land

Article

Application of Explainable Artificial Intelligence (XAI) in Urban Growth Modeling: A Case Study of Seoul Metropolitan Area, Korea

Minjun Kim [1], Dongbeom Kim [2], Daeyong Jin [3] and Geunhan Kim [1,*]

[1] Department of Environmental Planning, Korea Environment Institute, Sejong 30147, Republic of Korea
[2] Technical Research Institute NEGGA Co., Ltd., Seoul 07220, Republic of Korea
[3] Center for Environment Data Strategy, Korea Environment Institute, Sejong 30147, Republic of Korea
* Correspondence: ghkim@kei.re.kr; Tel.: +82-044-415-7752

Abstract: Unplanned and rapid urban growth requires the reckless expansion of infrastructure including water, sewage, energy, and transportation facilities, and thus causes environmental problems such as deterioration of old towns, reduction of open spaces, and air pollution. To alleviate and prevent such problems induced by urban growth, the accurate prediction and management of urban expansion is crucial. In this context, this study aims at modeling and predicting urban expansion in Seoul metropolitan area (SMA), Korea, using GIS and XAI techniques. To this end, we examined the effects of land-cover, socio-economic, and environmental features in 2007 and 2019, within the optimal radius from a certain raster cell. Then, this study combined the extreme gradient boosting (XGBoost) model and Shapley additive explanations (SHAP) in analyzing urban expansion. The findings of this study suggest urban growth is dominantly affected by land-cover characteristics, followed by topographic attributes. In addition, the existence of water body and high ECVAM grades tend to significantly reduce the possibility of urban expansion. The findings of this study are expected to provide several policy implications in urban and environmental planning fields, particularly for effective and sustainable management of lands.

Keywords: urban growth model; explainable artificial intelligence (XAI); extreme gradient boosting (XGBoost); Shapley additive explanations (SHAP)

check for updates

Citation: Kim, M.; Kim, D.; Jin, D.; Kim, G. Application of Explainable Artificial Intelligence (XAI) in Urban Growth Modeling: A Case Study of Seoul Metropolitan Area, Korea. *Land* **2023**, *12*, 420. https://doi.org/10.3390/land12020420

Academic Editors: Luka Rumora, Mario Miler and Damir Medak

Received: 31 December 2022
Revised: 1 February 2023
Accepted: 2 February 2023
Published: 6 February 2023

1. Introduction

Unplanned and rapid urban growth requires the reckless expansion of infrastructure including water, sewage, energy and transportation facilities [1,2], and thus causes environmental problems such as deterioration of old towns, reduction of open spaces, and air pollution [3]. In addition, the transition from forests and agricultural areas to urbanized areas can severely reduce the habitats and biodiversity of wild animals and plants [4]. To alleviate and prevent such problems induced by urban growth, the accurate prediction and management of urban expansion is crucial [5].

In this context, many researchers have conducted urban expansion modeling and prediction studies over several decades. In the early stages, cellular automata (CA) were the most representative methods used to predict urban expansion. The CA model focuses on the simulation of spatial patterns of urban expansion rather than spatiotemporal interpretation. However, it has been pointed out that this method cannot take into account socioeconomic and demographic features in predicting urban growth [6,7]. To overcome this limitation, researchers have utilized not only statistical methods including multiple [8,9] and logistic regression [10,11], but also several machine learning (ML)-based techniques such as decision trees [12,13], random forest [14], support vector machine [15,16], and neural network [17,18]. They predicted future urban growth of a given region or nation by learning urbanization

patterns from the past to the present and examined the effects of physical and socio-demographic characteristics on such urbanization [19].

This study proposes several research gaps from existing studies that dealt with the modeling and prediction of urban growth, including both theoretical and methodological aspects. First, the majority of existing studies have adopted the distances from each object (e.g., land cover) as major influencing factors in urban growth modeling [20,21]. This approach is intuitive, but is not suitable to reflect the areal and morphological characteristics of various features in explaining urban expansion [22]. In addition, since the influence of each object on urban expansion is not linear [23], it is necessary to define the optimal range of spatial extent that can maximize the accuracy of urban expansion modeling.

Second, there have been limitations in accuracy and explanatory power in machine learning methodologies used by previous works to predict urban expansion. White-box approaches, including regression and decision tree models, exhibited easier to understand outcomes of urban growth models, but sometimes the accuracy was not high enough to utilize them in predicting urban expansion [24]. Black-box model approaches (including Support Vector Machine (SVM) and Deep Neural Network (DNN)), on the other hand, tend to have higher prediction accuracy—but it is very difficult to interpret how such outcomes are derived [25,26]. To fully adopt AI techniques in urban expansion modeling, the criteria and process of determining how AI made such a judgement should be verified. To overcome this limitation, explainable artificial intelligence (XAI) has been highlighted in recent studies [27]. The XAI is a methodology that strengthens the interpretive aspects of machine learning algorithms so that humans can easily understand the model results [28].

This study aims at modeling and predicting urban expansion in Seoul metropolitan area (SMA), Korea, using GIS and XAI techniques. To this end, we examine the effects of land-cover, socio-economic, and environmental features within the optimal radius from a certain raster cell. For the optimization, a multiple buffer analysis is performed and the prediction accuracy of each urban expansion model is compared. Then, this study combines the extreme gradient boosting (XGBoost) model and Shapley additive explanations (SHAP) in analyzing urban expansion. The findings of the study are expected to provide several policy implications in urban and environmental planning fields, particularly for effective and sustainable management of lands.

In this paper, Section 2 describes the materials and methodology of the study. It covers the explanations of dependent and independent variables, and details of XGBoost–SHAP models. Section 3 provides training, validation, and test results for urban growth models developed in the study. In addition, the urban expansion in Seoul metropolitan area in the near future is predicted. Lastly, Section 5 discusses the findings of the study and concludes with several recommendations for future studies.

2. Materials and Methods

2.1. Study Area

The Seoul Metropolitan Area (SMA), which includes Seoul, Incheon, and Gyeonggi-do province, is the fifth largest metropolitan area in the world [29]. It is located in the northwestern part of South Korea and occupies the Han River, which crosses the region (Figure 1). As of 2020, the territorial area of SMA is about 12,685 km^2, and its population is approximately 26 million (60% of the national population) [30].

Throughout the region, the SMA has experienced unprecedented urban growth due to explosive population inflows during the last decades [27]. Since the mid-2000s, more than ten new towns were developed around Seoul city, and essential infrastructures such as roads, energy, and water facilities were constructed to support them [31]. The direction of urban growth in SMA has been skewed to the southwestern part, since the northern and eastern parts of the region area dominantly covered by dense mountainous area [32]. Currently, the SMA consists of 64 administrative districts, where 25, 8, and 31 districts are included in Seoul, Incheon, and Gyeonggi-do province, respectively [33].

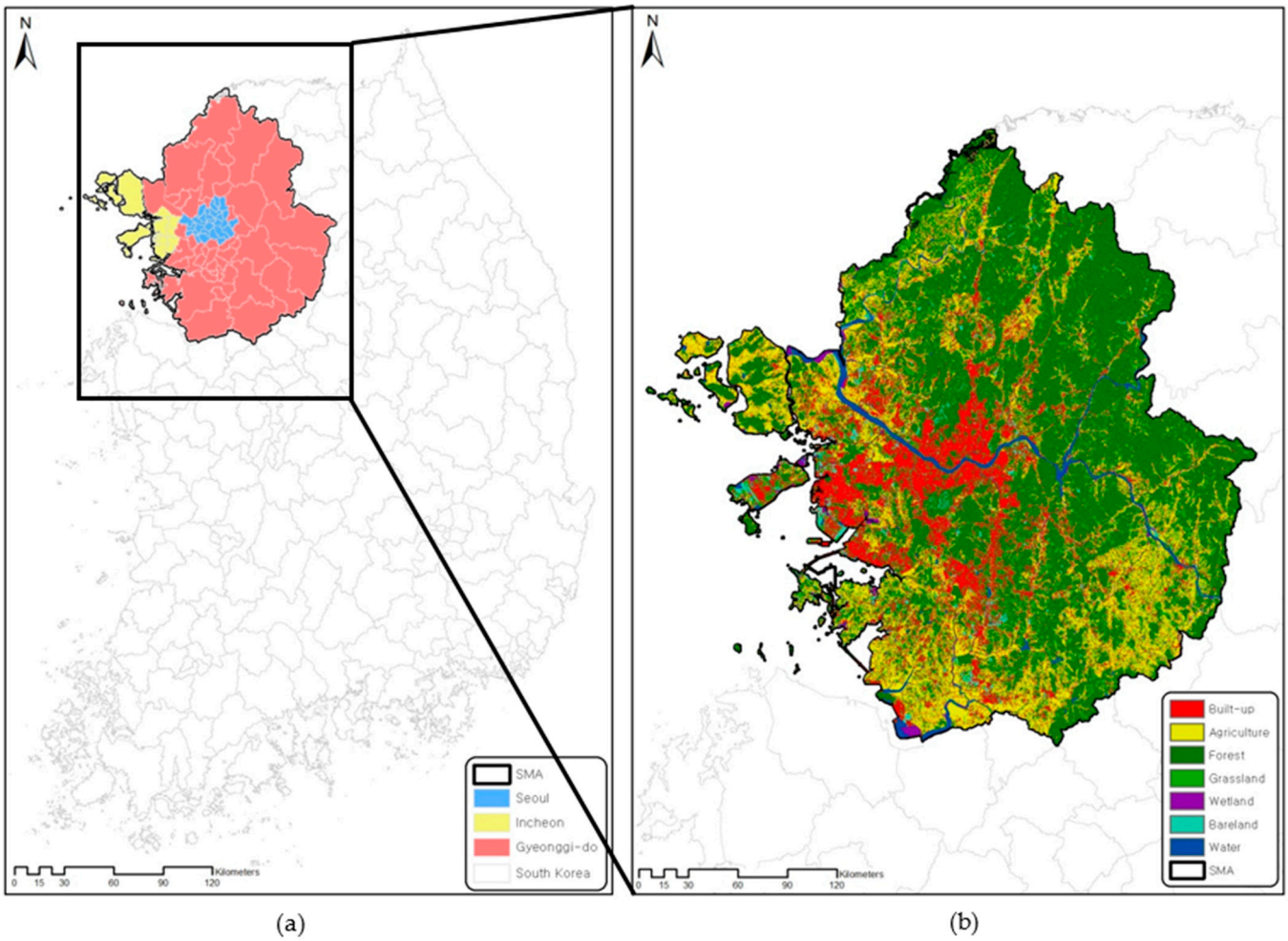


Figure 1. (**a**) Spatial extent and (**b**) land cover of study area (as of 2019).

2.2. Data

To develop the urban growth model in SMA, we constructed land-cover, topographic, socio-economic, and environmental features of the region for 2007 and 2019, respectively. Table 1 describes the dependent and independent variables used in the study.

Table 1. Description of variables used in the study.

	Data			Source (Year)
Dependent Variable	Dummy variable for urbanization from 2007 to 2019 (0: Non-urbanized area, 1: Urbanized area)			Land Cover Map (2007 and 2019)
Independent Variable	Topographic features		Elevation (m)	Digital Elevation Map (2007 and 2019)
			Slope (°)	
	Socio-economic features		Population density (person/km^2)	SGIS and KOSIS (2007 and 2019)
			GRDP per capita (1,000,000 won/person)	
	Land-cover features (within 50~1000 m buffer radius)	Urban areas	Residential area (m^2)	
			Commercial area (m^2)	
			Industrial area (m^2)	

Table 1. *Cont.*

Independent Variable	Land-cover features (within 50~1000 m buffer radius)	Non-urban areas	Recreational area (m^2)	
			Transportation area (m^2)	
			Public facility (m^2)	
			Rice paddy (m^2)	
			Farmland (m^2)	
			Facility cultivated area (m^2)	
			Orchard (m^2)	
			Other cultivated area (m^2)	
			Broadleaf forest (m^2)	
			Coniferous forest (m^2)	
			Mixed stand forest (m^2)	
			Natural grassland (m^2)	
			Artificial grassland (m^2)	
			Inland wetland (m^2)	
			Coastal wetland (m^2)	
			Natural bareland (m^2)	
			Artificial bareland (m^2)	
			Inland water (m^2)	
			Ocean (m^2)	
	Environmental features		Ecological ECVAM grade	ECVAM (2007 and 2019)
			Legislative ECVAM grade	

2.2.1. Dependent Variable

The dependent variable was a dummy variable that indicated whether a certain raster cell was urbanized or not from 2007 to 2019. In this study, 'urban areas' included residential, commercial, and industrial areas on the land-cover map, and the rest were defined as 'non-urban areas'. When a certain region that was a non-urban area in 2007 changed to urban area in 2019, we defined that the region was 'urbanized'. On the other hand, a certain region was considered as 'non-urbanized' when it remained a 'non-urban area' in both 2007 and 2019. We excluded raster cells that were already 'urban areas' in 2007 from study samples, because they would not be either 'urbanized' or 'non-urbanized' in 2019. For the analysis, this study extracted both 'urbanized' and 'non-urbanized' samples with the random sampling method. To avoid spatial autocorrelation issues, the minimum distance between each sample was selected as 50 m. Figure 2 illustrates the classification procedures for urbanized and non-urbanized areas in the study.

2.2.2. Independent Variable

The urban growth of a city is affected by various factors. To model and predict urban growth patterns, researchers have shown that urbanization was significantly associated with the city's land-cover [16,34], topographic [35], socio-economic [18,36], and environmental [3] features.

The independent variables in this study consist of (1) land-cover, (2) topographic, (3) socio-economic, and (4) environmental features. Land-cover and topographic features were derived from remotely sensed and digitized data, and socio-economic and environmental features were achieved from several national statistics databases. Figure 3 illustrates the independent variables (as of 2019) used in the study.

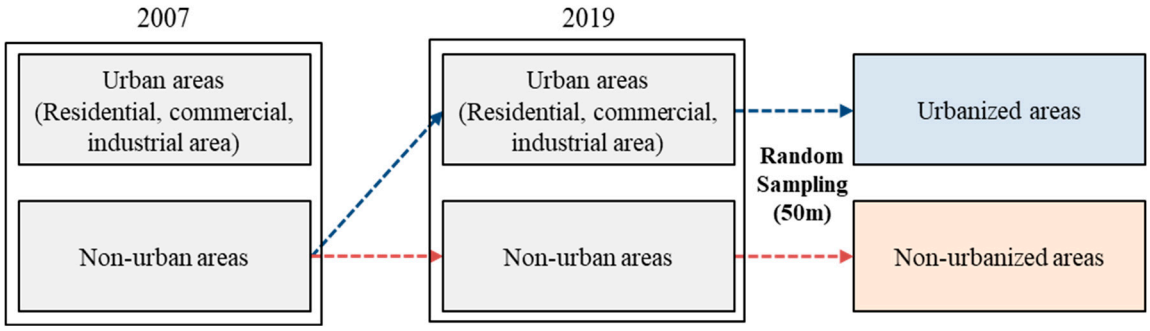


Figure 2. Classification of urbanized and non-urbanized areas.

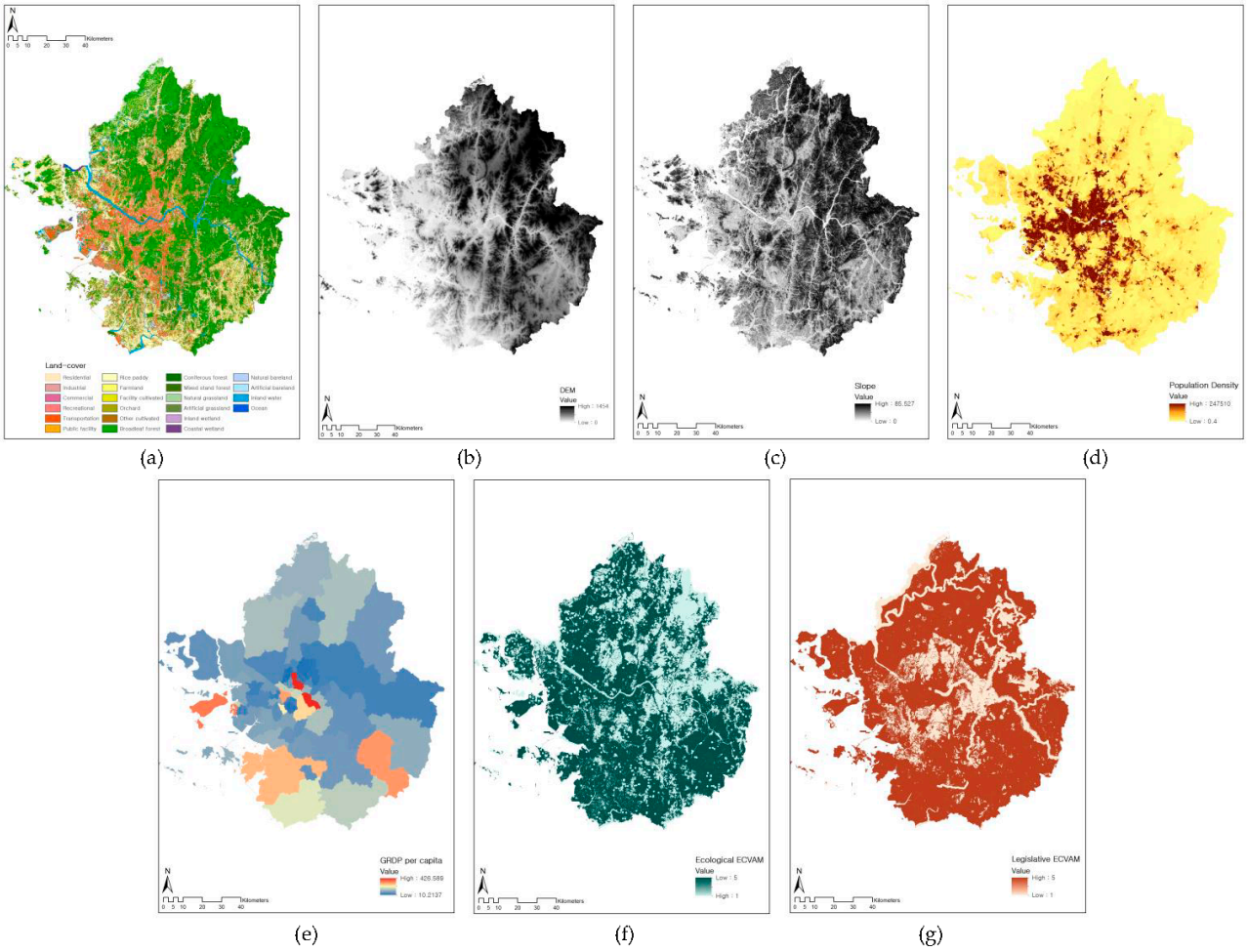


Figure 3. Variables used in the study—(**a**) Land-cover, (**b**) Elevation, (**c**) Slope, (**d**) Population density, (**e**) GRDP per capita, (**f**) Ecological ECVAM, (**g**) Legislative ECVAM.

First, this study utilized a 10 m resolution of national land-cover maps to derive the land-cover area within a radius from each raster cell. For 2007 and 2019, we calculated the area of 22 land-cover types within 50 m to 500 m buffer distances, by 100 m unit, from every single cell in SMA using ArcGIS software (ver. 10. 1). Second, this study achieved the elevation and slope of each raster cell in SMA by using a 10 m resolution of digital elevation map, which was provided by the National Spatial Data Infrastructure Portal (NSDI, http://www.nsdi.go.kr, accessed on 1 December 2022).

Third, we adopted population density and gross regional domestic product (GRDP) per capita as socio-economic features that affect urban growth patterns. These variables were provided as polygon data by Statistical Geographic Information Service (SGIS, https://

sgis.kostat.go.kr, accessed on 1 December 2022) and Korean Statistical Information Service (KOSIS, https://kosis.kr, accessed on 1 December 2022). The spatial unit of population density and GRDP per capita were census block group and county level, respectively. All raster cells that pertained to a certain census boundary were assigned its corresponding socio-economic values for 2007 and 2019.

Last, the Environment Conservation Value Assessment Map (ECVAM) was adopted as the environmental feature. It is provided annually by the Ministry of Environment (https://ecvam.neins.go.kr, accessed on 1 December 2022) and evaluates the environmental conservation value of national land in Korea [37]. The ECVAM is divided into two types: (1) the legislative and (2) ecological grade. Each grade is evaluated from grade 1 to 5, based on various environmental aspects of the entire nation. If the ECVAM grade of a certain region is close to 1, it means that the region has relatively high preservation value in terms of the environmental aspect and thus has low possibility of urbanization [38].

2.3. Methods

2.3.1. Research Process

The overall research process of the study is illustrated in Figure 4. First, we classified urbanized and non-urbanized areas from 2007 to 2019 and constructed an independent variable that corresponds with each dependent variable. To test the sensitivity of buffer distance on the model accuracy, all variables were calculated within 50 m to 1000 m radius from each raster cell in SMA.

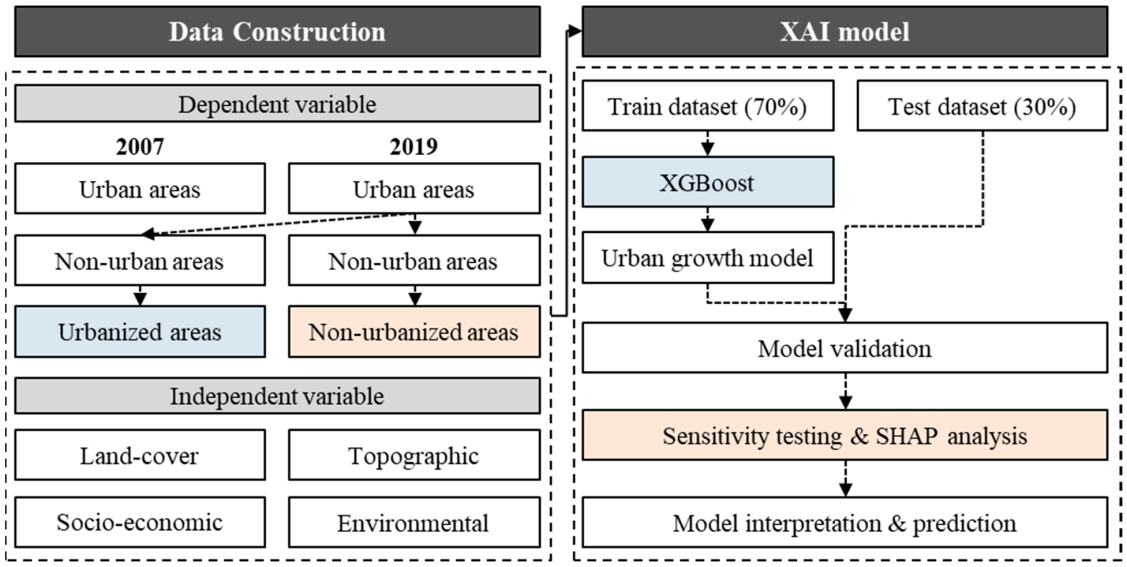


Figure 4. Research process.

Second, this study divided the dataset into training and testing parts, which account for 70% and 30% of total samples, respectively. Using the training dataset, the urban growth models were developed using XGBoost techniques. To optimize the model, several hyper-parameters were tuned using the 'Pycaret' package in Python. This package automatically adjusts many hyper-parameters when the number of folds is specified. We adopted a five-fold cross validation method to tune the hyper-parameters. To test the validity of the urban growth model, we utilized test dataset and predicted outcomes. Then, we tested the sensitivity of the XGBoost model with respect to buffer distance from raster cells and derived SHAP values for the optimal model. The SHAP analysis provides the relative importance and direction of independent variables in determining the possibility of urban growth.

Last, this study predicted the spatial patterns of SMA's urban growth in 2031, based on the constructed XGBoost model. While the process of urbanization is not linear, we use the urban growth tendency of previous 12 years to predict next 12 years. To this end, the

land-cover map in 2019 and surrounding land-cover, topographic, socio-economic, and environmental features were adopted as predictors. Outcomes include the probability of urban growth for every single raster cell in SMA, from 0 to 1.

2.3.2. XGBoost Model

The extreme gradient boosting (XGBoost) model is a decision tree-based algorithm that sequentially combines a number of weak learners to build a strong learner [39] and continuously reflects the residuals of the previous model into the next one to finally derive the optimal tree model [40]. It is also one of the most commonly used algorithms for solving problems with machine learning, and it is usually faster and more accurate than gradient boosting machines (GBMs) due to its mechanisms to prevent overfitting through regularization. Thanks to its high predictive accuracy and speed for both categorical and continuous variables, researchers in various fields have utilized the XGBoost model to predict their outcome which is also used in this paper [41].

The XGBoost algorithm consists of four main steps. First, the initialized tree model $\hat{y}_i$ for a given dataset $\{(x_i, y_i)\}_{i=1}^{N}$ is defined as follows:

$$\hat{y}_i = argmin_\gamma \sum_{i=1}^{N} L(x_i, \gamma) \quad (1)$$

where $argmin_\gamma$ indicates the constant value γ that minimizes the function, and $L(y, F(x))$ denotes a differentiable loss function of γ.

Second, for m number of iterations, the negative gradient of loss function $g_m(x_i)$ is calculated as:

$$g_m(x_i) = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}} \quad (2)$$

Here, $g_m(x_i)$ is a derivative of previous loss function $f_{m-1}(x)$.

Third, a base learner (or a weak learner) solves the optimization problem θ_m as:

$$\theta_m = argmin_\theta \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \theta t(x;\ \mu_m)) \quad (3)$$

where $L(y_i, F_{m-1}(x_i) + \theta t(x;\ \mu_m))$ denotes the loss function on each node i.

Last, the tree model is repeatedly updated as below:

$$f_m(x) = f_{m-1}(x) + \theta_m t(x;\ \mu_m) \quad (4)$$

Here, θ_m and $t(x;\ \mu_m)$ denote the learning rate and the selected node, respectively.

2.3.3. SHAP Values

To interpret the outcomes of AI-based models, several techniques including the local interpretable model-agnostic explanations (LIME) and the Shapley additive explanations (SHAP) were utilized [42]. While the LIME generates the surrogate model by randomly modifying input data and provides explanations, the SHAP provides the predictive ability of each variable [43,44]. Since the independent variables used in the study are diverse and nonlinear [45], we examined the relative importance of each feature through the SHAP model.

The Shapley Additive exPlanations (SHAP) is a methodology that provides explanations of results derived from machine learning models [46]. The SHAP value represents the average contribution of each attribute on predictor, by considering every possible combination [47]. For group $F\{i\}$, the SHAP value ϕ_i assigned to each feature i is calculated as below:

$$\phi_i = \sum_{S \in F\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_s(x_s)\right] \quad (5)$$

where $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_s(x_s)$ represents the differences of contribution between when feature i is used in the model or not.

3. Results

3.1. Descriptive Analysis

Table 2 summarizes the mean value of the independent variables of SMA in 2007 for areas which have been urbanized and non-urbanized by 2019. First, urbanized areas showed higher population density and GRDP per capita compared to non-urbanized areas. It seems reasonable in that densely populated areas with economic development are more advantageous for urban growth [48,49]. In addition, non-urbanized areas were found to have relatively higher elevation and slope than urbanized areas. This implies that harsh topographic environment is one of the powerful obstacles to urban growth [50].

Table 2. Descriptive statistics.

			Urbanized	Non-Urbanized
Number of samples			187,906	227,338
Socio-economic features (within 50 m buffer radius)		Population density (person/km^2)	1274.89	232.40
		GRDP per capita (1,000,000 won/person)	31.76	24.68
Topographic features (within 50 m buffer radius)		Elevation (m)	69.85	199.93
		Slope (°)	6.13	17.15
Land-cover features (within a 50 m buffer radius)	Urban areas	Residential area (m^2)	6703.92	1007.55
		Commercial area (m^2)	1828.96	281.35
		Industrial area (m^2)	928.06	121.83
		Recreational area (m^2)	85.30	44.95
		Transportation area (m^2)	3451.60	639.87
		Public facility (m^2)	1062.93	443.57
	Non-urban areas	Rice paddy (m^2)	34,875.41	24,232.09
		Farmland (m^2)	36,507.37	6652.93
		Facility cultivated area (m^2)	1123.34	202.86
		Orchard (m^2)	2730.18	722.37
		Other cultivated area (m^2)	5443.78	453.72
		Broadleaf forest (m^2)	6370.02	49,135.73
		Coniferous forest (m^2)	12,857.34	38,601.25
		Mixed stand forest (m^2)	8023.05	23,830.90
		Natural grassland (m^2)	1128.68	771.91
		Artificial grassland (m^2)	594.87	151.98
		Inland wetland (m^2)	1242.11	901.35
		Coastal wetland (m^2)	1505.63	1509.19
		Natural bareland (m^2)	60.29	24.00
		Artificial bareland (m^2)	11,949.39	1485.91
		Inland water (m^2)	2890.11	6484.35
		Ocean (m^2)	319.54	299.99
Environmental features (within 50 m buffer radius)		Ecological ECVAM grade	3.82	2.25
		Legislative ECVAM grade	3.28	2.35

In terms of land-cover features, areas that were urbanized from 2007 to 2019 showed higher residential, commercial, and industrial area with respect to those that were not urbanized. In addition, urbanized areas have relatively high agricultural, grassland, wetland, and bareland area, but low forest and water area compared to non-urbanized areas. The results suggest that urban growth is more prevalent in areas that are likely to be developed near urban areas [51,52]. Last, urbanized areas showed higher ECVAM grade for both ecological and legislative perspectives compared to non-urbanized areas. Since a higher ECVAM grade indicates lower environmental and ecological conservation value, the results imply that urban development is more active in areas with less legal regulations [53].

Figure 5 illustrates the spatial distribution of urban and non-urban areas in 2007 and 2019, as well as the urbanization from 2007 to 2019 in SMA. Over the 12 years, urban areas spread around Seoul city, particularly toward western and southern directions. This reflects

the spatial distribution of the new towns that were developed to alleviate the housing problems in Seoul city from mid-2000s [32].

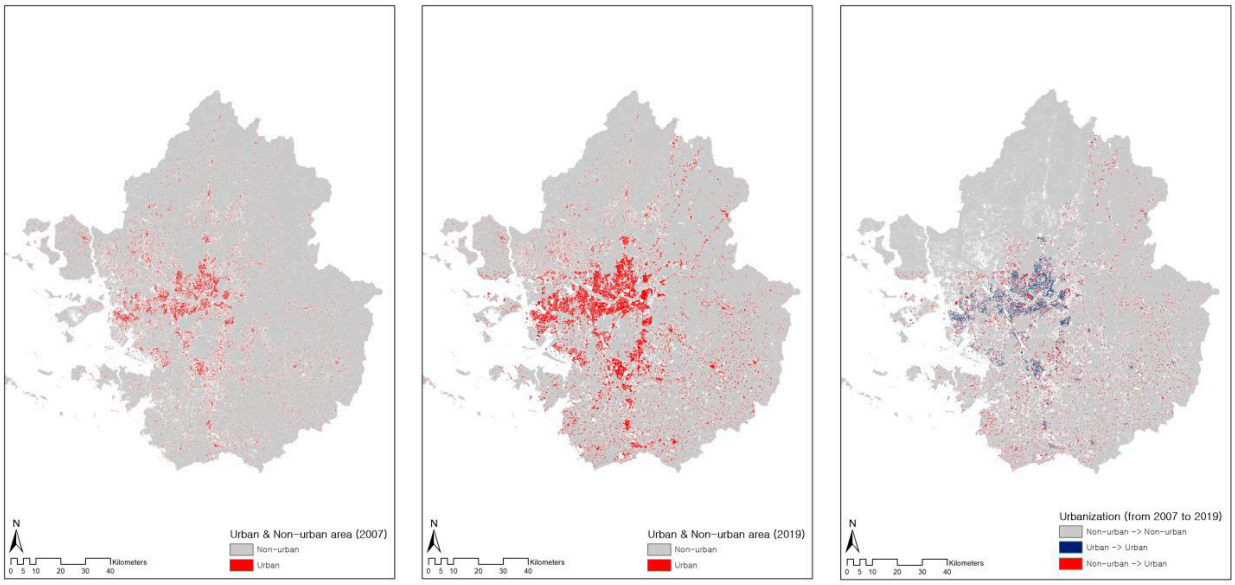


Figure 5. Spatial distribution of urban and non-urban areas in 2007 (**a**), 2019 (**b**) and urbanization from 2007 to 2019 (**c**).

3.2. Model Results

3.2.1. Hyper-Parameter Tuning

The hyper-parameters of the XGBoost model are summarized in Table 3. The optimal hyper-parameters vary as the buffer radius from each raster cell changes. However, Table 4 shows the hyper-parameter tuning results when a 50m buffer radius was adopted as a representative value. For the number of iterations, the model repeated 280 times to derive an optimized decision tree. In this model, the maximum depth of the decision tree is designated as 11, which indicates the level of complexity of a tree [54]. The ratio of the training dataset was 0.7, and the learning rate from the previous tree model was tuned to 0.2. The higher the learning rate is set, the more conservative the overall boosting process becomes [55]. In addition, the 'Alpha', 'Lambda', and 'Gamma' value of the XGBoost model was 1, 0, and 0, respectively, which controls the conservative level of the decision tree [56].

Table 3. Hyper-parameter tuning (XGBoost model with 50 m buffer radius).

Parameters	Values
Number of iterations	280
Max depth	11
Subsample ratio	0.7
Learning rate	0.2
Colsample_bytree	0.9
Alpha	1
Lambda	0
Gamma	0

Table 4. Confusion matrix.

		Predicted Value	
		TRUE	**FALSE**
Actual value	TRUE	TP	FN
	FALSE	FP	TN

3.2.2. Sensitivity Testing

Using the tuned hyper-parameters, we developed several urban growth models, by differentiating buffer radius from 50 m to 1000 m. To test the sensitivity of model accuracy, we calculated four accuracy metrics: accuracy, precision, recall, and F-1 score (Figure 6). Based on the confusion matrix (Table 4), each accuracy metrics can be calculated as the following:

$$Accuracy = \frac{(TN + TP)}{(TN + FP + FN + TP)} \quad (6)$$

$$Precision = \frac{TP}{(FP + TP)} \quad (7)$$

$$Recall = \frac{TP}{(FN + TP)} \quad (8)$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (9)$$

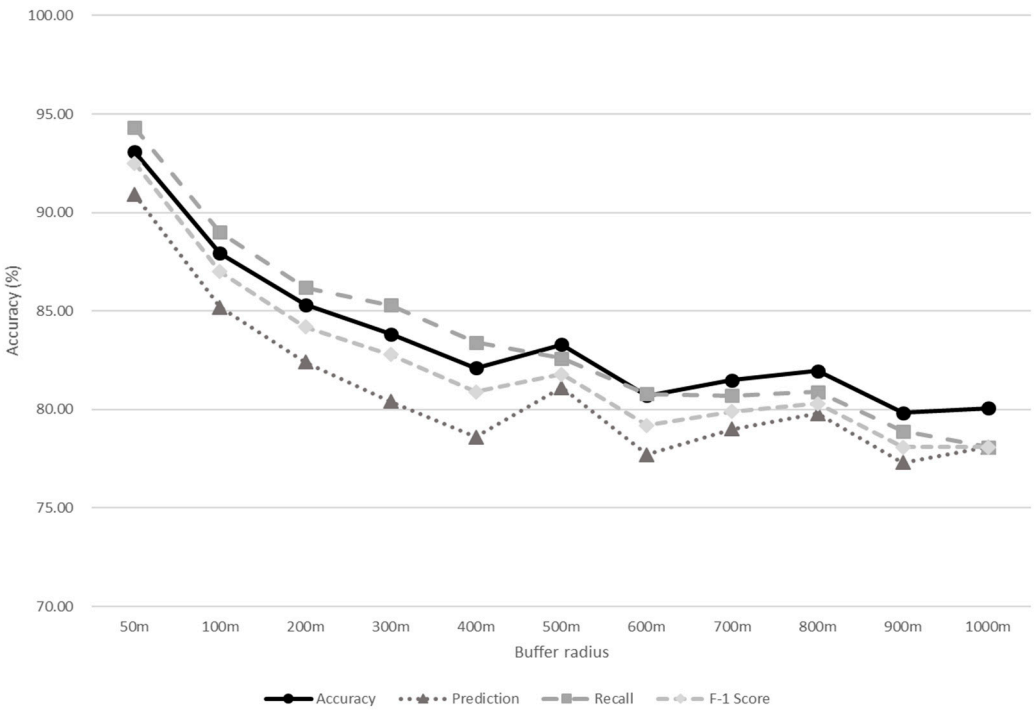
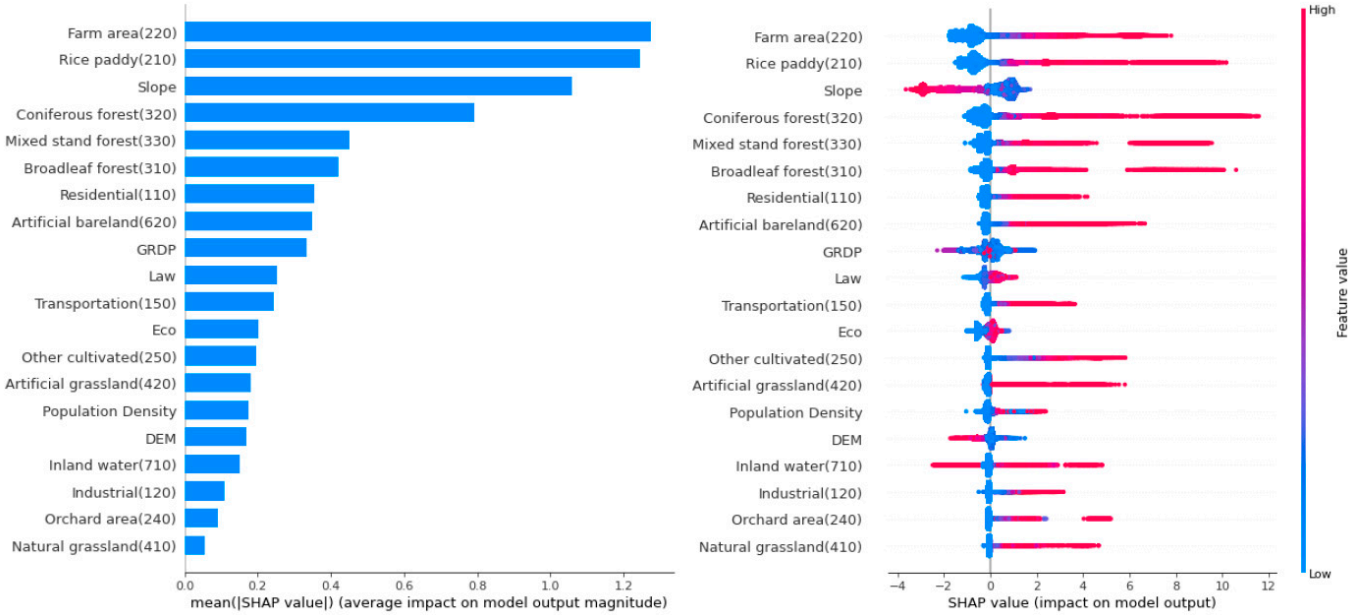


Figure 6. Sensitivity of XGBoost model accuracy by buffer radius (50~1000 m).

Overall, the prediction accuracy of the XGBoost model was higher than 75% regardless of buffer radius. The highest model accuracy was from approximately 90% to 95%, when a 50 m buffer from each raster cell was utilized as the independent variables. It is noteworthy that as the buffer radius from each raster cell increased, the overall accuracy of urban growth prediction decreased. This finding suggests that urban expansion is greatly influenced by the physical and socio-economic characteristics of immediate vicinity, rather than those of longer distances [57].

3.2.3. Factor Importance (SHAP)

Based on the XGBoost model developed in the study, we calculated SHAP values for each independent variable (Figure 7). Figure 7a illustrates the relative importance of variables, and Figure 7b shows the direction of variables in determining whether a certain area is being urbanized or not. The red- and blue-colored points indicate the high and low values of a certain variable, respectively, and the distribution of these points in the SHAP interval indicate the direction in which a variable contributes to the prediction [45].



(**a**) (**b**)

Figure 7. (**a**) Relative importance and (**b**) direction of SHAP values (50 m buffer radius).

First, agricultural areas including farm and rice paddy were the most influential factors of urban growth, followed by slope. This is line with previous studies' findings that flat and wide lands, such as agricultural areas, are more likely to be further developed [58]. In a similar vein, bareland and artificial grasslands were also found to be promoting factors of urbanization.

Second, developed areas including residential, transportation, and industrial areas within a 50 m buffer radius were positively associated with urban expansion. The inland water, however, showed more diverse impacts on whether a certain raster cell is urbanized or not. This implies that urban growth tends to occur near urbanized areas, while the existence of water may affect the level of urbanization [49]. Furthermore, the probability of urban growth of a certain raster cell was found to be negatively associated with the surrounding slope and elevation levels. It is not surprising that harsh topographic environment is one of the obvious obstacles in urban development [27].

Third, the socio-economic attributes including GRDP per capita and population density showed no significant effects on urban growth. The ecological and legislative ECVAM grades, on the other hand, showed slightly positive associations with urbanization. These findings suggest that urban growth is more dependent on land-use regulations than its economic and social driving factors [59].

3.3. Urban Growth Prediction in SMA

By applying the urban growth model of SMA from 2007 to 2019, we constructed the prediction map of urban growth in 2031 (Figure 8). The dependent variable of the urban growth model was whether a certain raster cell is urbanized or not from 2007 to 2019. In this kind of binary decision tree-based model, the predicted outcome is a number between 0 (non-urbanized at all) to 1 (100% sure of urbanization). The red-colored raster cells

indicate areas that were already urbanized in 2019, and the others indicate the probability of urbanization in SMA.

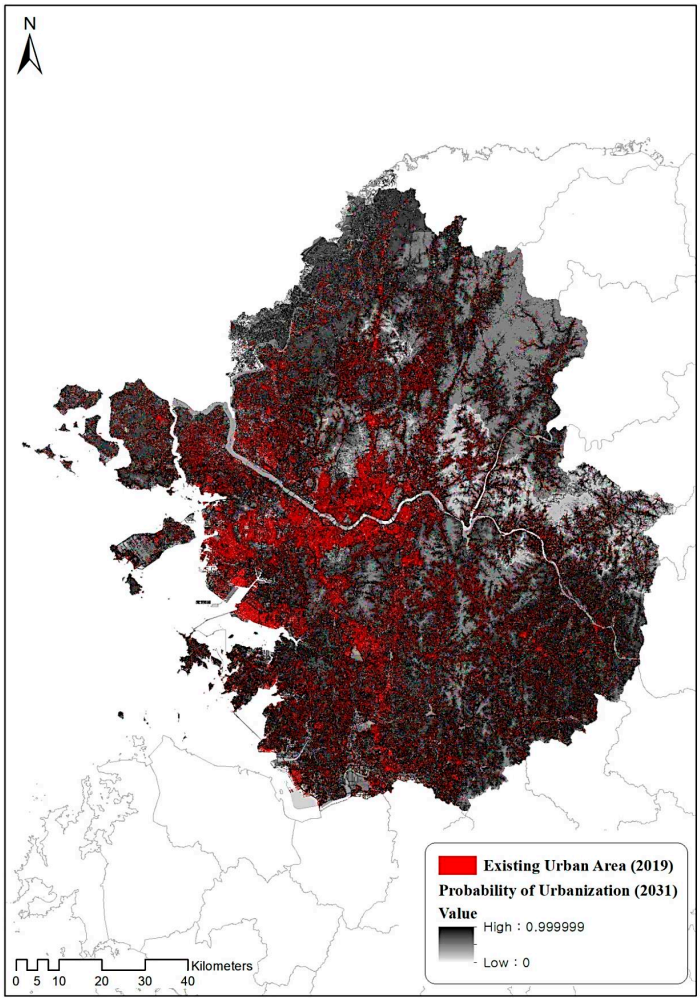


Figure 8. Prediction map of urban growth in SMA (2031).

For 95% and 90% probability, the predicted areas of urbanization in 2031 were 2514.14 km^2 and 4651.07 km^2, which accounts for 21.5% and 39.8% of the total areas in SMA, respectively. Considering that the proportion of urban areas in 2019 was only 4.7% of the total, our results suggest that future urban expansion may take place faster than before.

One of the noticeable points in the urban expansion map is that the southern part of SMA showed relatively high probability compared to northern and eastern parts of the region. It corresponds to the study's model results in that the majority of agricultural areas are currently distributed in the southern region (See Figure 1). In a similar vein, the northeastern part of the SMA dominantly consists of mountainous areas with high elevation and slope, and thus showed low probability of urbanization.

In addition, the probability of urban growth in 2031 seems to be spatially correlated with existing built-up areas in 2019, particularly for the distribution of urban sprawl area. It seems mainly due to the fact that that dense urban centers do not have enough land to be urbanized, but as they go to the outer areas, more developable areas such as farmlands and barelands are distributed around them [20].

4. Discussion

Urban growth is complex process in that various factors including physical, socio-economic, and political characteristics may affect the spatiotemporal changes of a city's

land-use patterns [60]. For this reason, analyzing and predicting the urban expansion has long been an area of interest in urban and environmental planning fields. However, from both methodological and theoretical perspectives, the existing literature had several limitations that required improvement.

In this context, the study takes a step forward in modeling urban growth in several aspects. First, the study adopted the XAI modeling techniques in examining and predicting urban expansion in SMA. More specifically, we integrated the XGBoost model and SHAP interpretations to interpret the relative importance and direction of various influencing factors on urban growth. It enabled the determination of the priorities in understanding the spatial patterns of urbanization, which had not been fully investigated in black-box models from previous literatures [61]. Furthermore, by utilizing the urban growth model developed in the study, we constructed a 10 m resolution map of urban growth prediction in 2031.

In addition, we compared several urban growth models with different buffer distances from raster cells, and the results showed that the overall accuracy of urban growth prediction is maximized (93%) when the physical and socio-economic attributes within a 50 m radius were used as predictors. When we developed additional urban growth models for a 10 m and 30 m radius, the prediction accuracy was 86% and 89%, respectively. This suggests that the effects of independent variables' spatial extents are not linearly associated with urban growth, and thus the optimal influencing distances need to be evaluated in prior to developing urban growth models.

From a theoretical perspective, the present study is novel in that it took into account land-cover, topographic, environmental, and socio-economic attributes in predicting urban growth patterns. The findings showed that urban growth was promoted when a certain area was close to agricultural and bareland areas with gentle elevation and slope. In addition, the existence of water body and high ECVAM grades tended to significantly reduce the possibility of urban expansion. It is noteworthy that both ecological and legislative regulations on land use were found to be significant factors in urban growth prediction. This suggests that the spatial patterns of urban expansion can be effectively controlled through institutional interventions [62].

Based on the study's findings, several policy implications in urban and environmental planning fields can be suggested. First, planners and practitioners in a given city (or nation) need to analyze the urban growth patterns and predict the spatial distributions of future urban areas. Such a prediction map of urban growth can help to estimate how cities will expand, and thus establish long-term strategies to prepare and mitigate problems [63]. Second, to control excessive urbanization and its spatial imbalance, the appropriate ecological and legislative restrictions on land-use development can be utilized. In order to apply such legal measures more effectively, the direction of policy measures need to be focused on public interest, such as designating national lands as a development restricted zone, rather than for individual benefit [64].

Third, the findings of the study showed that agricultural and forest areas that adjacent to built-up areas tended to be further urbanized. However, such tendency of urbanization may affect ecological systems within the city, such as a reduction of wildlife habitat and diversity [65]. In developing short- and long-term plans for urban growth, planners should consider not only its impact on humans, but also other species for sustainable urban development. Last, the application of XAI techniques can contribute to the development of both precise and interpretable urban growth models. The utility of XAI models is likely to increase in urban and environmental planning fields as it effectively supplemented the black-box features of AI, which has been one of the biggest obstacles.

5. Conclusions

This study developed an urban growth model in SMA, Korea by integrating XGBoost and SHAP models, and predicted future urban growth patterns in 2031. Results showed that urban growth is dominantly affected by land-cover characteristics, followed by topographic

and legal regulations. Based on these results, we suggested several policy measures that can be utilized in establishing and managing the sustainable urban development.

Despite the study's contribution in modeling and predicting urban growth, there is still some room for improvement in future research. First, as the study constructed the urban growth model by utilizing two cross-sectional attributes in 2007 and 2019, the process of urbanization during this period has not been fully reflected in the model. To overcome this in future studies, it will be necessary to construct a more precise urban growth model by collecting and analyzing time-series datasets regarding urbanization.

Second, the XGBoost–SHAP model used in this study does not guarantee the optimal prediction of urban growth among various XAI techniques. It is required to compare the accuracy of urban growth models by utilizing algorithms such as automated machine learning (AutoML). Thus, the prediction of urbanization can be further improved in future research [66].

Last, the effects of urban decline on the distribution of urban areas in future were not fully considered in this study. In South Korea, for example, urban decline was recently highlighted as one of the most urgent issues in urban planning fields [67]. While this study predicted urban growth in 2031 by reflecting the trends from 2007 to 2019, future works should also fully consider reduction in urban areas due to population decrease.

Author Contributions: Conceptualization, M.K. and G.K.; methodology, M.K., G.K. and D.J.; software, M.K. and D.K.; validation, M.K. and D.K.; formal analysis, M.K. and D.K.; investigation, M.K.; resources, G.K.; data curation, M.K. and D.K.; writing—original draft preparation, M.K.; writing—review and editing, D.J. and G.K.; visualization, M.K. and D.K.; supervision, G.K.; project administration, D.J. and G.K.; funding acquisition, D.J. and G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1C1C1013582).

Data Availability Statement: Not applicable.

Acknowledgments: This paper is based on the findings of the research project "A Study on Data-based Environment Inequality and Influence Analysis Techniques Using Machine Learning and Spatio-temporal Analysis", (2022-037(R)) which was conducted by the Korea Environment Institute (KEI).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gómez, J.A.; Patiño, J.E.; Duque, J.C.; Passos, S. Spatiotemporal modeling of urban growth using machine learning. *Remote Sens.* **2019**, *12*, 109. [CrossRef]
2. Moghadam, H.S.; Helbich, M. Spatiotemporal urbanization processes in the megacity of Mumbai, India: A Markov chains-cellular automata urban growth model. *Appl. Geogr.* **2013**, *40*, 140–149. [CrossRef]
3. Park, S.; Jeon, S.; Kim, S.; Choi, C. Prediction and comparison of urban growth by land suitability index mapping using GIS and RS in South Korea. *Landsc. Urban Plan.* **2011**, *99*, 104–114. [CrossRef]
4. Jiang, B.; Yao, X. Geospatial analysis and modeling of urban structure and dynamics: An overview. *Geospat. Anal. Model. Urban Struct. Dyn.* **2010**, *99*, 3–11.
5. Park, S.; Jeon, S.; Choi, C. Mapping urban growth probability in South Korea: Comparison of frequency ratio, analytic hierarchy process, and logistic regression models and use of the environmental conservation value assessment. *Landsc. Ecol. Eng.* **2012**, *8*, 17–31. [CrossRef]
6. Clarke, K.C.; Hoppen, S.; Gaydos, L. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environ. Plan. B Plan. Des.* **1997**, *24*, 247–261. [CrossRef]
7. Clarke, K.C.; Gaydos, L.J. Loose-coupling a cellular automaton model and GIS: Long-term urban growth prediction for San Francisco and Washington/Baltimore. *Int. J. Geogr. Inf. Sci.* **1998**, *12*, 699–714. [CrossRef]
8. Zhang, Q.; Su, S. Determinants of urban expansion and their relative importance: A comparative analysis of 30 major metropolitans in China. *Habitat Int.* **2016**, *58*, 89–107. [CrossRef]
9. Yu, W.; Zhou, W. The spatiotemporal pattern of urban expansion in China: A comparison study of three urban megaregions. *Remote Sens.* **2017**, *9*, 45. [CrossRef]
10. Cheng, J.; Masser, I. Urban growth pattern modeling: A case study of Wuhan city, PR China. *Landsc. Urban Plan.* **2003**, *62*, 199–217. [CrossRef]

11. Sarkar, A.; Chouhan, P. Modeling spatial determinants of urban expansion of Siliguri a metropolitan city of India using logistic regression. *Model. Earth Syst. Environ.* **2020**, *6*, 2317–2331. [CrossRef]
12. Pampoore-Thampi, A.; Varde, A.S.; Yu, D. Mining GIS data to predict urban sprawl. *arXiv* **2021**, arXiv:2103.11338.
13. Karimi, F.; Sultana, S.; Babakan, A.S.; Suthaharan, S. Urban expansion modeling using an enhanced decision tree algorithm. *GeoInformatica* **2021**, *25*, 715–731. [CrossRef]
14. Frimpong, B.F.; Molkenthin, F. Tracking urban expansion using random forests for the classification of landsat imagery (1986–2015) and predicting urban/built-up areas for 2025: A Study of the Kumasi Metropolis, Ghana. *Land* **2021**, *10*, 44. [CrossRef]
15. Mirbagheri, B.; Alimohammadi, A. Integration of local and global support vector machines to improve urban growth modelling. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 347. [CrossRef]
16. Karimi, F.; Sultana, S.; Babakan, A.S.; Suthaharan, S. An enhanced support vector machine model for urban expansion prediction. *Comput. Environ. Urban Syst.* **2019**, *75*, 61–75. [CrossRef]
17. Mohammady, S.; Delavar, M.R. Urban sprawl assessment and modeling using landsat images and GIS. *Model. Earth Syst. Environ.* **2016**, *2*, 155. [CrossRef]
18. Al Rifat, S.A.; Liu, W. Predicting future urban growth scenarios and potential urban flood exposure using Artificial Neural Network-Markov Chain model in Miami Metropolitan Area. *Land Use Policy* **2022**, *114*, 105994. [CrossRef]
19. Chaturvedi, V.; de Vries, W.T. Machine Learning Algorithms for Urban Land Use Planning: A Review. *Urban Sci.* **2021**, *5*, 68. [CrossRef]
20. Li, X.; Zhou, W.; Ouyang, Z. Forty years of urban expansion in Beijing: What is the relative importance of physical, socioeconomic, and neighborhood factors? *Appl. Geogr.* **2013**, *38*, 1–10. [CrossRef]
21. Mandal, J.; Ghosh, N.; Mukhopadhyay, A. Urban growth dynamics and changing land-use land-cover of megacity Kolkata and its environs. *J. Indian Soc. Remote Sens.* **2019**, *47*, 1707–1725. [CrossRef]
22. Domingo, D.; Palka, G.; Hersperger, A.M. Effect of zoning plans on urban land-use change: A multi-scenario simulation for supporting sustainable urban growth. *Sustain. Cities Soc.* **2021**, *69*, 102833. [CrossRef]
23. Jiang, L.; Deng, X.; Seto, K.C. Multi-level modeling of urban expansion and cultivated land conversion for urban hotspot counties in China. *Landsc. Urban Plan.* **2012**, *108*, 131–139. [CrossRef]
24. Javan, S.L.; Sepehri, M.M. A predictive framework in healthcare: Case study on cardiac arrest prediction. *Artif. Intell. Med.* **2021**, *117*, 102099. [CrossRef]
25. Pintelas, E.; Livieris, I.E.; Pintelas, P. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. *Algorithms* **2020**, *13*, 17. [CrossRef]
26. Herm, L.V.; Heinrich, K.; Wanner, J.; Janiesch, C. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *Int. J. Inf. Manag.* **2022**, 102538. [CrossRef]
27. Kim, M.; Kim, G. Modeling and Predicting Urban Expansion in South Korea Using Explainable Artificial Intelligence (XAI) Model. *Appl. Sci.* **2022**, *12*, 9169. [CrossRef]
28. Kim, M.; Kim, D.; Kim, G. Examining the Relationship between Land Use/Land Cover (LULC) and Land Surface Temperature (LST) Using Explainable Artificial Intelligence (XAI) Models: A Case Study of Seoul, South Korea. *Int. J. Environ. Res. Public Health* **2022**, *19*, 15926. [CrossRef]
29. Choi, C.G.; Lee, S.; Kim, H.; Seong, E.Y. Critical junctures and path dependence in urban planning and housing policy: A review of greenbelts and New Towns in Korea's Seoul metropolitan area. *Land Use Policy* **2019**, *80*, 195–204. [CrossRef]
30. Kim, S.; Kim, Y.J.; Peck, K.R.; Ko, Y.; Lee, J.; Jung, E. Keeping low reproductive number despite the rebound population mobility in Korea, a country never under lockdown during the COVID-19 pandemic. *Int. J. Environ. Res. Public Health* **2020**, *17*, 9551. [CrossRef]
31. Kim, H.; Kim, Y.K.; Song, S.K.; Lee, H.W. Impact of future urban growth on regional climate changes in the Seoul Metropolitan Area, Korea. *Sci. Total Environ.* **2016**, *571*, 355–363. [CrossRef]
32. Bae, S.; Chang, H. Urbanization and floods in the Seoul Metropolitan area of South Korea: What old maps tell us. *Int. J. Disaster Risk Reduct.* **2019**, *37*, 101186. [CrossRef]
33. Chang, H.; Kwon, W.T. Spatial variations of summer precipitation trends in South Korea, 1973–2005. *Environ. Res. Lett.* **2007**, *2*, 045012. [CrossRef]
34. Choi, Y.; Lim, C.H.; Chung, H.I.; Kim, Y.; Cho, H.J.; Hwang, J.; Kraxner, F.; Biging, G.S.; Lee, W.K.; Chon, J.; et al. Forest management can mitigate negative impacts of climate and land-use change on plant biodiversity: Insights from the Republic of Korea. *J. Environ. Manag.* **2021**, *288*, 112400. [CrossRef]
35. Dadashpoor, H.; Ahani, S. Explaining objective forces, driving forces, and causal mechanisms affecting the formation and expansion of the peri-urban areas: A critical realism approach. *Land Use Policy* **2021**, *102*, 105232. [CrossRef]
36. Liu, X.; Wei, M.; Li, Z.; Zeng, J. Multi-scenario simulation of urban growth boundaries with an ESP-FLUS model: A case study of the Min Delta region, China. *Ecol. Indic.* **2022**, *135*, 108538. [CrossRef]
37. Lyu, R.; Mi, L.; Zhang, J.; Xu, M.; Li, J. Modeling the effects of urban expansion on regional carbon storage by coupling SLEUTH-3r model and InVEST model. *Ecol. Res.* **2019**, *34*, 380–393. [CrossRef]
38. Kim, G.H.; Choi, H.S.; Kim, D.B.; Jung, Y.R.; Jin, D.Y. Urban sprawl prediction in 2030 using decision tree. *J. Korean Soc. Environ. Restor. Technol.* **2020**, *23*, 125–135.

39. Guo, R.; Zhao, Z.; Wang, T.; Liu, G.; Zhao, J.; Gao, D. Degradation state recognition of piston pump based on ICEEMDAN and XGBoost. *Appl. Sci.* **2020**, *10*, 6593. [CrossRef]
40. Parsa, A.B.; Movahedi, A.; Taghipour, H.; Derrible, S.; Mohammadian, A.K. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* **2020**, *136*, 105405. [CrossRef]
41. Zhang, Y.; Haghani, A. A gradient boosting method to improve travel time prediction. *Transp. Res. Part C Emerg. Technol.* **2015**, *58*, 308–324. [CrossRef]
42. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable [09.July]. 2020. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 1 February 2023).
43. Iban, M.C. An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants. *Habitat Int.* **2022**, *128*, 102660. [CrossRef]
44. Moon, J.; Rho, S.; Baik, S.W. Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values. *Sustain. Energy Technol. Assess.* **2022**, *54*, 102888. [CrossRef]
45. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **2022**, *96*, 101845. [CrossRef]
46. Park, J.; Lee, W.H.; Kim, K.T.; Park, C.Y.; Lee, S.; Heo, T.Y. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Sci. Total Environ.* **2022**, *832*, 155070. [CrossRef]
47. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
48. Richards, D.R.; Passy, P.; Oh, R.R. Impacts of population density and wealth on the quantity and structure of urban green space in tropical Southeast Asia. *Landsc. Urban Plan.* **2017**, *157*, 553–560. [CrossRef]
49. Li, G.; Li, F. Urban sprawl in China: Differences and socioeconomic drivers. *Sci. Total Environ.* **2019**, *673*, 367–377. [CrossRef]
50. Qian, S.; Qin, D.; Wu, X.; Hu, S.; Hu, L.; Lin, D.; Zhao, L.; Shang, K.; Song, K.; Yang, Y. Urban growth and topographical factors shape patterns of spontaneous plant community diversity in a mountainous city in southwest China. *Urban For. Urban Green.* **2020**, *55*, 126814. [CrossRef]
51. Hou, H.; Estoque, R.C.; Murayama, Y. Spatiotemporal analysis of urban growth in three African capital cities: A grid-cell-based analysis using remote sensing data. *J. Afr. Earth Sci.* **2016**, *123*, 381–391. [CrossRef]
52. Kafy, A.A.; Naim, N.H.; Khan, M.H.H.; Islam, M.A.; Al Rakib, A.; Al-Faisal, A.; Sarker, M.H.S. Prediction of urban expansion and identifying its impacts on the degradation of agricultural land: A machine learning-based remote-sensing approach in Rajshahi, Bangladesh. In *Re-Envisioning Remote Sensing Applications*; CRC Press: Boca Raton, FL, USA, 2021; pp. 85–106.
53. Hong, H.J.; Kim, C.K.; Lee, H.W.; Lee, W.K. Conservation, Restoration, and Sustainable Use of Biodiversity Based on Habitat Quality Monitoring: A Case Study on Jeju Island, South Korea (1989–2019). *Land* **2021**, *10*, 774. [CrossRef]
54. Zhou, S.; Liu, Z.; Wang, M.; Gan, W.; Zhao, Z.; Wu, Z. Impacts of building configurations on urban stormwater management at a block scale using XGBoost. *Sustain. Cities Soc.* **2022**, *87*, 104235. [CrossRef]
55. Hak Lee, E.; Kim, K.; Kho, S.Y.; Kim, D.K.; Cho, S.H. Estimating Express Train Preference of Urban Railway Passengers Based on Extreme Gradient Boosting (XGBoost) using Smart Card Data. *Transp. Res. Rec.* **2021**, *2675*, 64–76. [CrossRef]
56. Lin, L.; Liang, Y.; Liu, L.; Zhang, Y.; Xie, D.; Yin, F.; Ashraf, T. Estimating PM2. 5 Concentrations Using the Machine Learning RF-XGBoost Model in Guanzhong Urban Agglomeration, China. *Remote Sens.* **2022**, *14*, 5239. [CrossRef]
57. Jiao, L.; Liu, J.; Xu, G.; Dong, T.; Gu, Y.; Zhang, B.; Liu, Y.; Liu, X. Proximity Expansion Index: An improved approach to characterize evolution process of urban expansion. *Comput. Environ. Urban Syst.* **2018**, *70*, 102–112. [CrossRef]
58. Hu, Y.; Kong, X.; Zheng, J.; Sun, J.; Wang, L.; Min, M. Urban expansion and farmland loss in Beijing during 1980–2015. *Sustainability* **2018**, *10*, 3927. [CrossRef]
59. He, Q.; Zeng, C.; Xie, P.; Tan, S.; Wu, J. Comparison of urban growth patterns and changes between three urban agglomerations in China and three metropolises in the USA from 1995 to 2015. *Sustain. Cities Soc.* **2019**, *50*, 101649. [CrossRef]
60. Aburas, M.M.; Ho, Y.M.; Ramli, M.F.; Ash'aari, Z.H. The simulation and prediction of spatio-temporal urban growth trends using cellular automata models: A review. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 380–389. [CrossRef]
61. Rode, P.; Heeckt, C.; da Cruz, N.F. *National Transport Policy and Cities: Key Policy Interventions to Drive Compact and Connected Urban Growth*; Coalition for Urban Transitions: London, UK; Washington, DC, USA, 2019.
62. Boutaghane, H.; Boudjemaa, K.; Dehimi, S. Geospatial modelling of the future urban expansion map using AHP and GIS in Bordj Bou Arreridj, Algeria. *J. Degrad. Min. Lands Manag.* **2022**, *9*, 3733–3743. [CrossRef]
63. Xu, J.; Wang, X. Reversing Uncontrolled and Unprofitable Urban Expansion in Africa through Special Economic Zones: An Evaluation of Ethiopian and Zambian Cases. *Sustainability* **2020**, *12*, 9246. [CrossRef]
64. Cogan, C.B.; Davis, F.W.; Clarke, K.C. *Application of Urban Growth Models and Wildlife Habitat Models to Assess Biodiversity Losses*; University of California-Santa Barbara Institute for Computational Earth System Science; US Department of the Interior, US geological Survey, Biological Resources Division, Gap Analysis Program: Santa Barbara, CA, USA, 2001.
65. Yigitcanlar, T.; Kankanamge, N.; Regona, M.; Maldonado, A.; Rowan, B.; Ryu, A.; DeSouza, K.C.; Corchado, J.M.; Mehmood, R.; Li, R.Y.M. Artificial intelligence technologies and related urban planning and development concepts: How are they perceived and utilized in Australia? *J. Open Innov. Technol. Mark. Complex.* **2020**, *6*, 187. [CrossRef]

66. Guo, Y.; Quan, L.; Song, L.; Liang, H. Construction of rapid early warning and comprehensive analysis models for urban waterlogging based on AutoML and comparison of the other three machine learning algorithms. *J. Hydrol.* **2022**, *605*, 127367. [CrossRef]
67. Hwang, U.; Woo, M. Analysis of inter-relationships between urban decline and urban sprawl in city-regions of South Korea. *Sustainability* **2020**, *12*, 1656. [CrossRef]