filters), with a bottleneck size of 256 filters. In each convolutional block, a pair of 3x3 convolutions are first applied with ReLU activation followed by batch normalization, and then either a max pooling operation(encoder) or transposed convolution (decoder) is used to decrease/increase spatial resolution (64x64 pixels and higher). The network, consisting of nearly 1.95 million trainable parameters (as purposefully built as compactly as possible given hardware constraints) reflects previous observation that networks which tend to be much larger and hence achieve greater performance (section 2.2), require significantly more computation and larger datasets of labeled data, which we were able to accommodate for this thesis . In the encoder pathway, feature resolution is progressively decimated(128x128-x64x64x32x32x16x16at the bottleneck)and the number of filter channels is increased by a factor of 2 after each of the three encoding stages.

10

2

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

This simulation attempts to answer the original aim of this thesis: to project the future growth of the city (section 1.4, objective 6). This is an initial, exploration simulation based on trend. It was decided against building a complete generative model, e.g.

A diffusion model or GAN, which can, in principle, generate multiple plausible, probabilistically weighted future-growth scenarios rather than a single deterministic projection, due to computational limits and the project timeline.

A simple, clearly defined, trend-based simulation is used instead to provide a basic projection, and explicitly states what kind of model it is. The model extrapolates the observed average of historical annual growth rate coupled with a cellular-automata expansion rule, where only bare-ground pixels immediately adjacent to existing built-up land-defined through binary dilation, a typical morphology operation-are converted to built-up land on the observed rate for 2030 forward. The only pixels considered for conversion at each step of simulated time are those adjacent to existing land-only in the effort to capture the observed purely ring-like, fringe-dominated expansion of development, shown in chapter 4 (section 4.4).

In this approach, we make the stark assumption that, without policy, past historical rate of expansion and manner of radial development will continue unchanged through 2030. It does not yet take into account development constraints, service provision limits or policy measures that may, for instance be imposed as part of the Riyadh Region Plan and the outputs to the forecast are not qualified by a confidence interval due to the single scenario and deterministic format of the currently applied simulation.

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.

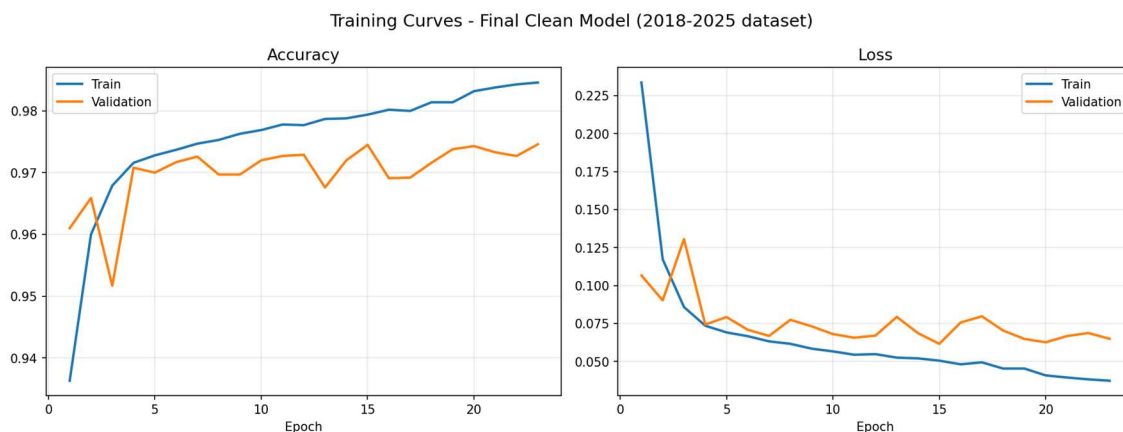


Figure 4.2. Training and validation accuracy/loss curves (23 epochs, EarlyStopping).

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.

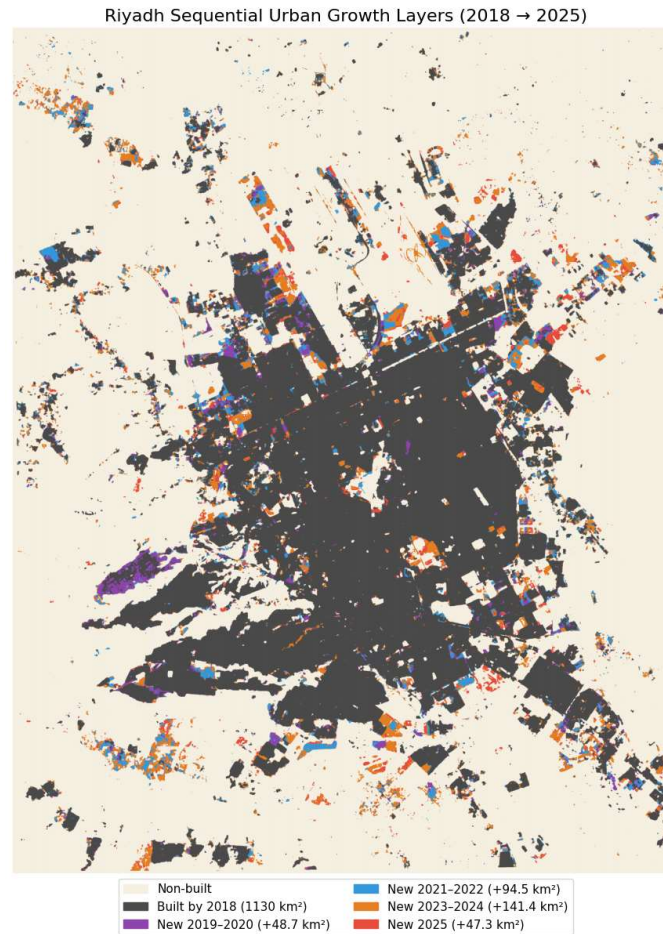


Figure 4.3. Sequential urban growth layers, 2018–2025, by period of first appearance as built-up.

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

The model, tested on recently acquired 2025 imagery in Wadi Hanifah (detailed in Section 3.11) had one drawback that the fixed test-set performance figure (Section 4.1) did not indicate; on one test patch 23.5% of the pixels were vegetated (according to the NDVI threshold detailed in Section 3.5), but the model assigned merely 3.7% to vegetation, and the remaining 69.3% to Built-up terrain. This is a far more pronounced occurrence of the built-up/bare-ground misclassification recorded in Section 4.1, but specific to this terrain, and driven by the fact that in the NDBI/BSI feature space (Section 2.3), exposed rock and built-up materials have virtually identical spectral characteristics which had not been learned sufficiently with the flat, built-up urban and desert datasets used for training.

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.

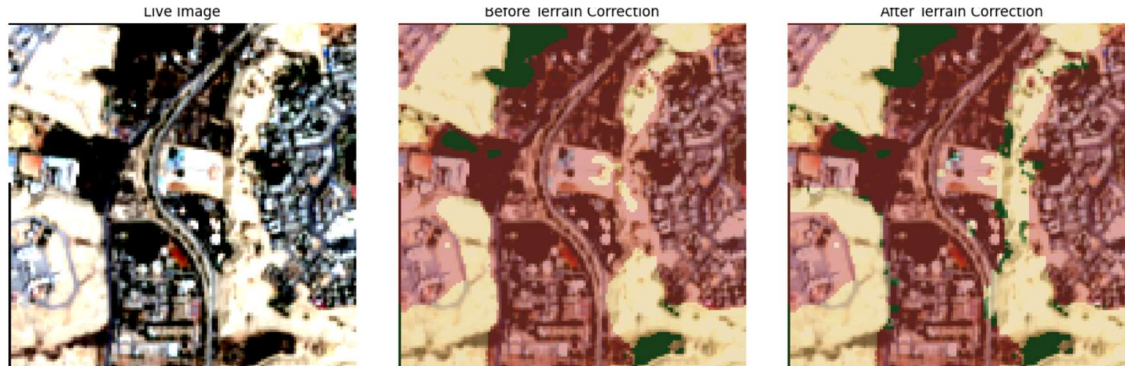


Figure 4.4. Terrain-aware correction on a Wadi Hanifah patch: live image (left), prediction before correction (center), and after correction (right).

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.

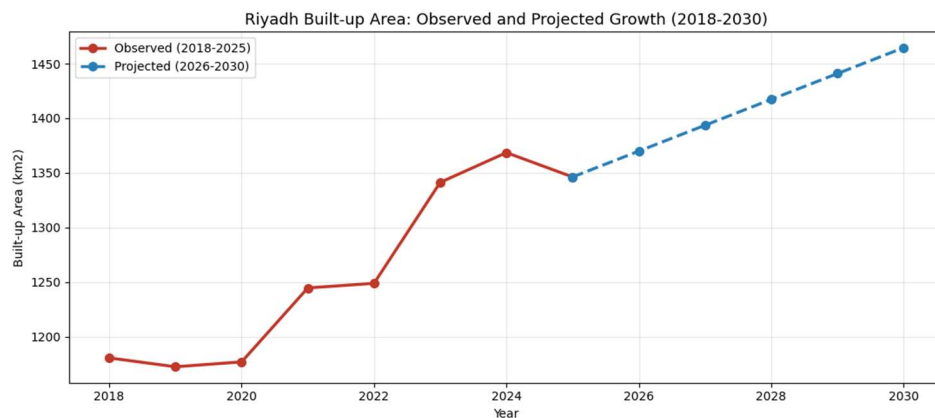


Figure 4.5. Observed built-up area (2018–2025) and projected growth (2026–2030).

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

61

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.