Genetic Programming for Amphibian Presence Prediction to Support Sustainable Land-Use Planning

Arvind Kumar[1*], Naveen Kumar Bonagiri[2]

[1*]Industrial Autonomy & Engineering Unit, Tata Consultancy Services Ltd, Noida, India.
mail: arvind.jki@gmail.com; Orcid-Id: 0000-0002-3808-7338
[2]Birla Institute of Technology and Science, Pilani, India.
E-mail: naveenkumar.bonagiri@gmail.com; Orcid-Id: 0009-0002-1726-4312

Abstract
The application of artificial intelligence-based techniques in spatial planning provides a low-cost and time-saving solution to extensive ecological field surveys. One spatial task is to predict the presence of various amphibian species while planning roads and bridges. At the place of fieldwork, we may use satellite imagery data to extract various features and then use these features to predict the presence of amphibians. Amphibians data set, available on the UCI repository, contains the features, created from satellite imagery and information collected from the inventories of two road projects in Poland. The objective is to forecast the presence of amphibians by utilizing these features. This work proposes a genetic programming-based approach for predicting amphibian species based on features generated from satellite imagery data. Experimental evaluation shows that the proposed method achieves better performance than SVM, standard GP, and existing literature across key metrics such as accuracy, sensitivity, specificity, and AUC. Furthermore, accurate amphibian-presence prediction can assist agricultural planners and environmental authorities in minimizing ecological disturbances during rural infrastructure development, contributing to sustainable land-use planning and biodiversity conservation in agricultural landscapes.

Keywords: Amphibian Presence Prediction, Genetic Programming (GP), Fitness Function, Imbalanced Classification, GIS, Agro-Ecological Planning, Species Distribution Modeling (SDM).

1. Introduction

Artificial intelligence (AI) facilitates better decision-making through data modelling across various fields. Geographic information systems (GIS) also utilize this facility by applying various AI-based modelling techniques. The applications of GIS-based AI techniques include identifying the biodiversity hot-spots [1], presence of various species [2], crop yields through precision agriculture [3], land or forest cover classification [4], to support crime-prevention efforts [5], classification of various objects on the map [6], flood [7]. In general, the ultimate goal of applying AI in GIS is that machines eventually figure out the factors or features without expert assistance by analyzing and processing the data. The intersection of AI and GIS is creating massive opportunities and facilitates spatial decision support [8]. These facilities save not only time but also resources and costs associated with various field activities. Amphibians play an important role in agricultural ecosystems by regulating insect populations, serving as bio-indicators of water quality, and contributing to ecological stability in crop landscapes. Changes in land use, irrigation practices, and rural infrastructure development often disrupt amphibian habitats, posing challenges for sustainable agriculture. Therefore, integrating species-distribution modelling into agricultural land-use planning is essential for maintaining biodiversity and ecosystem services [9]. By predicting amphibian presence using satellite-based features, agricultural planners can avoid or mitigate ecological disruptions when constructing canals, farm roads, irrigation structures, or rural connectivity projects. The synergy between AI-driven ecological prediction and agricultural planning aligns with the broader objective of sustainable land management [10]. In real-world scenarios, environmental impact assessment is required during the planning or construction of roads and bridges. Identification of amphibian breeding sites and their migration routes are also important while planning [11, 12]. However, doing fieldwork is time consuming and requires substantial resources and associated cost [13]. A near-optimal solution is to take satellite imagery data and apply AI-based techniques to extract features from this imagery. These extracted features may be used for modelling the presence of amphibians. Fig-1 gives a high-level overview of satellite data-based modelling approach [2]. Various species can be predicted by applying AI and GIS-based techniques. Such spatial-data-driven planning plays a crucial role in today's world [14].

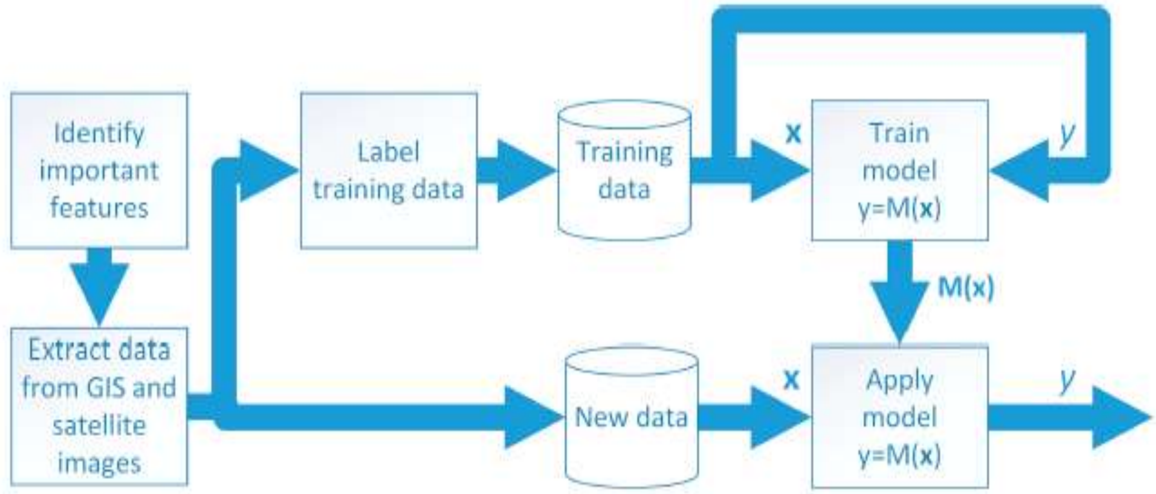

Fig. 1. GIS and satellite data-based ML modelling

Various algorithms like support vector machine (SVM) or decision trees (DT) or random forest (RF), or artificial neural networks (ANN), and evolutionary techniques are already used in AI-based GIS applications. Genetic programming is inspired by Darwin's theory of the 'survival of fittest', in which better individuals are generated, generation by generation [15]. In each generation, three nature-inspired operators, Reproduction, Crossover, and Mutation, are applied. All individuals are evaluated based on a mathematical criterion called 'fitness function', and the best individuals are selected based on their fitness. These individuals are called 'Program', and they represent a possible solution in a given problem context. GP is already utilized in many classification work in literature [16–18]. GP has inbuilt nature of feature selection, and therefore, it is also suitable for high dimensional data [19].

The amphibian data set is available publicly on the UCI repository. This data set has extracted features from satellite imagery. In this work, we use this amphibian data set to predict the existence of amphibians. The amphibian data set contains details about seven categories of amphibians. This multi-class classification is converted to seven binary-class classification problems by selecting one-versus-all. This one-versus-all conversion makes this classification problem an imbalanced-data classification problem. A dataset is termed imbalanced when it contains lessor samples from one class (called minority class) and higher samples from another class (called majority class) [20, 21].

Machine Learning (ML) algorithms produce the biased classification result when the data set is imbalanced [22–25]. The reason behind this is that the various algorithms focus on overall accuracy while learning. The uneven distribution of samples toward different classes is not given importance in the standard ML algorithm. As a solution, we proposed GP with the Euclidean distance and weight-based (EDWB) fitness function. We compared the results with original work [2] and the standard fitness-function-based GP. The result proves the superiority of the proposed approach. The proposed work incorporates predictive modelling of sensitive species into agricultural decision-support systems can significantly enhance environmentally responsible planning practices. Given the increasing overlap between agricultural expansion and semi-natural habitats, predictive modelling of amphibian presence provides valuable insights for environmentally responsible planning.

Section-II presents details about the amphibian data set used in this paper followed by section-III about GP. Section-IV provides results and discussion, and section-V gives conclusions.

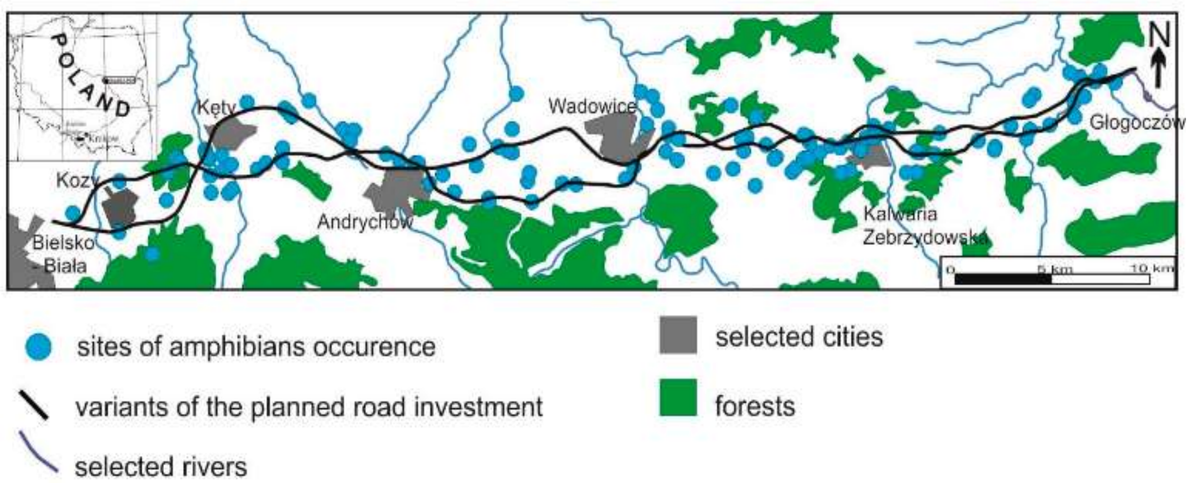

Fig. 2. Amphibian occurrence sites

2. Data set description

We used the amphibian data set in this work. This data set is available publicly at the UCI repository [26]. The amphibian data set contains the features extracted from satellite imagery. This data is accrued with the details collected from different wild inventories produced as a subset of environmental impact assessment (EIA) records. These records were created for the two road projects, Road-A and Road-B, located in Poland Fig-2.

Table 1: Data set description

Type of Amphibian	Count
Green Frogs	109
Brown Frogs	150
Common Toad	127
Fire-bellied Toad	62
Tree Frog	76
Common Newt	64
Great-crested Newt	28
Total Count	616

Following features are extracted from satellite imagery data:

1. Shore types (Categorical)
2. Maintenance status of the reservoir (Categorical)
3. The presence of fishing (Categorical)
4. Utilization of water reservoirs (Categorical)
5. Cover types of land enclosing the WRs (Categorical)
6. Presence of vegetation within WRs (Categorical)
7. Water reservoir types (Categorical)
8. Lowest distance to buildings (Ordinal)
9. Distance of routes from Water Reservoirs (Ordinal)
10. Percentage access to non-developed regions from the boundaries of WRs (Numerical)
11. Number of WRs (Numerical)
12. Surface of WRs (Numerical)

The number of amphibian species differs among groups. Table 1 provides a summary of these species counts. Seven amphibian species exist in this data set. Therefore, seven binary classification problems are created based on this data set. Although the dataset originates from non-agricultural road projects, many of the environmental variables such as vegetation cover, water-body characteristics, and landscape structure closely resemble those found in agricultural watersheds, making the findings transferable to rural land-use planning.

3. Genetic Programming

Genetic programming is an ML technique inspired by the natural evolution of species [15]. In nature, better and better species are generated based on their survival of the fittest. Thus, in each generation, all individuals are evaluated by their capabilities of survival in that environment. While simulating this behavior, this evaluation is done with a mathematical formula called fitness function. Based on this evaluation, individuals are selected for three processes called reproduction, crossover, and mutation. A higher fitness value indicates a higher probability of selecting any individual for these operations. In reproduction, an individual creates its copies for the next generation. In the crossover, two individuals are selected, and two children are created by swapping genes. Finally, in mutation, some random genes are randomly changed to create a new individual.

All three operations are similar to nature, in which there is potential to create new individuals generation after generation. However, it is possible that some individuals are generated, which is not good for a given problem context. In this case, those individuals' fitness value will be lower, so there is a very low chance that these individuals will go to the next generation. Thus, GP is a population-based search and optimization technique in which increasingly adapted species are generated, generation by generation. This GP framework may be summarized in Fig-3 [24]. A summary of the GP framework is given in the next subsections.

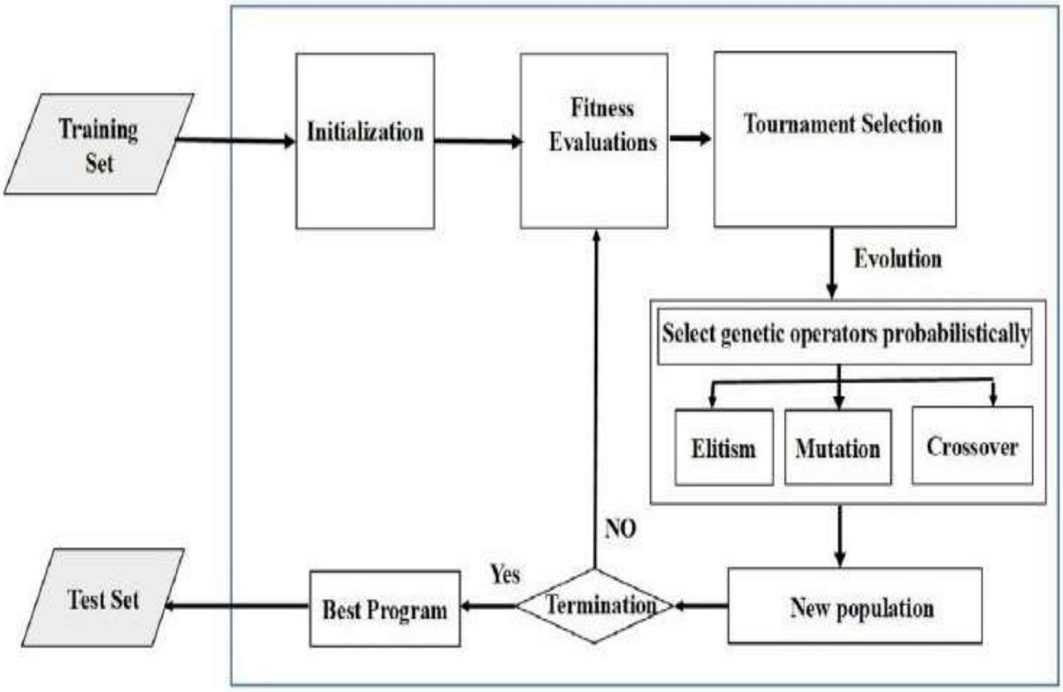

Fig. 3. GP framework

3.1 Initial Population

In the GP framework, each individual is represented as a tree. This tree is generated by randomly choosing elements of a predefined set of functions and terminals. The internal nodes are chosen from functions, while terminals make the leaf nodes. A sample GP tree is shown in Fig-4 [27]. Initialization is a very important step in any evolutionary algorithm. A good initial population can lead any GP algorithm to perform efficiently, whereas poor initialization may lead to local optima problems.

There are three popular ways to initialize the population in GP. In the Full method, the complete trees are constructed up to the defined depth using only functions. After that, only terminals are added as leaf nodes of trees. In the *Grow* method, maximum depth trees are constructed by arbitrary selection of nodes from functions and terminals. In the *Ramped half and half* method, the full method is used to initialize half of the population and the remaining with grow method to generate diversity in terms of sizes and shapes of trees in the population set [15].

3.2 Reproduction

A genetic operator is used in GP to direct the iterative process towards a solution. New individual programs are created in the population by applying genetic operators. For example, in the reproduction operator, a solution is selected and instantly copied to the new generation without being modified or changed.

3.3 Crossover

This operator randomly selects two subtrees from the parents and swaps them to create a new offspring. In this way, this crossover operator chooses genes from parent chromosomes and constructs a new solution. The easiest method to perform crossover is to select random crossover points for parents, and the first half is retrieved from the first parent, and the rest is copied in the child from the second parent.

3.4 Mutation

In mutation operation, parts of one individual are altered randomly. This random alteration brings diversity to the population and lowers the optimum local convergence. Mutation can be performed in three ways. Point mutation starts by selecting a node in the parent tree and replacing it with the same type selected arbitrarily. The subtree rooted at a random node is selected and replaced with a terminal node in Shrink mutation. In contrast, the Grow mutation replaces a randomly generated subtree with another subtree rooted at a random node.

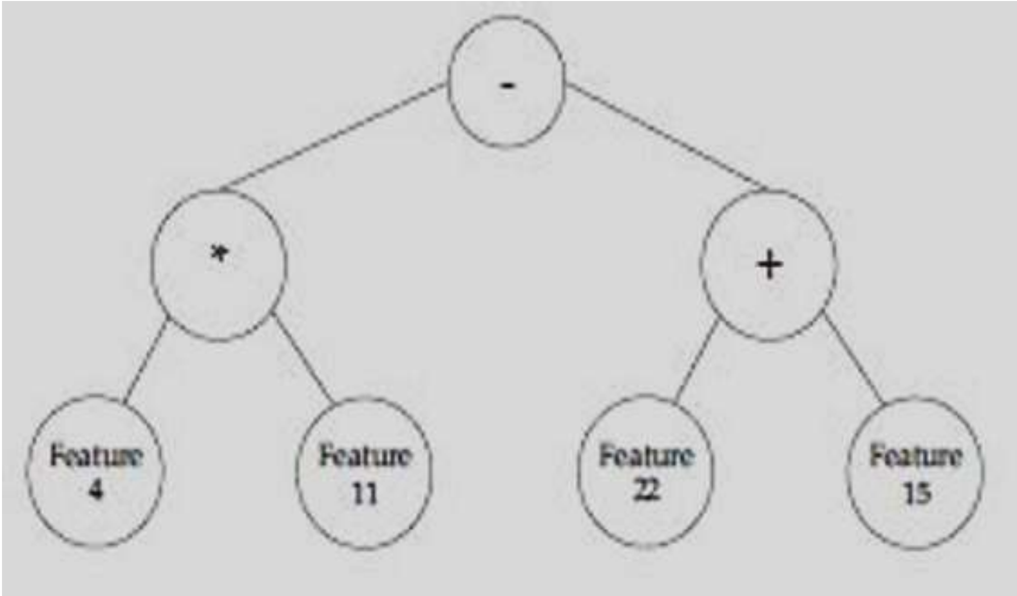

Fig. 4. Representation of a GP Program

3.5 Standard fitness function

In GP, classification accuracy (Eq. (1)), is utilized as the fitness function. Accuracy is defined as a percentage of correctly classified samples. The objective will be to converge the individual's fitness value close to 100% over generation-by-generation.

$$F_{acc} = 100 * \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

Where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. In Eq-1, the numerator part is total of correctly predicted counts. It doesn't discriminate whether this prediction belongs to the majority class samples or the minority class samples. As in an imbalanced data set, the majority class counts are higher, so leads to biased learning. In other words, the learning will be inclined toward the majority class. In an imbalanced set, this is seen as a drawback of the standard fitness function.

Also, in binary classification, the prediction is only considered true or false. Therefore, any value greater than a boundary value (generally 0.00 or 0.50), is treated as 'true', and less than that boundary value is treated as 'false'. If we take 0.50 as a boundary value, then predicted values of 0.59, 0.82, and 1.0 are translated as accurate for the positive class sample. But a classifier predicting 1.0 is more promising than classifiers predicting 0.82 and 0.59. Thus, the Eq-1 doesn't consider the fact of how close the model prediction is. This property is particularly important in ecological datasets, where minority classes often correspond to rare or threatened species whose accurate prediction carries high conservation value.

3.6 Proposed method

We utilized the Euclidean distance and weight-based (EDWB) fitness function depicted as Eq. (2). The given fitness function fixes both the problems described in the earlier section.

$$F_{edwb} = \sum_{i=1}^{N_m} \frac{|\Delta_{m_i}|^2}{2*N_m} + \sum_{i=1}^{N_M} \frac{|\Delta_{M_i}|^2}{2*N_M} \tag{2}$$

Where,
N_m: Total minority class samples count.
N_M: Total majority class samples count.
ΔM_i: For the ith majority class sample, Euclidean distance among the expected value and predicted value.
Δm_i: For the ith minority class sample, Euclidean distance among the expected value and predicted value.

The above Eq. (2) has two in-built properties. The first property is that both classes learning are calculated separately and given equal weightage. This equal weightage reduces the stronger influence of the majority class samples during training. The second is the build property of the Eq. (2) is that by calculating the Euclidean distance, it considers the fact of how close the prediction is. This factor resolves the under-fitting issue of the accuracy-based fitness function. Also, given the imbalanced nature of species-presence datasets, GP's evolutionary search is well suited for uncovering non-linear relationships in complex landscapes.

4. Results and discussions

This section evaluates the comparative performance of the proposed method against established baselines to validate its suitability for ecological prediction tasks. We implemented the proposed modelling approach in Python on a Windows-based system. Python Scikit-learn. NumPy and Pandas libraries were used to implement this. We also performed our experiments for the Support vector machine (SVM) and GP with an accuracy-based fitness function to compare our results in a similar environment. We divided our data set into two sets. Eighty percent of the samples are assigned for the training, and twenty percent samples are assigned for testing. For GP, reproduction probability, crossover probability, mutation probability, and maximum generation count are set as 0.10, 0.70, 0.20, and 100, respectively. For population initialization, we chose the 'ramped half and half' method. The initialization depth for the Program generation is set to (3,5). The function set contains: -, +, *, /, and sigmoid functions. We calculated the performance measurement matrices like accuracy, recall, specificity, and area under the ROC curve (AUC). The accuracy is not a good performance measurement for the imbalanced data set, so in this discussion, we focus mainly on AUC values, which are more appropriate metrics in this context.

The experiment results and comparison with original work in literature [2], are summarized in Table-2. Blachnik et al. (2019) [2] chose four machine learning techniques, i.e., Gradient Boosted Trees (GBT), AdaBoost (ADA), Random Forest (RF), and Decision Trees (DT). Blachnik et al. (2019) evaluated their models using the Accuracy and Area Under the ROC Curve (AUC) metrics. In contrast, we assessed our proposed techniques using Accuracy, Recall, Specificity, and AUC metrics. In our experimental evaluation, the proposed GP approach (GP_{edwb}) AUC values outperformed SVM and GP_{acc} for six classes out of seven classes. These six classes include Green-Frogs, Brown-Frogs, Fire-bellied-Toad, Tree-Frog, Common-Newt, and Great-crested-Newt amphibians. When we compare GP_{edwb} with original work, our proposed approach AUC values outperformed six classes out of seven classes. Sensitivity and specificity indicated the true positive rate (TPR) and true negative rate (TNR). Also, GP_{edwb} gives more balanced results on the matrices of sensitivity and specificity compared to other techniques.

Experimental results in Table-2 conclude that the proposed technique, which uses Euclidean distance and equal weight in the fitness function, superseded the original work done [2]. The reason behind this success is that the proposed technique assigns similar weightage to the minority and the majority class. Another reason behind this achievement is that incorporating Euclidean distance in the fitness function generates an improved generalized trained model. Thus, our proposed approach performed better for the classification of imbalanced nature amphibian data.

The improved predictive performance of the proposed GP_{EDWB} model highlights its suitability for ecological assessments that rely on remotely sensed environmental features. Such reliable species-presence predictions can support agricultural and rural planners in identifying ecologically sensitive zones, particularly around water bodies and field-margin habitats that directly influence agro-ecosystem stability. By integrating these AI-driven insights into land-use decision frameworks, stakeholders can better balance infrastructure development with biodiversity conservation in agricultural landscapes. These findings demonstrate that integrating EDWB-based GP models with satellite-derived environmental features can substantially improve species-presence prediction in real-world ecological contexts. As rural and agricultural regions undergo rapid infrastructure expansion, such predictive capability becomes increasingly crucial for balancing development with ecosystem conservation.

5. Conclusions

Identification of amphibian species is essential during urban and rural infrastructure planning because these species play a vital role in maintaining ecological balance and acting as sensitive bio-indicators of environmental quality in agricultural landscapes. However, traditional field surveys are time-consuming and require significant effort and associated costs. In this study, we proposed a Genetic Programming (GP)–based modelling approach that predicts amphibian presence using features extracted from satellite imagery, offering a low-cost and efficient alternative to extensive fieldwork. The proposed EDWB fitness function successfully addresses challenges of imbalanced datasets and improves the overall generalization capability of the model. Experimental results show that our approach outperforms SVM, standard GP, and prior literature across performance metrics such as accuracy, sensitivity, specificity, and AUC. Beyond technical performance, the ability to accurately predict amphibian occurrence has direct value for sustainable agricultural land-use planning. By integrating AI-based species-prediction tools into agricultural planning workflows, policymakers and land managers can better identify ecologically sensitive zones and minimize biodiversity loss during the construction of rural roads, irrigation channels, or other agricultural infrastructure.

Tab.2: Performance of the proposed method and its comparison with baseline methods

Amphibian Type	Reference Work	Method	Accuracy	Recall	Specificity	AUC
Green Frogs	Blachnik et al. (2019) [2]	GBT	66.51	–	–	0.688
		RF	67.44	–	–	0.759

Amphibian Type	Reference Work	Method	Accuracy	Recall	Specificity	AUC
		ADA	66.36	–	–	0.683
		DT	62.65	–	–	0.621
	Proposed Work	SVM	76.32	0.818	0.688	0.753
		GP_{acc}	78.95	0.773	0.813	0.793
		GP_{edwb}	81.58	0.773	0.875	0.824
Brown Frogs	Blachnik et al. (2019) [2]	GBT	60.58	–	–	0.636
		RF	54.61	–	–	0.560
		ADA	54.56	–	–	0.588
		DT	52.46	–	–	0.486
	Proposed Work	SVM	78.95	0.900	0.375	0.638
		GP_{acc}	86.84	1.000	0.375	0.688
		GP_{edwb}	81.58	0.800	0.875	0.838
Common Toad	Blachnik et al. (2019) [2]	GBT	64.57	–	–	0.716
		RF	60.73	–	–	0.635
		ADA	62.08	–	–	0.688
		DT	62.23	–	–	0.678
	Proposed Work	SVM	76.32	0.920	0.462	0.691
		GP_{acc}	76.32	0.800	0.692	0.746
		GP_{edwb}	68.42	0.680	0.692	0.686
Fire-bellied Toad	Blachnik et al. (2019) [2]	GBT	68.34	–	–	0.657
		RF	52.99	–	–	0.588
		ADA	54.93	–	–	0.568
		DT	56.64	–	–	0.532
	Proposed Work	SVM	68.42	0.333	0.846	0.590
		GP_{acc}	68.42	0.667	0.692	0.679
		GP_{edwb}	68.42	0.750	0.654	0.702
Tree Frog	Blachnik et al. (2019) [2]	GBT	60.24	–	–	0.672
		RF	57.43	–	–	0.631
		ADA	55.32	–	–	0.620
		DT	59.82	–	–	0.576
	Proposed Work	SVM	73.68	0.714	0.750	0.732
		GP_{acc}	71.05	0.500	0.833	0.667

Amphibian Type	Reference Work	Method	Accuracy	Recall	Specificity	AUC
		GP_{edwb}	73.68	0.786	0.708	0.747
Common Newt	Blachnik et al. (2019) [2]	GBT	61.44	–	–	0.672
		RF	54.80	–	–	0.631
		ADA	58.66	–	–	0.620
		DT	62.97	–	–	0.576
	Proposed Work	SVM	76.32	0.417	0.923	0.670
		GP_{acc}	71.05	0.583	0.769	0.676
		GP_{edwb}	76.32	0.667	0.808	0.737
Great-crested Newt	Blachnik et al. (2019) [2]	GBT	67.56	–	–	0.831
		RF	54.76	–	–	0.870
		ADA	68.15	–	–	0.775
		DT	51.79	–	–	0.510
	Proposed Work	SVM	78.95	0.250	0.853	0.551
		GP_{acc}	89.47	0.250	0.971	0.610
		GP_{edwb}	86.84	1.000	0.853	0.926

Data availability
All benchmark data sets used in this study were available from the UCI repository. [https://archive.ics.uci.edu/ml/index.php].

Disclosure of interest
The authors declare that they have no conflict of interest.

Disclosure of funding
No funding was received.

References
[1] F. Sangermano, J. R. Eastman, “A GIS framework for the refinement of species geographic ranges,” *International Journal of Geographical Information Science*, vol. 26, no. 1, pp. 39–55, 2012.
[2] M. Blachnik, M. Soltysiak, D. Dabrowska, “Predicting presence of amphibian species using features obtained from gis and satellite images,” *ISPRS International Journal of Geo-Information*, vol. 8, no. 3, p. 123, 2019.
[3] A. Chlingaryan, S. Sukkarieh, B. Whelan, “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review,” *Computers and electronics in agriculture*, vol. 151, pp. 61–69, 2018.
[4] A. Kumar, T. Choudhary, “A machine learning approach for the land-type classification,” in *Innovations in Electrical and Electronic Engineering*, Springer, 2021, pp. 647–656.
[5] S. Brayne, A. Christin, “Technologies of crime prediction: The reception of algorithms in policing and criminal courts,” *Social Problems*, vol. 68, no. 3, pp. 608–624, 2021.
[6] M. Lopes, M. Fauvel, S. Girard, D. Sheeren, “Object-based classification of grasslands from high resolution satellite image time series using gaussian mean map kernels,” *Remote Sensing*, vol. 9, no. 7, p. 688, 2017.
[7] H. Mojaddadi, B. Pradhan, H. Nampak, N. Ahmad, A. H. b. Ghazali, “Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS,” *Geomatics, Natural Hazards and Risk*, vol. 8, no. 2, pp. 1080–1102, 2017.
[8] G. Andrienko, N. Andrienko, P. Jankowski, D. Keim, M.-J. Kraak, A. MacEachren, S. Wrobel, “Geovisual analytics for spatial decision support: Setting the research agenda,” *International journal of geographical information science*, vol. 21, no. 8, pp. 839–857, 2007.
[9] C. S. Reddy, “Remote sensing of biodiversity: what to measure and monitor from space to species?,” *Biodiversity and Conservation*, vol. 30, no. 10, pp. 2617–2631, 2021.

[10] T. A. Gardner, J. Barlow, C. A. Peres, "Paradox, presumption and pitfalls in conservation biology: the importance of habitat change for amphibians and reptiles," *Biological conservation*, vol. 138, no. 1-2, pp. 166–179, 2007.

[11] J. Baker, T. Beebee, J. Buckley, T. Gent, D. Orchard, "Amphibian habitat management handbook," *Amphibian and Reptile Conservation, Bournemouth*, 2011.

[12] Z. K. Senturk, "Amphibian species detection in water reservoirs using artificial neural networks for ecology-friendly city planning," *Ecological Informatics*, vol. 69, p. 101640, 2022.

[13] A. Kumar, P. Maurya, S. M. Tiwari, A. Ali, H. Vasisht, A. S. Baghel, "Classification of forest cover-type using ensemble of decision tree, random forest and K nearest neighbor," *JIMS8I International Journal of Information Communication and Computing Technology*, vol. 10, no. 2, pp. 615–619, 2022.

[14] A. Kumar, A. Kakkar, R. Majumdar, A. S. Baghel, "Spatial Data Mining: Recent Trends and Techniques," in *2015 International Conference on Computer and Computational Sciences (ICCCS)*, 2015, pp. 39–43, IEEE.

[15] J. Koza, "On the programming of computers by means of natural selection," *Genetic Programming*, 1992.

[16] D. Acharya, S. Goel, R. Asthana, A. Bhardwaj, "A novel fitness function in genetic programming to handle unbalanced emotion recognition data," *Pattern Recognition Letters*, vol. 133, pp. 272–279, 2020.

[17] A. Bhardwaj, A. Tiwari, R. Krishna, V. Varma, "A novel genetic programming approach for epileptic seizure detection," *Computer methods and programs in biomedicine*, vol. 124, pp. 2–18, 2016.

[18] A. M. Napiórkowska, *The use of Genetic Programming for detecting the incorrect predictions of Classification Models*. Master's thesis, Universidade NOVA de Lisboa (Portugal)), 2020.

[19] J. R. Koza, "Human-competitive results produced by genetic programming," *Genetic programming and evolvable machines*, vol. 11, no. 3-4, pp. 251–284, 2010.

[20] C. Beyan, R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653–1672, 2015.

[21] J. Zhai, S. Zhang, C. Wang, "The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 3, pp. 1009–1017, 2017.

[22] K. N. Rao, "A novel class imbalance learning method using subset filtering," *International Journal of Scientific and Engineering Research*, 2012.

[23] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern recognition*, vol. 46, no. 12, pp. 3460–3471, 2013.

[24] W. Pei, B. Xue, L. Shang, M. Zhang, "Genetic programming for development of cost-sensitive classifiers for binary high-dimensional unbalanced classification," *Applied Soft Computing*, p. 106989, 2020.

[25] M. A. U. H. Tahir, S. Asghar, A. Manzoor, M. A. Noor, "A classification model for class imbalance dataset using genetic programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019.

[26] D. Dua, C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[27] B. Tran, B. Xue, M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, no. 1, pp. 3–15, 2016.