Applied Soft Computing Journal 155 (2024) 111468

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Classification Assessment Tool: A program to measure the uncertainty of classification models in terms of class-level metrics

Szilárd Szabó [a], Imre J. Holb [b,c], Vanda Éva Abriha-Molnár [d], Gábor Szatmári [e], Sudhir Kumar Singh [f], Dávid Abriha [a,*]

[a] Department of Physical Geography and Geoinformatics, Faculty of Science and Technology, University of Debrecen, Egyetem tér 1, Debrecen 4032, Hungary
[b] Institute of Horticulture, University of Debrecen, Böszörményi u. 138, Debrecen 4032, Hungary
[c] HUN-REN, Centre for Agricultural Research, 1022 Budapest, Herman Ottó út 15, Hungary
[d] HUN-REN-UD, Anthropocene Ecology Research Group, University of Debrecen, Egyetem tér 1, Debrecen H-4032, Hungary
[e] HUN-REN Centre for Agricultural Research, Institute for Soil Sciences, Herman Ottó út 15, Budapest H-1022, Hungary
[f] K. Banerjee Centre of Atmospheric & Ocean Studies, IIDS, Nehru Science Centre, University of Allahabad, Prayagraj, Uttar Pradesh, India

HIGHLIGHTS

- Accuracy assessments are biased by the testing dataset.
- Repetitions help to quantify the uncertainty of accuracy measures.
- The developed tool determines the class level accuracies with the uncertainties.
- We pointed on the accuracy measures biased by many true negative cases.
- F1, IOU and Matthews correlation performed well in all experiments.

ARTICLE INFO

Keywords:
Model evaluation
Model stability
Testing
Repetitions
Python

ABSTRACT

Accuracy assessments are important steps of classifications and get higher relevance with the soar of machine and deep learning techniques. We provided a method for quick model evaluations with several options: calculate the class level accuracy metrics for as many models and classes as needed; calculate model stability using random subsets of the testing data. The outputs are single calculations, summaries of the repetitions, and/or all accuracy results per repetitions. Using the application, we demonstrated the possibilities of the function and analyzed the accuracies of three experiments. We found that some popular metrics, the binary Overall Accuracy, Sensitivity, Precision, and Specificity, as well as ROC curve, can provide false results when the true negative cases dominate. F1-score, Intersection over Union and the Matthews correlation coefficient were reliable in all experiments. Medians and interquartile ranges (IQR) of the repeated sampling from the testing dataset showed that IQR were small when a model was almost perfect or completely unacceptable; thus, IQR reflected the model stability, reproducibility. We found that there were no general, statistically justified relationship with the median and IQR, furthermore, correlations of accuracy metrics varied by experiments, too. Accordingly, a multi-metric evaluation is suggested instead of a single metric.

1. Introduction

Modelling is a common task both in scientific and practical parts of data science and became popular with machine learning and deep learning algorithms. Models aim to help to understand the environment, its features, processes, changes, and the consequences of changes. Especially in geosciences, agricultural, biological and even medical sciences, models are used to identify target objects, such as land cover units (e.g. forest fires, grassland mapping), roofing materials, species (plant species, habitats), diseases (lung cancer, melanoma) etc. Abdollahi et al., [19,2,25,29,3,32,37,41,42,44,53,8]. Accuracy measures reveal the efficacy of model predictions, which can limit the

* Corresponding author.
E-mail address: abriha.david@science.unideb.hu (D. Abriha).

https://doi.org/10.1016/j.asoc.2024.111468
Received 28 April 2023; Received in revised form 11 February 2024; Accepted 1 March 2024
Available online 12 March 2024

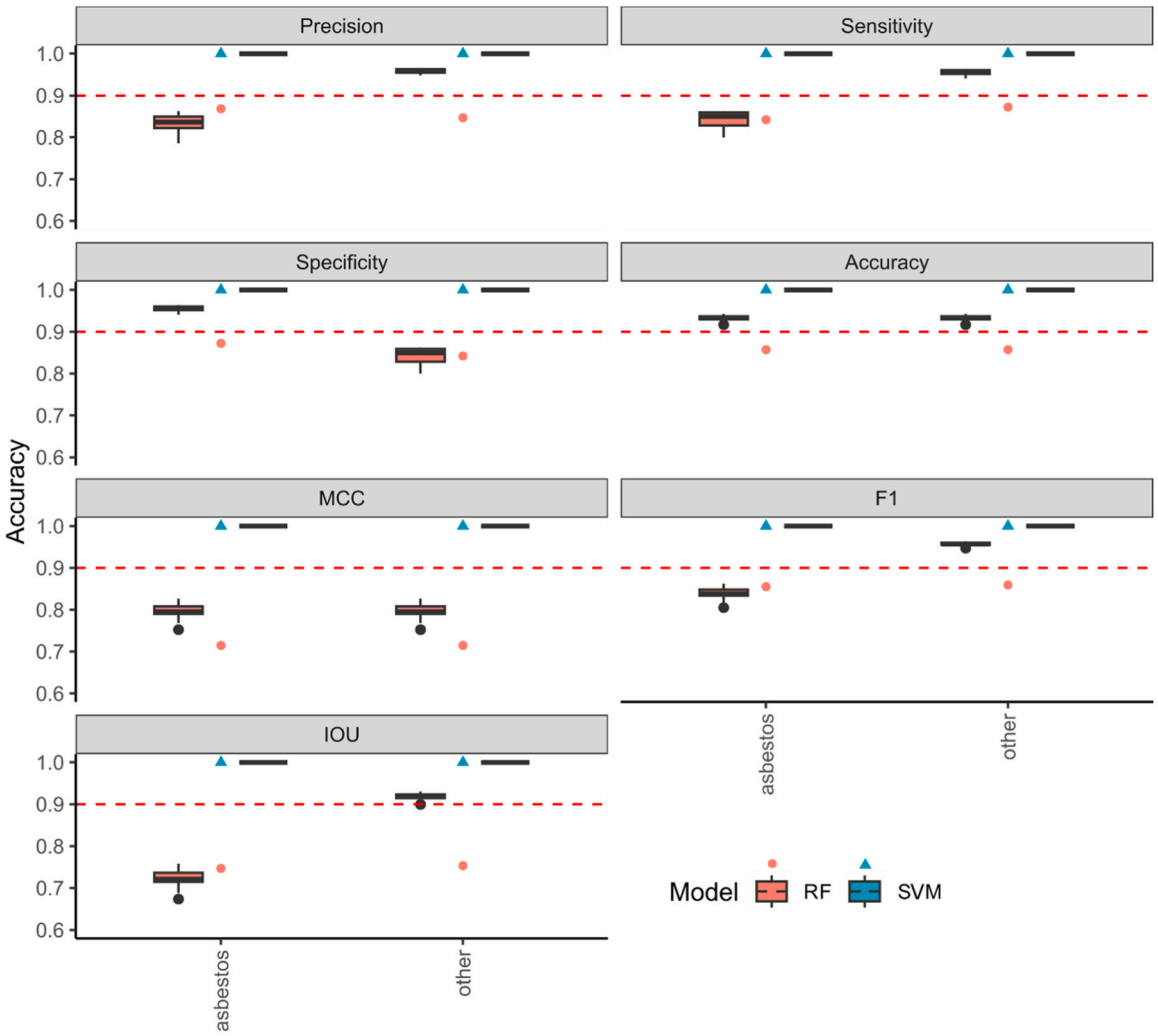

Fig. 1. Accuracy metrics of the roofing classifications (RF: Random Forest, SVM: Support Vector Machine; boxplots are derived from the repetitions of the randomly sampled testing dataset: median, quartiles, 1.5 × interquartile ranges, black points: outliers; red and green symbols: accuracy metrics derived from the whole testing dataset; red dashed line: 90% accuracy benchmark). MCC: Matthews correlation coefficient, IOU: Intersection over Union, F1: F1-score.

applicability of the given approach [50]. Although even a bivariate linear regression is a model and widely used in studies, statistical models have to follow rules: there should be model training/building, a validation (to calculate model parameters), and finally the testing of the predictions. Models should be developed with a 'training dataset' and the testing should be performed with an independent dataset (i.e., the part of the reference data which was not used for training) [51]. Simple linear regression models are usually performed on the whole dataset, but this is only acceptable when the model is used to reveal the contribution of the independent variable(s) and not used for prediction. Accuracy of predictions is biased when the model is tested with the same data as the model was developed.

Testing dataset is not always a separated one, it can be only a temporarily removed part of the reference database, e.g. in case of the cross-validation (CV) technique. CV can be 'leaving-one-out', removing only one case from the reference data at a time, or 'k-fold' cross-validation (KCV) splitting randomly the dataset into 'k' folds, and during the calculations one fold is hold out for testing, the remaining ones are used for training the models, then the holdout fold is replaced and another fold is taken out; the procedure stops when all folds were used as testing data. KCV can be extended with repetitions, i.e. 'repeated k-fold' cross-validation (RKCV) when we repeat the random split of the reference data [26]. The output will be as many accuracy metrics as many folds (and repetitions) we applied, the optimal number for 'k' is between 2 and 10, while the repetitions depend on the folds, the final model number is 30–100 (e.g., 3 folds with 10 repetitions with small dataset, or 10 folds with 3 repetitions with large datasets) [38]. CV-methods provide an insight of the models with a distribution of Root Mean Square Errors (RMSEs), R-squares, or Overall Accuracies (OAs), can be described with means, standard deviations (SDs), or quartiles. The larger the range (maximum – minimum) or interquartile range (upper quartile – lower quartile), the larger the uncertainty of the model, which also reflects the reliability of the reference database [11].

Although CV sounds an efficient alternative to independent testing with a testing dataset, but, indeed, independent testing cannot be replaced with the RKCV. CV helps in validating the model; thus, we do not need to separate a third, 'validation' part of the reference data (e.g. for hyperparameter tuning), because CV provides it within the training data with the folds, also can describe the model reliability, but not applicable for testing the predictions [1]. Furthermore, especially for classifications, the output is only an overall accuracy without the details of class level metrics. However, the idea of having a range of accuracy instead of a single number is a reasonable demand to judge our final results, which can be solved with repetitions. But unlike in case of CV-approach where folds and repetitions are used for model training, in our case repetition is performed on the testing data. Thus, finally, we can have a similar output like in the training phase, however, in this case we tested the reliability of the predictions.

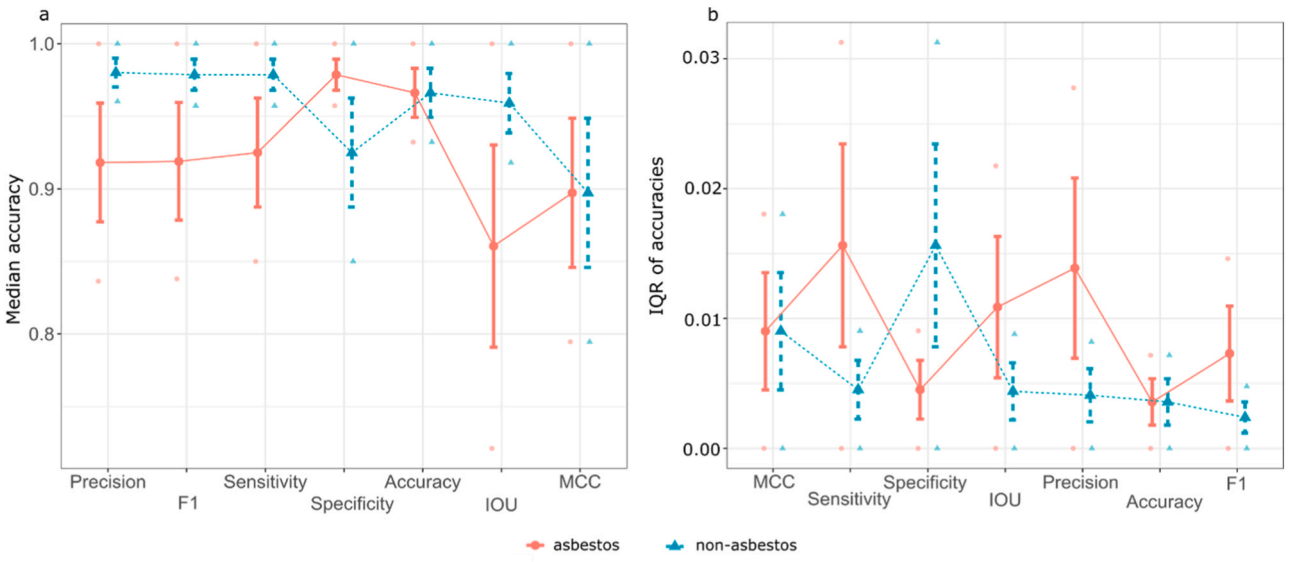

Fig. 2. Medians (a) and interquartile ranges (IQR, b) calculated from the repeated model metrics of roof type (asbestos and non-asbestos) classifications (10 repetitions from 60% stratified random samples; ●,▲: median; whiskers: lower and upper quartiles). MCC: Matthews correlation coefficient, IOU: Intersection over Union, F1: F1-score.

A wide literature exists regarding the details of the above metrics; however, authors involve the whole testing dataset and results rely on a single calculation [21,27,47,49]; usually, the confusion matrix is evaluated. Supposing that the testing dataset does not necessarily represent the statistical population; thus, the testing dataset can be regarded as a random subset of the population. Using only the part of all possible data (i.e. random subset) from the testing dataset and repeating the random sampling we can have several outputs with mean and variance and can draw better conclusions: if accuracy metrics have a low variance, it ensure model replicability, but high variance means eventuality [14]. This type of accuracy evaluation tool is missing; accordingly, we aimed to develop a simple toolbox to determine the most important metrics, both for calculating class metrics and to measure the uncertainty of the thematic accuracies.

The developed code works in Python environment and is freely available (see https://github.com/AbrihaDavid/Classification-Assessment-Tool) and is accordance with Barnes [4], i.e., codes help reproducibility and reliability of results. Four different types of models have been evaluated, and based on the extracted accuracy metrics we tested the reasonability of the theory of repeated accuracy assessment approach along the following hypotheses: (i) single accuracy metrics calculated from the whole testing dataset can have higher or lower values than ranges determined from repetitions, i.e., multiple models; (ii) accuracy metrics with repetitions provide information on model stability, (iii) model stability is in correlation with the accuracy, and (iv) correlations of accuracy metrics vary by the success of the trained models.

2. Materials and methods

We developed a method to compute the overall and class level accuracy metrics which uses repetitions. The calculation procedure is developed in Python programming environment with the aim to provide a method, which is available even for those users who are not familiar with scripting. The function requires a table with the observed data should be in the first column in number format, i.e., coded as numeric variable, and in other columns there should be the similarly coded predictions. Predictions should be named after the models (see the example in the supplementary materials). We developed a simple and advanced tool for the calculations. The basic version calculates the metric using the whole testing dataset, and the output is a table arranged by models and classes. The advanced version calculates random datasets taken from the testing data, i.e., uses repetitions.

We provided the source code, and also developed a web-based application in the Streamlit platform. Streamlit runs Python in a remote virtual machine, calls the program codes from the GitHub, and use the uploaded data only for a session, but does not store permanently. Its advantage is that the platform provides a graphical user interface, users can easily upload their own files, and the output is immediately downloaded into the user's computer. As data is not stored, there is no concern about the data policy, after finishing a task, all data is deleted from the server. Furthermore, we also developed an executable version (see https://zenodo.org/records/10646420), that can be run on the computer and has the same features as the source code and the web-based application.

2.1. The basic function (Extract Accuracy)

First option is to use the whole testing dataset, which corresponds to the traditional calculations form the confusion matrix. The advantage is that this application calculates all metrics at a time and performs the calculations for several models. The output is a single value for all metrics per models.

2.2. The advanced function (Extract Accuracy with repetitions)

The advanced version is very similar to the basic version, but this procedure takes subsamples from the whole testing dataset, and two parameters should be set:

- *fraction*: the percentage of cases from the dataset to run the subsample (the default is 0.6, i.e., 60%, can be changed between $0 < x <= 1$, it ensures a stratified random data selection representing all classes with a balanced structure;
- *iterations*: number of repetitions of the calculations (the default is 10, can be changed but too many repetitions will not ensure better results whilst the computation time can be long).

There are two options for the model evaluation, and the corresponding output:

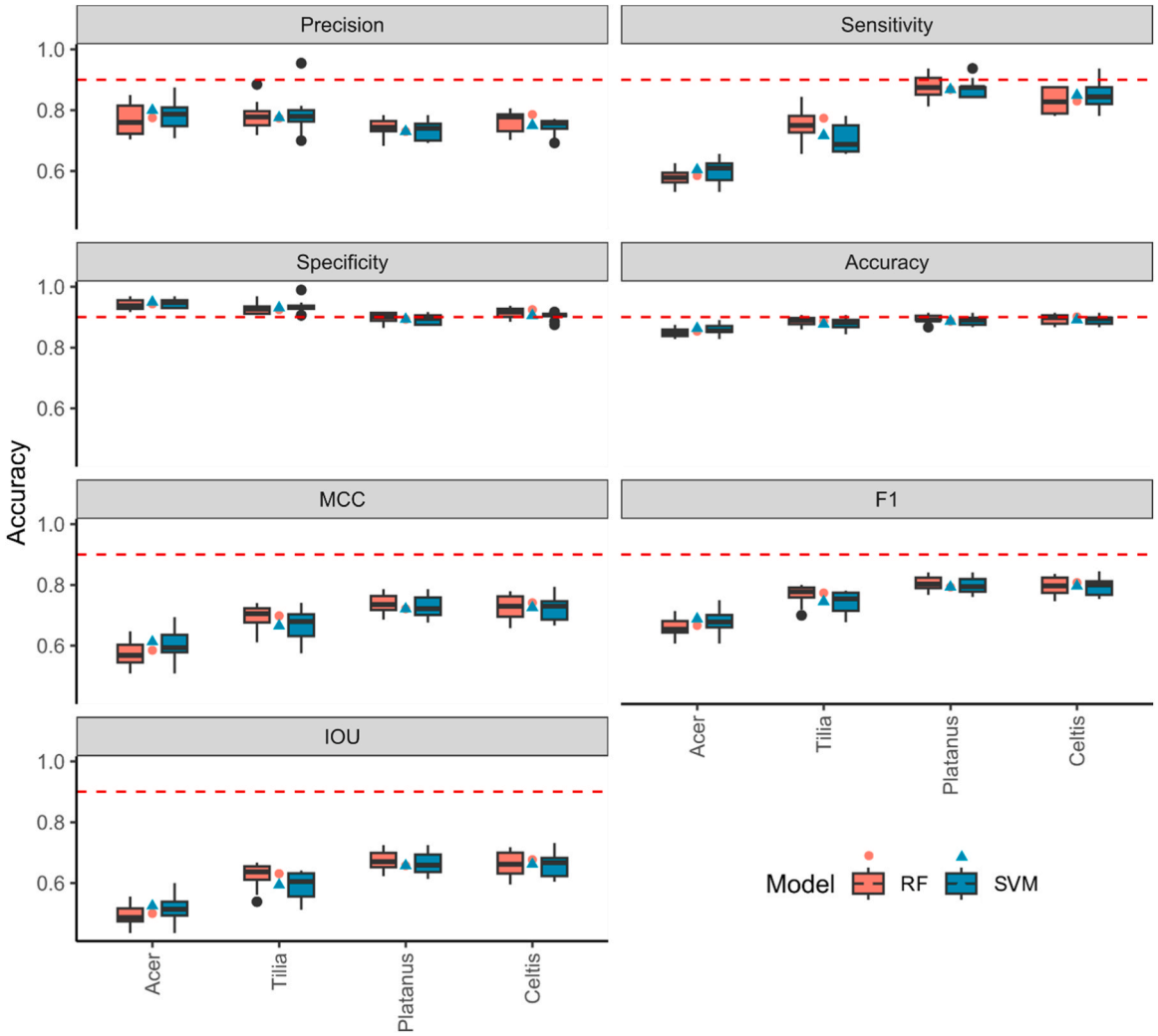

Fig. 3. Accuracy metrics of the tree species classifications (RF: Random Forest, SVM: Support Vector Machine; boxplots are derived from the repetitions of the randomly sampled testing dataset: median, quartiles, 1.5 × interquartile ranges, black points: outliers; red and green symbols: accuracy metrics derived from the whole testing dataset; red dashed line: 90% accuracy benchmark). Forest tree species are Acer: *Acer platanoides*, Tilia: *Tilia x europaea*, Platanus: *Platanus x hybrida* and Celtis: *Celtis occidentalis*. MCC: Matthews correlation coefficient, IOU: Intersection over Union, F1: F1-score.

- a table with the means and standard deviations (SDs) of the accuracy metric of repeated subsamples
- or choosing the "All data" option, a table with all the results with as many outputs as had been set in the code as repetitions. In this case, two types of figures are generated: (i) a boxplot diagram to visualize differences among the classes, and (ii) scatterplots to visualize the accuracy of classes by models (which is useful when there are several models and classes).

Both outputs have their own roles. The summary of means and SDs provides direct information about the variances of the accuracies, and the table with all the repetitions can be the basis of further calculations.

2.3. Information about the model reliability (Extract model stability)

Changes in test data sets may result in a different performance of the model, and the more possibilities are evaluated, the greater the likelihood of a more realistic assessment of accuracy. Accordingly, we have developed an automated method for extracting metrics from a variety of inputs. It allows for the definition of numerous fractions and repetitions, and the output is a table in CSV format with color tables representing the chosen metrics, enabling users to verify the accuracy range. If values are in a narrow range, regardless of model accuracy, the model performance is stable, but if small and large values also appear, models' accuracy is not reliable, because the performance is biased by the input data. The output is the average accuracy of all classes and can be used to find the settings (random fraction and number of repetitions) of the lowest or highest values settings. Next, using these settings, the relating class level metrics can be extracted with the "Extract Accuracy with reps" tool, depending on the aim of analysis.

2.4. Class level metrics

The script calculates several class level accuracy indices derived from the confusion matrix. The most common indices are the Overall Accuracy (OA). OA provides a general insight into the thematic accuracy, but nothing about the class level information. Class level metric area calculated as one vs. all other classes, as they were binary (e.g., forest and non-forest, presence and absence). Precision (in remote sensing User's Accuracy, UA), Sensitivity (in remote sensing Producer's Accuracy, PA) [5,9]. Precision, as a class level metric shows the level of commission error, while Sensitivity is the level of omission error. There are preferred metrics in different fields of science: Precision is important when false positives are not allowed due to high cost of being wrong (such as SPAM filters, we do not want to lose any important emails), but in medical science false positives mean that instead of losing a patient,

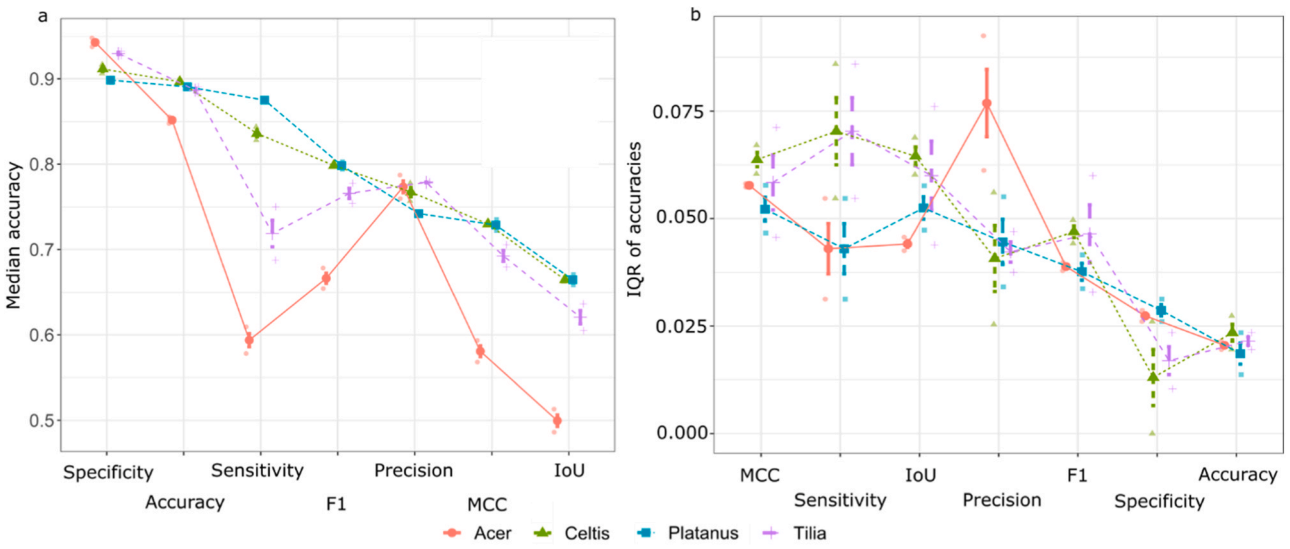

Fig. 4. Medians (a) and interquartile ranges (IQR, b) calculated from the repeated model metrics of tree species classifications (10 repetitions from 60% stratified random samples; ●,▲: median; whiskers: lower and upper quartiles). Forest tree species are Acer: *Acer platanoides*, Tilia: *Tilia x europaea*, Platanus: *Platanus x hybrida* and Celtis: *Celtis occidentalis*.

further investigations help to decide e.g., the presence of a cancer and finding the best treatment in time, thus, finally the cost of misclassification is lower. Another point of view is the Sensitivity and Precision from the error term: a high Sensitivity can be the consequence of a low Precision due to high rate of commission error [46]. A lesser-known metric is the Specificity, known also as the True Negative Rate, which is used when the false positive cases are highly disregarded, possibly causing costs or inconvenience.

$$OA = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

Although the above metrics can reflect the thematic accuracy, there is a need to express the reliability of the outcomes with one number. Accordingly, as a harmonic mean of Sensitivity and Precision, we can express the F1-scores (also known as Dice Similarity Coefficient), which now takes into consideration both the FP and FN cases. Intersection over Union (IOU; also known as Jaccard index) is calculated as the ratio of the correctly predicted cases (TP) and the errors, too [39]. Although F1 and IOU seems similar, F1 is closer to the average performance, while IOU is closer to a worst case scenario (Willem, 2017); accordingly, F1 calculates larger scores for models where Precision and Sensitivity is similar [20].

$$F1 = \frac{2TP}{2TP + FP + FN}$$

$$IOU = \frac{TP}{TP + FP + FN}$$

Matthews correlation coefficient (MCC) is a tool for binary classifiers to calculate the accuracy in a different way: it calculates the correlation between the predicted and the observed data. Unlike to the pervious metrics, MCC's outputs are ranges between −1 and +1 where +1 is the perfect prediction and −1 indicates that none of the cases were classified correctly. This metric correctly shows the accuracy even in case of high imbalance of the categories [6,7].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

2.5. Probability based accuracy (Extract ROC)

Receiver Operation Characteristic (ROC) curves provide a tool to evaluate the performance of models considering all classification thresholds. The ROC curve visualizes the true positive rate (i.e., Sensitivity) and false positive rate (1- Specificity) with different classification thresholds: lower thresholds result in more positive classifications, which also increases the false positives. Area Under the ROC curve (AUC) is an overall diagnostic value to quantify the accuracy of classification models and makes possible to compare different models [23,30]. If models are run with the option of saving the probabilities with the predictions, ROC curves and AUCs also can be derived from the data in the application. Uncertainty is also involved by setting the random fraction of the data and the number of repetitions (important note that low random fraction results in an empty ROC curve diagram due to possible lack of data in a given class). As ROC curve is for binary data, multiclass models are evaluated in a binary approach, a class versus the rest of all other classes. Figures are generated as separated and stacked plots. Stacked plots help to compare classes (for multiclass classification), and different algorithms (for binary tasks).

2.6. Experiments

2.6.1. Model building

We applied the same models for all experiments. Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB) models were run using the training data with 3-fold cross validation with 10 repetitions (RKCV) to ensure similar conditions for all experiments. Accordingly, we did not have to specify a third portion of the reference data for parameter tuning, which was performed in the RKCV phase. Model buildings were conducted in R 4.2.2 with the caret and rpart packages [28,40,48]. The aim was to train a model regardless of the thematic accuracy, and to ensure varied outputs for the evaluation.

We then performed the predictions on the testing database and applied the Extract Accuracy Rep function with the "All data" option. For case studies #1-#5 we calculated the predictions with the RF and

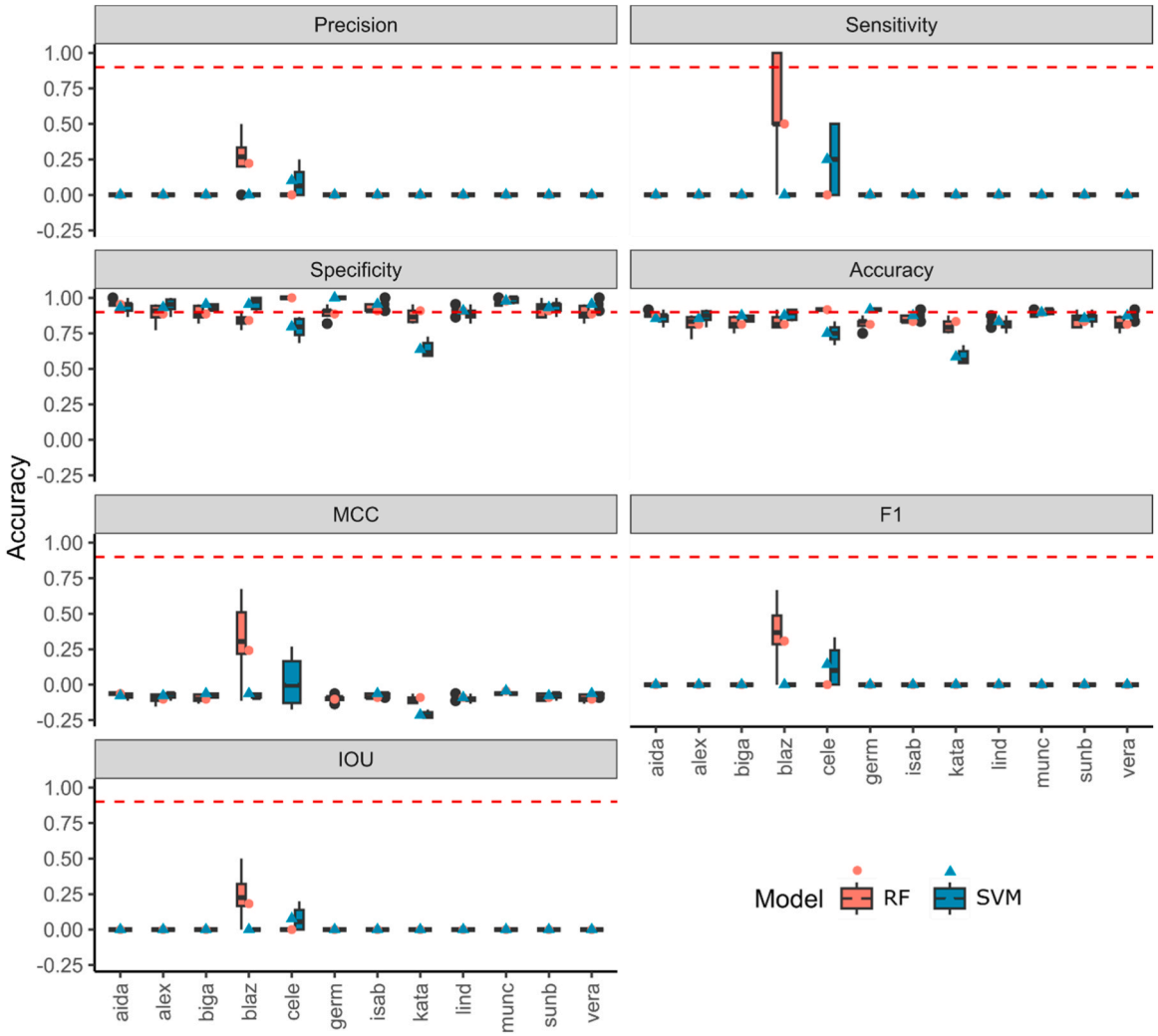

Fig. 5. Accuracy metrics of the cultivar classifications (RF: Random Forest, SVM: Support Vector Machine; boxplots are derived from the repetitions of the randomly sampled testing dataset: median, quartiles, 1.5 × interquartile ranges, black points: outliers; red and green symbols: accuracy metrics derived from the whole testing dataset; red dashed line: 90% accuracy benchmark). Cultivars are: aida: 'Aida', alex: 'Axel', biga: 'Biggareau burlat', blaz: 'Blaze Star', cele: 'Celeste', germ: 'Germersdorfi 3', isab: 'Izabella', kata: 'Katalin', lind: 'Linda', munc: 'Early Münchebergi', sunb: 'Sunburst', and vera: 'Vera'. MCC: Matthews correlation coefficient, IOU: Intersection over Union, F1: F1-score.

SVM algorithms, and for case studies #4-#5 the probabilities were also determined using only the RF and XGB. Results were evaluated by prediction models and classes.

2.6.2. Study sites and experiments

2.6.2.1. Case study #1. The example of roofs represented a binary case ('asbestos' and 'other' classes). Roofing data was collected by field observations combined with visual interpretation of the image. In the training dataset, the number asbestos and other types of pixels were not equal (368 versus 1161), but we ensured equal groups for testing (133 pixels for each group) to avoid results biased by different number of true positive and true negative cases. The involved image was a hyperspectral image taken by an AISA Eagle II sensor in 2013 August with 368 narrow bands (400 and 2400 nm), and 1 m spatial resolution. The study area was in Debrecen (NE-Hungary), in a district with older detached houses where the asbestos was a general roofing material (the most common roofing material is red, brown, and grey tiles, but asbestos cement sheet roofing is also present in large numbers). Although formerly asbestos cement was very popular due to its relatively low cost and good thermal insulation properties, it has since been discovered that asbestos fibers are carcinogenic and, thus, pose a serious health risk. Asbestos was tested against all other roofing materials during the classifications. We conducted a field survey and collected reference data using a Stonex S9 RTK GNSS device. The reference data was vectorized and split in 70–30% ratio into training and testing subsets.

2.6.2.2. Case study #2. The example of tree species identification represented a multiclass classification with 150–200 pixels of training data per species and 50 pixels of testing data per species. The study area was also in Debrecen, the selected species were planted as street trees. A set of four species were chosen: Norway maple (*Acer platanoides* L.), common linden (*Tilia x europaea* L.), London planetree (*Platanus x hybrida* Brot.) and common hackberry (*Celtis occidentalis* L.). Then, during the reference data collection, the coordinates of 943 individual tree crowns were recorded, each point representing one pixel in the satellite image for classification. The processed image was a WorldView-2 satellite image with 8 bands (400–1040 nm), and the 2 m multispectral geometric resolution was improved to 0.5 m with pan-sharpening (with the Gram-Schmidt method). The points from each category were randomly divided into 70% training and 30% testing subsets, which is a common allocation ratio for similar sample sizes.

2.6.2.3. Case study #3. The third experiment represented a multiclass

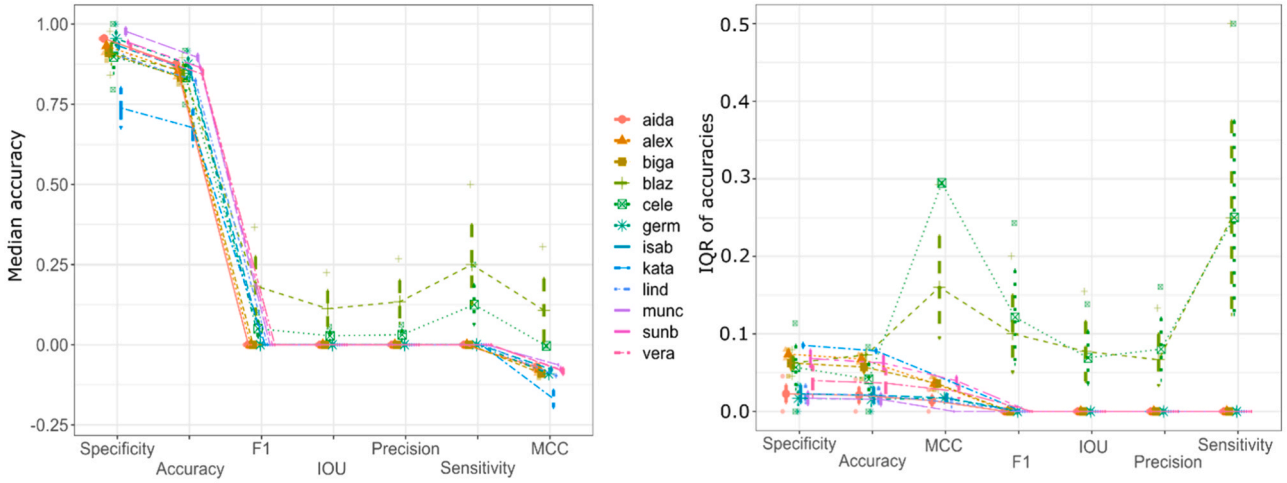

Fig. 6. Medians (a) and interquartile ranges (IQR, b) calculated from the repeated model metrics of cultivars classification (10 repetitions from 60% stratified random samples; ●,▲: median; whiskers: lower and upper quartiles). Cultivars are aida: 'Aida', alex: 'Axel', biga: 'Biggareau burlat', blaz: 'Blaze Star', cele: 'Celeste', germ: 'Germersdorfi 3', isab: 'Izabella', kata: 'Katalin', lind: 'Linda', munc: 'Early Münchebergi', sunb: 'Sunburst', and vera: 'Vera'. MCC: Matthews correlation coefficient, IOU: Intersection over Union, F1: F1-score.

classification, and the example was an experimental study with intentionally limited amount of data of two 3-year campaigns (2004, 2005, 2006 and 2013, 2014, 2015) and we aimed to discriminate the species of a sweet cherry orchard near to Debrecen, in the Experimental Fruit Station (University of Debrecen). This example represented an extreme case with many classes against low number of training data, and a resulting in a poor model. The orchard was established in 2000. Trees were pruned to super and free spindle for the two pruning systems. The orchards consisted of twelve cultivars: 'Aida', 'Axel', 'Biggareau burlat', 'Blaze Star', 'Celeste', 'Germersdorfi 3', 'Izabella', 'Katalin', 'Linda', 'Early Münchebergi', 'Sunburst', and 'Vera'. We assessed the disease incidence of cherry leaf spot (CLS) in each cultivar sub-plot in all six years, involving both pruning systems. We divided trees into four quadrants and collected 25 terminal shoots from each quadrant. We determined the average CLS incidence of all (12) cultivars at six consecutive dates in Septembers of all years and the measurements were averaged by the quadrants. Finally, we had 144 data, which we split into 60–40% into training and testing subsets with stratified random sampling to ensure the equal number of cultivars. The dependent variable was the 'cultivars', and the independent variables were the 'years', 'pruning systems', and 'CLS incidence', and, according to the data collection, only 8 and 4 data per cultivar were in the training and testing subsets, respectively. Our aim was to present an example with several target objects with only few instances.

2.6.2.4. Case study #4. The fourth experiment was performed with the 'Diabetes' benchmark dataset of the National Institute of Diabetes and Digestive and Kidney Diseases [13]. 8 variables were used as predictors (Number of times pregnant; Plasma glucose concentration; Diastolic blood pressure; Triceps skin fold thickness; 2-Hour serum insulin; Body mass index; Diabetes Pedigree Function; Age), and the target variable was the diagnosed diabetes (binary: yes/no). Results were evaluated with the confusion matrix-based metrics and the ROC curves and the AUC including the repetitions.

2.6.2.5. Case study #5. We used the iris database for the fifth case study [16]. Species (*Iris setosa, Iris versicolor*, and *Iris virginica*) were considered as the target variable, and the sepal and petal length and width were the predictors; thus, the iris dataset represented the multiclass case of ROC curve and AUC based evaluation. Furthermore, we also evaluated the classifications using the confusion matrix-based metrics.

2.7. Model evaluations

Finally, we derived three types of accuracy outputs using the developed functions: (i) single data determined from the whole testing dataset; (ii) mean and SD calculated from 10 repetitions where each represented the 60% of the testing dataset with stratified random sampling (based on the roofing type); and (iii) raw accuracy data of the 10 repetitions. Next, results were summarized on diagrams, and in tables. Beside the mean and SD, lower quartile (LQ), median, upper quartile (UQ), and interquartile range (IQR) were also determined. We emphasize that in this case the aim was not to reach the highest accuracy, but to ensure varied outputs to show the relevance of deeper accuracy assessments.

Model stability was based on the assumption that if repetitions performed on the testing data do not influence the accuracy, i.e., the IQR or SD has a narrow range, the model represents a consistent solution, regardless of the accuracy. In this term, stability is not equal to reliability, a weak model can be stable, but at a lower range of accuracy. However, if the IQR or SD has a wide range, it indicates that the outputs vary by the resampling of the testing data; accordingly, the output is not consistent. We calculated the Spearman correlation coefficients between the model medians and IQRs of the repetitions, and among the accuracy metrics. Model medians were analyzed with a General Linear Model (GLM) as a two-way factorial ANOVA including the statistical interactions (H0: there is no difference in the means of the source of the accuracy metrics, i.e., the experiments #1-#3, and the types of accuracy metrics, and there is no interaction between the two factors). We also reported the effect sizes (ω^2) as a standardized magnitude of the effect of the variables on the model [15].

3. Results

3.1. Confusion matrix-based metrics – the binary case

3.1.1. Identification of asbestos roofing - Case study #1

Asbestos classification was the binary case and according to the results, the asbestos can be discriminated from the other types of roofing materials with high accuracy. Based on the training data, the median Overall Accuracies (OAs) were 0.99 with the Random Forest (RF) and 1.00 (i.e., perfect) with the Support Vector Machine (SVM) and the testing with the independent data revealed that although trained model was good, there can be relevant number of misclassifications. Usually, all class level metrics showed that RF underperformed SVM. According

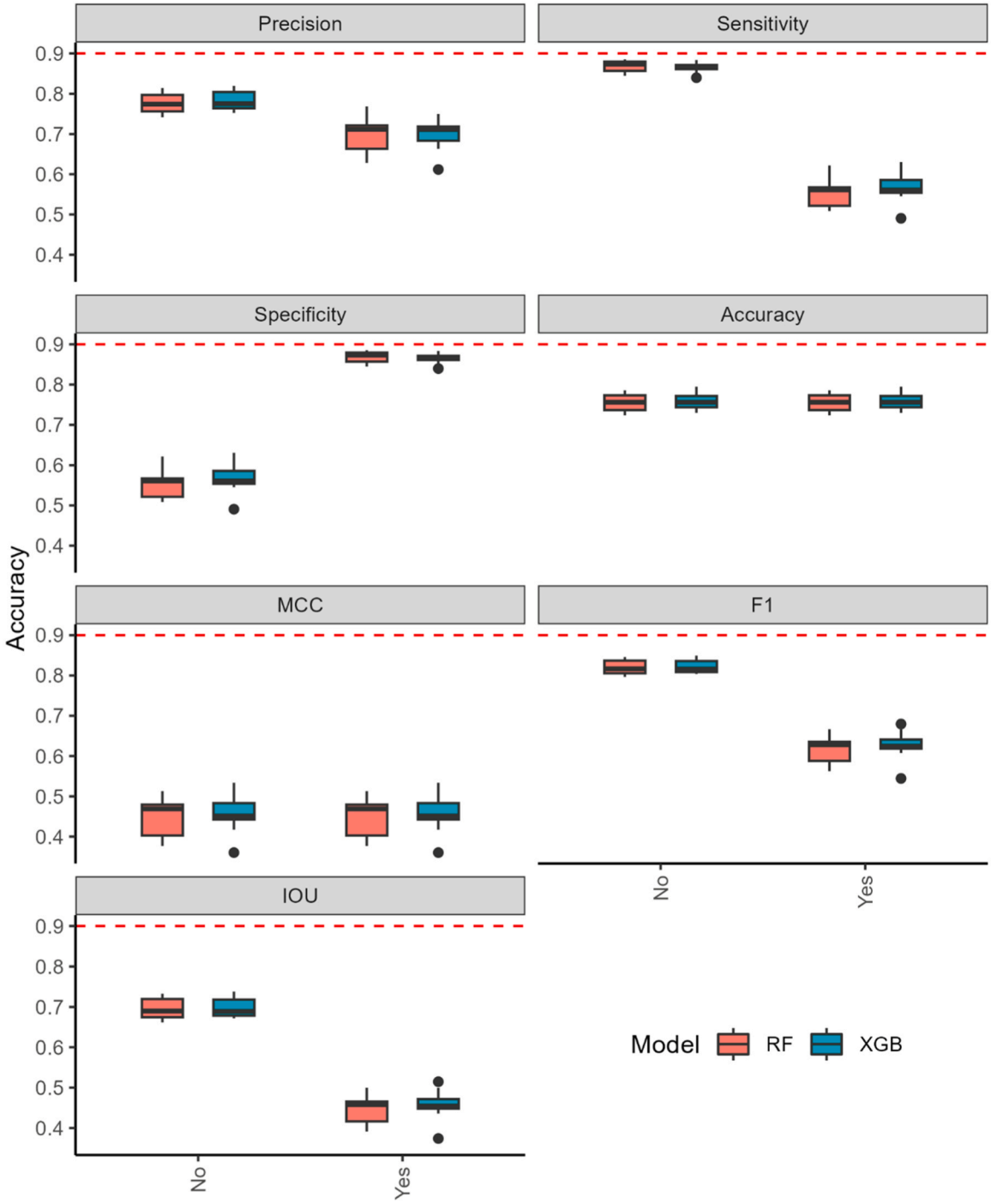

Fig. 7. Accuracy metrics of the *diabetes* classifications (boxplots are derived from the repetitions of the randomly sampled testing dataset: median, quartiles, 1.5 × interquartile ranges, black points: outliers; red and green symbols: accuracy metrics derived from the whole testing dataset; red dashed line: 90% accuracy benchmark).

to Precision, Sensitivity, and the binary class level Accuracy, only the asbestos had issues in correct classifications. Additionally, Specificity, Matthews correlation coefficient (MCC) and Intersection over Union (IOU) revealed that the 'other' (non-asbestos) class also had problems (Fig. 1). Calculations based on the whole database provided a single number, which was usually in the range of repeated metrics, but there were differences. In case of RF outputs, single numbers were above the upper quartile (UQ) of Precision, F1 and IOU, and for Sensitivity, MCC and Accuracy the metrics were pessimistic, below the lower quartile (LQ) of the repetitions. Standard deviations (SDs) and interquartile ranges (IQRs) were usually in a narrow range, largest IQR belonged to the Sensitivity and Specificity metrics (RF model, asbestos roofs), but only with 3.1% value, and the SDs were <1%, thus, the models were stable.

Correlation between the medians of repetitions and the IQRs were −0.96 ($p<0.001$), i.e., the better models had narrower IQRs, but even the wider IQRs were within 1% (Fig. 2).

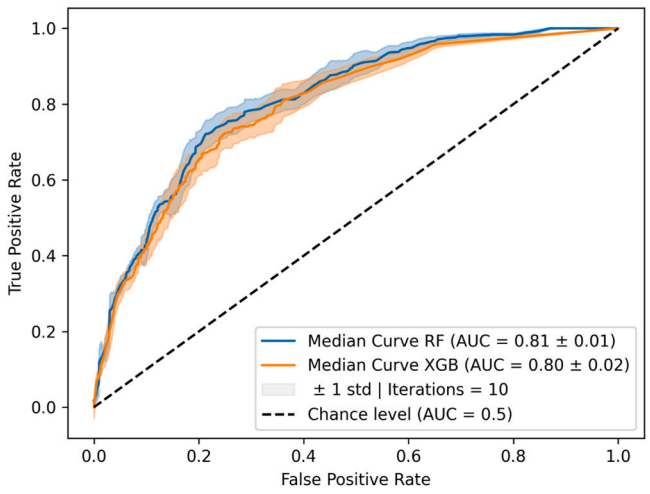

Fig. 8. ROC curve of the diabetes prediction.

3.2. *Confusion matrix-based metrics – the multiclass case*

3.2.1. *Species level identification of tree species - Case study #2*

Accuracies of tree species classification was not high, median OAs calculated from the 30 model of the cross-validation was 74% with the SVM, and 70% with the RF models. Class level metrics helped to reveal the level inaccuracies: Specificity and the binary Accuracy showed higher values (around 90%), but the other metrics, except the Precision, called the attention of possible misclassifications: due to higher level commission error (1-Precision), Precision was not able to point on the real underlying problems. All other metrics correctly identified that especially the species *Acer platanoides* and *Tilia x europaea* had concerns regarding the omission error (1-Sensitivity). In this case, the metrics derived from the whole testing dataset usually fell within the IQR range of the repetitions (Fig. 3). IQRs varied between 3% and 9%, relative SDs were between 1% and 10%. We found that, although, the SVM had better OA in the training phase, RF's MCCs' IQR values were more favorable (i.e., lesser with 1–3%). Model stabilities were similar, in a ~10% rate.

Accuracies varied by the metrics: based on the medians of the repetitions MCC and IOU had lower values, while Specificity and Accuracy showed an optimistic result (Fig. 4). Accordingly, Specificity and Accuracy had the lowest, while MCC and IOU had the highest IQRs. The correlation was −0.63 ($p<0.001$) between the model medians and IQRs.

3.2.2. *Cultivar identification - Case study #3*

Discriminating the cultivars with the given independent variables was not successful. Median OAs of the training phase were low (OA_{RF} = 0.058, OA_{SVM} = 0.028). According to the binary approach, only two cultivars ('Blaze Star' and 'Celeste') were different than the others, and the most metrics justified the weak performance. However, Specificity and the binary Accuracy showed falsely favorable results. Usually, the IQRs had a narrow range, except for the two mentioned cultivars, ensuring a case for the coincidence of low model performance and high model stability. Cultivars 'Blaze Star' and 'Celeste' had wide ranges of accuracies, which showed that these species had a chance to be identified with the input variables, but only with a low reliability (Fig. 5). The single data derived from the whole testing dataset was accordance with the repetitions (tended to 0).

The correlation between the medians of repetitions and the IQRs was low, r=0.35 ($p<0.001$), in this case it meant that low medians had narrow IQR ranges, and higher medians had wider IQRs between 2% and 10% (Fig. 6). The result highlighted that cultivars 'Blaze Star' and 'Celeste' were the only ones where a small difference could be observed, but the stability was very low due to high IQRs.

3.3. *Probability-based metrics – the binary case*

3.3.1. *Detecting diabetes disease – the binary case of ROC*

In case of the diabetes dataset, we determined both the confusion matrix-based, and the probability based metrics. Class level metrics did not vary by the models, both XGB and RF had similar performance (with <1 differences). While metrics indicated better accuracies for the non-diabetes class, MCC highlighted that both classes had issues (Fig. 7). The ROC curve and the AUC were 0.81 ± 0.01 and 0.80 ± 0.02 for the RF and the XGB, respectively, indicating a 'good' model. The related uncertainties (SDs) were small for both models, and the maximum can be identified at the average TPR and FPR cutpoints (Fig. 8). However, compared with the class level metrics, AUC seemed too optimistic.

We chose the F1-scores to evaluate the stability of the model, and the findings showed that the number of iterations and varied data fractions had no bearing on the change. In case of RF, the range was 0.706–0.729, and for XGB, the range was 0.707–0.724. Both ranges were narrow and proved that the accuracy was reliable (Fig. 9).

3.3.2. *Detecting Iris species – the multiclass case of ROC*

Iris species can be discriminated successfully based on the sepal and petal length and width. Especially, *I. setosa* can be distinguished, its classification accuracy metrics (medians) were the highest (close to 1),

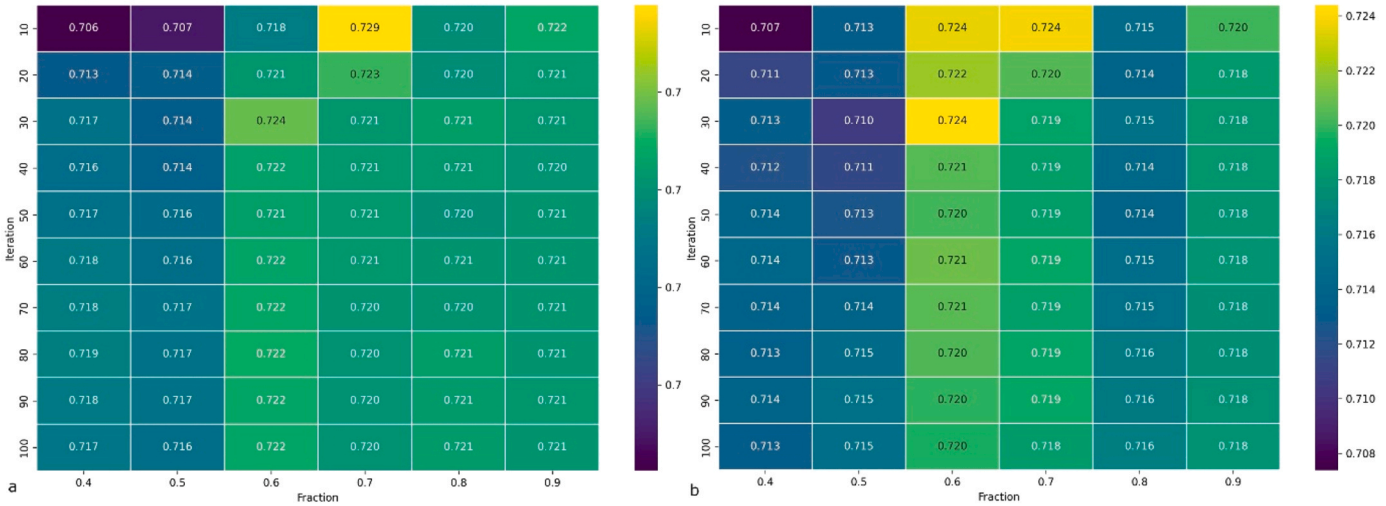

Fig. 9. Stability of the diabetes group classification by fractions and iterations for the RF (a) and XGB (b) models.

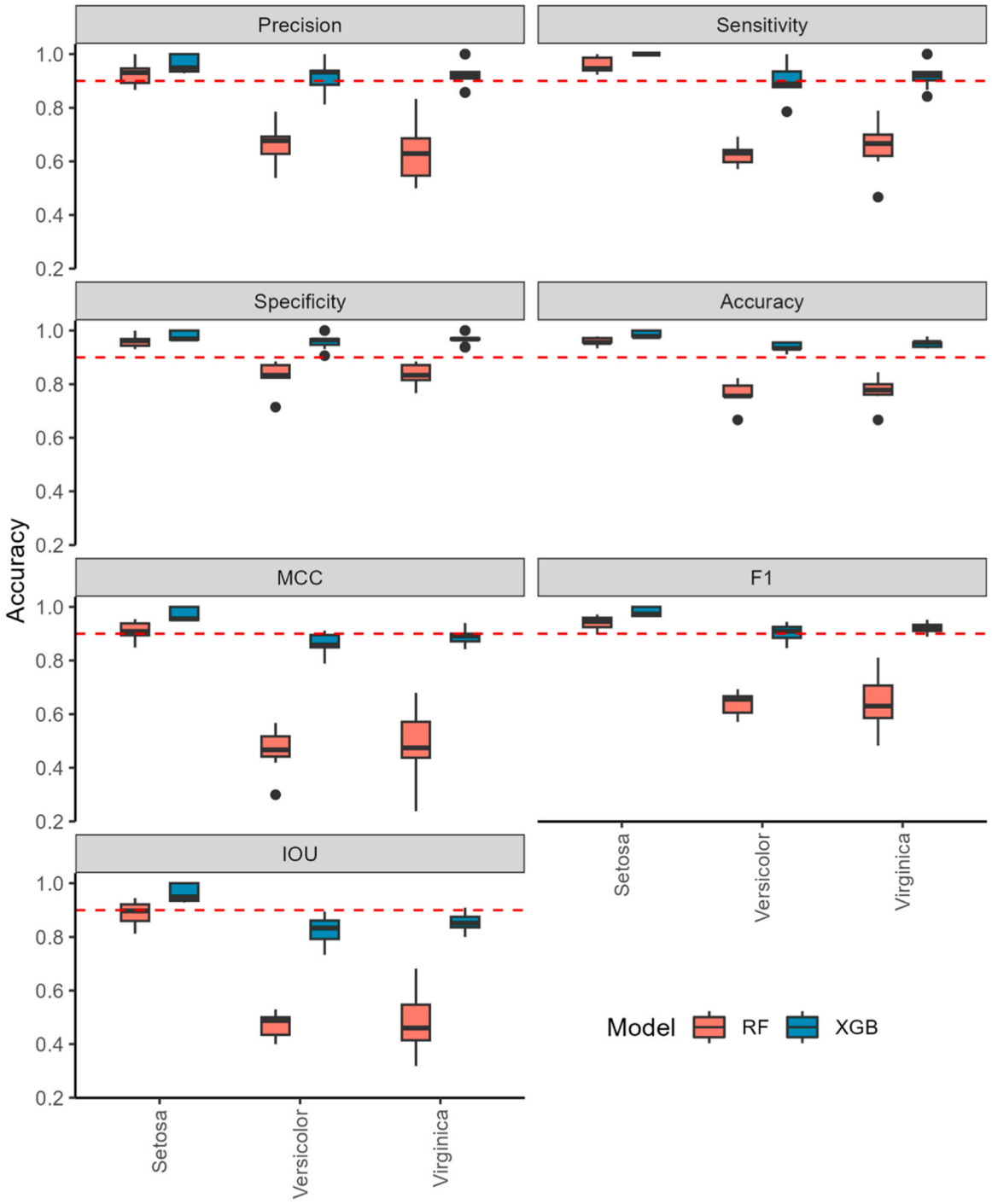

Fig. 10. Accuracy metrics of the *Iris* species classifications (boxplots are derived from the repetitions of the randomly sampled testing dataset: median, quartiles, 1.5 × interquartile ranges, black points: outliers; red and green symbols: accuracy metrics derived from the whole testing dataset; red dashed line: 90% accuracy benchmark).

while the *I. versicolor* and *I. virginica* had lower, but similar values (~0.5). IQRs varied, usually had the same range for all species except the MCC, F1, and IOU (Fig. 10). Regarding the AUC, we found similar results, but in this case, the AUC was 1.00±0.00 for the *I. setosa*, and 0.86–0.85 with ±0.04–0.02 SD for the *I. versicolor* and *I. virginica*, respectively; accordingly, it was the most optimistic result among all metrics.

Stabilty analysis, using the F1-score showed similar result to the diabetes dataset, i.e., the ranges of the outcomes were narrow for both models, 0.760–0.770 and 0.920–0.936 in case of RF and XGB, respectively (Fig. 12). Changes with the varied inputs were below 2%, thus, the accuracies can be considered reliable.

3.4. Common evaluation of the metrics

The GLM justified that both the source of the data (i.e., experiments) and the metric types, as well as their interaction were significant ($p<0.001$) and explained 98% of the variance ($F=383$, $df=20$, $SS=45.6$). Based on the effect sizes, the largest effect belonged to the source of the data ($\omega^2=0.403$, large), the metric type had the smallest

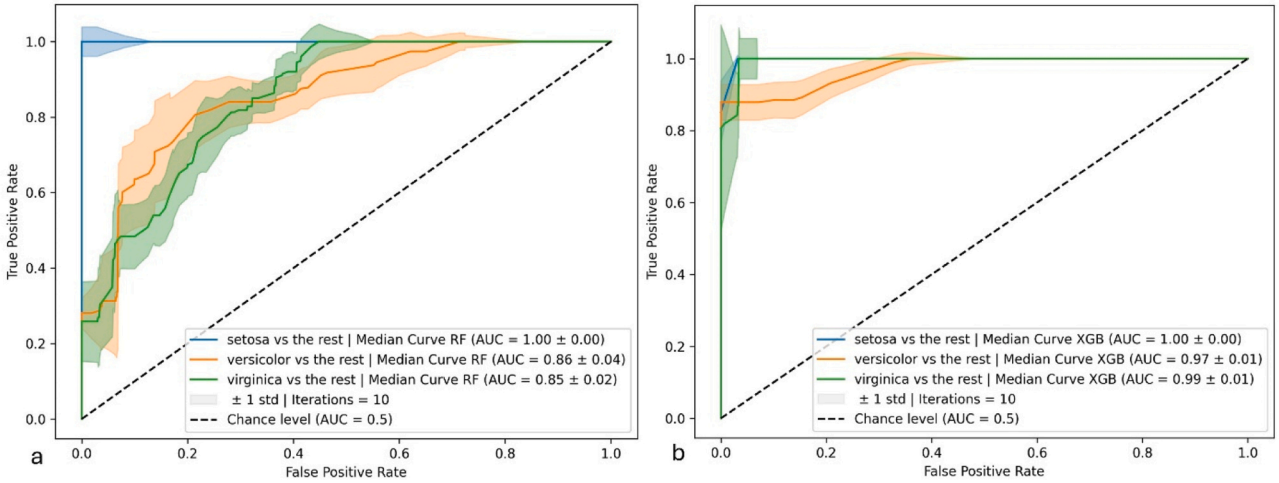

Fig. 11. ROC curve of the species prediction (a: RF model; b: XGB model).

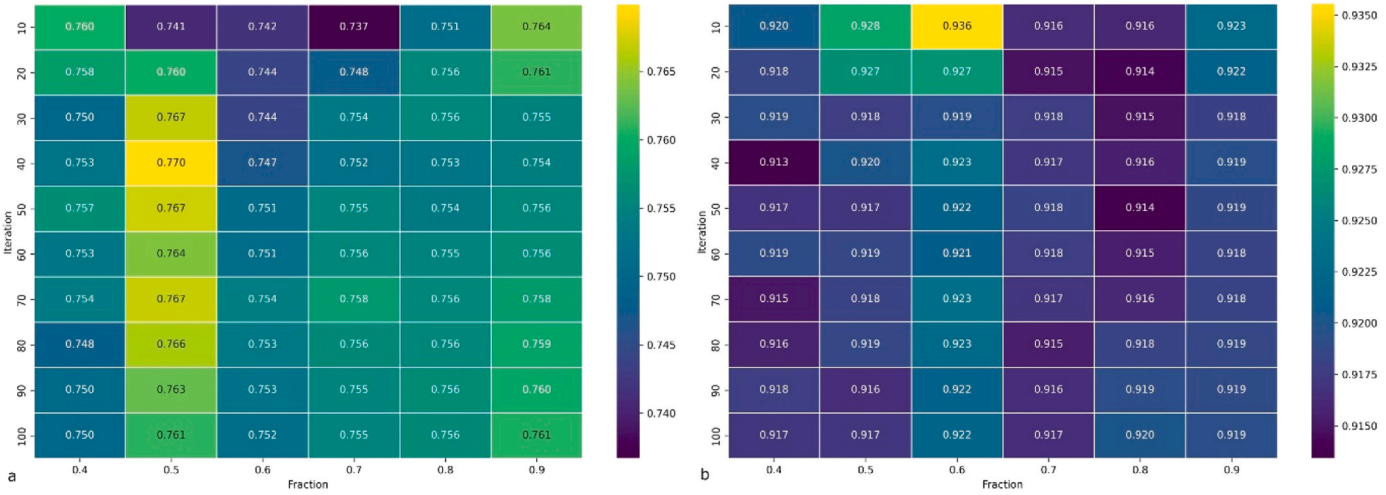

Fig. 12. Stability of the species classification by fractions and iterations for the RF (a) and XGB (b) models.

(ω^2=0.089, small), and the interaction had a medium effect (ω^2=0.137).

The correlation between the accuracy medians and IQRs, involving all experiments, was only weak, r=0.27 ($p<0.001$). Correlations among the accuracy metrics were different by the experiments (Fig. 13). In case of roof types, the minimal correlation was 0.95, and some metric-pairs were in perfect correlation. Correlations were lower among the metrics of tree species, and for some pairs the relationships were non-significant; nevertheless, we also found perfect correlations (e.g., F1 and IOU. The experiment with the cultivars provided the lowest correlations.

4. Discussion

We aimed to develop a tool that would guarantee a deeper understanding of accuracy assessments. Measuring the performance of predictive models is a crucial task, and have a wide literature in all fields of science [10,24,31,34,36,52], but usually the output is a single value calculated from the testing dataset. Although k-fold cross-validation ensures a complex evaluation based on several models, but only in the training phase, not with a completely independent dataset. Our developed application directly works with the testing data comparing the observed and predicted values and can calculate accuracy metrics using repetitions of the models; thus, provides information on model stability.

As most accuracy metrics use the TP, TN, FP, and FN values, their sensitivity to the model performance varies on the given equation (i.e., which type of prediction condition is used) [22]. As most class level metrics can be calculated as a binary index, first, we recoded the classes (if it was needed), and then we evaluated the accuracy of all classes versus all other classes merged into one. It raised the issue of imbalanced design, and when the initial number of classes were higher, the larger the imbalance was. Experiment #1 was a binary approach; therefore, the balanced set of testing data remained the same, but in case of experiment #2, with 4 classes, the imbalance was 3 times (53 vs 159 data), and in experiment #3 it was 11 times (4 vs 44). Experiment #3 shows as a special case that structured random resampling is not possible to ensure the same number of data, because 4 is a very few data, fewer than the classes we had. There are research fields where the number of reference data can be large (e.g. remote sensing); however, in other fields such as in geology, agronomy or medical sciences, data collection can have large costs or the sampling can be limited by occurrence or availability; consequently, the number of data is limited. It was the case with our agricultural experiment (#3), too. Thus, the imbalance due to non-binary classes is a natural phenomenon, and the solution is to use metrics being neutral for this experiment design. Chicco & Jurman [7] and Chatterjee et al. (2022) also pointed on this robust feature of MCC in clinical research; however, it is not used in geo-, bio or agrisciences: according to the Scopus database, in January of 2023, there is no occurrence of MCC in any journal paper in these scientific fields based

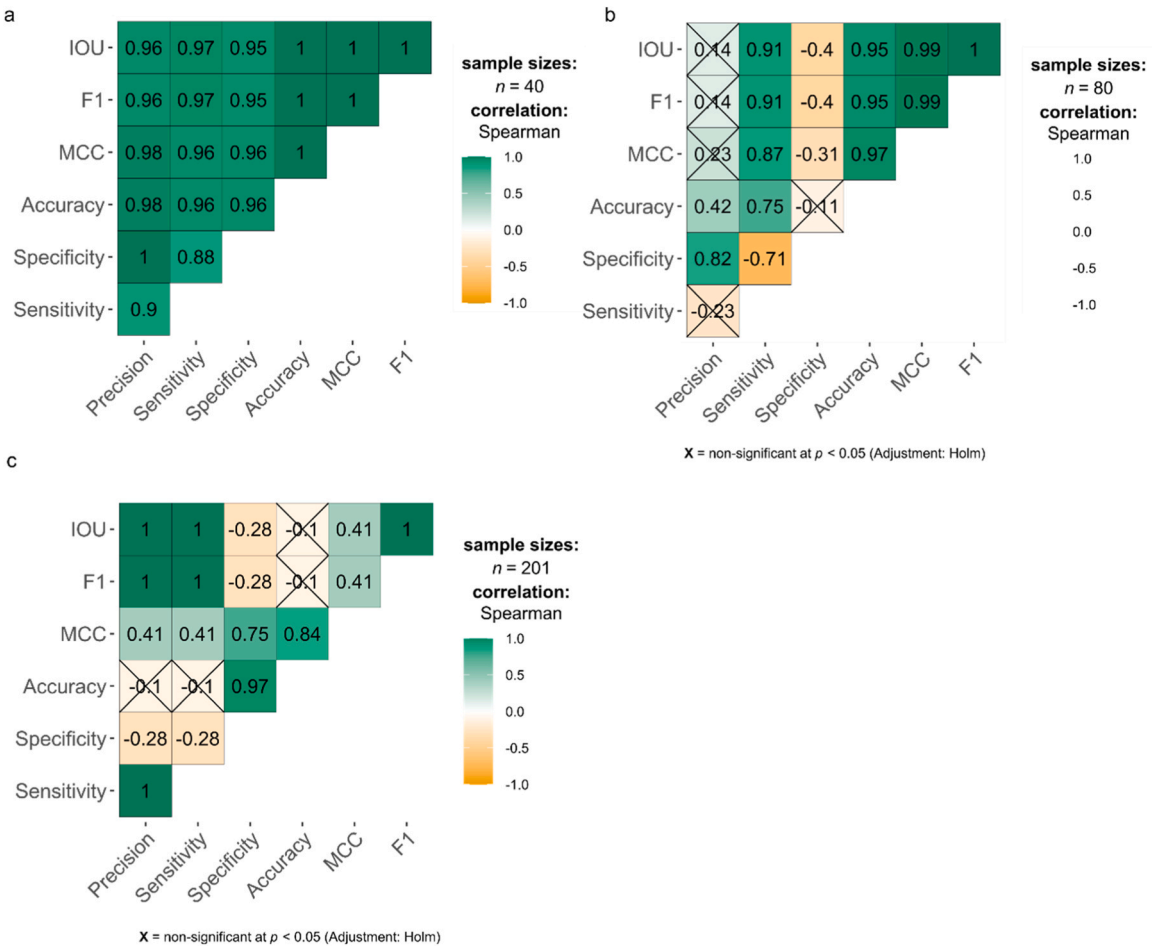

Fig. 13. Correlations among the accuracy metrics by experiments (a: roof types, Exp. #1; b: tree species, Exp. #2; c: cultivars, Exp.#3; Spearman correlation with the Holm adjustment; MCC: Matthews correlation coefficient, IOU: Intersection over Union, F1: F1-score).

on the titles, keywords or abstracts (with term of 'Matthews correlation').

We supposed that a single calculation from the testing dataset can be misleading, and when the random subsets are used in the testing, the IQRs provide a more reliable information. This hypothesis was true with the experiment #1, single values were outside the IQRs, the differences from the medians were even >10% (Fig. 1). In the other experiments, we also observed differences, but even with the worst models, the magnitudes were within the IQRs, and usually close to the medians. According to our results, Precision, Sensitivity, Specificity, or the binary Accuracy can be criticized with the distortions coming from the special type of errors when a class is misclassified and becomes dominant on the area. Consequently, the Precision will be high, there can be cases where the Sensitivity is lower, because if the model performance is only medium, the misclassifications bias these metrics. These two metrics, as the inverse of error of commission and omission, are in relationship: (i) if the commission error is low, the omission error can be high, while it is true form the other side, too: (ii) with a large commission error rate, the ratio of omissions will be low. However, we must keep in mind that FP and FN values are in relation with the number of possible cases. Therefore, while in experiment #1 the design was balanced, in the other cases (#2, and #3), the higher number of 'others' (3 times and 11 times) raised the possible issues of the testing phase. F1, IOU and MCC provided a more realistic output, values were lower, both types of errors were involved in a single number (while Precision and Sensitivity should be evaluated together). It was true in all experiments but was obvious in experiment #2 and #3 where the models had not ~99% accuracies. While in case of experiment #2 the MCC, IOU and F1 provided more reliable accuracies, in case of experiment #3 these metrics were the only acceptable solutions, because the high level of TN data, both Specificity and Accuracy were high, relevantly distorting the real results. Although in remote sensing the UA (Precision) and PA (Sensitivity) are basic class level metrics, F1, IOU and MCC would be desirable in model evaluations, [12] and Sokolova et al. [45] also suggested to use the F1. Chicco, Jurman [7] proved that MCC is not biased by the imbalance and F1, and IOU can be biased by the different number of data in the classes, but in our experiments F1 and IOU also showed realistic outcomes.

The common evaluation of the three experiments showed that based on these experiments we are not able to draw general conclusions: model medians of the repetitions and the IQRs had varying correlation by experiments (-0.96; −0.63; +0.35), but the overall correlation was low, r=0.27; thus, our third hypothesis was not confirmed. It means that the applied extreme examples (roofs with 99%, and the cultivars with almost zero accuracies) and the acceptable tree species classification with the 70–80 accuracies cannot be interpreted as an ideal set of data and calls the attention of the relevance of experiment-based characteristics. The high correlation in case of roof classification was the consequence of the good model: regardless of the random subsets of the testing data, the accuracies were high. In other words, both for the asbestos and non-asbestos roofing the trained models ensured a stable basis for the predictions, and the correlation indicated when the accuracy is high, the IQR is low (model stability is good). But it is not true in general terms, because, in case of cultivars, the models were weak, therefore, the IQRs stably showed with the narrow range that the model performances were consistently unacceptable. The correlation was positive because the small accuracies were paired with the narrow IQRs.

The outputs of the tree species experiment can be regarded a general case when the model is not perfect, 5–10% misclassification is a common phenomenon, thus, the −0.63 correlation can be a general figure, too. The IQR can be even 10%, but it only shows that both the trained model and the testing data include uncertainty and more and/or better predictors, larger training and testing dataset, better sample of the statistical population can help to reach better outputs. The example of cultivars showed that some cherry tree variants differed from the others, but with the given predictors were not appropriate for model building. Accordingly, in the testing data, using the single accuracy metrics, only one cultivar was classified accurately, but not all the 4 cases (twice with the RF, and once with the SVM model). Repetitions revealed that with other sets of data, although, not accurately, but cultivars can be found. Generally, our findings on the relation and model stability confirmed our second hypothesis.

The application also provides the option to calculate the ROC curve and the AUC with the option to perform the repetitions using the randomly selected fractions of the dataset with repetitions. AUC values of the diabetes datasets were more optimistic related to the class level metrics derived from the confusion matrix. Differences are based on the calculations: ROC-AUC is calculated from the probabilities of the classes; the class metrics are the outputs of the classifications (based on the probabilities in the background). According to [33], ROC can be misleading with imbalanced data, and for diabetes dataset we had 350 negative and 188 positive cases. However, for the iris dataset, classes had 25 cases per species (i.e. perfectly balanced), and the AUC values corresponded with the class metrics. ROC-AUC is an important measure of accuracy but should be applied only for balanced designs.

All models have uncertainty, and also the considered fractions and number of repetitions can influence the output. In our example, we determined the F1-scores, and we found that there were only slight differences in the results, the default values (60% fraction, 10 repetitions) ensured an acceptable solution. Foody [17] called the attention to the possible issues of confusion matrix, and called it the "weakest link" of accuracy assessment. The more training data we have the more reliable model can be built and the more testing data ensure a reliable accuracy assessment. Our tool provides an alternative to test if our testing data is from the statistical population (regardless of the fraction of randomly selected data and the number of repetitions provides similar accuracy metrics) or just a distorted section of the possible values and the IQRs and stability ranges are wide indicating uncertain outputs. As this solution was missing from the accuracy assessment solutions, can be a help for the users in classification tasks.

Correlations among the accuracy metrics, similarly to the median and IQR correlations, showed a varied, sometimes contradicting result, which confirmed our fourth hypothesis (correlations of accuracy metrics change by experiments). The best models with the roof type classification had 99–100% overall accuracies and most class level accuracy metrics were also around 98–100%; accordingly, the inter-correlations were high, too. Metrics of the tree species experiment were lower because the models were not perfect, especially the identification of the *Acer platanoides* yielded in ~50% accuracy, and generally the metrics varied between 70% and 90%; thus, not all the correlations were strong; moreover, there were non-significant metric-pairs. The negative sign of IOU, F1 and MCC with the Specificity indicated issues with the TN values: due to the imbalance, the metric started to show falsely good results. This phenomenon was enhanced in the experiment with the cultivars. Generally, the only metric-pair which was in perfect correlation in all the three experiments was the one with the F1 and IOU, all the other metrics had 0.8 range in their correlations through the three experiments indicating the influence of input data. This influence calls the attention to use more than one metric in the model evaluation: if correlations vary, model performance needs to check several aspects, too. Congalton [9] pointed on the relevance of eligible data for accuracy assessment, and [18] provided a method to calculate the minimum number of data as a function of standard error of the chosen confidence level and the planned accuracy. Small datasets can have the issue of non-representativeness, accordingly, the testing ideally performed with an appropriate size of data can ensure the more reliable output, which is accordance with the results of [14]. The MCC found to be a reliable metric, which provides high values only when the model performance is good. Advantage of MCC was also confirmed by [43], too, in a study of image segmentation.

The application calculates the most of the important accuracy metrics with the possibility of determining the model stability which describes the reliability of the models on class level. In the future we plan to implement further options for the repetitions (e.g., bootstrapping) and other metrics (balanced accuracy, Fowlkes-Mallows Index, markedness, informedness, prevalence threshold, [35]) to ensure a wider range of calculations.

5. Conclusions

Class level accuracy metrics are crucial factors in evaluating the model performance, and we aimed to develop a tool to automatize the calculations of several popular metrics, furthermore, we studied the main characteristics of the metrics. Our tool provided the data for the evaluations, and summarizes the accuracies by metrics, classes, and models. The tool is flexible, can be extended with further features, and ensures different requirements from single solutions to repetitions; provides bulk calculations for model stability assessment with the option to define the proportion of randomly selected data and the number of repetitions. ROC curves are also can be calculated if the users have probability data from the classification.

We confirmed that in some cases the single solutions over- or underpredict the real accuracies, which can be described as a range (e.g., IQR); ranges can reflect the model stability (the similarity of the outputs). Correlation of the model median accuracies and the IQRs was influenced by the experiments: good or weak models also can be stable; thus, the correlation can also be positive or negative. A general dataset usually provides negative correlation, i.e., the higher the accuracy, the narrower is the IQR; the model stability is higher. We justified that MCC, F1 and IOU are the most appropriate in the model performance evaluations, because these metrics provided reliable results in extreme conditions, too, the in spite of high class imbalance, and true negative cases provided realistic outputs.

CRediT authorship contribution statement

Szilárd Szabó: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Imre J. Holb:** Data curation, Formal analysis, Resources, Validation, Writing – review & editing. **Vanda Éva Abriha-Molnár:** Data curation, Project administration, Software, Validation, Visualization, Writing – original draft. **Gábor Szatmári:** Writing – original draft, Investigation, Validation. **Sudhir** Kumar **Singh:** Methodology, Writing – original draft. **Dávid Abriha:** Conceptualization, Data curation, Methodology, Software, Writing – original draft.

Declaration of Competing Interest

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

Data availability

I have shared my data and code at the attach file step.

Acknowledgements

The research was financed by the NKFI K138079. SS and DA was supported by the NKFI KKP 144068 and K138503, and IJH by the NKFI K131478. GS was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.asoc.2024.111468.

References

[1] C. A. Ramezan, T. A. Warner, A. E. Maxwell, Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification, Remote Sens 11 (2019) 185, https://doi.org/10.3390/rs11020185.
[2] A. Abdollahi, Y. Liu, B. Pradhan, A. Huete, A. Dikshit, N. Nguyen Tran, Short-time-series grassland mapping using Sentinel-2 imagery and deep learning-based architecture, Egypt. J. Remote Sens. Space Sci. 25 (2022) 673–685, https://doi.org/10.1016/j.ejrs.2022.06.002.
[3] S. Balogh, T.J. Novák, Trends and hotspots in landscape transformation based on anthropogenic impacts on soil in Hungary, 1990–2018, Hung. Geogr. Bull. 69 (2020) 349–361, https://doi.org/10.15201/hungeobull.69.4.2.
[4] N. Barnes, Publish your computer code: it is good enough, 753–753, Nature 467 (2010), https://doi.org/10.1038/467753a.
[5] Á. Barsi, Zs Kugler, I. László, Gy Szabó, H.M. Abdulmutalib, Accuracy Dimensions in remote sensing, Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII–3 (2018) 61–67, https://doi.org/10.5194/isprs-archives-XLII-3-61-2018.
[6] Cao, C., Chicco, D., Hoffman, M.M., 2020. The MCC-F1 curve: a performance evaluation technique for binary classification. https://doi.org/10.48550/arXiv.2006.11278.
[7] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.
[8] N. Codella, Q.B. Nguyen, S. Pankanti, D.A. Gutman, B. Helba, A. Halpern, J. Smith, Deep learning ensembles for melanoma recognition in dermoscopy images, IBM J. Res. Dev. (2017), https://doi.org/10.1147/JRD.2017.2708299.
[9] R.G. Congalton, A review of assessing the accuracy of classifications of remotely sensed data, Remote Sens. Environ. 37 (1991) 35–46, https://doi.org/10.1016/0034-4257(91)90048-B.
[10] R. Costache, A. Arabameri, H. Moayedi, Q.B. Pham, M. Santosh, H. Nguyen, M. Pandey, B.T. Pham, Flash-flood potential index estimation using fuzzy logic combined with deep learning neural network, naïve Bayes, XGBoost and classification and regression tree, Geocarto Int. 37 (2022) 6780–6807, https://doi.org/10.1080/10106049.2021.1948109.
[11] Z. Csatáriné Szabó, T. Mikita, G. Négyesi, O.G. Varga, P. Burai, L. Takács-Szilágyi, S. Szabó, Uncertainty and overfitting in fluvial landform classification using laser scanned data and machine learning: a comparison of pixel and object-based approaches, Remote Sens 12 (2020) 3652, https://doi.org/10.3390/rs12213652.
[12] Czakon, J., 2022. F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? [WWW Document]. neptune.ai. URL ⟨https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc⟩ (accessed 2.8.24).
[13] Diabetes Dataset [WWW Document], n.d. URL ⟨https://www.kaggle.com/datasets/mathchi/diabetes-data-set⟩ (accessed 2.8.24).
[14] B.G. Farrar, K. Voudouris, N.S. Clayton, Replications, comparisons, sampling and the problem of representativeness in animal cognition research, Anim. Behav. Cogn. 8 (2021) 273–295, https://doi.org/10.26451/abc.08.02.14.2021.
[15] Field, F., 2022. Discovering Statistics Using IBM SPSS Statistics [WWW Document]. SAGE Publ. Ltd. URL ⟨https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-ibm-spss-statistics/book257672⟩ (accessed 9.9.22).
[16] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188, https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
[17] G.M. Foody, Status of land cover classification accuracy assessment, Remote Sens. Environ. 80 (2002) 185–201, https://doi.org/10.1016/S0034-4257(01)00295-4.
[18] G.M. Foody, Sample size determination for image classification accuracy assessment and comparison, Int. J. Remote Sens. 30 (2009) 5273–5291, https://doi.org/10.1080/01431160903130937.
[19] C.I. Gedeon, M. Árvai, G. Szatmári, E.C. Brevik, T. Takáts, Z.A. Kovács, J. Mészáros, Identification and Counting of European Souslik Burrows from UAV Images by pixel-based image analysis and random forest classification: a simple, semi-automated, yet accurate method for estimating population size, Remote Sens 14 (2022) 2025, https://doi.org/10.3390/rs14092025.
[20] M. Grandini, E. Bagli, G. Visani, Metr. Multi-Cl. Classif.: Overv. (2020), https://doi.org/10.48550/arXiv.2008.05756.
[21] A. Gudmann, L. Mucsi, Pixel and object-based land cover mapping and change detection from 1986 to 2020 for Hungary using histogram-based gradient boosting classification tree classifier, Geogr. Pannonica 26 (2022), https://doi.org/10.5937/gp26-37720.
[22] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. 12 (2022) 5979, https://doi.org/10.1038/s41598-022-09954-8.
[23] Z.H. Hoo, J. Candlish, D. Teare, What is an ROC curve? Emerg. Med. J. (2017) https://doi.org/10.1136/emermed-2017-206735.
[24] G. Hu, W. He, C. Sun, H. Zhu, K. Li, L. Jiang, Hierarchical belief rule-based model for imbalanced multi-classification, Expert Syst. Appl. 216 (2023) 119451, https://doi.org/10.1016/j.eswa.2022.119451.
[25] E. Isleyen, S. Duzgun, R. McKell Carter, Interpretable deep learning for roof fall hazard detection in underground mines, J. Rock. Mech. Geotech. Eng. 13 (2021) 1246–1255, https://doi.org/10.1016/j.jrmge.2021.09.005.
[26] James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R, 1st ed. 2013, Corr. 7th printing 2017 edition. ed. Springer, New York.
[27] A. Kornejady, A. Goli Jirandeh, H. Alizadeh, A. Sarvarinezhad, A. Bameri, L. Lombardo, C. Conoscenti, A. Alizadeh, M. Karimi, M. Samadi, E. Silakhori, Chapter 38 - Doing more with less: A comparative assessment between morphometric indices and machine learning models for automated gully pattern extraction (A case study: Dashtiari region, Sistan and Baluchestan Province), in: H. R. Pourghasemi (Ed.), Computers in Earth and Environmental Sciences, Elsevier, 2022, pp. 523–534, https://doi.org/10.1016/B978-0-323-89861-4.00007-5.
[28] Kuhn, M., 2022. caret: Classification and Regression Training.
[29] S.B. Likó, L. Bekő, P. Burai, I.J. Holb, S. Szabó, Tree species composition mapping with dimension reduction and post-classification using very high-resolution hyperspectral imaging, Sci. Rep. 12 (2022) 20919, https://doi.org/10.1038/s41598-022-25404-x.
[30] J.N. Mandrekar, Receiver Operating Characteristic Curve in Diagnostic Test Assessment, J. Thorac. Oncol. 5 (2010) 1315–1316, https://doi.org/10.1097/JTO.0b013e3181ec173d.
[31] C. Martinello, C. Cappadonia, C. Conoscenti, E. Rotigliano, Landform classification: a high-performing mapping unit partitioning tool for landslide susceptibility assessment—a test in the Imera River basin (northern Sicily, Italy), Landslides 19 (2022) 539–553, https://doi.org/10.1007/s10346-021-01781-8.
[32] M. Mohajane, R. Costache, F. Karimi, Q. Bao Pham, A. Essahlaoui, H. Nguyen, G. Laneve, F. Oudija, Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area, Ecol. Indic. 129 (2021) 107869, https://doi.org/10.1016/j.ecolind.2021.107869.
[33] F. Movahedi, R. Padman, J.F. Antaki, Limitations of receiver operating characteristic curve on imbalanced data: assist device mortality risk scores, J. Thorac. Cardiovasc. Surg. 165 (2023) 1433–1442.e2, https://doi.org/10.1016/j.jtcvs.2021.07.041.
[34] S. Nageswaran, G. Arunkumar, A.K. Bisht, S. Mewada, J.N.V.R.S. Kumar, M. Jawarneh, E. Asenso, Lung cancer classification and prediction using machine learning and image processing, BioMed. Res. Int. 2022 (2022) 1755460, https://doi.org/10.1155/2022/1755460.
[35] S. Okyay, S. Aygun, Experimental interpretation of adequate weight-metric combination for dynamic user-based collaborative filtering, PeerJ Comput. Sci. 7 (2021) e784, https://doi.org/10.7717/peerj-cs.784.
[36] C. Oşlobanu, M. Alexe, Built-up area analysis using sentinel data in metropolitan areas of Transylvania, Romania, Hung. Geogr. Bull. 70 (2021) 3–18, https://doi.org/10.15201/hungeobull.70.1.1.
[37] M. Onishi, T. Ise, Explainable identification and mapping of trees using UAV RGB image and deep learning, Sci. Rep. 11 (2021) 903, https://doi.org/10.1038/s41598-020-79653-9.
[38] K. Phinzi, D. Abriha, S. Szabó, Classification Efficacy Using K-Fold cross-validation and bootstrapping resampling techniques on the example of mapping complex gully systems, Remote Sens 13 (2021) 2980, https://doi.org/10.3390/rs13152980.
[39] Powers, D., 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach Learn Technol 2.
[40] R Core Team, 2022. R: a language and environment for statistical computing [WWW Document]. URL ⟨https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing⟩ (accessed 9.9.22).
[41] K. Rhodes, V. Sagan, Integrating remote sensing and machine learning for regional-scale habitat mapping: advances and future challenges for desert locust monitoring, IEEE Geosci. Remote Sens. Mag. 10 (2022) 289–319, https://doi.org/10.1109/MGRS.2021.3097280.
[42] S. Saha, B. Bera, P.K. Shit, S. Bhattacharjee, N. Sengupta, Prediction of forest fire susceptibility applying machine and deep learning algorithms for conservation priorities of forest resources, Remote Sens. Appl. Soc. Environ. 29 (2023) 100917, https://doi.org/10.1016/j.rsase.2022.100917.
[43] A.W. Setiawan, Image Segmentation Metrics in Skin Lesion: Accuracy, Sensitivity, Specificity, Dice Coefficient, Jaccard Index, and Matthews Correlation Coefficient. 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Presented at the 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), 2020, pp. 97–102, https://doi.org/10.1109/CENIM51130.2020.9297970.
[44] A. Shimazaki, D. Ueda, A. Choppin, A. Yamamoto, T. Honjo, Y. Shimahara, Y. Miki, Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method, Sci. Rep. 12 (2022) 727, https://doi.org/10.1038/s41598-021-04667-w.
[45] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, in: A. Sattar, B. Kang (Eds.), AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2006, pp. 1015–1021, https://doi.org/10.1007/11941439_114.
[46] M. Story, Accuracy assessment: a user's perspective, Photogramm. Eng. Remote Sens. 52 (1986) 397–399.

[47] L. Szabó, D. Abriha, K. Phinzi, S. Szabó, Urban vegetation classification with high-resolution PlanetScope and SkySat multispectral imagery, Landsc. Environ. 15 (2021) 66–75, https://doi.org/10.21120/LE/15/1/9.
[48] Therneau, T., Atkinson, B., 2022. rpart: Recursive Partitioning and Regression Trees.
[49] R.H. Topaloğlu, G.A. Aksu, Y.A.G. Ghale, E. Sertel, High-resolution land use and land cover change analysis using GEOBIA and landscape metrics: a case of Istanbul, Turkey, Geocarto Int 0 (2021) 1–27, https://doi.org/10.1080/10106049.2021.2012273.
[50] O.G. Varga, Z. Kovács, L. Bekő, P. Burai, Z. Csatáriné Szabó, I. Holb, S. Ninsawat, S. Szabó, Validation of visually interpreted corine land cover classes with spectral values of satellite images and machine learning, Remote Sens 13 (2021) 857, https://doi.org/10.3390/rs13050857.
[51] Williams, G., 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, 2011th edition. ed. Springer, New York.
[52] H. Yang, L. Chen, Z. Cheng, M. Yang, J. Wang, C. Lin, Y. Wang, L. Huang, Y. Chen, S. Peng, Z. Ke, W. Li, Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study, BMC Med 19 (2021) 80, https://doi.org/10.1186/s12916-021-01953-2.
[53] G. Zhang, M. Wang, K. Liu, Deep neural networks for global wildfire susceptibility modelling, Ecol. Indic. 127 (2021) 107735, https://doi.org/10.1016/j.ecolind.2021.107735.