Predicting the Presence of Amphibians with GIS Data Through the Use of Multi-Label Machine Learning Classifiers

Fady Gouda, Abhi Jha, Griffin Noe, Utkrist P. Thapa

Computer Science 297a: Machine Learning Methods

Professor Cody Watson

November 21, 2020

Washington and Lee University

Introduction

The dataset used for this research project was taken from the paper *Predicting Presence of Amphibian Species Using Features Obtained from GIS and Satellite Images* by Blachnik, Sołtysiak, Dąbrowska[1]. In their paper, the researchers conducted data collection from GIS and satellite images; manually tagging features such as the presence of fishing, the type of reservoir, and the proximity to buildings (full list of features found in Appendix A) then adding labels for the presence of seven distinct amphibians. The data was collected from two strips of proposed motorway in Poland where environmental assessment was needed (maps of motorways from original paper in Appendix B). The goal of this classification is to conduct an environmental impact assessment without having to physically assess these often vast environments.

In the original paper, researchers used four models: C4.5 decision tree, AdaBoost, random forest and gradient-boosted trees to try and predict the multilabel classification problem. The authors stated that "Other methods like neural networks, distance-based algorithms or kernel-based methods (like support vector machines) require an initial preprocessing stage which transforms non numerical attributes into numerical ones. Unfortunately, the preprocessing introduces extra variance to the system and can decrease the overall performance." We decided to test this claim by conducting a variety of feature selection and feature extraction methods before running an array of classifiers and comparing our results to those of the researchers.

Related Works

As satellite imagery tools have greatly improved in the past few years, research concerning the predictions of animal presences based off of these images has proliferated. In Hollings, Tracey,

[1] Blachnik, M.; Sołtysiak, M.; Dąbrowska, D. Predicting Presence of Amphibian Species Using Features Obtained from GIS and Satellite Images. ISPRS Int. J. Geo-Inf. 2019, 8, 123.

et al.[2], the current methods, most prominent issues, and future developments of remotely sensed imagery to detect and count animals is discussed in length. In Ray, Lehmann, and Joly[3] a GIS approach is used for spatial distributions of amphibians, but no machine learning classifiers are utilized in their prediction processes. In Lenhardt, Patrick, et al.[4] landscape features such as land type and the presence of reservoirs are considered but in order to estimate the movements and migration paths of animals not their presence in a given environment.

Methodology

EDA/Feature Extraction

The proper use of feature selection methods to enhance the understanding of the model is essential to the overall process of classification. The original dataset contained 27 columns (Appendix A), a number that expanded substantially after one-hot encoding the categorical variables. Since it's plausible that there would be features that may not explain the final classification, the selection of key features through a method of ranking in terms is one of the data techniques we applied. Specifically, we pursued a univariate feature selection strategy and used the chi-squared function to calculate the score for each feature. We then utilized the correlation matrix to determine the collinearity that existed within the features (a zoom-in of the correlation heatmap highlighting all of the collinear features in Appendix C). Furthermore, the elimination of features was performed on the basis of their individual chi-squared scores.

[2] Hollings, Tracey, et al. "How Do You Find the Green Sheep? A Critical Review of the Use of Remotely Sensed Imagery to Detect and Count Animals." Methods in Ecology and Evolution, vol 9, 2018, 881–892.
[3] Ray, N., Lehmann, A. & Joly, P. Modeling spatial distribution of amphibian populations: a GIS approach based on habitat matrix permeability. Biodiversity and Conservation 11, 2143–2165 (2002).
[4] Lenhardt, Patrick P., et al. "An Expert-Based Landscape Permeability Model for Assessing the Impact of Agricultural Management on Amphibian Migration." Basic and Applied Ecology, vol 14, 2013, 442–451.

Model Optimization

For this project, we utilized five distinct machine learning algorithms in an attempt to optimally make predictions: k-nearest neighbors, logistic regression, naive bayes, random forest, and support vector machines. We attempted various scalers throughout the frameworks but found that the standard scaler did not always optimize the accuracy but in almost every occasion maximized the F1 and ROC AUC scores so that was selected for all of the models. As there were relatively few examples in the dataset, we decided to use a standard train/test split of 80/20 to prioritize training examples for the models. For every model that used hyperparameters, we set up a parameter grid of potential permutations with a wide array of potential hyperparameters. After running a grid search on the param grid to determine the general range of ideal hyperparameters, we did manual small-scale tweaking to find the truly optimal values.

Evaluation

For each machine learning algorithm, we conducted at least four permutations for evaluation purposes: no feature extraction, linear discriminant analysis, principal component analysis, and any combination of the aforementioned feature selection/extraction methods. We show the best performing models below based on 10 runs with the average of the scores reported. The accuracy score referenced looks at the average accuracy of all the labels, not the accuracy in predicting all seven labels correctly.

In the research paper that used this dataset originally, the only metrics recorded were ROC AUC scores and balanced accuracy. We used accuracy, precision, f1, and ROC AUC in order to prepare our results to those of the researchers (Appendix E). The best performing model by every metric was the logistic regression using pca with an accuracy of 73.8%. When compared to the researcher's results, our ROC AUC score would place fourth out of five models, but it is impossible to truly assess which was better. It should also be noted that the researchers seemed to approach this issue as a multi single label classification problem as opposed to a single multi label classification problem as we did so that also affects reported accuracy.

Model	Accuracy Score	Precision	ROC AUC Score	F1 Score
KNN_LDA	0.662	0.477	0.559	0.515
KNN_PCA	0.696	0.485	0.565	0.502
NB_LDA	0.675	0.509	0.593	0.541
NB_PCA	0.645	0.495	0.570	0.495
RF_LDA	0.641	0.510	0.558	0.517
RF_PCA	0.703	0.488	0.563	0.468
LR_LDA	0.666	0.468	0.565	0.474
LR_PCA	**0.738**	**0.515**	**0.602**	**0.520**
LR_UNI	0.713	0.491	0.555	0.426
SVM_UNI_P_L	0.728	0.367	0.519	0.356

Threats to Validity

The original dataset contains 189 rows and the multilabel classification task requires the model to classify instances into seven target classes. Since every instance has a binary classification for seven target classes. This is essentially taking the same set of features in order to make seven different binary classifications. Hence, it should be stated that 189 instances may not be enough data to train a model that performs reasonably well across seven multilabel classifications.

Our approach to this multilabel classification task treats the seven target class labels as independent from one another. In Appendix D, we show that there is some correlation between the class labels (for instance L4 and L6; L7 and L6). This would imply that the presence or absence of one type of amphibian may affect the chances of another type of amphibian being either present or absent. However, we ignore the impact this may have on our model's performance and treat labels as being independent from one another.

The quality of features in the dataset is important to building robust models, regardless of the classifier or data preprocessing techniques. Characteristics that are most relevant to the target value may sometimes not be included in the dataset due to various reasons. This lowers the classifier's ability to make accurate predictions. That is to say, we are not certain if the features in our dataset are most relevant for maximizing accuracy, or if there exist other features that would have produced better metrics.

Conclusion/Future Work

The purpose of this paper is to demonstrate the difference in results produced by the aforementioned authors (Blachnik, Sołtysiak, Dąbrowska) and our model. In their paper, the authors claim that the extra steps required by preprocessing increases variance and thus decreases overall performance of the model. We applied several feature selection and feature extraction techniques to produce a subset of features that we believed would enhance the results of our models. We found that performing principal component analysis on our feature space, followed by the use of a logistic regressor produced the best results. Based on the ROC scores, our results would rank fourth compared to that of the authors. However, we are unable to determine and compare the actual effectiveness of the models due to vastly different approaches to the problem.

The dataset used in this study is not sufficiently large for a seven-class multilabel classification problem. Future work in this regard could improve upon the study by collecting a larger amount of data from geographic information systems. Furthermore, the features in the dataset used in this study were manually collected by the researchers mentioned earlier. A better approach could be to find ways of automatically extracting relevant features from images collected from GIS resources. To this end, deep learning approaches could vastly increase model performance. However, DL techniques require a lot more data and we leave this task for future considerations in related research.

Lastly, finding ways of accounting for the correlation between the target labels in the dataset presents possibilities for improvement in model performance. For example, classifiers can be chained with each classifier using the prediction of the previous label to inform its classification for the next.

Appendices

A) Feature List[1]:

1) ID - vector ID (not used in the calculations)
2) MV - motorway (not used in the calculations)
3) SR - Surface of water reservoir numeric
4) NR - Number of water reservoirs in habitat - Comment: The larger the number of reservoirs, the more likely it is that some of them will be suitable for amphibian breeding.
5) TR - Type of water reservoirs:
   a. reservoirs with natural features that are natural or anthropogenic water reservoirs (e.g., subsidence post-exploited water reservoirs), not subjected to naturalization
   b. recently formed reservoirs, not subjected to naturalization
   c. settling ponds
   d. water reservoirs located near houses
   e. technological water reservoirs
   f. water reservoirs in allotment gardens
   g. trenches
   h. wet meadows, flood plains, marshes
   i. river valleys
   j. streams and very small watercourses
6) VR - Presence of vegetation within the reservoirs:
   a. no vegetation
   b. narrow patches at the edges
   c. areas heavily overgrown
   d. lush vegetation within the reservoir with some part devoid of vegetation
   e. reservoirs completely overgrown with a disappearing water table
   Comment: The vegetation in the reservoir favors amphibians, facilitates breeding, and allows the larvae to feed and give shelter. However, excess vegetation can lead to the overgrowth of the pond and water shortages.
7) SUR1 - Surroundings; the dominant types of land cover surrounding the water reservoir
8) SUR2 - Surroundings; the second most dominant types of land cover surrounding the water reservoir
9) SUR3 - Surroundings; the third most dominant types of land cover surrounding the water reservoir
   Comment: The surroundings feature was designated in three stages. First, the dominant surroundings were selected. Then, two secondary types were chosen.
   a. forest areas (with meadows) and densely wooded areas
   b. areas of wasteland and meadows
   c. allotment gardens
   d. parks and green areas
   e. dense building development, industrial areas
   f. dispersed habitation, orchards, gardens
   g. river valleys
   h. roads, streets
   i. agricultural land
   The most valuable surroundings of water reservoirs for amphibians are areas with the least anthropopressure and proper moisture.
10) UR - Use of water reservoirs:
   a. unused by man (very attractive for amphibians)
   b. recreational and scenic (care work is performed)
   c. used economically (often fish farming)
   d. technological
11) FR - The presence of fishing:
   a. lack of or occasional fishing
   b. intense fishing
   c. breeding reservoirs

Comment: The presence of a large amount of fishing, in particular predatory and intense fishing, is not conducive to the presence of amphibians.

12) OR - Percentage access from the edges of the reservoir to undeveloped areas (the proposed percentage ranges are a numerical reflection of the phrases: lack of access, low access, medium access, large access to free space):

a. 0-25; lack of access or poor access
b. 25-50; low access
c. 50-75; medium access,
d. 75-100; large access to terrestrial habitats of the shoreline is in contact with the terrestrial habitat of amphibians.

13) RR Minimum distance from the water reservoir to roads:

a. <50 m
b. 50-100 m
c. 100-200 m
d. 200-500 m
e. 500-1000 m
f. >1000 m

Comment: The greater the distance between the reservoir and the road, the more safety for amphibians.

14) BR - Building development - Minimum distance to buildings:

a. <50 m
b. 50-100 m
c. 100-200 m
d. 200-500 m
e. 500-1000 m
f. >1000 m

Comment: The more distant the buildings, the more favorable the conditions for the occurrence of amphibians.

15) MR - Maintenance status of the reservoir:

a. Clean
b. slightly littered
c. reservoirs heavily or very heavily littered

Comment: Trash causes devastation of the reservoir ecosystem. Backfilling and leveling of water reservoirs with ground and debris should also be considered.

16) CR - Type of shore

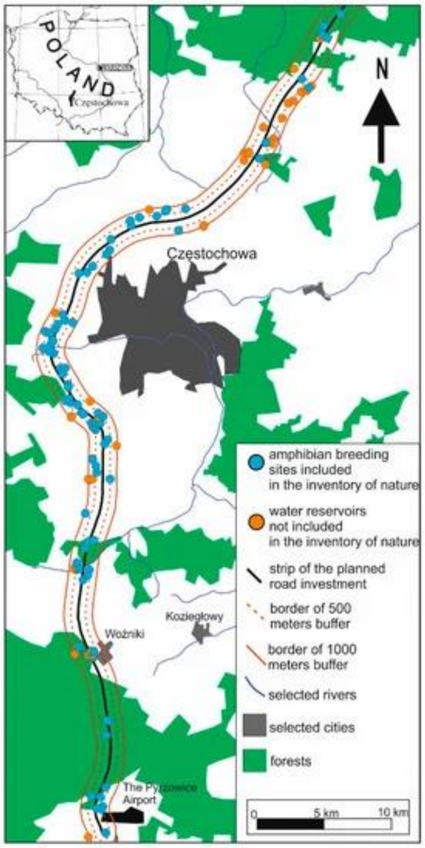
a. Natural
b. Concrete

Comment: A concrete shore of a reservoir is not attractive for amphibians. A vertical concrete shore is usually a barrier for amphibians when they try to leave the water.

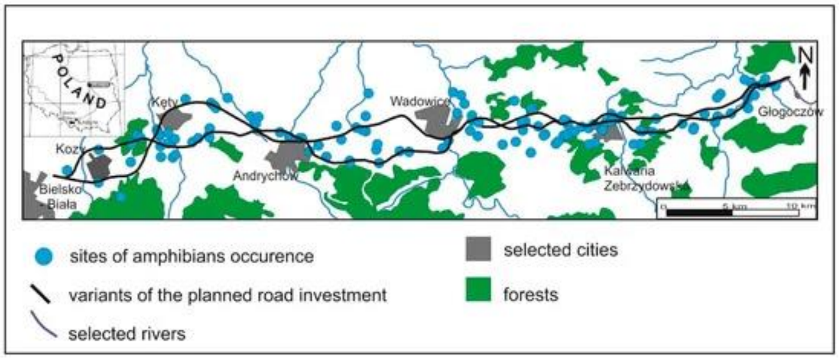
17) Label 1 - the presence of Green frogs
18) Label 2 - the presence of Brown frogs
19) Label 3 - the presence of Common toad
20) Label 4 - the presence of Fire-bellied toad
21) Label 5 - the presence of Tree frog
22) Label 6 - the presence of Common newt
23) Label 7 - the presence of Great crested newt

B) Motorway Maps[1]:

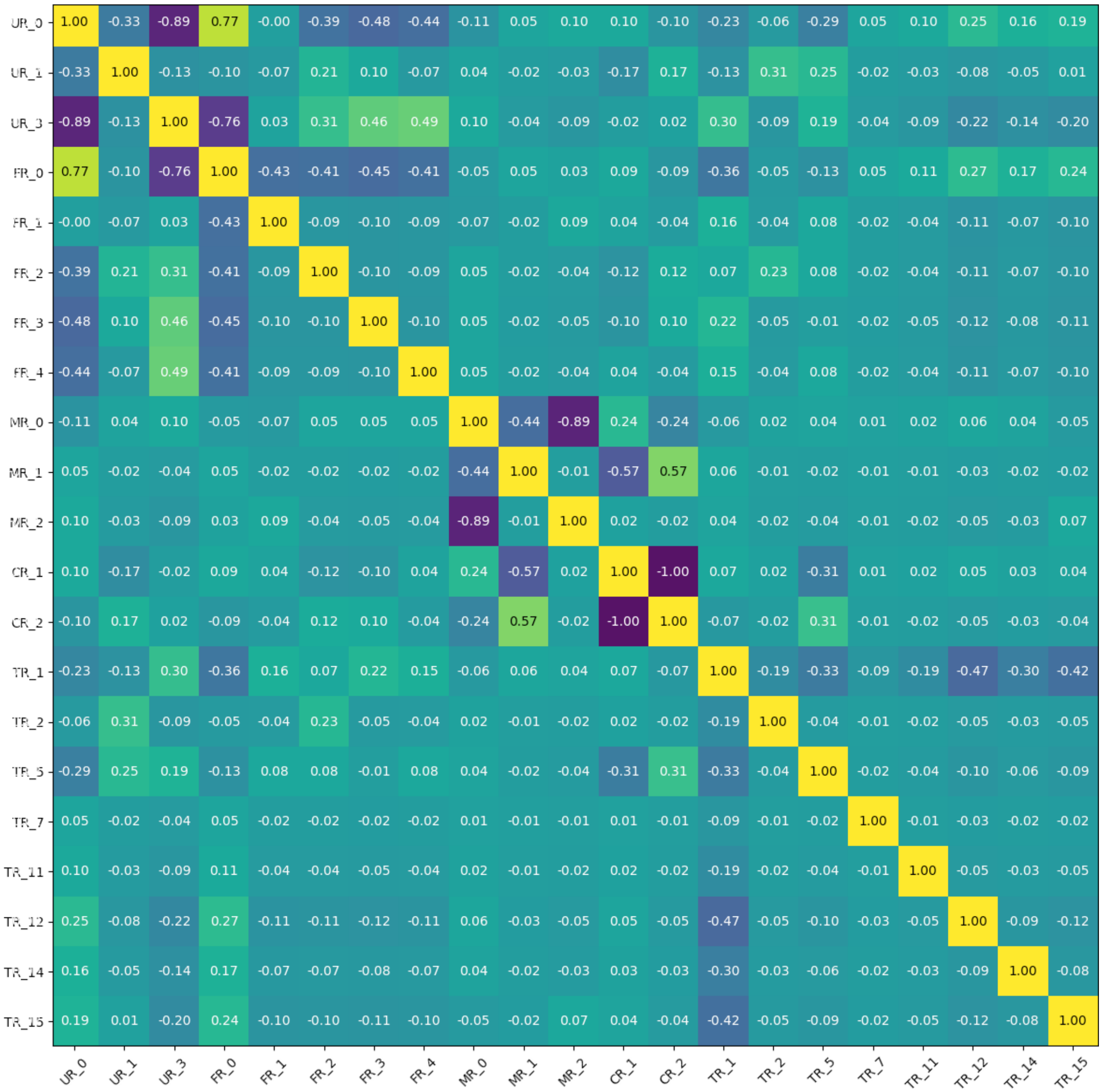
Proposed A1 Motorway:


POLAND
Częstochowa
N
Częstochowa
Woźniki
Koziegłowy
The Pyrzowice Airport
amphibian breeding sites included in the inventory of nature
water reservoirs not included in the inventory of nature
strip of the planned road investment
border of 500 meters buffer
border of 1000 meters buffer
selected rivers
selected cities
forests
0
5 km
10 km

Proposed S52 Motorway:

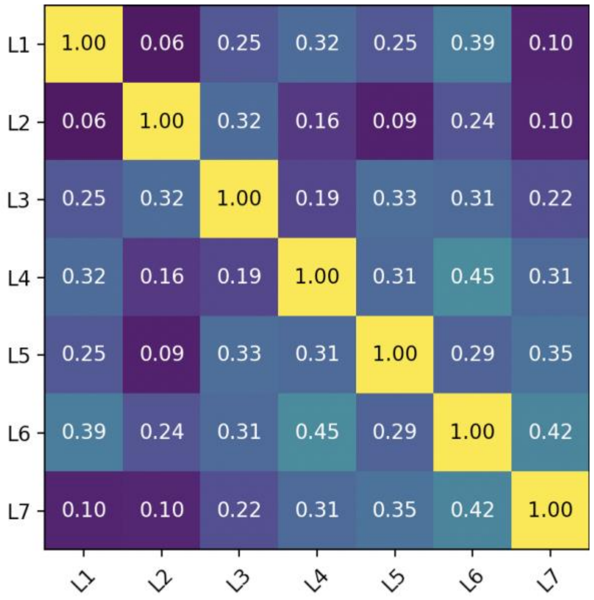


C) Collinear Features Heatmap:



	UR_0	UR_1	UR_3	FR_0	FR_1	FR_2	FR_3	FR_4	MR_0	MR_1	MR_2	CR_1	CR_2	TR_1	TR_2	TR_5	TR_7	TR_11	TR_12	TR_14	TR_15
UR_0	1.00	-0.33	-0.89	0.77	-0.00	-0.39	-0.48	-0.44	-0.11	0.05	0.10	0.10	-0.10	-0.23	-0.06	-0.29	0.05	0.10	0.25	0.16	0.19
UR_1	-0.33	1.00	-0.13	-0.10	-0.07	0.21	0.10	-0.07	0.04	-0.02	-0.03	-0.17	0.17	-0.13	0.31	0.25	-0.02	-0.03	-0.08	-0.05	0.01
UR_3	-0.89	-0.13	1.00	-0.76	0.03	0.31	0.46	0.49	0.10	-0.04	-0.09	-0.02	0.02	0.30	-0.09	0.19	-0.04	-0.09	-0.22	-0.14	-0.20
FR_0	0.77	-0.10	-0.76	1.00	-0.43	-0.41	-0.45	-0.41	-0.05	0.05	0.03	0.09	-0.09	-0.36	-0.05	-0.13	0.05	0.11	0.27	0.17	0.24
FR_1	-0.00	-0.07	0.03	-0.43	1.00	-0.09	-0.10	-0.09	-0.07	-0.02	0.09	0.04	-0.04	0.16	-0.04	0.08	-0.02	-0.04	-0.11	-0.07	-0.10
FR_2	-0.39	0.21	0.31	-0.41	-0.09	1.00	-0.10	-0.09	0.05	-0.02	-0.04	-0.12	0.12	0.07	0.23	0.08	-0.02	-0.04	-0.11	-0.07	-0.10
FR_3	-0.48	0.10	0.46	-0.45	-0.10	-0.10	1.00	-0.10	0.05	-0.02	-0.05	-0.10	0.10	0.22	-0.05	-0.01	-0.02	-0.05	-0.12	-0.08	-0.11
FR_4	-0.44	-0.07	0.49	-0.41	-0.09	-0.09	-0.10	1.00	0.05	-0.02	-0.04	0.04	-0.04	0.15	-0.04	0.08	-0.02	-0.04	-0.11	-0.07	-0.10
MR_0	-0.11	0.04	0.10	-0.05	-0.07	0.05	0.05	0.05	1.00	-0.44	-0.89	0.24	-0.24	-0.06	0.02	0.04	0.01	0.02	0.06	0.04	-0.05
MR_1	0.05	-0.02	-0.04	0.05	-0.02	-0.02	-0.02	-0.02	-0.44	1.00	-0.01	-0.57	0.57	0.06	-0.01	-0.02	-0.01	-0.01	-0.03	-0.02	-0.02
MR_2	0.10	-0.03	-0.09	0.03	0.09	-0.04	-0.05	-0.04	-0.89	-0.01	1.00	0.02	-0.02	0.04	-0.02	-0.04	-0.01	-0.02	-0.05	-0.03	0.07
CR_1	0.10	-0.17	-0.02	0.09	0.04	-0.12	-0.10	0.04	0.24	-0.57	0.02	1.00	-1.00	0.07	0.02	-0.31	0.01	0.02	0.05	0.03	0.04
CR_2	-0.10	0.17	0.02	-0.09	-0.04	0.12	0.10	-0.04	-0.24	0.57	-0.02	-1.00	1.00	-0.07	-0.02	0.31	-0.01	-0.02	-0.05	-0.03	-0.04
TR_1	-0.23	-0.13	0.30	-0.36	0.16	0.07	0.22	0.15	-0.06	0.06	0.04	0.07	-0.07	1.00	-0.19	-0.33	-0.09	-0.19	-0.47	-0.30	-0.42
TR_2	-0.06	0.31	-0.09	-0.05	-0.04	0.23	-0.05	-0.04	0.02	-0.01	-0.02	0.02	-0.02	-0.19	1.00	-0.04	-0.01	-0.02	-0.05	-0.03	-0.05
TR_5	-0.29	0.25	0.19	-0.13	0.08	0.08	-0.01	0.08	0.04	-0.02	-0.04	-0.31	0.31	-0.33	-0.04	1.00	-0.02	-0.04	-0.10	-0.06	-0.09
TR_7	0.05	-0.02	-0.04	0.05	-0.02	-0.02	-0.02	-0.02	0.01	-0.01	-0.01	0.01	-0.01	-0.09	-0.01	-0.02	1.00	-0.01	-0.03	-0.02	-0.02
TR_11	0.10	-0.03	-0.09	0.11	-0.04	-0.04	-0.05	-0.04	0.02	-0.01	-0.02	0.02	-0.02	-0.19	-0.02	-0.04	-0.01	1.00	-0.05	-0.03	-0.05
TR_12	0.25	-0.08	-0.22	0.27	-0.11	-0.11	-0.12	-0.11	0.06	-0.03	-0.05	0.05	-0.05	-0.47	-0.05	-0.10	-0.03	-0.05	1.00	-0.09	-0.12
TR_14	0.16	-0.05	-0.14	0.17	-0.07	-0.07	-0.08	-0.07	0.04	-0.02	-0.03	0.03	-0.03	-0.30	-0.03	-0.06	-0.02	-0.03	-0.09	1.00	-0.08
TR_15	0.19	0.01	-0.20	0.24	-0.10	-0.10	-0.11	-0.10	-0.05	-0.02	0.07	0.04	-0.04	-0.42	-0.05	-0.09	-0.02	-0.05	-0.12	-0.08	1.00

D) Label Correlation Heatmap:



	L1	L2	L3	L4	L5	L6	L7
L1	1.00	0.06	0.25	0.32	0.25	0.39	0.10
L2	0.06	1.00	0.32	0.16	0.09	0.24	0.10
L3	0.25	0.32	1.00	0.19	0.33	0.31	0.22
L4	0.32	0.16	0.19	1.00	0.31	0.45	0.31
L5	0.25	0.09	0.33	0.31	1.00	0.29	0.35
L6	0.39	0.24	0.31	0.45	0.29	1.00	0.42
L7	0.10	0.10	0.22	0.31	0.35	0.42	1.00

E) Blachnik, Sołtysiak, Dąbrowska Results:

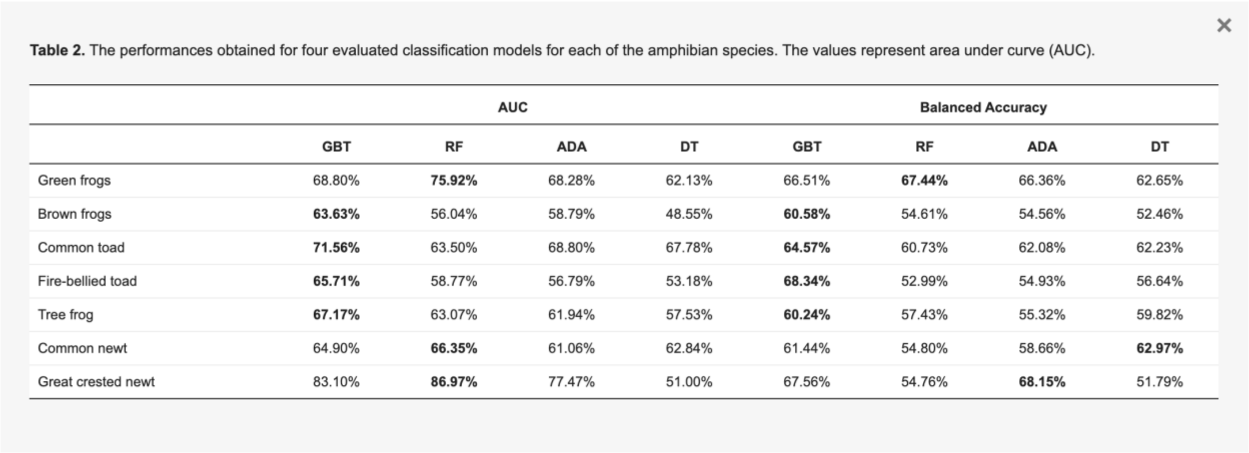


Table 2. The performances obtained for four evaluated classification models for each of the amphibian species. The values represent area under curve (AUC).

	AUC				Balanced Accuracy			
	GBT	RF	ADA	DT	GBT	RF	ADA	DT
Green frogs	68.80%	**75.92%**	68.28%	62.13%	66.51%	**67.44%**	66.36%	62.65%
Brown frogs	**63.63%**	56.04%	58.79%	48.55%	**60.58%**	54.61%	54.56%	52.46%
Common toad	**71.56%**	63.50%	68.80%	67.78%	**64.57%**	60.73%	62.08%	62.23%
Fire-bellied toad	**65.71%**	58.77%	56.79%	53.18%	**68.34%**	52.99%	54.93%	56.64%
Tree frog	**67.17%**	63.07%	61.94%	57.53%	**60.24%**	57.43%	55.32%	59.82%
Common newt	64.90%	**66.35%**	61.06%	62.84%	61.44%	54.80%	58.66%	**62.97%**
Great crested newt	83.10%	**86.97%**	77.47%	51.00%	67.56%	54.76%	**68.15%**	51.79%

Average: 69.27 67.23 64.73 57.53

References

1. Blachnik, M.; Sołtysiak, M.; Dąbrowska, D. Predicting Presence of Amphibian Species Using Features Obtained from GIS and Satellite Images. ISPRS Int. J. Geo-Inf. 2019, 8, 123.

2. Hollings, Tracey, et al. "How Do You Find the Green Sheep? A Critical Review of the Use of Remotely Sensed Imagery to Detect and Count Animals." *Methods in Ecology and Evolution*, vol. 9, no. 4, 2018, pp. 881–892., doi:10.1111/2041-210x.12973.

3. Ray, N., Lehmann, A. & Joly, P. Modeling spatial distribution of amphibian populations: a GIS approach based on habitat matrix permeability. Biodiversity and Conservation 11, 2143–2165 (2002). https://doi.org/10.1023/A:1021390527698

4. Lenhardt, Patrick P., et al. "An Expert-Based Landscape Permeability Model for Assessing the Impact of Agricultural Management on Amphibian Migration." Basic and Applied Ecology, vol. 14, no. 5, 2013, pp. 442–451., doi:10.1016/j.baae.2013.05.004.