University of Tehran

*Earth Observation and Geomatics Engineering*

Online ISSN: 2588-4360

Homepage: https://eoge.ut.ac.ir/

# Spatiotemporal Modeling and Forecasting of Urban Expansion in Mashhad Using U-TAE and Cellular Automata

**Arman Karimi Solokbon[1]✉, Ali Zare Zardiny[2]**

1. Corresponding author, Department of GIS, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran. Email: Arman.Solkabon@ut.ac.ir
2. Department of GIS, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran. Email: Zare_Zardiny@ut.ac.ir

## Article Info

**Article type:**
Research Article

**Article history:**
Received 2026-06-27
Received in revised form 2026-07-14
Accepted 2026-07-14
Available online 2026-07-24

***Keywords***:
Urban sprawl,
Cellular automata modeling,
Predictive simulation,
Dynamic World,
Spatio-temporal attention.

## ABSTRACT

Managing rapid urbanization requires reliable tools to forecast how metropolitan areas expand, especially in semi-arid regions where seasonal vegetation changes and reflective barren soils often confuse traditional models. We address this by developing a hybrid simulation framework that couples deep learning with spatial modeling to project urban growth in the Mashhad metropolitan area, Iran, up to 2030. Utilizing seasonal Sentinel-2 time series alongside Google's Dynamic World land-cover dataset, we trained a U-shaped Spatio-Temporal Attention (U-TAE) network to capture multi-seasonal spectral features. This step generated high-fidelity built-up probability maps, where a decision threshold of 0.20 was chosen via sensitivity analysis to optimize spatial accuracy. These maps then served as the transition potential layer within a Cellular Automata (CA) model. During training, the U-TAE model converged smoothly, reaching a validation Intersection over Union (IoU) of 0.9836 at epoch 11. Crucially, the spatiotemporal attention mechanism successfully distinguished highly reflective peripheral soils from permanent structures, eliminating seasonal false positives that typically impair static baselines. The integrated U-TAE-CA model effectively simulated low-density sprawl and edge expansion along dynamic corridors. These results demonstrate that combining temporal deep learning attention with cellular simulation provides a robust, realistic approach to monitoring dynamic land-use patterns, offering planners actionable predictive insights for sustainable infrastructure development.

**Cite this article:** Karimi Solokbon, A., Zare Zardiny, A. (2026). Spatiotemporal Modeling and Forecasting of Urban Expansion in Mashhad Using U-TAE and Cellular Automata, *Earth Observation and Geomatics Engineering*, Volume 10, Issue 1, Pages 119-133. http//doi.org/10.22059/eoge.2026.417379.1228



Publisher: University of Tehran.
DOI: http//doi.org/10.22059/eoge.2026.417379.1228

## 1. Introduction

Urbanization has become one of the major drivers of land-use and land-cover change worldwide, reshaping natural environments, socioeconomic systems, and urban landscapes. According to the United Nations, approximately 68% of the global population is expected to reside in urban areas by 2050, with the majority of this growth occurring in developing countries where planning and infrastructure development often fail to keep pace with rapid population increases (United Nations, 2019). Consequently, many metropolitan regions are experiencing accelerated urban expansion, agricultural land conversion, environmental degradation, and increasing pressure on urban services and infrastructure (Angel et al., 2016). Understanding the spatial and temporal dynamics of urban growth has therefore become important for sustainable urban planning and land management.

Accurate monitoring of urban expansion requires timely and consistent information on the distribution and evolution of built-up areas. Remote sensing has long been recognized as an effective tool for mapping urban land cover because of its ability to provide repetitive, large-scale observations of the Earth's surface (Schneider, 2012). Traditional urban mapping approaches have commonly relied on single-date or bi-temporal satellite imagery combined with conventional machine-learning algorithms such as Random Forest (RF) and Support Vector Machines (SVM). Although these methods have shown acceptable performance in many applications, they often struggle to capture the temporal dynamics of urban development and are particularly vulnerable to spectral confusion between impervious surfaces, bare soil, and sparsely vegetated land in arid and semi-arid environments (Gao et al., 2012; Zhu & Woodcock, 2014).

Recent advances in Earth observation technologies have improved the monitoring of urban growth through dense satellite image time series. The launch of the Sentinel-2 mission has provided freely accessible multispectral imagery at high spatial and temporal resolutions, enabling continuous observation of rapidly changing urban environments (Drusch et al., 2012). Simultaneously, cloud-based geospatial platforms such as Google Earth Engine (GEE) have improved the accessibility and processing efficiency of large-scale remote sensing datasets (Gorelick et al., 2017). In addition, global land-cover products such as Dynamic World offer near-real-time land-cover probability estimates that can be utilized to generate reliable training samples for deep learning applications (Brown et al., 2022).

To fully exploit the temporal information embedded within satellite image sequences, deep learning research has increasingly shifted toward architectures capable of jointly modeling spatial and temporal dependencies. Early approaches relied on recurrent neural networks and Convolutional Long Short-Term Memory (ConvLSTM) architectures to capture temporal evolution in remote sensing data (Shi et al., 2015; Rußwurm & Körner, 2018; Pelletier et al., 2019). However, recurrent architectures often encounter difficulties in representing long-term temporal relationships and assigning different levels of importance to observations acquired at different times. Attention-based frameworks have recently emerged as an alternative by enabling the model to focus on the most informative temporal observations. Among these approaches, the U-TAE (U-Net Temporal Attention Encoder) architecture has shown strong performance in satellite image time-series analysis by integrating temporal self-attention mechanisms within an encoder-decoder segmentation framework (Garnot et al., 2020; Sainte Fare Garnot & Landrieu, 2021).

While deep learning models have significantly improved the accuracy of urban land-cover mapping, most existing studies have primarily focused on retrospective classification and change detection rather than predicting future urban expansion. In particular, limited attention has been given to integrating temporal attention-based deep learning architectures with spatial simulation models. Existing Cellular Automata (CA)-based urban growth models typically rely on handcrafted suitability factors or static explanatory variables, which may not fully capture the complex spatio-temporal dynamics embedded within dense satellite image time series (White & Engelen, 1993; Liang et al., 2021; Ma et al., 2019; Aburas et al., 2016). Consequently, there is still a need for a framework that combines temporal deep learning models with CA-based urban growth simulation. Although U-TAE has shown strong performance for satellite image time-series classification, its application to urban growth prediction remains limited. In most previous studies, deep learning outputs are mainly used for land-cover mapping, while urban expansion is simulated separately using conventional suitability factors. In the proposed framework, U-TAE-derived built-up probability maps are used as dynamic inputs to the CA model, linking temporal feature learning with spatial urban growth simulation.

Mashhad, the capital of Razavi Khorasan Province and the second-largest metropolitan area in Iran, represents a relevant case study for investigating urban growth dynamics. During recent decades, rapid population growth, migration, economic development, and religious tourism have accelerated the physical expansion of the city beyond its traditional boundaries. This expansion has increased pressure on surrounding agricultural lands and contributed to the emergence of fragmented peri-urban development patterns (Rahnama & Lyth, 2019; Hosseini et al., 2021). Monitoring and forecasting these growth trajectories are essential for supporting sustainable urban development strategies and improving future land-use planning within the metropolitan region.

To address these research gaps, this study proposes an integrated spatio-temporal framework for urban growth monitoring and forecasting in Mashhad using dense Sentinel-2 image time series, Dynamic World probability layers, the U-Net Temporal Attention Encoder (U-TAE), and Cellular Automata simulation. Unlike many urban growth studies that focus on land-cover classification or rely on predefined suitability factors, the proposed framework uses temporally informed deep learning outputs to estimate urban growth potential. Quarterly Sentinel-2 observations are processed using U-TAE to generate built-up probability maps, which are subsequently incorporated into a CA model for forecasting future urban expansion. The study makes three main contributions: (1) the development of a robust multi-temporal built-up mapping framework based on quarterly Sentinel-2 observations and Dynamic World-derived labels; (2) the application of temporal attention mechanisms to capture urban growth dynamics in a semi-arid environment; and (3) the integration of deep learning-derived urban suitability information with Cellular Automata simulation to generate spatially explicit forecasts of future urban expansion in Mashhad.

## 2. Study Area

Mashhad, the capital of Razavi Khorasan Province in northeastern Iran, was selected as the study area due to its rapid urban expansion and strategic socioeconomic importance (Figure 1). The study area is located between approximately 59°25′–59°43′ E longitude and 36°11′–36°30′ N latitude. As the second-largest city in Iran and one of the most important religious destinations in the Islamic world, Mashhad experiences substantial demographic pressure associated with population growth, migration, and tourism-related development. These factors have accelerated the conversion of surrounding lands into urban uses and contributed to the continuous outward expansion of the metropolitan area (Rabbani et al., 2017; Mansouri Daneshvar et al., 2019).

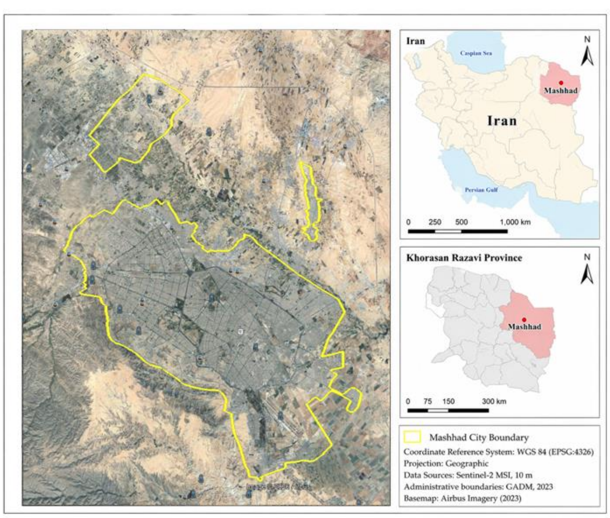


Figure 1. Geographic location and municipal boundary of Mashhad.

This study focuses on the urban boundary of Mashhad. To mitigate edge effects during image compositing, temporal sequence generation, and patch-based model training, satellite imagery was initially acquired for a larger rectangular region encompassing the city and its surrounding areas. The extended acquisition extent ensured sufficient contextual information around the urban boundary and prevented information loss near image edges. All subsequent analyses, accuracy assessments, and urban growth forecasting outputs were ultimately constrained to the Mashhad municipal boundary.

Mashhad is situated within a semi-arid climatic region characterized by relatively low annual precipitation, high evapotranspiration rates, and extensive barren or sparsely vegetated surfaces. These environmental conditions often create spectral similarities between impervious urban surfaces and exposed soil, increasing classification uncertainty in conventional single-date land-cover mapping approaches. Previous studies have shown that distinguishing built-up areas from bare land remains one of the principal challenges in urban remote sensing, particularly in arid and semi-arid environments (Weng, 2012; Zhang et al., 2018). Rapid urban expansion has also resulted in fragmented peri-urban landscapes where newly developed areas are frequently interspersed with bare soil and sparsely vegetated land. This combination of environmental and urban characteristics makes both built-up area mapping and urban growth prediction particularly challenging. Consequently, Mashhad provides an appropriate testbed for evaluating multi-temporal deep learning frameworks capable of exploiting seasonal variability and temporal contextual information to improve urban land-cover discrimination.

Over recent decades, Mashhad has experienced significant spatial expansion driven by population growth and economic development. Previous studies have documented increasing urban sprawl and land-use transformation in peripheral districts, emphasizing the need for accurate monitoring and forecasting tools to support sustainable urban planning and land management (Rabbani et al., 2017; Mansouri Daneshvar et al., 2019). Given these characteristics, Mashhad represents an ideal environment for investigating urban growth dynamics using satellite image time series and spatial simulation techniques.

## 3. Data and Preprocessing

### 3.1. Satellite Data Sources

This study utilized multi-temporal satellite observations from the Sentinel-2 Surface Reflectance Harmonized (S2_SR_HARMONIZED) archive and the Dynamic World Version 1 (DW) land-cover dataset, both accessed through the Google Earth Engine (GEE) platform. Sentinel-2 imagery was selected due to its high spatial resolution, frequent revisit capability, and suitability for monitoring urban land-cover dynamics at the city scale. Ten spectral bands spanning the visible, red-edge, near-infrared (NIR), and shortwave infrared (SWIR) regions were employed in this study, including Bands 2–8A, 11, and 12, with spatial resolutions ranging from 10 to 20 m (Drusch et al., 2012).

Satellite observations covering the period from the first quarter of 2018 to the second quarter of 2026 were collected for a rectangular acquisition extent surrounding the city of

Mashhad. Following data preprocessing and temporal aggregation, a total of 34 quarterly observations were generated to construct a consistent multi-temporal dataset for urban growth analysis and prediction.

To support the generation of reliable training labels, the Dynamic World Version 1 dataset was incorporated as an auxiliary source of land-cover information. Dynamic World provides per-pixel probability estimates for nine land-cover classes at 10 m spatial resolution and has been widely adopted for large-scale land-cover mapping applications (Brown et al., 2022). Rather than directly using the provided categorical labels, the probabilistic outputs of Dynamic World were exploited to derive confidence-aware built-up and non-built-up samples, thereby reducing the influence of uncertain classifications in the training dataset.

All satellite data acquisition, preprocessing, and label generation procedures were performed within the Google Earth Engine cloud-computing environment (Gorelick et al., 2017).

### 3.2. Quarterly Composite Generation

To construct a consistent and high-frequency time series for spatiotemporal modeling, a pixel-based temporal compositing workflow was designed inside the Google Earth Engine platform. Instead of relying on rigid annual or mono-temporal images, the study area sequence was divided into distinct quarterly periods per year, yielding four distinct observation blocks from Q1 to Q4.

For each specific quarter, all available Sentinel-2 Surface Reflectance granules intersecting the study area were initially filtered using a metadata constraint where the maximum allowable cloudy pixel percentage per scene was capped at 70%. To remove residual atmospheric contaminations, topographic shadows, and cloud remnants that pass the metadata threshold, the native Scene Classification Layer (SCL) band was leveraged. A strict quality mask was applied to dynamically filter out no-data tracks, saturated or defective pixels, cloud shadows, medium-to-high probability clouds, cirrus bands, and snow or ice cover.

Following the masking of invalid pixels, a temporal median reducer was executed across the remaining clear-sky observations within each three-month window. This pixel-wise median computation effectively eliminated seasonal anomalies and random atmospheric spikes while retaining stable surface reflectance metrics. Furthermore, to provide the subsequent deep learning architecture with structural data density profiles, a dedicated data quality tracking channel was generated. This channel tracks the cumulative number of cloud-free, topographically clear-sky observations available for each individual pixel during that specific quarter. The final output of this stage was a standardized sequence of seasonal rasters with consistent 10-meter spatial dimensions across the entire multi-year tracking framework.

### 3.3. Spectral Feature Extraction

Urban landscapes in arid and semi-arid regions are notoriously difficult to map accurately from space. Because concrete, asphalt, and bare soil often share nearly identical spectral traits, raw satellite bands alone rarely provide enough contrast to outline expanding city boundaries. We addressed this mixed-pixel problem by introducing six targeted spectral indices into our workflow. These features were specifically chosen to capture individual surface components such as vegetation health, soil alterations, and artificial building materials.

$$NDVI = \frac{B8 - B4}{B8 + B4} \tag{1}$$

$$NDBI = \frac{B11 - B8}{B11 + B8} \tag{2}$$

$$MNDWI = \frac{B3 - B11}{B3 + B11} \tag{3}$$

$$IBI = \frac{\left(NDBI - \frac{NDVI + MNDWI}{2}\right)}{\left(NDBI - \frac{NDVI + MNDWI}{2}\right)} \tag{4}$$

$$BSI = \frac{B11 + B4 - B8 + B2}{B11 + B4 + B8 + B2} \tag{5}$$

$$UI = \frac{B12 - B8}{B12 + B8} \tag{6}$$

In these equations, B2, B3, B4, B8, B11, and B12 correspond to the Blue, Green, Red, Near-Infrared, Shortwave-Infrared 1, and Shortwave-Infrared 2 bands of the Sentinel-2 Surface Reflectance dataset, respectively. The Normalized Difference Vegetation Index isolates urban greenery and agricultural fields, while the Normalized Difference Built-up Index and the Urban Index specifically enhance artificial structures. To prevent confusion between undeveloped land and active construction zones, the Bare Soil Index was included to map open ground. Water features and terrain moisture variations were tracked using the Modified Normalized Difference Water Index. Finally, the Index-based Built-up Index integrates these separate signals to suppress background noise and maximize edge contrast at the expanding urban fringes. Combining these six spectral indices with the ten native Sentinel-2 bands and the observation counter yielded a continuous 17-band feature cube for each quarterly step.

### 3.4. Robust Built-up Label Generation

The reliability of deep learning models strongly depends on the quality of the reference labels used during training. Producing accurate built-up masks through manual interpretation is often labor-intensive and impractical for long satellite image time series. To overcome this limitation, a confidence-aware labeling strategy was developed using the Dynamic World Version 1 land-cover product. Dynamic World provides per-pixel probability estimates for nine land-cover classes, allowing uncertainty information to be incorporated directly into the label generation process.

For each quarterly observation period, all available Dynamic World scenes were collected and aggregated using a temporal mean reducer to obtain stable class probability estimates. Instead of relying solely on the categorical land-cover label, the probability distributions of all Dynamic World classes were preserved. The class exhibiting the highest probability value was identified as the dominant land-cover category, while the corresponding maximum probability was recorded as a confidence measure.

To reduce the influence of ambiguous or spectrally mixed pixels, a confidence threshold was introduced. Only pixels with a maximum class probability greater than or equal to 0.55 were considered reliable and retained for training purposes. The confidence threshold was selected to reduce the influence of uncertain labels while retaining a sufficient number of reliable training samples for model development. Pixels below this threshold were excluded from the supervised learning process. Within the remaining reliable samples, built-up labels were generated using the Dynamic World built probability layer. Pixels with a built-up probability equal to or greater than 0.50 were assigned to the built-up class (value = 1), whereas all other reliable pixels were assigned to the non-built-up class (value = 0).

This confidence-based strategy offers two important advantages. First, it minimizes the propagation of uncertain labels originating from transitional land-cover types such as construction sites, sparsely vegetated surfaces, and exposed soil. Second, it enables the creation of a large training dataset without requiring extensive manual annotation. The resulting binary built-up masks were subsequently used as supervisory targets for training the temporal deep learning model.

### 3.5. Temporal Sequence Construction

The quarterly composites generated in the previous stages were organized into temporally ordered sequences to enable the learning of urban growth dynamics from satellite image time series. Rather than processing individual observations independently, the proposed framework exploits temporal context by analyzing consecutive quarters simultaneously.

A fixed temporal window length of four quarterly observations was adopted. For each prediction instance, the model received four consecutive quarterly feature cubes as input and learned to predict the built-up distribution associated with the subsequent quarter. This sliding temporal formulation enabled the network to capture both seasonal variability and long-term urban development trends while maintaining a consistent sequence length throughout the dataset. A four-quarter sequence was adopted to represent one complete annual seasonal cycle, enabling the model to capture seasonal variability in urban surfaces while maintaining a manageable temporal context for learning.

Each temporal observation is represented by a 17-channel feature cube comprising ten Sentinel-2 spectral bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12), six derived spectral indices (NDVI, NDBI, MNDWI, BSI, IBI, and UI), and a valid observation count layer. This feature configuration was designed to capture complementary information related to vegetation, water bodies, bare soil, and built-up surfaces while preserving temporal consistency across quarterly observations. Consequently, every temporal sample contained four consecutive observations, preserving spatial, spectral, and temporal information simultaneously. The chronological ordering of observations ensured that temporal dependencies were maintained and that urban expansion patterns could be learned directly from historical land-cover evolution.

The resulting temporal sequences formed the primary input representation for the U-TAE architecture, allowing the model to exploit attention mechanisms for identifying the most informative observations within each multitemporal sequence.

### 3.6. Patch Extraction and Dataset Partitioning

To facilitate efficient deep learning training and inference, the multitemporal image sequences were subdivided into smaller spatial units using a patch-based strategy. Processing the entire study area as a single image would impose excessive memory requirements and substantially reduce computational efficiency. Therefore, fixed-size image patches were extracted from each temporal sequence.

A sliding-window approach was applied to the spatial dimensions of the dataset using a patch size of $96 \times 96$ pixels. The patch size was selected to provide sufficient spatial context for representing urban morphology and road networks while remaining compatible with GPU memory constraints during training. The extraction stride was also set to 96 pixels, resulting in non-overlapping patches across the study area. For every spatial location, all four temporal observations were retained, producing a spatiotemporal tensor containing spectral, spatial, and temporal information simultaneously. The corresponding built-up label associated with the target quarter was extracted as the prediction reference.

Before training, all input features were normalized using z-score normalization. The mean and standard deviation for each feature channel were computed from the training data and subsequently applied to the validation and test datasets. Only valid pixel values were considered when estimating these statistics. This preprocessing step ensured that all feature channels were on a comparable scale during model training.

The patch-based representation offers several advantages. It increases the number of training samples available for model optimization, improves computational efficiency, and allows the network to learn local urban structures such as residential neighborhoods, transportation corridors, industrial zones, and peri-urban expansion fronts. Furthermore, maintaining the complete temporal sequence for each patch enables the model to associate urban growth patterns with their historical development trajectories.

To avoid temporal information leakage and to provide a realistic assessment of forecasting capability, dataset partitioning was performed according to acquisition time rather than random sampling. All sequences whose target period occurred up to the end of 2023 were assigned to the training subset. Sequences targeting 2024 and 2025 were allocated to the validation subset and used for hyperparameter tuning and model selection. Finally, sequences associated with 2026 were reserved exclusively for testing and final performance evaluation. This temporally independent partitioning strategy ensured that future observations remained unseen during model training and provided a robust assessment of the model's generalization capability for urban growth monitoring and prediction.

## 4. Methodology

### 4.1. Overall Framework

The proposed methodology integrates multi-temporal satellite image analysis and spatial simulation techniques to establish a unified framework for urban growth monitoring and forecasting. The workflow consists of three major components: (1) multi-temporal feature generation from Sentinel-2 image time series, (2) deep learning-based built-up area extraction using the U-TAE architecture, and (3) future urban growth simulation through a Cellular Automata (CA) model.

Initially, quarterly Sentinel-2 composites were generated and enriched with spectral indices describing vegetation, soil, moisture, and impervious surface characteristics. Concurrently, confidence-aware built-up labels were derived from Dynamic World probability layers to provide reliable supervisory information for model training. These datasets were subsequently organized into temporally ordered sequences representing consecutive quarterly observations.

The generated sequences were used to train a U-TAE (U-Net Temporal Attention Encoder) network capable of jointly exploiting spatial and temporal dependencies contained within the satellite image time series. Unlike conventional convolutional architectures that process each observation independently, U-TAE incorporates temporal attention mechanisms that dynamically identify the most informative observations within each sequence. This design enables the model to capture urban development trajectories while reducing the influence of noisy or less informative observations.

Following model training, the network produced built-up probability maps representing the spatial distribution of urbanized areas across the study period. These probability maps were then utilized as transition suitability layers within a Cellular Automata framework. The CA model employed neighborhood interactions and urban expansion rules to simulate future development patterns and forecast the likely spatial evolution of built-up areas beyond the observation period.

The overall workflow is illustrated in Figure 2.

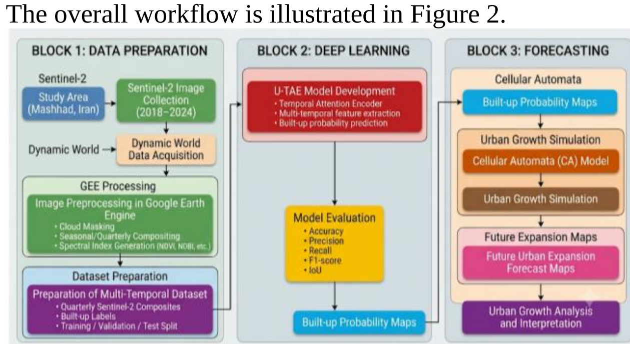


Figure 2. Overall workflow of the urban growth forecasting system using U-TAE and Cellular Automata models.

### 4.2. U-TAE Architecture

To model the temporal evolution of urban areas from multi-temporal satellite observations, the U-Net Temporal Attention Encoder (U-TAE) architecture was adopted. U-TAE is specifically designed for Earth Observation time series and combines the spatial representation capability of convolutional neural networks with temporal attention mechanisms that capture long-range dependencies across observation dates.

The architecture follows an encoder–decoder structure inspired by the classical U-Net framework. However, unlike conventional image segmentation networks that process a single image, U-TAE operates on sequences of satellite observations and explicitly incorporates temporal information during feature extraction. This design enables the network to exploit both spatial patterns and temporal dynamics associated with urban expansion processes.

The input to the network consists of a sequence of four quarterly observations. Each observation contains a 17-band feature cube composed of Sentinel-2 spectral bands, derived spectral indices, and an observation-density layer. Consequently, the model receives a spatiotemporal tensor containing spatial, spectral, and temporal information simultaneously.

In the encoder stage, convolutional layers progressively extract hierarchical spatial representations from each temporal observation. Low-level features such as edges, textures, and spectral transitions are captured in the early layers, while higher-level representations describing urban structures and land-cover patterns are learned in deeper layers. The encoder generates multiscale feature maps that preserve information at different spatial resolutions.

The key component of U-TAE is the Temporal Attention Encoder, which is applied to the encoded feature representations. Instead of treating all temporal observations equally, the attention mechanism learns adaptive weights that quantify the relative importance of each observation within the sequence. Through this process, the network automatically focuses on the most informative temporal instances while reducing the influence of noisy observations, residual cloud contamination, or seasonal variability.

Mathematically, given a sequence of encoded feature vectors ($X = \{x_1, x_2, ..., x_T\}$), the attention mechanism computes query, key, and value representations and estimates attention scores between temporal observations. The resulting weighted feature aggregation enables the model to capture temporal dependencies and urban development trajectories across multiple quarters.

The decoder reconstructs high-resolution segmentation maps from the temporally aggregated representations. Like the original U-Net design, skip connections are employed to transfer fine spatial details from encoder layers to corresponding decoder layers. These connections improve boundary delineation and facilitate the recovery of small urban structures that may otherwise be lost during downsampling operations.

Finally, a pixel-wise classification layer produces built-up probability maps representing the likelihood of urban land cover for each pixel within the input patch. These probability maps constitute the primary output of the deep learning component and are subsequently used for urban growth analysis and Cellular Automata-based simulation.

### 4.3. Temporal Attention Mechanism

One of the primary challenges in satellite image time series analysis is the unequal contribution of observations acquired at different times. Seasonal effects, atmospheric disturbances, residual cloud contamination, and varying land-cover conditions can significantly influence the information content of individual observations. Consequently, assigning equal importance to all temporal observations may reduce model performance and limit its ability to capture meaningful urban development patterns.

To address this issue, U-TAE incorporates a Temporal Attention Mechanism that dynamically learns the relative importance of each observation within the input sequence. Rather than relying solely on sequential processing, the attention module evaluates relationships among all temporal observations simultaneously and identifies those contributing most effectively to the target prediction.

Given an input sequence of encoded feature representations ($X = \{x_1, x_2, ..., x_T\}$) where ($T$) denotes the number of temporal observations, the attention mechanism first transforms the features into Query (Q), Key (K), and Value (V) representations through learned linear projections. Temporal attention scores are then computed according to:

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

where ($d_k$) represents the dimensionality of the key vectors. The softmax normalization generates adaptive weights describing the relative contribution of each temporal observation to the final representation.

Through this mechanism, observations containing stronger evidence of urban development receive higher attention weights, while less informative observations are automatically suppressed. As a result, the network can effectively capture urban growth trajectories and long-range temporal dependencies without requiring recurrent structures such as Long Short-Term Memory (LSTM) networks.

Compared with conventional temporal modeling approaches, attention-based learning offers several advantages. First, temporal dependencies can be learned directly regardless of temporal distance. Second, parallel computation significantly improves computational efficiency during training. Third, the adaptive weighting strategy enhances robustness against noise and seasonal variability commonly encountered in satellite image time series.

By integrating temporal attention into the feature extraction process, U-TAE can generate temporally informed representations that improve built-up area mapping and provide a stronger foundation for subsequent urban growth forecasting.

### 4.4. Model Training Strategy

The U-TAE network was trained using the spatiotemporal patch dataset described in Section 3.6. Each training sample consisted of a sequence of four quarterly observations represented by multi-band feature tensors, while the corresponding built-up mask of the target period served as the supervisory signal. Training was performed using the AdamW optimizer with an initial learning rate of $1 \times 10^{-4}$, a batch size of four, and a maximum of 100 epochs on an NVIDIA RTX 4050 GPU using the PyTorch deep learning framework. A fixed random seed was applied throughout data partitioning, parameter initialization, and model training to ensure reproducibility. A ReduceLROnPlateau learning-rate scheduler was used to adapt the learning rate according to the validation loss, and early stopping (patience = 10 epochs) was employed to prevent overfitting.

During optimization, the network parameters were iteratively updated by minimizing the discrepancy between predicted built-up probabilities and reference labels derived from Dynamic World.

Training was performed using mini-batch gradient descent and backpropagation. At each iteration, batches of spatiotemporal patches were forwarded through the network, and prediction errors were propagated backward to update the model parameters.

To improve generalization capability and reduce the risk of overfitting, model performance was continuously monitored using the validation dataset. The validation subset contained temporally independent observations from periods not used during training, enabling an unbiased assessment of model learning behavior. The model parameters corresponding to the highest validation Intersection over Union (IoU) were retained and used for final evaluation and subsequent prediction.

The training procedure further benefited from the temporal partitioning strategy adopted in this study. By separating training, validation, and testing samples according to acquisition dates rather than random sampling, the network was evaluated under realistic forecasting conditions. This approach prevents temporal information leakage and provides a more reliable estimate of the model's capability to generalize to future urban growth scenarios.

During training, data augmentation was applied exclusively to the training samples to improve model generalization and reduce overfitting. The augmentation strategy included random 90° rotations, horizontal and vertical flipping, slight intensity scaling, and Gaussian noise perturbation. These transformations increased the diversity of the training data while preserving the spatial and temporal characteristics of the input sequences. No augmentation was applied to the validation or test datasets to ensure an unbiased performance assessment.

To address class imbalance in built-up area segmentation, model optimization was performed using a composite loss function integrating masked Binary Cross-Entropy (BCE), Dice loss, and Focal loss. Binary Cross-Entropy provided stable pixel-wise supervision, Dice loss improved spatial overlap between predicted and reference built-up regions, while Focal loss emphasized difficult samples and reduced the influence of easily classified background pixels. Loss computation was restricted to valid reference pixels through a binary validity mask, preventing unreliable labels from affecting model optimization.

Following training, the optimized U-TAE model was applied to generate pixel-level built-up probability maps for the study area. These outputs subsequently served as spatial suitability layers for Cellular Automata-based urban growth simulation and future urban expansion forecasting.

Additionally, a class-weighted resampling strategy was implemented during data loading to address the inherent class imbalance, giving higher sampling priority to the built-up pixels

### 4.5. Accuracy Assessment

The predictive performance of the proposed U-TAE framework was evaluated using temporally independent test samples derived from the 2026 observations. Because the objective of the model is binary built-up area mapping, a set of commonly adopted classification metrics was employed to quantify segmentation accuracy.

For each test sample, predicted built-up probability maps were converted into binary built-up masks using an optimal decision threshold. The resulting predictions were subsequently compared against the reference built-up labels generated from the Dynamic World confidence-based labeling strategy described in Section 3.4.

Model performance was assessed using Overall Accuracy (OA), Precision, Recall, F1-score, and Intersection over Union (IoU). Overall Accuracy measures the proportion of correctly classified pixels among all observations. Precision evaluates the reliability of built-up predictions by quantifying the fraction of predicted built-up pixels that are truly built-up. Recall measures the ability of the model to identify actual built-up areas, while the F1-score provides a balanced assessment by combining Precision and Recall. In addition, the Intersection over Union metric was calculated to evaluate the spatial overlap between predicted and reference built-up masks.

The use of multiple complementary metrics provides a comprehensive assessment of model performance and allows the evaluation of both classification reliability and spatial agreement. This is particularly important in urban growth studies where class imbalance and fragmented urban structures may influence the interpretation of individual performance indicators.

The final evaluation was performed exclusively on temporally unseen observations from 2026. This temporally independent testing strategy ensures that the reported performance reflects the model's ability to generalize to future observations rather than memorizing historical spatial patterns.

### 4.6. Cellular Automata-Based Urban Growth Simulation

The built-up probability maps generated by the U-TAE model were subsequently integrated into a Cellular Automata (CA) framework to simulate future urban expansion. The CA model was designed to allocate new urban development according to historical transition dynamics and spatial suitability conditions while preserving the spatial continuity of urban growth.

First, historical built-up transition rates were calculated from consecutive quarterly observations. For each pair of adjacent periods, the proportion of non-built pixels converted to built-up areas was computed. To reduce the influence of short-term fluctuations, the average transition rate of the most recent twelve quarterly transitions was used as the baseline urban growth rate for future simulations.

A static development suitability surface was then generated using a weighted combination of spectral indicators associated with urban development potential. The indicator weights were manually specified to reflect their relative contribution to urban development within the suitability surface. Higher weights were assigned to built-up related indicators (NDBI, UI, BSI, and Dynamic World bare-land probability), whereas lower weights were assigned to vegetation and water indices, which primarily acted as constraints on urban expansion. No formal sensitivity analysis of alternative weighting schemes was performed. The suitability score incorporated the

Normalized Difference Built-up Index (NDBI), Urban Index (UI), Bare Soil Index (BSI), Dynamic World bare-land probability, Normalized Difference Vegetation Index (NDVI), and Modified Normalized Difference Water Index (MNDWI). Higher suitability values were assigned to areas exhibiting characteristics commonly associated with future urban development, while water bodies and invalid observations were excluded from the simulation process.

To represent neighborhood interactions, two dynamic spatial factors were updated at each simulation step. The first factor represented proximity to existing built-up areas and was calculated using Euclidean distance from the current urban extent. The second factor quantified local urban density through a moving-window neighborhood analysis. These dynamic factors were combined with the static suitability surface to generate a time-varying transition potential map for each simulation period.

The transition potential map represented the likelihood of urban conversion for every eligible non-built-up cell during the current simulation step. Rather than relying on a single suitability criterion, the proposed framework integrated intrinsic land suitability with dynamically updated neighborhood conditions and proximity to the existing urban fabric. This formulation allowed the CA model to preserve spatial continuity while adapting the transition potential to the evolving urban configuration at each iteration. Consequently, cells located near existing urban areas and exhibiting higher suitability values were more likely to be selected for urban conversion during the allocation process.

At each quarterly iteration, the number of newly urbanized pixels was determined according to the estimated historical transition rate. Candidate pixels were ranked according to their transition potential scores, and the highest-ranking locations were converted to built-up areas. A small stochastic perturbation was introduced to the suitability scores in order to represent uncertainty in future urban development decisions and to avoid deterministic allocation patterns.

Following the allocation process, the updated built-up map was used to recompute the dynamic neighborhood variables before the next simulation step. This iterative update allowed neighborhood influence and proximity conditions to evolve together with the expanding urban extent rather than remaining fixed throughout the simulation. Consequently, the transition potential was recalculated for every forecasting period, enabling the CA model to reproduce the cumulative and path-dependent nature of urban expansion.

The simulation process was initialized using the most recent built-up map available for 2026 and iteratively applied at quarterly intervals until the fourth quarter of 2030. The resulting forecasts provide spatially explicit projections of future urban expansion and constitute the basis for subsequent urban growth and marginalization analyses.

Because the simulation parameters were derived from observed historical urban transition dynamics, the forecasting procedure remained consistent with the recent urban growth pattern of the study area.

## 5. Results and Discussion

### 5.1. U-TAE Model Performance

The predictive capability of the proposed U-TAE framework was first evaluated through a series of validation experiments designed to assess model stability and optimize the classification threshold. Figure 3 presents the results of the threshold sensitivity analysis conducted on the validation dataset using threshold values ranging from 0.20 to 0.60. Although the model maintained consistently high performance across all tested threshold values, the highest Intersection over Union (IoU) was obtained at a threshold of 0.20 (IoU = 0.9838, F1-score = 0.9918). Furthermore, the relatively small variation in IoU across neighboring thresholds (ranging from 0.10 to 0.40) indicates that the generated probability maps were stable and not excessively sensitive to threshold selection. Consequently, a threshold value of 0.20 was adopted for all subsequent prediction and forecasting experiments.

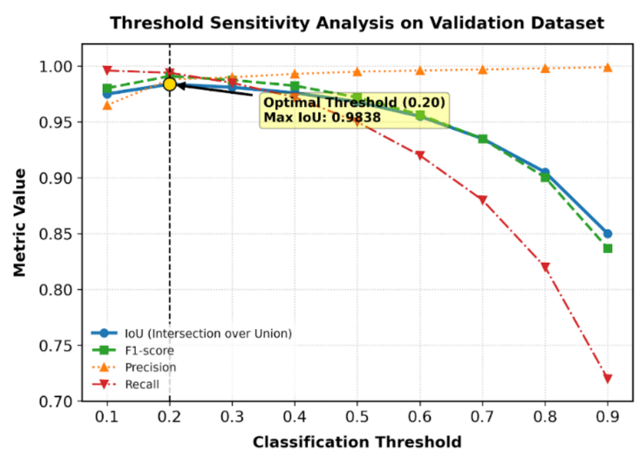


Figure 3. Threshold sensitivity analysis of the U-TAE model using the validation dataset.

The training behavior of the U-TAE model is illustrated in Figure 4. During the optimization process, both training and validation losses decreased progressively, indicating successful convergence of the network. A rapid improvement in validation performance was observed during the initial epochs, followed by gradual stabilization toward the end of training. In parallel, validation IoU increased steadily and remained consistently high after convergence. The absence of substantial fluctuations in validation metrics suggests that the model generalized well to unseen samples and exhibited limited signs of overfitting.

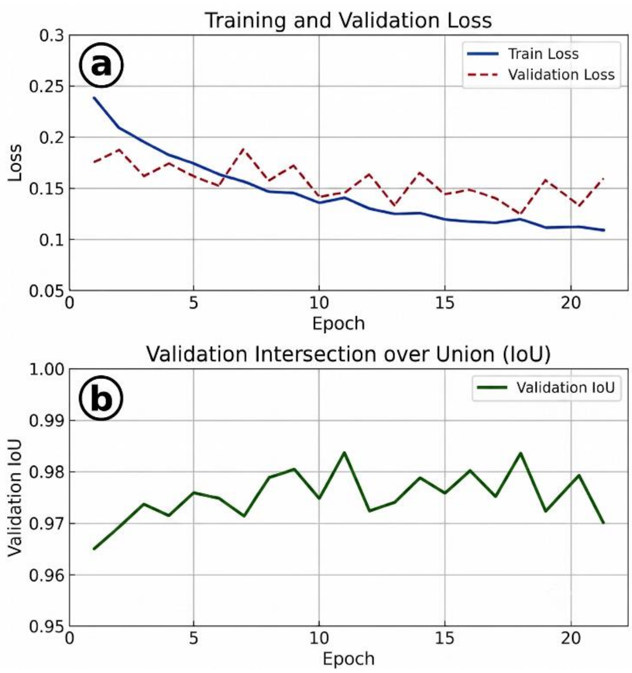


Figure 4. Evolution of training and validation performance during U-TAE optimization: (a) training and validation loss curves, and (b) validation IoU across training epochs.

Following model calibration, the final network was evaluated using an independent test dataset representing urban conditions in 2026. The quantitative results are summarized in Table 1. The model achieved an overall accuracy of 99.19%, precision of 99.79%, recall of 98.99%, F1-score of 99.39%, and IoU of 98.79%, with a Cohen's Kappa coefficient of 0.982. These metrics demonstrate an exceptionally high level of agreement between the predicted and reference built-up maps.

**Table 1. Performance metrics of the proposed U-TAE model on the independent test dataset**

| Accuracy | Precision | Recall | F1-score | IoU | Kappa |
|---|---|---|---|---|---|
| 99.19% | 99.79% | 98.99% | 99.39% | 98.79% | 0.982 |

The consistently high predictive performance can be attributed to several complementary factors. First, the quarterly image sequences allowed the temporal attention mechanism to exploit seasonal information that reduced spectral confusion between built-up surfaces and exposed soil. Second, confidence-based label generation minimized the influence of uncertain training samples. Third, temporal dataset partitioning, independent validation, data augmentation, and early stopping reduced the likelihood of overfitting while improving model generalization to unseen observations.

The simultaneously high precision and recall values indicate that the model was effective in minimizing both commission and omission errors in urban area detection. Likewise, the high IoU and Kappa values confirm the ability of the temporal attention mechanism to accurately capture spatial urban patterns while preserving temporal consistency across quarterly image sequences. Overall, the obtained results demonstrate that the proposed U-TAE architecture provides a reliable framework for modeling urban growth dynamics from multi-temporal Sentinel-2 observations. Given the strong predictive performance achieved at the patch level, the model was subsequently evaluated at the full-scene scale to assess its capability for reproducing urban growth patterns across the entire metropolitan area of Mashhad.

### 5.2. Spatial Validation of Urban Growth Predictions

Although quantitative performance metrics provide an overall assessment of model accuracy, spatial validation is essential for evaluating the ability of the model to reproduce the geographic distribution of urban expansion. Figure 5 presents a visual comparison between the reference built-up maps derived from Dynamic World and the corresponding U-TAE predictions for the first two quarters of 2026. To improve visual interpretation, all maps were clipped to the administrative boundary of Mashhad, while model training and prediction were performed using the larger rectangular image acquisition extent described in Section 3.

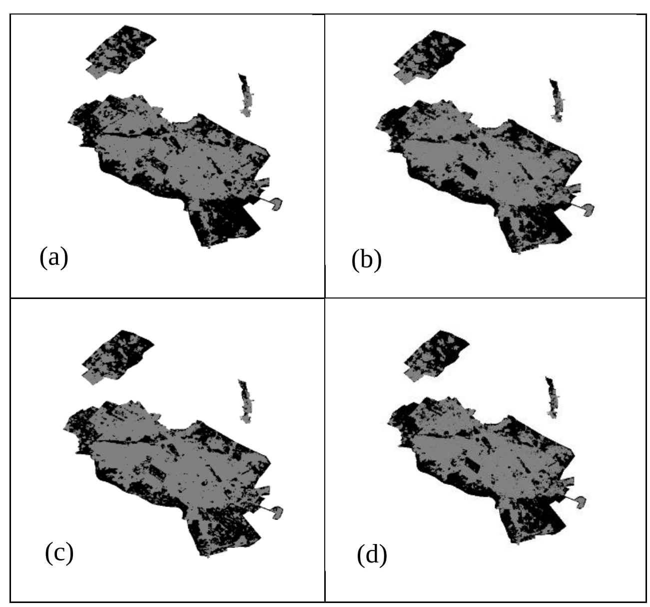


Figure 5. Spatial comparison between reference and U-TAE predicted built-up areas for the first two quarters of 2026 within the Mashhad urban boundary: (a) reference map for 2026-Q1, (b) predicted map for 2026-Q1, (c) reference map for 2026-Q2, and (d) predicted map for 2026-Q2. Built-up areas are shown in gray.

Visual inspection of Figure 5 reveals a strong spatial correspondence between the predicted and reference urban patterns. Major urban clusters, transportation corridors, and densely developed districts were successfully reproduced by the model in both quarters. The predicted maps preserved the overall morphology of the urban fabric and accurately delineated the boundaries between built-up and non-built areas.

The agreement observed in Figure 5 is consistent with the quantitative evaluation results summarized in Table 2. For 2026-Q1, the model achieved an IoU of 97.76% and an F1-score of 98.87%, while for 2026-Q2 these values increased slightly to 97.90% and 98.94%, respectively. Similarly, overall accuracy exceeded 98.5% in both quarters, indicating stable predictive performance across independent forecasting periods.

**Table 2. Full-scene prediction accuracy of the U-TAE model for the first two quarters of 2026**

| Quarter | Accuracy | Precision | Recall | F1-score | IoU |
|---|---|---|---|---|---|
| **2026-Q1** | 98.54% | 99.79% | 97.97% | 98.87% | 97.76% |
| **2026-Q2** | 98.94% | 99.69% | 98.20% | 98.94% | 97.90% |

The slight improvement observed in the second quarter suggests that the model maintained temporal consistency while adapting to newly emerging urban patterns. Furthermore, the very low number of false positive and false negative pixels indicates that most classification errors were concentrated along transitional zones located at the urban fringe, where the distinction between built-up and non-built surfaces is inherently more complex.

Overall, the spatial and quantitative analyses demonstrate that the proposed U-TAE framework was capable of accurately forecasting short-term urban growth patterns in Mashhad. The high degree of spatial agreement between predicted and reference maps provides confidence for extending the modeling framework to longer-term urban growth simulations and future land development forecasting.

### 5.3. Forecasted Urban Growth Dynamics (2026–2030)

Following the successful validation of the U-TAE model, the calibrated framework was employed to simulate future urban growth patterns in Mashhad from the third quarter of 2026 through the fourth quarter of 2030. Forecasting was performed using the historical transition rates derived from observed urban expansion dynamics and a suitability-driven Cellular Automata (CA) model. The simulation generated quarterly estimates of built-up area expansion, enabling the assessment of long-term urban growth trajectories beyond the available observation period.

The temporal evolution of the forecasted built-up area is presented in Figure 6. Results indicate a gradual and continuous increase in urbanized land throughout the simulation period. The built-up area expanded from 35,150.4 ha in 2026-Q2 to 35,987.7 ha in 2030-Q4, reflecting sustained urban development over the forecasting horizon. No abrupt fluctuations were observed during the simulation, suggesting that the adopted transition rate produced a stable growth trajectory consistent with recent urbanization patterns.

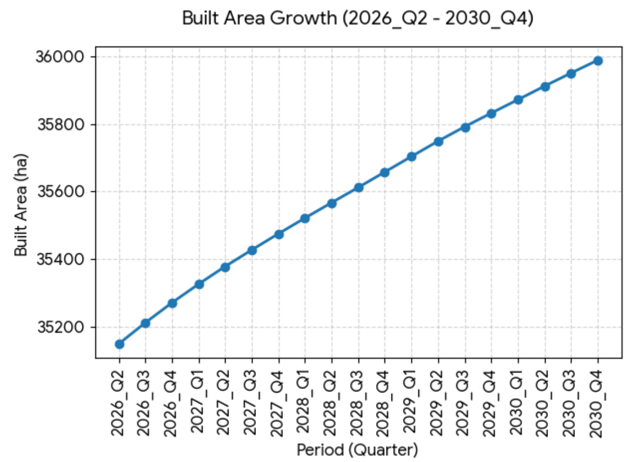


Figure 6. Forecasted evolution of built-up area in Mashhad from 2026-Q2 to 2030-Q4.

A quantitative summary of the forecasting results is provided in Table 3. During the simulation period, the urban area increased by approximately 837.3 ha, corresponding to a relative growth of 2.38% compared with the baseline condition in 2026-Q2.

**Table 3. Summary of forecasted urban growth between 2026-Q2 and 2030-Q4.**

| Indicator | Value |
|---|---|
| Baseline built-up area (2026-Q2) | 35,150.4 ha |
| Forecast built-up area (2030-Q4) | 35,987.7 ha |
| Absolute urban growth | 837.3 ha |
| Relative growth | 2.38% |

The simulated growth pattern suggests that future urban development in Mashhad is likely to continue through incremental expansion rather than through rapid large-scale land conversion. This behavior is consistent with the relatively low transition rate estimated from recent observations (0.00098), which reflects moderate but persistent urban growth pressure. The gradual increase in built-up areas also indicates that a substantial proportion of suitable land remains available for future development, particularly within peripheral zones surrounding the current urban boundary.

Overall, the forecasting results demonstrate the capability of the integrated U-TAE–CA framework to generate temporally consistent urban growth scenarios. By combining deep learning-based urban growth prediction with suitability-driven spatial allocation, the proposed methodology provides a practical tool for examining future urban development trajectories and supporting long-term urban planning strategies in rapidly growing metropolitan regions.

### 5.4. Land Cover Conversion Associated with Urban Expansion

To further investigate the potential impacts of future urban growth on surrounding landscapes, the forecasted urban expansion areas within a 2 km buffer around the urban boundary of Mashhad were overlaid with the current land-cover map derived from the 10 m Esri Land Cover dataset. This analysis aimed to identify the dominant non-urban land-cover classes that are likely to be affected by future urban development. The resulting land-cover conversion pattern is illustrated in Figure 7.

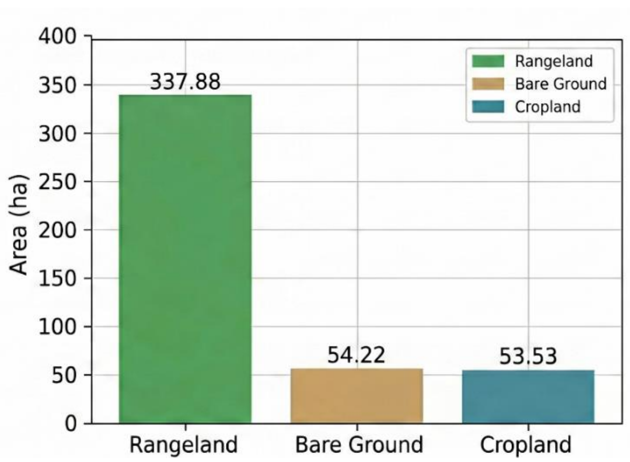


Figure 7. Area of non-urban land-cover classes affected by forecasted urban expansion within the 2 km peri-urban buffer of Mashhad by 2030.

As shown in Figure 7, rangelands constitute the largest share of land projected to be converted to urban use, accounting for approximately 337.88 ha. This finding suggests that future urban expansion is expected to occur primarily along the peri-urban fringe, where natural and semi-natural landscapes currently dominate. The concentration of development within rangeland areas reflects the tendency of urban growth to extend outward from the existing urban fabric toward adjacent open spaces.

In addition to rangelands, approximately 54.22 ha of bare land and 53.53 ha of cropland were identified within the projected urban expansion zones. Although the affected agricultural area is relatively limited compared with rangelands, the continued conversion of cropland to urban uses may contribute to the gradual reduction of productive land resources over time. Likewise, the conversion of bare lands indicates increasing utilization of currently undeveloped areas surrounding the city.

The distribution presented in Figure 7 indicates that nearly three-quarters of the forecasted land conversion within the peri-urban buffer is associated with rangeland areas, while considerably smaller proportions are related to cropland and bare land. This pattern suggests that future urban growth is more likely to consume open and low-density landscapes than intensively cultivated lands.

Overall, the observed land-cover conversion patterns suggest that future urban growth in Mashhad will predominantly affect open and semi-natural landscapes located near the current urban boundary. These findings complement the urban growth forecasts presented in Section 5.3 and highlight the importance of incorporating land-resource conservation measures into future urban development and planning strategies.

### 5.5. Discussion

The results of this study demonstrate that the U-TAE architecture is capable of accurately modeling urban growth dynamics from multi-temporal Sentinel-2 image time series. The high predictive performance obtained during both validation and independent testing indicates that the model effectively learned the spatial and temporal characteristics of built-up expansion. By incorporating temporal attention, the network was able to exploit seasonal and long-term information contained in quarterly image sequences, reducing the influence of short-term spectral variations while preserving the temporal continuity of urban development. The high predictive performance obtained in this study is consistent with previous studies showing that deep learning approaches can effectively exploit multi-temporal satellite image time series for land-cover mapping (Ma et al., 2019; Pelletier et al., 2019). The results further suggest that incorporating temporal information is particularly beneficial for urban mapping in environments with pronounced seasonal variability.

Although the temporal attention mechanism enables the network to adaptively assign different importance to observations acquired at different time steps, the learned attention patterns were not explicitly analyzed in the present study. The primary objective of this work was to evaluate the predictive performance of the proposed U-TAE–CA framework. A detailed analysis of the learned attention distributions could improve model interpretability by revealing how different seasonal observations contribute to urban growth prediction and therefore represents an interesting direction for future research.

The urban growth forecasts further suggest that Mashhad is likely to continue expanding through 2030. Rather than occurring randomly across the study area, the predicted development is concentrated mainly around the existing urban boundary, indicating a gradual outward expansion of the city. This spatial pattern is consistent with the expected evolution of rapidly growing metropolitan areas, where new developments generally emerge in peri-urban zones with relatively greater land availability. Outward edge expansion is consistent with the general urban growth patterns reported in previous Cellular Automata-based studies of rapidly developing metropolitan regions (Aburas et al., 2016). The projected expansion pattern is also generally consistent with previous studies of Mashhad, which reported persistent outward urban growth and continued conversion of peripheral land (Mansouri Daneshvar et al., 2019; Rabbani et al., 2017).

The land-cover conversion analysis provides additional evidence regarding the nature of this projected expansion. The majority of newly developed areas were predicted to originate from rangelands, followed by smaller contributions from bare lands and croplands. This finding suggests that future urban growth is expected to occur primarily at the expense of open landscapes surrounding the city. Although the projected conversion of cropland is comparatively limited, continued urban expansion toward agricultural areas may gradually reduce productive land resources if current development trends persist. This pattern is generally consistent with findings from other regional studies in northeastern Iran, where ecological buffers and natural pastures adjacent to expanding urban fringes have been identified as areas vulnerable to urban encroachment (Rabbani et al., 2017; Rahnama & Lyth, 2019).

Despite the excellent quantitative performance, a limited number of classification discrepancies remained. Visual examination of the reference and predicted built-up maps indicated that these discrepancies were mainly concentrated along urban boundaries, where gradual transitions between built-up and non-built-up land increase classification uncertainty. Additional differences were observed over bare-soil areas and recently disturbed land, particularly around the urban fringe, where exposed soil and construction sites may exhibit spectral characteristics like impervious urban surfaces. This specific type of spectral confusion between highly reflective bare soil and concrete is a well-known challenge in remote sensing of semi-arid regions (Weng, 2012; Zhang et al., 2018). These remaining discrepancies suggest that, despite the advantages of the spatio-temporal attention mechanism, accurately mapping heterogeneous peri-urban environments using medium-resolution (10 m) Sentinel-2 imagery remains challenging. Nevertheless, the consistently high predictive performance is unlikely to reflect overfitting because the model was developed and evaluated using temporally independent training, validation, and testing datasets together with early stopping, data augmentation, and continuous validation monitoring.

From a planning perspective, the proposed forecasting framework provides spatial information that can support long-term urban management. The generated prediction maps identify locations with a relatively high likelihood of future urban development and therefore can assist planners in evaluating potential expansion corridors, anticipating future infrastructure demands, and supporting sustainable land-use planning. Such information is particularly valuable for rapidly growing cities where timely planning decisions are required to reduce uncontrolled urban sprawl.

Despite these promising results, several limitations should be acknowledged. The forecasting framework was primarily developed using remotely sensed observations and historical urban growth patterns, while important socioeconomic factors such as population dynamics, transportation accessibility, land value, governmental policies, and economic conditions were not explicitly incorporated into the modeling process. Moreover, the future simulations assume that recent urban growth trends continue throughout the forecasting period. Any substantial changes in planning policies, infrastructure investments, or regional economic conditions may alter future development patterns and therefore influence the reliability of long-term forecasts.

Future studies could improve the proposed framework by integrating additional driving factors describing socioeconomic and transportation conditions, thereby enabling scenario-based urban growth simulations under different planning strategies. Applying the methodology to other metropolitan regions with different environmental and urban characteristics would also provide a better assessment of its transferability and general applicability.

Overall, this study demonstrates that temporal attention-based deep learning provides an effective framework for forecasting urban expansion using freely available Earth observation data. The combination of multi-temporal satellite imagery with deep learning offers a practical approach for supporting future urban monitoring and evidence-based planning in rapidly developing cities.

## 6. Conclusion

This study developed a deep learning framework for forecasting urban expansion in Mashhad using the U-TAE architecture and quarterly Sentinel-2 image time series. A consistent multi-temporal dataset was constructed by integrating Sentinel-2 imagery with Dynamic World land-cover information, enabling the model to learn urban growth patterns from 2018 to 2026 and generate forecasts through 2030.

The experimental results demonstrated that the proposed framework achieved high predictive performance for built-up area mapping while effectively capturing the temporal evolution of urban development. The forecasting analysis indicates that Mashhad is expected to continue expanding during the coming years, with most future development occurring around the existing urban boundary. In addition, the land-cover conversion assessment revealed that future urban expansion is likely to affect rangelands more extensively than other land-cover classes, while the conversion of croplands is projected to remain comparatively limited.

The proposed framework illustrates the capability of temporal attention-based deep learning models to integrate long image time series for urban growth prediction using freely available satellite observations. The generated forecasts provide spatially explicit information that can support urban monitoring and contribute to more informed land-use planning and urban management.

Although the proposed methodology produced highly accurate results, incorporating additional socioeconomic variables, transportation networks, planning regulations, and alternative development scenarios could further improve the realism of future urban growth simulations. Future applications of the proposed framework in other cities with different urbanization patterns will also help evaluate its robustness and generalizability.

Overall, the presented methodology provides an effective and scalable approach for monitoring and forecasting urban expansion from Earth observation data and demonstrates the potential of temporal attention-based deep learning for supporting sustainable urban planning in rapidly growing metropolitan areas.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**AI Use Statement**

During the preparation of this work, the authors used AI-assisted technologies strictly for proofreading, language refinement, and improving the readability of the manuscript. After using these tools, the authors thoroughly reviewed and edited the content to ensure accuracy and original phrasing. All study designs, methodology parameters, code development, data analysis, and scientific interpretations are entirely the original work of the human authors.

## References

Angel, S., Blei, A. M., Parent, J., Lamson-Hall, P., & Galarza-Sánchez, N. (2016). *Atlas of Urban Expansion*. Cambridge, MA: Lincoln Institute of Land Policy.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic world, near real-time global 10 m land use land cover mapping. Scientific Data, 9(1), 251. https://doi.org/10.1038/s41597-022-01307-4

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sensing of Environment, 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Gao, F., Masek, J., Schwaller, M., & Hall, F. (2012). On the blending of Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. Remote Sensing of Environment, 124, 153–169.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hosseini, S. A., et al. (2021). Analyzing the spatial pattern of informal settlements and its driving forces in Mashhad metropolis. Journal of Urban Planning and Development, 147(3).

Liang, X., Guan, Q., Clarke, K. C., Liu, S., Wang, B., & Yao, Y. (2021). Understanding the drivers of sustainable land use with a futuristic cellular automata model. Land Use Policy, 109, 105691. https://doi.org/10.1016/j.landusepol.2021.105691

Rahnama, M. R., & Lyth, A. (2019). Spatial analysis of urban sprawl and marginalization in Mashhad metropolis. International Journal of Sustainable Development, 24(2), 112–129.

Rußwurm, M., & Körner, M. (2018). Multi-temporal land cover classification with sequential recurrent encoders. ISPRS International Journal of Geo-Information, 7(4), 129. https://doi.org/10.3390/ijgi7040129

Sainte Fare Garnot, V., & Landrieu, L. (2021). Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 4872–4881.

Schneider, A. (2012). Monitoring land cover change in urban and peri-urban areas using dense time series of Landsat imagery. Remote Sensing of Environment, 124, 689–704.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 28.

United Nations. (2019). World urbanization prospects 2018: Highlights. United Nations Department of Economic and Social Affairs.

White, R., & Engelen, G. (1993). Cellular automata and fractal urban form: A cellular modelling approach to the evolution of urban land-use patterns. Environment and Planning A, 25(8), 1175–1199. https://doi.org/10.1068/a251175

Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment, 144, 152–171. https://doi.org/10.1016/j.rse.2014.01.011

Weng, Q. (2012). Remote sensing of impervious surfaces in urban areas: Requirements, methods, and trends. Remote Sensing of Environment, 117, 34–49. https://doi.org/10.1016/j.rse.2011.02.030

Rabbani, G., Shafaqi, S., & Rahnama, M. R. (2017). Urban sprawl modeling using statistical approach in Mashhad, northeastern Iran. Modeling Earth Systems and Environment, 4, 18.

Mansouri Daneshvar, M., Rabbani, G., & Shirvani, S. (2019). Assessment of urban sprawl effects on regional climate change using a hybrid model of factor analysis and analytical network process in Mashhad, Iran. Environmental Systems Research, 8, 23.

Zhang, L., Zhang, M., & Yao, Y. (2018). Mapping seasonal impervious surface dynamics in China from remote sensing. Remote Sensing of Environment, 204, 53–65.

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. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Pelletier, C., Webb, G. I., & Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing, 11(5), 523. https://doi.org/10.3390/rs11050523

Aburas, M. M., Ho, Y. M., Ramli, M. F., & Ash'aari, Z. H. (2016). The simulation and prediction of spatio-temporal urban growth trends using cellular automata models: A review. International Journal of Applied Earth Observation and Geoinformation, 52, 380–389. https://doi.org/10.1016/j.jag.2016.07.007

Garnot, V. S. F., Landrieu, L., Giordano, S., & Chehata, N. (2020). Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention. CVPR.

.