PNIE PROCEEDINGS OF THE NATIONAL INSTITUTE OF ECOLOGY OF THE REPUBLIC OF KOREA

Original Article

PNIE 2021;2(1):1-14
https://doi.org/10.22920/PNIE.2021.2.1.1
pISSN 2765-2203, eISSN 2765-2211

Application of Statistical and Machine Learning Techniques for Habitat Potential Mapping of Siberian Roe Deer in South Korea

Saro Lee[1,2*], Fatemeh Rezaie[1,2]

[1]Geoscience Platform Division, Korea Institute of Geoscience and Mineral Resources (KIGAM), Daejeon, Korea
[2]Department of Geophysical Exploration, Korea University of Science and Technology, Daejeon, Korea

ABSTRACT

The study has been carried out with an objective to prepare Siberian roe deer habitat potential maps in South Korea based on three geographic information system-based models including frequency ratio (FR) as a bivariate statistical approach as well as convolutional neural network (CNN) and long short-term memory (LSTM) as machine learning algorithms. According to field observations, 741 locations were reported as roe deer's habitat preferences. The dataset were divided with a proportion of 70:30 for constructing models and validation purposes. Through FR model, a total of 10 influential factors were opted for the modelling process, namely altitude, valley depth, slope height, topographic position index (TPI), topographic wetness index (TWI), normalized difference water index, drainage density, road density, radar intensity, and morphological feature. The results of variable importance analysis determined that TPI, TWI, altitude and valley depth have higher impact on predicting. Furthermore, the area under the receiver operating characteristic (ROC) curve was applied to assess the prediction accuracies of three models. The results showed that all the models almost have similar performances, but LSTM model had relatively higher prediction ability in comparison to FR and CNN models with the accuracy of 76% and 73% during the training and validation process. The obtained map of LSTM model was categorized into five classes of potentiality including very low, low, moderate, high and very high with proportions of 19.70%, 19.81%, 19.31%, 19.86%, and 21.31%, respectively. The resultant potential maps may be valuable to monitor and preserve the Siberian roe deer habitats.

Keywords: Convolutional neural network algorithm, Frequency ratio method, Habitat potential map, Long short-term memory algorithm, Machine learning algorithms, Siberian roe deer

Introduction

The roe deer as a meso-mammals of Palearctic distribution can be classified into two distinct species: the European roe deer (*Capreolus capreolus*) and the Siberian roe deer (*Capreolus pygargus*). The habitats of *Capreolus pygargus* is continental Asia and some regions of Eastern Europe, from the Khoper River and Don River bend to the Ural Mountains and across southern Siberia. It is distributed throughout northern Mongolia and east to the

Received December 28, 2020; **Revised** January 13, 2021; **Accepted** January 13, 2021

***Corresponding author:** Saro Lee
e-mail leesaro@kigam.re.kr
https://orcid.org/0000-0003-0409-8263

coastlines of the East Sea, and the Yellow Sea, including the Korean Peninsula especially Jeju Island (Koh & Randi, 2001; Lee *et al*., 2015; 2016; Sokolov *et al*., 1982). The European roe deer travel from Europe to Western Russia and the Siberian roe deer migrate from Russia to Korea (Park *et al*., 2014). The differences between these species are their body sizes and morphometric characteristics (Danilkin & Hewison, 1996). Abundance of roe deer species as an inevitable food source is one of the main factor to control many raptors population (Jiang *et al*., 2015). Although International Union for the Conservation of Nature declared that Siberian roe deer is in the Red List as the least concern specie (Lovari *et al*., 2016), their population have considerably declined owing to trapping and overhunting. Therefore, habitat potential map can play a prominent role in developing conservation plans and maintaining biodiversity of roe deer as a game species.

In recent years, because of rapid increasing of the quantity and quality of geospatial data, attempts have been made to generate more accurate habitat potential maps for a variety of species. To achieve this goal, geographic information system (GIS) as a powerful instrument can be applied to assess the spatial correlations between species occurrences and spatial variables. Based on the literature review, various types of GIS-based models have been employed to develop habitat potential maps for different plant and animal species. These methods can be categorized into four groups, namely bivariate, multivariate, multi-criteria decision-making (MCDM), and artificial intelligence (AI) algorithms (Kadirhodjaev *et al*., 2020). The first two methods are statistical modeling which can be implemented easily and the outcomes can be interpreted simply. In spite of the uncomplicated performance of these data-driven methods, the size and precision of input data can affect the predictive accuracy of employed models. Additionally, selection effective parameters among multicollinearity variables can cause challenges in statistical models (Farrell *et al*., 2019). The most representative of bivariate and multivariate approaches are the frequency ratio (FR) (Choi *et al*., 2011a), weights-of-evidence model (Choi *et al*., 2011b), logistic regression (Imam & Kushwaha, 2013; Pereira & Itami, 1991), and evidential belief function (Lee *et al*., 2019).

Multi-criteria decision-making (MCDM) algorithms are a type of knowledge-driven methods applied to assign weights to conditioning factors based on evidence of varying quality, guideline, and expert opinions (Al-Abadi *et al*., 2016). Analytical hierarchy process is the most widely used MCDM method for habitat mapping (Aini *et al*., 2015; Imam & Tesfamichael, 2013; Mesfin & Berhan, 2016; Sanare *et al*., 2015).

Nowadays, machine learning methods received growing attention of researchers in a wide range of scientific fields due to their predictive ability to identify patterns in historical data and find the nonlinear connection between variables. The most prevalent machine learning approaches implemented for species geographic distributions modeling includes artificial neural network (Lee *et al*., 2013; 2017), random forest (Ng *et al*., 2018; Robert *et al*., 2016), maximum entropy model (Park *et al*., 2018; Zhang *et al*., 2020a), support vector machine (Cui *et al*., 2020), multivariate adaptive regression spline (Leathwick *et al*., 2005), generalized additive models (Kosicki, 2020; Sanchez *et al*., 2008; Schmiing *et al*., 2013) and classification and regression tree (Garzón *et al*., 2006). In addition, Oh *et al*. (2019) showed the effectiveness of support vector machine and naïve bayes techniques to produce potential ruditapes philippinarum habitat map in the Geunso Bay, South Korea. Rahimian Boogar *et al*. (2019) used support vector machine and maximum entropy methods to predict habitat suitability of Juniperus spp. in the Southern part of Zagros Mountains, Iran. Farrell *et al*. (2019) developed habitat suitability maps of wild turkeys in Mississippi using three machine learning algorithms, namely maximum entropy, random forests, and support vector machines.

With the advent of big data and in order to overcome the drawbacks of mentioned techniques, deep learning methods were introduced as a subfield of machine learning techniques powered by AI which outperformed the prediction precision. The most popular deep learning algorithms include convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM) network, and deep belief networks. As emphasized by a great number of studies, these algorithms are able to find relationship between training sample data and apply it to new sets. In facts, deep learning algorithms have promising capacity of processing unstructured data which contains multiple features. Moreover, these approaches can enhance the robustness and accuracy of analyzing.

The main objective of the present study is to investigate applicability of CNN and LSTM algorithms to produce Siberian roe deer habitat distributions map in South Korea for the first time in the field of species distribution model. Furthermore, the accuracy of predictions are compared with FR method as a bivariate statistical model. Finally, the spatial relationships between the several habitat distributions and various environmental influencing factors will be elucidated which can be used for defining long-term programs to conserve this specie.

Materials and Methods

Materials

Study area

The Korean Peninsula is situated in Northeast Asia. The country and all its islands with an approximately area of 100,302 km^2 lies between latitudes 33° and 39° N, and

longitudes 124° and 130° E (Fig. 1). The country reaches to Democratic People's Republic of Korea in the north and it is surrounded by the East Sea to the east, the Korea Strait and the East China Sea to the south, and the Yellow Sea to the west. There are large coastal plains in the southern and western regions and mountains covered majority parts of South Korea (70%) are located in the eastern and northern parts. Owing to these topographical structures, only 30% of the total land in South Korea, which is located in the west and southeast, are suitable for agricultural purposes.

South Korea has more than three thousand islands which most of them are small and uninhabited located in the western and southern coasts of South Korea. Jeju, the largest Island in this country, is about 1,825 km^2 and 100 km off the southern coast of South Korea. Jeju is warmer than the rest of South Korea because of its humid and subtropical climate. Winters are cool and dry while summers are hot, humid, and even sometimes rainy. The Halla-san Mountain as an extinct shield volcano is one of the three main mountains of South Korea; with elevation 1,950 m above sea level is in Jeju which makse this Island to be the highest point in South Korea. The two other high mountains are Jirisan and Seoraksan.

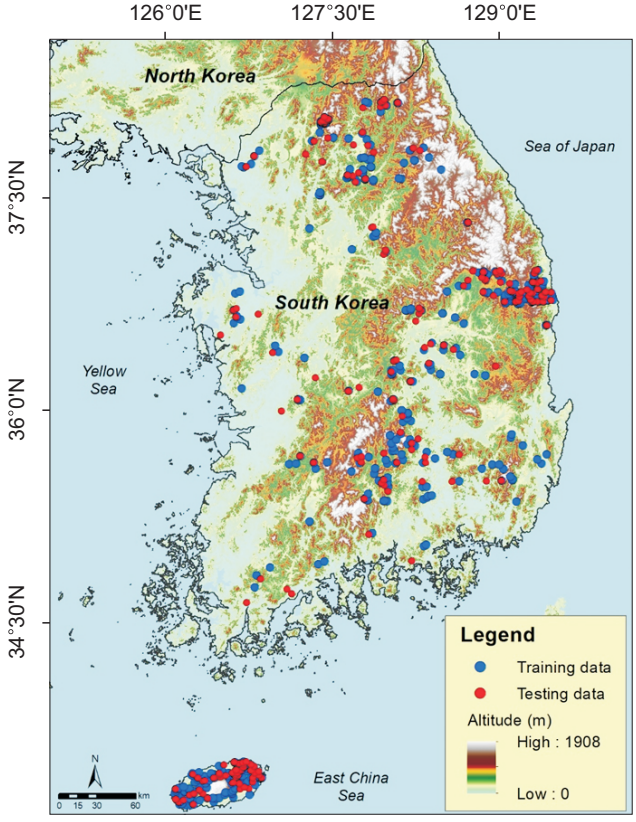

Fig. 1. Study area with roe deer habitat locations from field observations.

The longest river in South Korea is Nakdong (521 kilometers) and Han and Geum rivers comes after. These major rivers respectively flow from the north to south and the east to west, and they ended up the Yellow Sea or the Korea Strait. Their water flow fluctuates seasonally and they are broad and shallow.

South Korea climate can be classified as the humid–temperate zone with both continental and oceanic characters. The maximum and minimum annual temperature during 2010-2020 took the values between 19.1 to 17.6°C and 7.8 to 9°C, respectively. The sum of annual precipitation were in the interval 95.8 to 126.2 mm/yr, approximately 37% of which falls in the summer season, between June and Aguste (Korea meteorological administration: KMA, 2020).

Dataset preparation for spatial modeling

A full-scale investigation into the species distributions have been launched since 1997 by the Korean National Institute for Environmental Studies to monitor species populations particularly for endangered ones. The survey was carried out based on extensive field observations to identify special spreading of wildlife habitats. The observed or detected locations were documented using hand-held global positioning system. As a result, roe deer was discerned at 741 points in the study area according to experts report (Fig. 1). For the purpose of building model and validating its performance, both habitat and non-habitat locations are needed. Therefore, the same number of habitat locations (741 points) were identified as non-habitat ones. Afterwards, the yield data as the habitat and non-habitat locations were randomly divided into two groups including training and validation in the ratio of 70:30, respectively (Chen *et al*., 2019; Jaafari *et al*., 2019; Kamali Maskooni *et al*., 2020; Oh *et al*., 2019; Razavi Termeh *et al*., 2018). These points (training: validation) were selected utilizing 'Create Random Points' function in ArcGIS 10.8 software (Esri, Redlands, CA, USA). Then, 70% of the habitat and non-habitat locations were merged to build training dataset (1,038 points) which was applied for constructing model. The remaining 30% of points were amalgamated to make the testing dataset employed for model validation (444 Points) (Dodangeh *et al*., 2020; Panahi *et al*., 2020a). Finally, the attributes of each point was extracted by overlaying both training and validation datasets with habitat influential factors.

Habitat-related factors

Habitat selection is an optimization procedure which interaction between diverse environmental variables can affect habitat distributions of wild animals. Indeed, determination of contributing factors exerts undue influence on the accuracy of habitat potential mapping and there are no comprehensive guidelines for selecting appropri-

ate affecting factors. In the current study, 15 factors were opted depending upon the literature review and availability of data. Afterwards, FR model was applied to assess the correlation between roe deer habitats and the affecting factors. Finally, According to the results of FR method, the potential maps were developed based on the integration of 10 evaluated thematic maps with FR values greater than 0.50. Results implied that the highest FR value belongs to slope height (FR=0.94), followed by drainage density (0.86), altitude (0.85), road density (0.76), topographic wetness index (TWI; 0.73), valley depth (0.69), morphological feature (0.68), normalized difference water index (NDWI; 0.65), radar intensity (0.59) and topographic position index (TPI, 0.52) (Fig. 2). The mentioned maps were prepared and analyzed using ArcGIS 10.8 software. First, a digital elevation model (DEM) with a pixel size of 30×30 m was prepared using topographical maps published by the National Geographic Information Institute. Afterwards, DEM applied to derive topographic factors, namely valley depth, slope height, TPI, TWI and morphological features.

Field observations show that the roe deer population increase at higher elevations distant from urban areas and transport infrastructure. TPI involves the analyzing of topographic landscape terrain units into the upper, middle and lower parts. By declining elevation of each cell in a DEM, the TPI values decrease. The TPI value for flat ground surface located in the foot of hills or mid sloppy region is near zero (Arya *et al*., 2020). TWI is another key parameter which greatly affect animal distribution and their abundance. IWI indicates the soil moisture which impacts on vegetation patterns of an area. The habitat preferences is determined fundamentally by the existence of sufficient food in a territory.

Water availability is another decisive factor in the distribution of animal and plant species. The higher availability

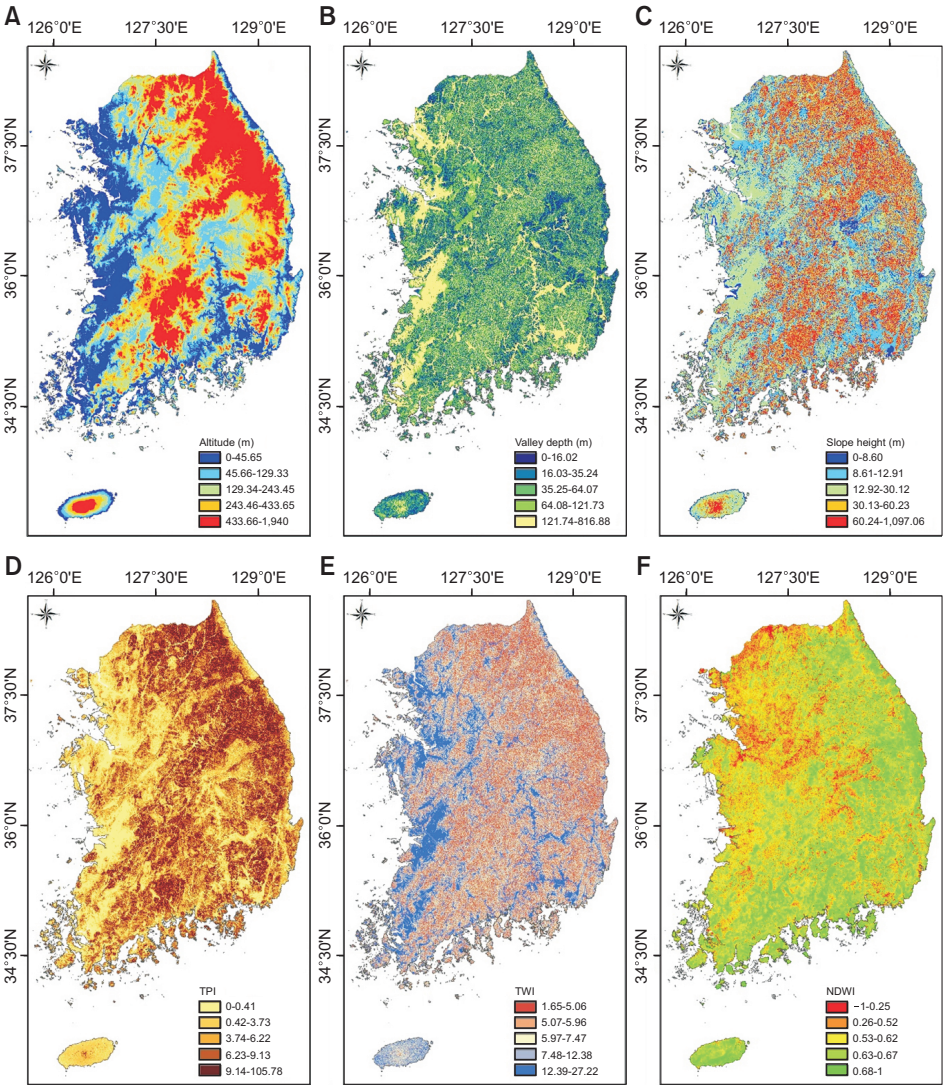

Fig. 2. Influential factors: (A) altitude, (B) valley depth, (C) slope height, (D) TPI, (E) TWI, (F) NDWI, (G) drainage density, (H) road density, (I) radar intensity, and (J) morphological feature. TPI, topographic position index; TWI, topographic wetness index; NDWI, normalized difference water index.

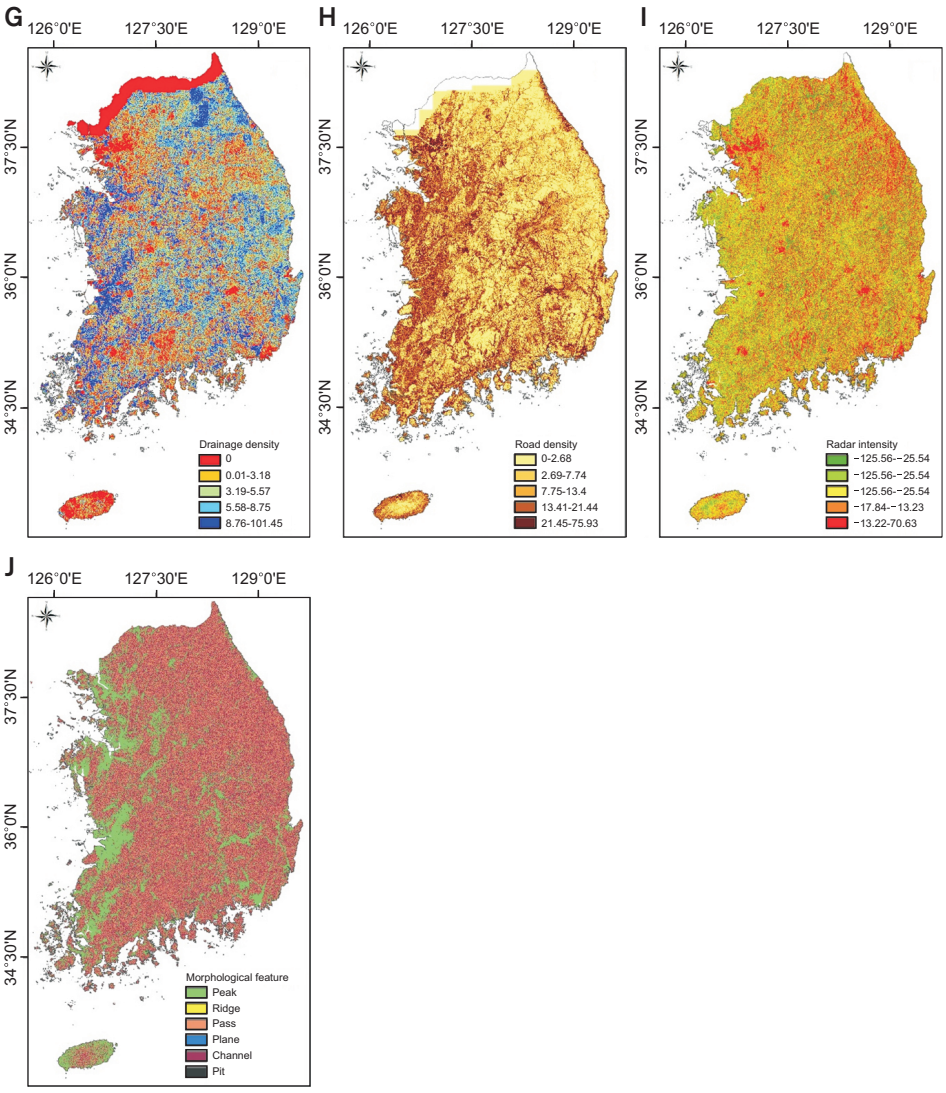

Fig. 2. Continued.

promotes plant growth and, thus, supports more food sources for roe deer (Ng *et al*., 2018). Water body has low radiation and strong absorption in the range from visible to infrared wavelengths; therefore, NDWI map provides information about water body in a region. NDWI map at the spatial resolution of 30 m was produced using the green and Near Infrared bands of Sentinel-2 images. The value of NDWI for water bodies is positive while vegetative cover or soil mass gain negative values (Biswas *et al*., 2020). The land cover map was prepared and published by the Ministry of Environment was employed to develop drainage and road density maps. Several studies clarified the effects of barriers such as road infrastructure and water surfaces on roe deer displacements and dispersal movements (Duarte *et al*., 2010; Rosell *et al*., 1996).

Landforms can be considered as an important parameter influences population distribution and patterns of habitat selection by animals. Morphometric features, which characterize the landform of a terrain, was extracted from DEM using SAGA–GIS software (http://saga-gis.org). The inventory map was categorized into plane, pit, ridge, channel, peak, and pass. The concentration of roe deer habitats is in areas with moderate altitudes and gentle slopes, such as river valleys and flat areas. They prefer to reduce their movements in areas with steep topography (ridges) which confirms the studies carried out by Acevedo *et al*. (2011), López-Martín *et al*. (2009), Loro *et al*. (2016), and Pays *et al*. (2012). In fact, intermediate and concave reliefs lead to decrease of habitat quality. Convex and convex/concave structure have positive impact on it in both summer and winter (Reimoser *et al*., 2009).

Satellite radar intensity image produced by processing Sentinel-1 (ESA operating) images of the Korean Peninsula in 2020. Each pixel in an intensity image indicates the proportion of microwave backscattering strength from that area on the ground. The radar intensity map was

utilized to determine the surface roughness and the moisture level of the area. These parameters can play a vital role in the distribution of wild animal populations (Evcin *et al*., 2019).

Methods

Description of models

To identify roe deer potential habitats, three big data analysis algorithms including FR, CNN, and LSTM were employed in the current study. Additionally, the receiver operating characteristic (ROC) curve was applied to determine the accuracy rate of prediction. The methodologies of study were clarified in the following sections and the flowchart is shown the different processing steps of adopted approaches in present study (Fig. 3).

Frequency ratio model: The FR is a type of bivariate statistical methods with the capability of deducing the potential probabilistic correlation between an incident and each related variable. This ability leads to prophesy an event rest on an assumption that the conditions will not be altered (Lee & Talib, 2005). FR of each effective factors in mapping roe deer potential habitats can be expressed as follows:

$$FR_{ij}=\frac{H_{ij}/T_H}{F_{ij}/T_f}=\frac{\%\ Habitats}{\%\ Pixels} \quad (1)$$

where FR_{ij} denotes the FR of a i-th class for the j-th factor, H_{ij} is the number of pixels contains habitat location in the i-th class of the j-th factor and T_H shows the total number of habitats. F_{ij} represents the number of pixels in the i-th class of the j-th factor and T_f indicates the total number of pixels of each factor. Finally, the habitat potential maps is prepared by summation of FR model result for each contributing factor. The value of FR method signifies the amount of correlation between the class and habitat potentiality. In other words, the higher value of FR specifies the higher probability of consideration a location as a possible habitat, and vice versa (Altafi Dadgar *et al*., 2017).

Convolutional neural network algorithm: Deep learning algorithms are one of the artificial neural networks (ANN) model which are structurally and numerically different from ANN. CNN is a class of deep learning neural network method in which at least one of its layers is allocated to the convolution operation (Tekerek, 2021). The main structure of CNN model composed of input layer, convolution layer, pooling or subsampling layer, fully connected layer, and output layers (Panahi *et al*., 2020b).

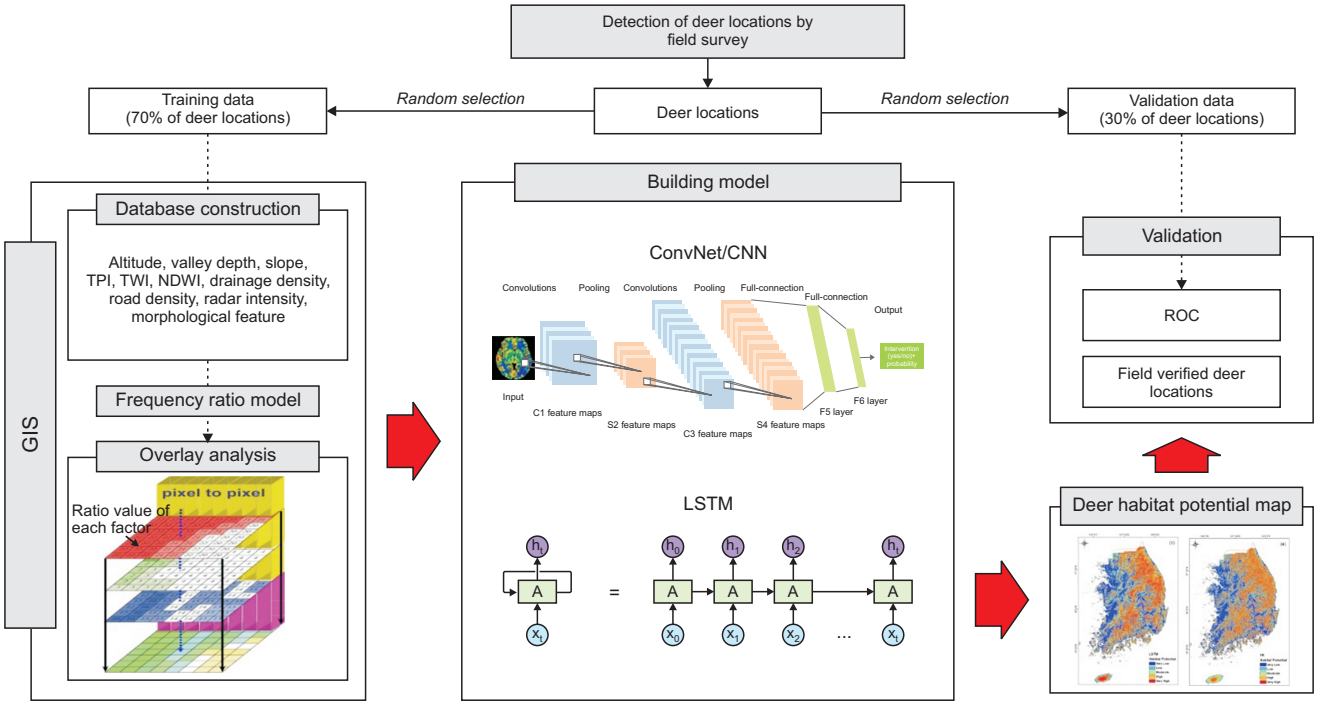

Fig. 3. Overview of applied methodology for delineation of roe deer potential habitats. GIS, geographic information system; ROC, receiver operating characteristic; CNN, convolutional neural network; LSTM, long short-term memory; TPI, topographic position index; TWI, topographic wetness index; NDWI, normalized difference water index.

The input layer is designed for converting the input data to numerical format which had been stored in a matrix. Based on the input properties, matrix property may be vary. Convolutional layer is the major feature which causes the CNN to be dissimilar to other standard neural network approaches. The sum of matrix multiplication process between kernel and input will be applied in order to extract new features in convolutional layer, and then this process will be followed by mathematical operations over the input that is called activation functions. Activation functions are necessary for introducing non-linearity into neural networks because they help to catch up non-linear features from the input data (Zhang *et al*., 2019).

Reducing the numerous parameters as well as the feature selection is being done by pooling layers. There are two common types of pooling function that are called maximum pooling and average pooling. Maximum or average pooling functions are employed to calculate maximum or average value for each patch of the feature map, respectively (Zhang *et al*., 2020b).

Fully connected layer normally comes after all convolutional and pooling layers when high-level features were extracted. This layer has the function to manage the classification settings likes prevailing feed-forward ANN (Kumar & Hati, 2020). Fully connected layer consists of all neurons that are connected to the past layer and mathematically can be defined by the following equation:

$$y=f(X_i \times a+u) \tag{2}$$

where y shows the output of the fully connected layer. X_i denotes the input vector, a is the weight, u stands for bias of the fully connected layer, and the activation function is f(.) (Tang *et al*., 2020). Softmax function is the activation function of fully connected layer and computes the probabilities of i-th target class ($\sigma(X_i)$). Activation function can be expressed as follows:

$$\sigma(z_i)=\frac{e^{Z_i}}{\sum_{i=1}^{N} e^{z_i}} \qquad i=1,2,3, \ldots, N \tag{3}$$

where z_i is the i-th output neuron and N indicates the number of possible target classes for the specific inputs (Zare & Ayati, 2020). The last fully connected layer is interpreted as an output layer in which every neuron is responsible for representing final probabilities of each class.

Long short-term memory algorithm: RNNs has various classes that one of them are LSTM which has been developed to process time-dependent variables presented in sequential data (Vu *et al*., 2020). In comparison with ordinary RNNs, in the hidden layer of LSTM there is the exception of the summation units that will be replaced by memory blocks (Wei, 2020). Each LSTM unit comprises of three gates including forget gate (f_t), input gate (i_t), and output gate (o_t) which all of them remove or add information to each cell state. Since LSTM has the ability to learn hidden long-term sequential dependencies, it becomes popular among scholars (Yang *et al*., 2017). For a defined time point t, the input of a LSTM cell is x_t as well as its previous output (h_{t-1}). $\tilde{C}_t$ and C_t represent the cell input and output state, and C_{t-1} is its previous state. Based on the LSTM architecture, C_t and h_t will be passed to the next cell in the network. C_t and h_t can be determined by the following equations (Tin *et al*., 2019):

$$f_t=\sigma(W_f.[h_{t-1},x_t]+b_f) \tag{4}$$

$$i_t=\sigma(W_i.[h_{t-1},x_t]+b_i) \tag{5}$$

$$\tilde{C}_t=tanh(W_C.[h_{t-1},x_t]+b_C) \tag{6}$$

$$C_t=f_t \odot C_{t-1}+i_t \odot \tilde{C}_t \tag{7}$$

where W, b, σ and *tanh* are weighted matrices, bias vector, sigmoid function, and hyperbolic tangent function. The forget gate has the role of reserving or forgetting information. New information will be added with proper scale by input gate. In Eqs.5 and 6, the values updates by sigmoid activation function and new candidate values generates by *tanh* function. The updated new candidate can be computed according to Eq.7 (Supreetha *et al*., 2020). A *tanh* layer, a sigmoid layer, and a pointwise multiplication operation ($\odot$) make the output gate (o_t). Finally, the output gate takes a decision which information will be output according to the cell state C_t as well as the input x_t and h_{t-1}, by using the following equations (Shi *et al*., 2021):

$$o_t=\sigma(W_o.[h_{t-1},x_t]+b_o) \tag{8}$$

$$h_t=o_t \odot tannh(C_t) \tag{9}$$

Evaluation of models

ROC curve analysis is the most common statistical method employed to assess quantitatively the prediction capability of developed model for both training and testing phase. The ROC curve plots false positive rate on the X-axis (1-specificity) versus true positive rate on the Y-axis (sensitivity) (Al-Abadi *et al*., 2017). In current research, the 1-specificity displays the portion of pixels inaccurately predicted by the presence or absence of habitat potential and the sensitivity is the portion of pixels classified accurately (Pradhan *et al*., 2020). In fact, the area under the ROC curve (AUC) demonstrates the ability of the method to predict where an event happens or not. The AUC value can vary between 0 and 1. The AUC equals to 0.5 speci-

Table 1. The effect of each influential factor subclasses on habitat distributions using FR model

Influential factors	Classes	Number of pixels	Number of habitat	FR
Altitude (m)	0-45.65	22,173,780	32	0.31
	45.66-129.33	22,198,200	34	0.33
	129.34-243.45	22,214,775	69	0.66
	243.46-433.65	22,038,609	151	1.46
	433.66-1,940	22,014,492	233	2.26
Valley depth (m)	0-16.02	20,315,890	227	2.38
	16.03-35.24	23,898,669	131	1.17
	35.25-64.07	22,805,822	80	0.75
	64.08-121.73	22,041,141	49	0.47
	121.74-816.88	21,578,334	32	0.32
Slope height (m)	0-8.60	16,874,116	31	0.39
	8.61-12.91	19,399,497	42	0.46
	12.92-30.12	38,234,399	123	0.69
	30.13-60.23	18,843,481	142	1.61
	60.24-1,097.06	17,288,363	181	2.23
TPI	0-0.41	23,970,287	37	0.33
	0.42-3.73	21,751,996	144	1.41
	3.74-6.22	21,640,415	123	1.21
	6.23-9.13	21,647,975	129	1.27
	9.14-105.78	21,629,183	86	0.85
TWI	1.65-5.06	18,901,824	159	1.79
	5.07-5.96	25,030,432	130	1.11
	5.97-7.47	22,989,581	99	0.92
	7.48-12.38	22,105,070	96	0.93
	12.39-27.22	21,612,949	35	0.35
NDWI	–1-0.25	21,374,407	19	0.19
	0.26-0.52	21,839,899	80	0.78
	0.53-0.62	19,426,220	117	1.28
	0.63-0.67	21,359,487	154	1.53
	0.68-1	26,219,650	149	1.21
Drainage density	0	25,265,819	176	1.48
	0.01-3.18	22,500,741	159	1.50
	3.19-5.57	22,665,901	82	0.77
	5.58-8.75	20,907,668	63	0.64
	8.76-101.45	18,878,276	39	0.44
Road density	0-2.68	33,123,325	230	1.45
	2.69-7.74	19,301,104	85	0.92
	7.75-13.4	18,128,869	100	1.15
	13.41-21.44	18,395,607	68	0.77
	21.45-75.93	18,936,390	33	0.36

Table 1. Continued

Influential factors	Classes	Number of pixels	Number of habitat	FR
Radar intensity	–125.56–-25.54	18,047,153	68	0.79
	–25.53–-21.7	22,669,814	123	1.14
	–21.69–-17.85	24,776,910	149	1.26
	–17.84–-13.23	21,357,501	107	1.05
	–13.22–-70.63	21,898,743	72	0.69
Morphological features	Peak	25,575,559	81	0.68
	Ridge	150,186	0	0.00
	Pass	39,090,576	135	0.74
	Plane	444,837	3	1.44
	Channel	45,271,226	299	1.41
	Pit	107,472	1	1.98

FR, frequency ratio; TPI, topographic position index; TWI, topographic wetness index; NDWI, normalized difference water index.

fies random prediction and a higher value, which is closer to one, is as an indicator of better model predictability (Chen *et al*., 2019; Nguyen *et al*., 2020). Therefore, values between 0.5-0.6 indicates low accuracy, 0.6-0.7 moderate, 0.7-0.8 high, 0.8 to 0.9 very high accuracy (Arabameri *et al*., 2019). If AUC takes a value bigger than 0.9, the model prediction accuracy is perfect. AUC is calculated by training and validation dataset separately as called success and prediction rate curves, respectively. The success rate curve illustrates the model ability to fit the observed events and prediction rate curve displays the model prediction quality.

Results

Impact of influential factor subclasses on habitat distributions

The FR method was applied to determine the effect of influential factors on roe deer habitat distribution. With regard to the results presents in Table 1, the higher elevations raised the possibility of roe deer concentration due to providing a safe habitat away from human activities, especially from hunting. In the valley depth parameter, the class of 0-16.02 and 16.03-35.24 had the most impact on choosing a place as a habitat with the FR values 2.38 and 1.17, respectively. Observation showed that the population of roe dear increased with the increase of slope height. Highest value was found in case of 60.24-1,097.06 m. The FR values had been computed for the TPI factor as a proxy for topographic landscape terrain units, showed that higher values were related to the class of 0.42-3.73 (FR=1.41) followed by 3.74-6.22 and 6.23-9.13 classes (FR=1.21 and 1.27, respectively). For TWI, the estimated FR values decreased with increase this index. The class of 1.65-5.06 had the highest FR value (1.79), followed by the class of 5.07-5.96 (1.11). The spatial analysis revealed that NDWI class of 0.63-0.67 had the higher FR value (FR=1.92) and significantly affected roe deer habitat distribution. The highest weighted class was observed in drainage density, the second class with a FR value 1.50, followed by the road density class of 0-2.68 (FR=1.45) and the third class of the variable radar intensity (FR=1.26). Greater road density correlated with low habitat suitability for roe deer species. In case of morphological feature, roe deer tend to live in regions located in a convex/concave land formation.

Habitat potential maps

Roe deer habitat potential map based on FR model was generated by the summation of FR values of each influential factor and their subclasses using ArcGIS 10.8 through the following equation:

$$Potential\ Habitat\ Areas=\sum FR_i \qquad i=1,2,3,...,N \quad (10)$$

where FR_i is the FR value of each factor's class and N denotes the total number of influential factors. As shown in Fig. 4, the resultant map was classified into five classes using quantile method including very low, low, moderate, high and very high habitat potential zones, covering about 19.71%, 19.83%, 20.17%, 20.11%, and 20.19% of the study area, respectively. The two other maps were prepared by applying CNN and LSTM models. The CNN model predicted that approximately 19.87%, 19.85%, 19.25%, 19.47%, and 21.57% areas were designated to the very low, low, moderate, high and very high habitat potential zones whereas theses values for the LSTM model were 19.70%, 19.81%, 19.31%, 19.86%, and 21.31%, respectively. The models prediction accuracy were concluded by using AUC analysis. During the training phase the AUC of FR, CNN, and LSTM models were 0.74, 0.73 and 0.76,

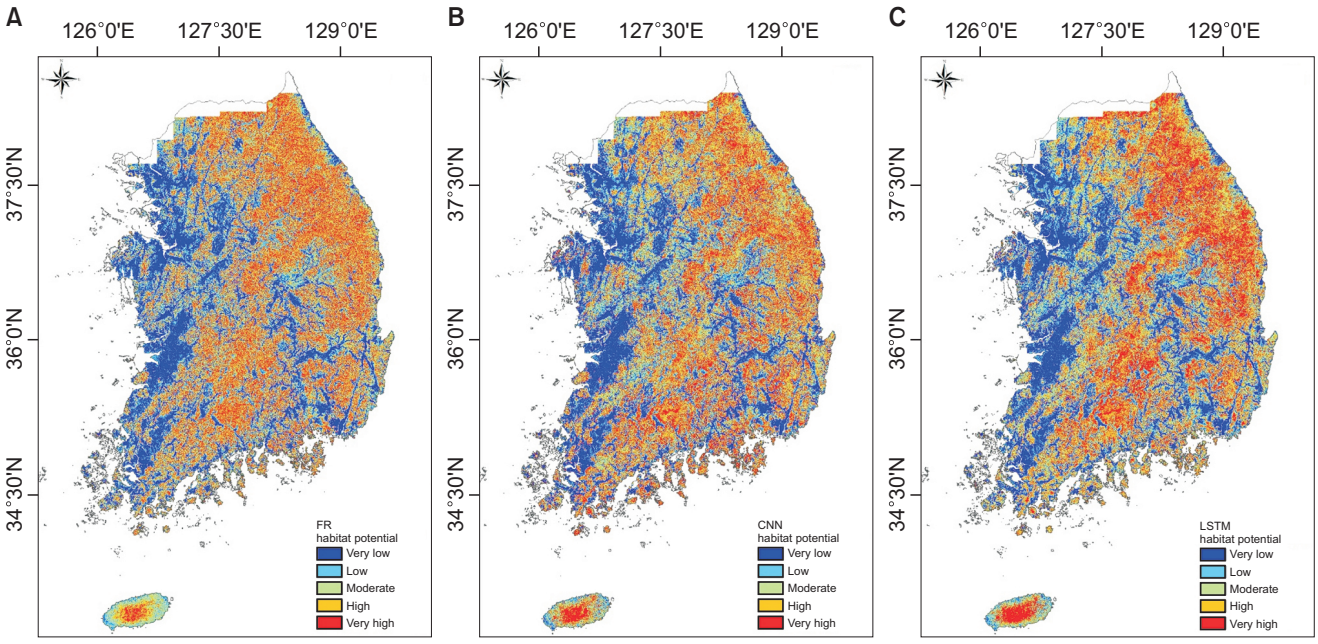

Fig. 4. Habitat potential maps developed using (A) FR, (B) CNN, (C) LSTM models. FR, frequency ratio; CNN, convolutional neural network; LSTM, long short-term memory.

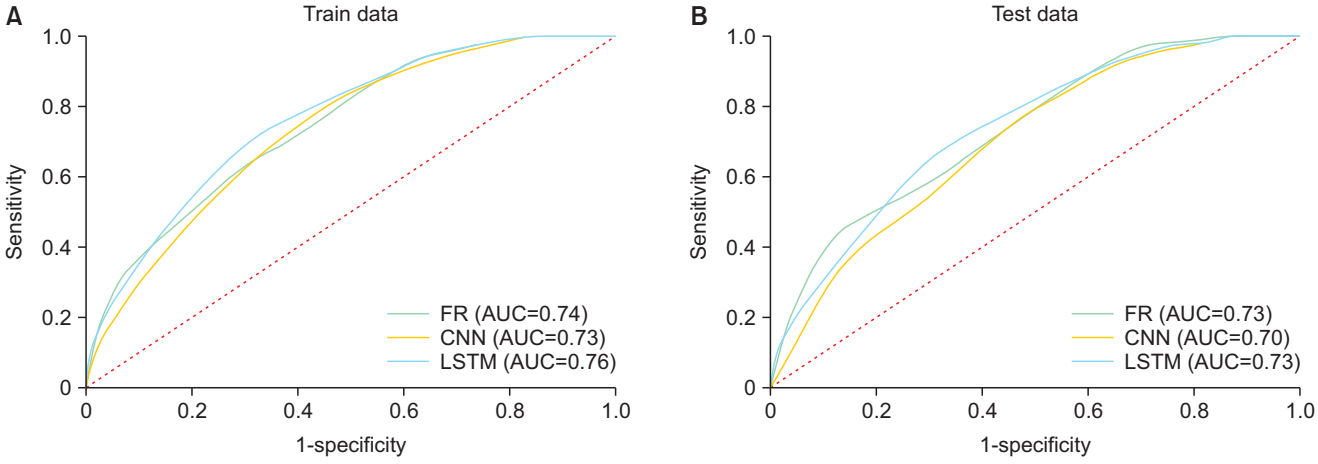

Fig. 5. Comparison of models prediction power using AUC (A) training phase, (B) testing phase. AUC, area under the receiver operating characteristic curve; FR, frequency ratio; CNN, convolutional neural network; LSTM, long short-term memory.

respectively (Fig. 5). Moreover, The FR, CNN, and LSTM models had AUC of 0.73, 0.70 and 0.73 during the testing phase.

Discussion

The present study was executed to assess the ability of deep learning algorithms (CNN and LSTM) to identify the spatial relationships between the possible roe deer habitat and related environmental factors for the first time. In fact, the prediction ability of deep learning approaches directly depends on input data accuracy and imprecise habitat location data can lead to inaccurate prediction. Concerning the performance evaluation, LSTM model had relatively higher prediction ability in comparison to FR and CNN models according to its AUC values during training and validation process (AUC=0.76% and 0.73%). Percentage of roe deer habitat locations within each class of habitat potential map (Fig. 6) display that the most numbers of locations fell into very high class; therefore, models have shown acceptable performances.

The results of variable importance analysis of influential

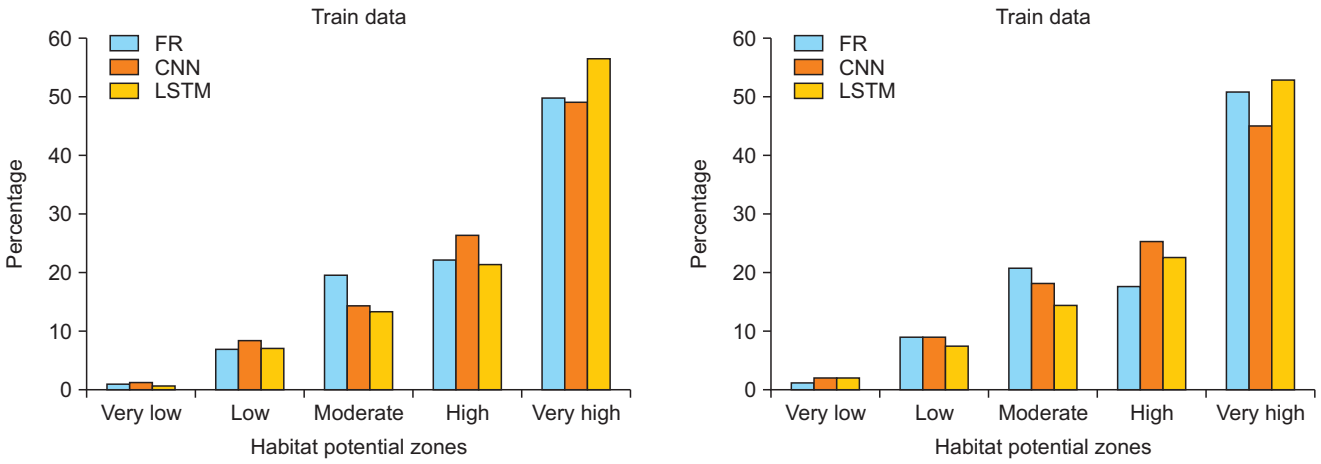

Fig. 6. Percentage of roe deer habitat locations within each class of habitat potential map. FR, frequency ratio; CNN, convolutional neural network; LSTM, long short-term memory.

factors denoted that TPI, TWI, altitude and valley depth had the greatest contribution in the modelling procedure with a mean effective weight of 63.3%, 58.1%, 57.8%, and 54.8%, respectively. On the other contrary, morphological feature, drainage density and radar intensity had lowest impact on habitat potential mapping in descending order. Moreover, the final maps reveal that there are slight differences between stability and robustness of developed models and two algorithms are a class of same family. Additionally, not only the values of AUC were close but also the distribution of the potential classes had remarkable similar pattern. The northeastern and central parts of the inland area of South Korea and the central part of Jeju Island have very high potential. However, LSTM model performance was superior to CNN and FR model. It should be mentioned that although FR model is easy and the result can be interpreted simply but has limitation in assessing relationships between an event occurrence and effective factors because of defining statistical assumptions prior to any study in comparison with machine learning methods. Finally, the resulting habitat potential map can make an outstanding contribution to define effective conservation plans and manage the roe deer habitats.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgments

This research was supported by the Basic Research Project of the Korea Institute of Geoscience and Mineral Resources (KIGAM) and Project of Environmental Business Big Data Platform and Center Construction funded by the Ministry of Science and ICT.

References

Acevedo, P., Real, R., and Gortázar, C. (2011). [Favorabilidad ecogeográfica para el corzo: distribución y abundancia]. *Pirineos*, 166, 9-27. Spanish.

Aini, S., Sood, A.M., and Saaban, S. (2015). Analysing elephant habitat parameters using GIS, remote sensing and analytic hierarchy process in Peninsular Malaysia. *Pertanika Journal of Science and Technology*, 23, 37-50.

Al-Abadi, A.M., Al-Temmeme, A.A., and Al-Ghanimy, M.A. (2016). A GIS-based combining of frequency ratio and index of entropy approaches for mapping groundwater availability zones at Badra-Al Al-Gharbi-Teeb areas, Iraq. *Sustainable Water Resources Management*, 2, 265-283.

Al-Abadi, A.M., Pourghasemi, H.R., Shahid, S., and Ghalib, H.B. (2017). Spatial mapping of groundwater potential using entropy weighted linear aggregate novel approach and GIS. *Arabian Journal for Science and Engineering*, 42, 1185-1199.

Altafi Dadgar, M., Zeaieanfirouzabadi, P., Dashti, M., and Porhemmat, R. (2017). Extracting of prospective groundwater potential zones using remote sensing data, GIS, and a probabilistic approach in Bojnourd basin, NE of Iran. *Arabian Journal of Geosciences*, 10, 114.

Arabameri, A., Rezaei, K., Cerda, A., Lombardo, L., and Rodrigo-Comino, J. (2019). GIS-based groundwater potential mapping in Shahroud plain, Iran. A comparison among statistical (bivariate and multivariate), data mining and MCDM approaches. *Science of the Total Environment*, 658, 160-177.

Arya, S., Subramani, T., and Karunanidhi, D. (2020). Delineation of groundwater potential zones and recommendation of artificial recharge structures for augmentation of groundwater resources in Vattamalaikarai Basin, South India. *Environmental Earth Sciences*, 79, 102.

Biswas, S., Mukhopadhyay, B.P., and Bera, A. (2020). Delineating groundwater potential zones of agriculture dominated landscapes using GIS based AHP techniques: a case study from

Uttar Dinajpur district, West Bengal. *Environmental Earth Sciences*, 79, 302.
Chen, W., Panahi, M., Khosravi, K., Pourghasemi, H.R., Rezaie, F., and Parvinnezhad, D. (2019). Spatial prediction of groundwater potentiality using ANFIS ensembled with teaching-learning-based and biogeography-based optimization. *Journal of Hydrology*, 572, 435-448.
Choi, J.K., Oh, H.J., Koo, B.J., Ryu, J.H., and Lee, S. (2011a). Crustacean habitat potential mapping in a tidal flat using remote sensing and GIS. *Ecological Modelling*, 222, 1522-1533.
Choi, J.K., Oh, H.J., Koo, B.J., Ryu, J.H., and Lee, S. (2011b). Spatial polychaeta habitat potential mapping using probabilistic models. Estuarine, *Coastal and Shelf Science*, 93, 98-105.
Cui, X., Liu, H., Fan, M., Ai, B., Ma, D., and Yang, F. (2021). Seafloor habitat mapping using multibeam bathymetric and backscatter intensity multi-features SVM classification framework. *Applied Acoustics*, 174, 107728.
Danilkin, A., and Hewison, A.J.M. (1996). *Behavioural Ecology of Siberian and European Roe Deer*, London: Chapman and Hall.
Dodangeh, E., Panahi, M., Rezaie, F., Lee, S., Bui, D.T., Lee, C.W., *et al*. (2020). Novel hybrid intelligence models for flood-susceptibility prediction: meta optimization of the GMDH and SVR models with the genetic algorithm and harmony search. *Journal of Hydrology*, 590, 125423.
Duarte, J., Farfán, M.A., and Vargas, J.M. (2010). [Selección primaveral de hábitat del corzo andaluz (Capreolus capreolus) en un borde de su área de distribución]. *Ecología*, 23, 177-192. Spanish.
Evcin, O., Kucuk, O., and Akturk, E. (2019). Habitat suitability model with maximum entropy approach for European roe deer (Capreolus capreolus) in the Black Sea Region. *Environmental Monitoring and Assessment*, 191, 669.
Farrell, A., Wang, G., Rush, S.A., Martin, J.A., Belant, J.L., Butler, A.B., *et al*. (2019). Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. *Ecology and Evolution*, 9, 5938-5949.
Garzón, M.B., Blazek, R., Neteler, M., De Dios, R.S., Ollero, H.S., and Furlanello, C. (2006). Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. *Ecological Modelling*, 197, 383-393.
Imam, E., and Kushwaha, S.P.S. (2013). Habitat suitability modelling for Gaur (Bos gaurus) using multiple logistic regression, remote sensing and GIS. *Journal of Applied Animal Research*, 41, 189-199.
Imam, E., and Tesfamichael, G.Y. (2013). Use of remote sensing, GIS and analytical hierarchy process (AHP) in wildlife habitat suitability analysis. *Journal of Materials and Environmental Science*, 4, 460-467.
Jaafari, A., Panahi, M., Pham, B.T., Shahabi, H., Bui, D.T., Rezaie, F., *et al*. (2019). Meta optimization of an adaptive neuro-fuzzy inference system with grey wolf optimizer and biogeography-based optimization algorithms for spatial prediction of landslide susceptibility. *CATENA*, 175, 430-445.
Jiang, G., Qi, J., Wang, G., Shi, Q., Darman, Y., Hebblewhite, M., *et al*. (2015). New hope for the survival of the Amur leopard in China. *Scientific Reports*, 5, 15475.
Kadirhodjaev, A., Rezaie, F., Lee, M.J., and Lee, S. (2020). Landslide susceptibility assessment using an optimized group method of data handling model. *ISPRS International Journal of Geo-Information*, 9, 566.
Kamali Maskooni, E., Naghibi, S.A., Hashemi, H., and Berndtsson, R. (2020). Application of advanced machine learning algorithms to assess groundwater potential using remote sensing-derived data. *Remote Sensing*, 12, 2742.
Koh, H.S., and Randi, E. (2001). Genetic distinction of roe deer {Capreolus pygargus Pallas) sampled in Korea. *Mammalian Biology*, 66, 371-375.
Korea Meteorological Administration (KMA). (2020). *Article title*. Retrieved December 11, 2020 from https://www.weather.go.kr/weather/main.jsp.
Kosicki, J.Z. (2020). Generalised Additive Models and Random Forest approach as effective methods for predictive species density and functional species richness. *Environmental and Ecological Statistics*, 27, 273-292.
Kumar, P., and Hati, A.S. (2020). Deep convolutional neural network based on adaptive gradient optimizer for fault detection in SCIM. *ISA Transactions*. doi:10.1016/2020.10.052.
Leathwick, J.R., Rowe, D., Richardson, J., Elith, J., and Hastie, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, 50, 2034-2052.
Lee, S., and Talib, J.A. (2005). Probabilistic landslide susceptibility and factor effect analysis. *Environmental Geology*, 47, 982-990.
Lee, S., Lee, S., Song, W., and Lee, M.J. (2017). Habitat potential mapping of marten (Martes flavigula) and leopard cat (Prionailurus bengalensis) in South Korea using artificial neural network machine learning. *Applied Sciences*, 7, 912.
Lee, S., Park, I., Koo, B.J., Ryu, J.H., Choi, J.K., and Woo, H.J. (2013). Macrobenthos habitat potential mapping using GIS-based artificial neural network models. *Marine Pollution Bulletin*, 67, 177-186.
Lee, S., Syifa, M., Koo, B.J., Lee, C.W., and Oh, H.J. (2019). Spatial macrobenthos habitat on Ganghwa tidal flat, Korea: Part II- habitat potential mapping of Potamocorbula laevis using probability models. *Journal of Coastal Research*, 90(SI), 401-408.
Lee, Y.S., Markov, N., Argunov, A., Voloshina, I., Bayarlkhagva, D., Kim, B.J., *et al*. (2016). Genetic diversity and phylogeography of Siberian roe deer, *Caproulus pygargus*, in central and peripheral populations. *Ecology and Evolution*, 6, 7286-7297.
Lee, Y.S., Markov, N., Voloshina, I., Argunov, A., Bayarlkhagva, D., Oh, J.G., *et al*. (2015). Genetic diversity and genetic structure of the Siberian roe deer (Capreolus pygargus) populations from Asia. *BMC Genetics*, 16, 100.
Loro, M., Ortega, E., Arce, R.M., and Geneletti, D. (2016). Assessing landscape resistance to roe deer dispersal using fuzzy set theory and multicriteria analysis: a case study in Central Spain. *Landscape and Ecological Engineering*, 12, 41-60.
Lovari, S., Masseti, M., and Lorenzini, R. (2016). *Capreolus pygargus*. Retrieved December 14, 2020 from https://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T42396A22161884.en.
López-Martín, J.M., Martínez-Martínez, D., and Such, A. (2009). Supervivencia, dispersión y selección de recursos de corzos

Capreolus capreolus (Linnaeus, 1758) reintroducidos en un hábitat mediterráneo. *Galemys*, 21, 143-164.

Mesfin, Y., and Berhan, G. (2016). Geospatial approach for Grevy's zebra suitable habitat analysis, Allidegi wildlife reserve, Ethiopia. International *Journal of Applied Remote Sensing and GIS*, 3, 34-42.

Ng, W.T., Cândido de Oliveira Silva, A., Rima, P., Atzberger, C., and Immitzer, M. (2018). Ensemble approach for potential habitat mapping of invasive Prosopis spp. in Turkana, Kenya. *Ecology and Evolution*, 8, 11921-11931.

Nguyen, P.T., Ha, D.H., Avand, M., Jaafari, A., Nguyen, H.D., Al-Ansari, N., *et al*. (2020). Soft computing ensemble models based on logistic regression for groundwater potential mapping. *Applied Sciences*, 10, 2469.

Oh, H.J., Syifa, M., Lee, C.W., and Lee, S. (2019). Ruditapes philippinarum habitat mapping potential using SVM and Naïve Bayes. *Journal of Coastal Research*, 90(sp1), 41-48.

Panahi, M., Gayen, A., Pourghasemi, H.R., Rezaie, F., and Lee, S. (2020a). Spatial prediction of landslide susceptibility using hybrid support vector regression (SVR) and the adaptive neuro-fuzzy inference system (ANFIS) with various metaheuristic algorithms. *Science of the Total Environment*, 741, 139937.

Panahi, M., Sadhasivam, N., Pourghasemi, H.R., Rezaie, F., and Lee, S. (2020b). Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR). *Journal of Hydrology*, 588, 125033.

Park, J., Kim, D.S., Song, K.H., Jeong, T.J., and Park, S.J. (2018). Mapping potential habitats for the management of exportable insects in South Korea. *Journal of Asia-Pacific Biodiversity*, 11, 11-20.

Park, Y.S., Kim, B.J., Lee, W.S., Kim, J.T., Kim, T.W., and Oh, H.S. (2014). Molecular phylogenetic status of Siberian roe deer (Capreolus pygargus) based on mitochondrial cytochrome b from Jeju Island in Korea. *Chinese Science Bulletin*, 59, 4283-4288.

Pays, O., Fortin, D., Gassani, J., and Duchesne, J. (2012). Group dynamics and landscape features constrain the exploration of herds in fusion-fission societies: the case of European roe deer. *PloS One*, 7, e34678.

Pereira, J.M.C., and Itami, R.M. (1991). GIs-based habitat modeling using logistic multiple regression: a study of the Mt. Graham red squirrel. *Photogrammetric Engneering and Remote Sensing*, 57, 1475-1486.

Pradhan, A., Kim, Y.T., Shrestha, S., Huynh, T.C., and Nguyen, B.P. (2020). Application of deep neural network to capture groundwater potential zone in mountainous terrain, Nepal Himalaya. *Environmental Science and Pollution Research International*. doi:10.1007/11356-020-10646-x.

Rahimian Boogar, A., Salehi, H., Pourghasemi, H.R., and Blaschke, T. (2019). Predicting habitat suitability and conserving Juniperus spp. habitat using SVM and maximum entropy machine learning techniques. *Water*, 11, 2049.

Razavi Termeh, S.V., Kornejady, A., Pourghasemi, H.R., and Keesstra, S. (2018). Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. *Science of the Total Environment*, 615, 438-451.

Reimoser, S., Partl, E., Reimoser, F., and Vospernik, S. (2009). Roe-deer habitat suitability and predisposition of forest to browsing damage in its dependence on forest growth- model sensitivity in an alpine forest region. *Ecological Modelling*, 220, 2231-2243.

Robert, K., Jones, D.O.B., Roberts, J.M., and Huvenne, V.A.I. (2016). Improving predictive mapping of deep-water habitats: considering multiple model outputs and ensemble techniques. *Deep Sea Research Part I: Oceanographic Research Papers*, 113, 80-89.

Rosell, C., Carretero, M.A., Cahill, S., and Pasquina, A. (1996). Seguimiento de una reintroducción de corzo (Capreolus capreolus) en ambiente mediterráneo. Dispersión y área de campeo. *Donana Acta Vertebrata*, 23, 109-122. Spanish.

Sanare, J.E., Ganawa, E.S., and Abdelrahim, A.M.S. (2015). Wildlife habitat suitability analysis at Serengeti National Park (SNP), Tanzania case study Loxodonta sp. *Journal of Ecosystem and Ecography*, 5, 164.

Sanchez, P., Demestre, M., Recasens, L., Maynou, F., and Martin, P. (2008). Combining GIS and GAMs to identify potential habitats of squid Loligo vulgaris in the Northwestern Mediterranean. In V.D., Valavanis (Eds.), *Essential Fish Habitat Mapping in the Mediterranean* (pp. 91-98). Dordrecht: Springer.

Schmiing, M., Afonso, P., Tempera, F., and Santos, R.S. (2013). Predictive habitat modelling of reef fishes with contrasting trophic ecologies. *Marine Ecology Progress Series*, 474, 201-216.

Shi, Y., Song, X., and Song, G. (2021). Productivity prediction of a multilateral-well geothermal system based on a long short-term memory and multi-layer perceptron combinational neural network. *Applied Energy*, 282, 116046.

Sokolov, V.E., Danilkin, A.A., and Dulamtseren, S. (1982). Contemporary distribution and populations of the forest ungulates in Mongolia. *Zoological Studies in the People's Republic of Mongolia*, 8, 37-56.

Supreetha, B.S., Shenoy, N., and Nayak, P. (2020). Lion algorithm-optimized long short-term memory network for groundwater level forecasting in Udupi District, India. *Applied Computational Intelligence and Soft Computing*, 2020, 8685724.

Tang, J., Su, Q., Su, B., Fong, S., Cao, W., and Gong, X. (2020). Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition. *Computer Methods and Programs in Biomedicine*, 197, 105622.

Tekerek, A. (2021). A novel architecture for web-based attack detection using convolutional neural network. *Computers and Security*, 100, 102096.

Tin, T.C., Chiew, K.L., Phang, S.C., Sze, S.N., and Tan, P.S. (2019). Incoming Work-In-Progress prediction in semiconductor fabrication foundry using long short-term memory. *Computational Intelligence and Neuroscience*, 2019, 8729367.

Vu, M.T., Jardani, A., Massei, N., and Fournier, M. (2020). Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network. *Journal of Hydrology*. doi:10.1016/2020.125776.

Wei, C.C. (2020). Development of stacked long short-term memory neural networks with numerical solutions for wind velocity predictions. *Advances in Meteorology*, 2020, 5462040.

Yang, H., Pan, Z., and Tao, Q. (2017). Robust and adaptive

online time series prediction with long short-term memory. *Computational Intelligence and Neuroscience*, 2017, 9478952.

Zare, S., and Ayati, M. (2020). Simultaneous fault diagnosis of wind turbine using multichannel convolutional neural networks. *ISA transactions*. doi:10.1016/2020.08.021.

Zhang, G., Zhu, A.X., He, Y.C., Huang, Z.P., Ren, G.P., and Xiao, W. (2020a). Integrating multi-source data for wildlife habitat mapping: a case study of the black-and-white snub-nosed monkey (Rhinopithecus bieti) in Yunnan, China. *Ecological Indicators*, 118, 106735.

Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y., and Yu, B. (2019). Recent advances in convolutional neural network acceleration. *Neurocomputing*, 323, 37-51.

Zhang, X., Wu, F., and Li, Z. (2020b). Application of convolutional neural network to traditional data. *Expert Systems with Applications*. doi:10.1016/2020.114185.