COMPARISON OF MACHINE LEARNING CLASSIFIERS FOR MULTITEMPORAL AND MULTISENSOR MAPPING OF URBAN LULC FEATURES

Y. Ouma[1,*], B. Nkwae[1], D. Moalafhi[1], P. Odirile[1], B. Parida[2], G. Anderson[1], J. Qi[3]

[1] University of Botswana, Botswana – (oumay, nkwaeb, moalafhid, odirilep, andersong)@ub.ac.bw
[2] Department of Civil and Environmental Engineering, BIUST, Botswana - paridab@biust.ac.bw
[3] Center for Global Change and Earth Observations, Michigan State University, USA - qi@msu.edu

KEY WORDS: Landsat sensors, Urban LULC features, Multitemporal, Multisensor, Machine learning, Classification feature fusion

ABSTRACT:

This study compares four machine-learning algorithms comprising of Classification And Regression Trees (CART), Random Forest (RF), Gradient Tree Boosting (GTB) and Support Vector Machine (SVM) for the classification of urban land-use and land-cover (LULC) features. Using multitemporal and multisensor Landsat data from 1984-2020 at 5-year intervals for the Greater Gaborone Planning Area (GGPA) in Botswana, the aim of the study is to determine the performance of the classifiers in the extraction of different urban LULC features as built-up, bare-soil, water, grass, shrubs and forest. The results show that for mapping built-up areas, RF and SVM presented the best results with overall accuracy of 85%. Bare soil is best mapped using RF and CART with accuracy of up to 98%, while SVM and GTB were most suitable for mapping water bodies. The suitable classifiers for mapping the vegetation classes were RF for grass (94.5%), SVM for shrubland (81.5%) and GTB for forest (84.3%). In terms of class specific accuracy, RF achieved the highest performance with average overall accuracy (OA) of 95.9%, SVM (95.8%), GTB (95.6%) and CART (95.1%). The same performance pattern was observed from the F1-score, True Positive Rate (TPR), False Positive Rate (FPR) and Area under ROC curve (AUC) metrices for the class classification accuracies. The overall accuracy for the eight-epoch years were RF (87.8%), SVM (87.5%), GTB (86.4%) and CART (85.3%). To improve on the urban LULC mapping, the study proposes the post-classification feature fusion of the best classifier results.

1. INTRODUCTION

The accurate extraction of land-use and land-cover (LULC) information in urban environments is important in the provision of the critical input for environmental planning and ecological management (Fan et al., 2007). For urban LULC mapping and change detection, remote sensing data provides the optimal spatial and temporal data sources. However, the extraction of urban LULC features is often a challenging task due to the high degree of interactions and complexities within the urban LULC classes or features in terms of their spectral, spatial and textural properties (Ouma and Tateishi, 2008; Blaschke et al., 2014). Due to these factors, the applications of traditional pixel-based classification approaches in urban LULC mapping often leads to unsatisfactory results (Johnson and Xie 2013).

To overcome the drawbacks in pixel-wise classifications, Blaschke et al. (2014) proposed the Geographic Object-Based Image Analysis (GEOBIA) focusing on the segmentation of very high-spatial resolution (VHR) image data. Through GEOBIA segmentation, pixels are grouped into similar and semantically independent image segments or objects for feature extraction and classification. However, the GEOBIA approach only performs well only in VHR image data (Johnson and Xie 2013).

At medium-resolution and low-resolution image data, methods comprising of unsupervised algorithms, parametric supervised and machine learning methods have been proposed for LULC mapping (Orieschnig et al., 2021). Amongst others, the supervised classifiers comprise of maximum likelihood classifier, Mahalanobis distance, k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest (RF), Decision Trees (DT), Spectral Angle Mapper (SAM), fuzzy logic, fuzzy Adaptive Resonance Theory-Supervised Predictive Mapping (Fuzzy-ARTMAP), Radial Basis Function (RBF), Artificial Neural Networks (ANN), Naive Bayes (NB), etc (Shih et al., 2019). The unsupervised classifiers include among other methods: fuzzy *c*-means, *k*-means algorithm, Affinity Propagation clustering algorithm, ISODATA techniques (Maxwell et al., 2018).

In particular, the application of machine leaning (ML) algorithms for LULC mapping have recently attracted considerable research interests. This is mainly because machine learning algorithms do not require hypotheses on the input data distribution and tend to yield better results than the traditional parametric classifiers (Johnson and Xie 2013). Different ML algorithms have been used for LULC mapping and modelling (e.g., Talukdar et al., 2020) and have also been compared (e.g., Camargo et al., 2019). However, each machine learning algorithm will yield different accuracy levels for specific case study and data and more so for specific LULC class or feature. Further, in addition to the quality and quantity of the imagery, the choice of the suitable machine learning classifier is still a challenge as the classifier and its implementation and hyperparametrization may influence the LULC mapping results and so will the temporal variabilities and sensor characteristics (Nichols et al., 2019).

For urban LULC classification, different studies have compared different machine leaning classifiers for their accuracy, but not necessarily in terms of their mathematical and functional approach and for the extraction of specific classes within a scene (Camargo et al., 2019). For example, Ghosh and Joshi (2014), in classifying urban landscapes using Landsat data, indicated that SVM and RF produced similar classification

* Corresponding Author

results. Khatami et al. (2016) also found that SVM, RF and kNN outperformed the traditional supervised classifiers. Pouteau et al. (2011) compared RF, SVM, kNN, Naïve Bayes, C4.5 and Boosted Regression Tree for different datasets and concluded that kNN performed better for the classification of urban LULC from Landat-7 ETM+ data for different Landsat sites. Further, Heydari and Mountrakis (2018) compared five ML classification algorithms: SVM, kNN, Naïve Bayesian, Tree ensemble and artificial Neural Networks (ANN), with the conclusion that SVM and kNN were the best classifiers in the classification of Landsat data. From the review, RF, SVM and ANN models have been reported to provide higher overall accuracy in LULC modelling as compared to the traditional classification techniques (Carranza-García et al., 2019). The nonparametric ML algorithms are considered superior as they do not relay on a priori hypotheses of the input data distribution (Nery et al., 2016). However, results from different case studies have demonstrated that the performance of a given ML classifier is not only specific to the case study, but also influenced by the setup of the ML model and the quality of the training data.

Further, while several studies have been conducted on urban LULC mapping using machine algorithms (Dutta et al., 2019), the performance of the models cannot be replicated from one case study to another, and most studies do not focus on classifier-class performance rather the emphasis is on the overall LULC classification accuracy. Secondly most of the studies are based on single-date imagery and not on the multitemporal imagery with varied sensor characteristics. Besides, previous studies have also pointed out that the performance of the ML classifiers in LULC classification are affected by the limitations in the spectral and spatial resolutions of the sensors especially at medium- and low-resolutions (Pal and Talukdar et al., 2018).

With focus on the open-source solutions, this study evaluates the performs of CART, RF, gradient decision tree boosting (GTB) as decision-tree based machine learning classifiers, with SVM classifier as the benchmark ML classifier implemented in the Google Earth Engine (GEE) platform. ANN was not compared in the current study as its implementation requires external model training on the TensorFlow platform (Abadi et al., 2016), which requires additional cost-based components of Google Cloud. Thus SVM is considered for comparison in the current study as its results have been comparable to ANN (Carranza-García et al., 2019).

For the mapping of built-up, water, grass, shrubs, forest and bare-soil urban LULC classes, the objectives of this study are: (1) to implement decision-tree based CART and ensemble RF and GTB classifiers and compare the results with SVM for urban LULC mapping from multitemporal and multisensor Landsat data from 1984 to 2020 at 5-year intervals, and (2) to evaluate the performance of the classifiers for different urban LULC classes at different temporal intervals from multisensor Landsat. Using medium-spatial resolution data, the main contribution of the current study is in the determination of the suitability of decision-tree based and SVM machine learning classifiers in the extraction in individual urban LULC classes from multisensor and multitemporal image data.

2. STUDY AREA AND DATA

2.1 Study area

The GGPA is located between 20° 30′S and 24° 45′S and 25° 50′E and 26° 12′E (Figure 1) and occupies an area of 961.73 km^2. Within the commuting radius of the Gaborone city, a dormitory of suburbs are rapidly developing which are mostly characterized by centripetal movement of rural–urban migrations.

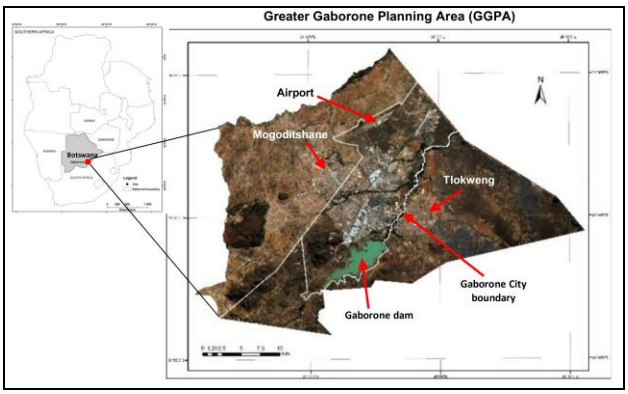


Figure 1. Location of study area and RGB image of the Greater Gaborone Planning Area (GGPA) in Botswana.

2.2 Data

Multitemporal Landsat data from Landsat 4 (L4-MSS), (Landsat 5 (L5-TM), Landsat 7 (L7-ETM+) and Landsat 8 (L8 OLI) acquired between 1984-2020 were acquired for the study area at 5-year temporal intervals. The study utilized the blue, green, red, NIR, SWIR1 and SWIR2 multispectral imagery as available for the respective study years from the USGS data portal (https://earthexplorer.usgs.gov/). The multitemporal and multisensor Landsat imagery were atmospherically corrected using the ATCOR2 tool and histogram equalization in ERDAS Imagine. The timeseries Landsat bands were mosaiced, composited, resampled to 30 m resolution, and clipped to the study area.

The urban LULC classes comprised of built-up (residential, commercial, industrial and impervious surfaces); bare-land (soil cover), water, and vegetation cover (grass, shrubs and forest). Figure 2 presents the spectral reflectance trend for the six urban LULC classes in the Landsat's visible, NIR and SWIR bands. The training and testing data samples were collected from visual identification and interpretations from the Landsat imagery, Google Earth high-resolution imagery and the historical LULC maps and based on the size and homogeneity of the study area. For each year, the training samples were collected in polygons with each polygon comprising of 50 pixels. For all the LULC classes except water, 120 polygons were used for training and 50 polygons for validation of the results. For the water class, due to its being smaller in size, the training and validation comprised of 70 and 30 polygons respectively.

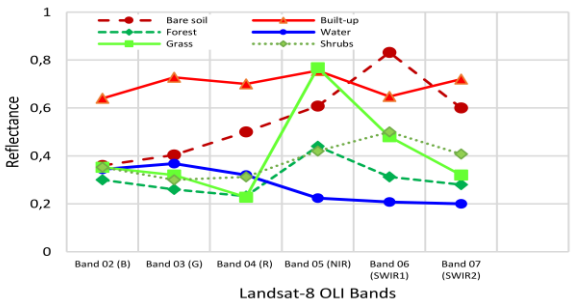


Figure 2. Spectral profiles for urban LULC classes based on the Landsat sensors.

3. METHODS

3.1 Methods

Decision-tree based machine learning classifiers have been considered among the best classifiers. To increase their accuracy, the combination of decision-trees, ensembles, have been preferred. Random forests and boosting are among two strategies for combining decision trees. This section presents a brief background to the CART decision-tree, RF and GTB ensemble DTs and SVM machine learning algorithms.

3.1.1 Classification and Regression Trees (CART): CART is a typical decision-tree (DT) model which explores the structure of data, while evolving to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome. The decision at each internal node (Figure 3), is assessed by information gain or entropy to compare the value of attributes in the data from the root to each of the leaves. The nodes of a DT tree have multiple levels with the top node as the root node. Depending on the test outcome, the classification algorithm branches towards the appropriate child node where the process of test and branching repeats until it reaches the leaf node. Figure 3 shows the classification tree structure for sample three class labels, two predictors within a rectangular partition feature space *X*. At the intermediate node, a classification goes to the left child if and only if the condition is satisfied. The leaf or terminal nodes correspond to the decision outcomes or the predicted class. CART has several advantages as simplicity in interpretation, fast to execute and shows better accuracy for image classification. However, the algorithm suffers from overfitting in the decision-tree. CART can also create over-complex trees which cannot generalize the data well. Because of their low variance and high predictive accuracy, in many domains the use of CART has largely been supplanted by resampling ("ensemble") methods that address CART's potential instability by averaging the results of many trees.

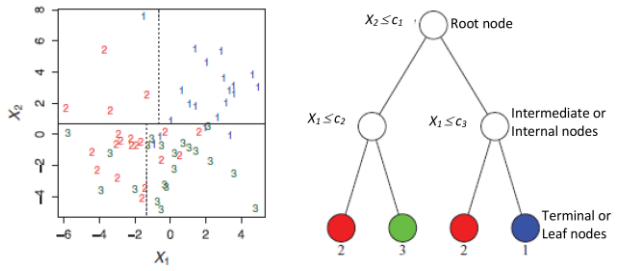


Figure 3. CART: partitions (left) and decision-tree structure (right) for a classification tree model with three classes *c1*, *c2* and *c3*.

3.1.2 Random Forest (RF): RF are an ensemble of k untrained decision trees with only a root node and M bootstrap samples. In its implementation, different parts of the training datasets are used to train the different DTs (Figure 4). For the classification of a new sample, the input vector of the sample is required to pass down with each DT of the forest. Each DT then considers a different part of that input vector and gives a classification outcome. The forest then chooses the classification of having the most 'votes' (for discrete classification outcome) or the average of all trees in the forest (for numeric classification outcome). To reduce the correlation between the estimators, RF are trained using a variant of the random subspace method, which is a method of training multiple RF models by randomly sampling the initial feature space. Since the RF algorithm considers the outcomes from many different DTs, it can reduce the variance resulting from the consideration of a single DT for the same dataset

The advantage of RF is that it can produce stable, robust and accurate results even with minimal tuning of the hyperparameters. The algorithm is easy to parameterize, insensitive to overfitting and deals with outliers in training data, reporting the classification error and variable significance. Further, RF is able to process multidimensional features from both continuous and categorical datasets. The biggest disadvantage of random forests is that the analysis, which aggregates over the results of many bootstrap trees, does not produce a single, easily interpretable tree diagram. Deep DTs can cause overfitting of the training data, resulting into the variation in the classification outcome for any small change in the input training data. The This implies the DTs are sensitive to the training data, which makes them error-prone to the test dataset.

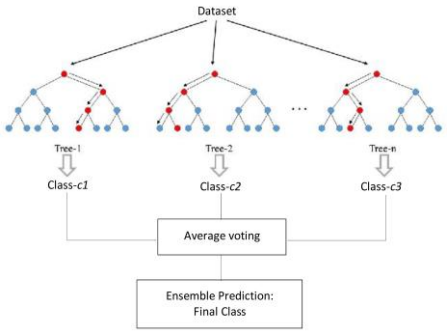


Figure 4. Classification concept based on RF ensemble.

3.1.3 Gradient Tree Decision Boosting (GTB): The algorithm attains its classification accuracy by the iterative combination of weak learner ensembles into stronger ensemble of trees through stepwise minimization of the loss function based on the gradient descent optimization (Friedman 2002). GTB like RF, aggregates an ensemble of decision trees (Figure 5). GTB however confines individual trees to a weaker prediction model hence limiting the complexity of the decision trees. As shown in Figure 5, the model constructed by weaker prediction F_m can be modified to become stronger by adding new tree (F_{m+1}). In the next iteration step a new model F_{m+1} is constructed using m+1 trees and it corrects its predecessor F_m. F_{m+1} is then boosted to model F_{m+2} in next iteration step and the consecutive error correction ultimately leads to a model providing the most accurate classification.

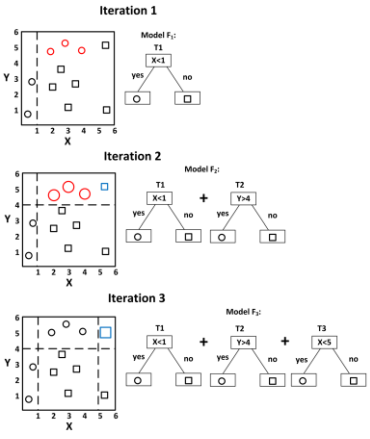


Figure 5. Visualizing gradient tree decision boosting.

The difference between GTB and other ensemble learning algorithms, is that it fits the residual of the regression tree at each iteration using negative gradient values of loss. GTB complements the weak learning DTs, thus improves the ability of representation, optimization, and generalization. GTB can capture higher-order information, is invariant to scaling of sample data and can effectively avoid overfitting by weighting combination scheme.

3.1.4 Support vector machine (SVM): In classifying linear and non-linear data, the SVM algorithm first maps the n-feature data items into an n-dimensional feature space. An optimal decision hyperplane that separates the data items into two classes is established such that the marginal distance between classes is maximized and the classification errors are minimized. The class marginal distance is the distance between the decision hyperplane and its nearest instance which is a member of that class, and the classification is performed when the hyperplane differentiates any two classes by the maximum margin as illustrated in Figure 6.

If x is the input feature vector, w is the weight vector and b is the bias, the aim of training in SVM model is to determine the w and b so that the hyperplane separates the data and maximizes the margin $1/\|w\|^2$. Vectors x_i for which $|y_i|\left(wx_i^T+b\right)=1$ will be termed support vector. The main advantage of the SVM is in the ability to overcome the high dimensionality problem, with a high discriminative power for classification.

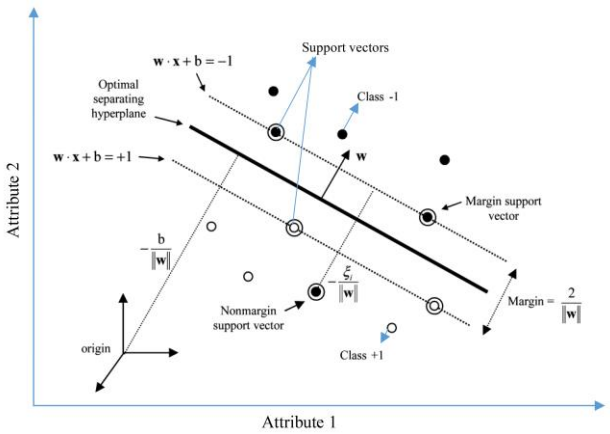


Figure 6: Maximum margin-minimum norm classifier in SVM with optimal decision hyperplane for non-separable classes.

3.2 Performance Evaluation

From the classification confusion matrix, the Producer Accuracy (PA), User Accuracies (UA), True Positive Rate (Recall/Sensitivity (TPR)), False Negative Rate (FNR), True Negative Rate (Specificity (TNR)) and False Positive Rate (FPR) are used to determine the feature extraction accuracies. The overall accuracy (OA), Kappa coefficient and F1-score (F-measure) were used to compare the overall classification performances by the machine learning algorithms.

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FNR = \frac{FN}{TP + FN} \quad (2)$$

$$TNR = \frac{TN}{TN + FP} \quad (3)$$

$$FPR = \frac{FP}{TN + FP} \quad (3)$$

$$F1 = 2 * \frac{UA * PA}{UA + PA} = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} \quad (4)$$

In addition, the Receiver Operating Characteristic (ROC) curve is derived to evaluate the prediction performance based on the sensitivity and specificity, and the area under the ROC curve (AUC) is also calculated. For statistical significance detection, the differences in the classification accuracy between the classifiers was evaluated using pairwise *z*-score test. The *z*-test was applied to the OA results for testing statistical significance at a significance level of 5%. If $z > 1.96$, the test is significant, leading to the conclusion that the obtained results from the compared classifiers differ from each other.

4. RESULTS

4.1 PA and UA for urban LULC class mapping

4.1.1 Urban built-up: Figure 7 shows that the PA values are generally higher and more consistent than the UA measures. From 1984 to 2020, the average PA measures for the built-up areas were determined as: RF had the highest average of 98.5%, followed by SVM (96.5%), GTB (96.3%) and lastly CART (95.8%). Considering the UA measure, SVM and RF had the highest values of 85.3% and 85% respectively, while CART and GTB also had nearly the same UA values of 79.9% and 79.6% respectively. The results imply that the average PA was 14.3% higher than the average UA, with SVM and RF as the best classifiers for mapping urban built-up areas.

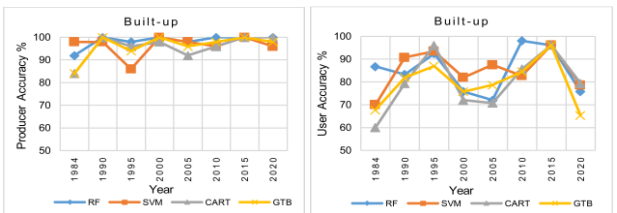


Figure 7. PA and UA results for urban built-up.

4.1.2 Bare soil: In mapping the bare-soil cover, all the classifiers achieved lower PA as compared to the UA values (Figure 8). This could be attributed to the spectral confusion of bare-soil with the impervious surfaces including buildings and roads. For mapping bare-soil, RF attained average PA of 87.8% which was 1.5%, 3.5% and 6.3% respectively higher than SVM, GTB and CART. The UA for bare soil was highest for RF (98.4%), followed by CART (98%), SVM (96.5%) and GTB (95.3%). From the UA results, bare-soil within urban areas can be accurately mapped using RF and CART classifiers.

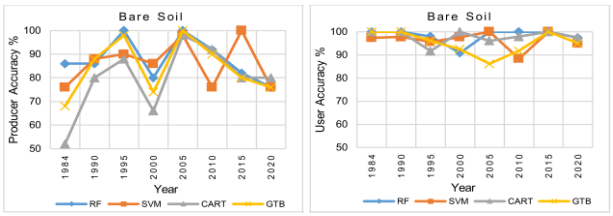


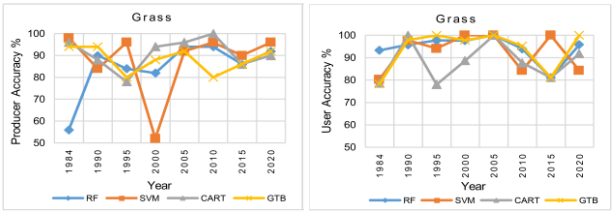
Figure 8. PA and UA results for mapping of bare-soils.

4.1.3 Water: The water bodies were classified with a consistently higher average PA accuracy of 95.4% using RF (Figure 9). From the PA accuracy results in the extraction of water bodies, GTB performed better than CART and SVM by 2.5%, whose performances were averagely equal at 89%. According to UA measures, results depict SVM (99.6%) and GTB (98.6%) as having stable and higher accuracies compared to RF (97%) and CART (96.4%). Thus, for mapping of water bodies within urban areas the UA results shows that SVM is the most suitable machine learning classifier. Comparatively, the UA results are observed to be more consistent and higher with time and sensor.

Figure 9: PA and UA results for water body mapping.

4.1.4 Vegetation classes: The results for the mapping of the vegetation classes (grass, shrubland and forest cover) are presented in Figure 10. Grass was mapped with the highest PA average of 88% and UA of 92.3%. For the mapping of grass within the urban scene, CART represented the highest PA of 91% while RF had the highest UA of 94.5%. This implies that in reference to ground truth, RF is considered as the most suitable classifier for the extraction of grass cover within the case study area.

Shrubland and forest cover were respectively mapped with average PA of 80.5% and 82.1%. The corresponding UA were 80.2% and 83.2%. For mapping of shrublands, SVM consistently had the highest average PA (83.8%) and UA (81.5%). To map forest cover within the urban environment, the results in Figure 10 shows that RF had the highest PA (83.3%) while GTB had the highest UA (84.3%). Compared in terms of the UA measures, RF was the most suitable for the extraction of grass, SVM was the most optimal for mapping shrubland while GTB was more suitable for detecting forest cover.



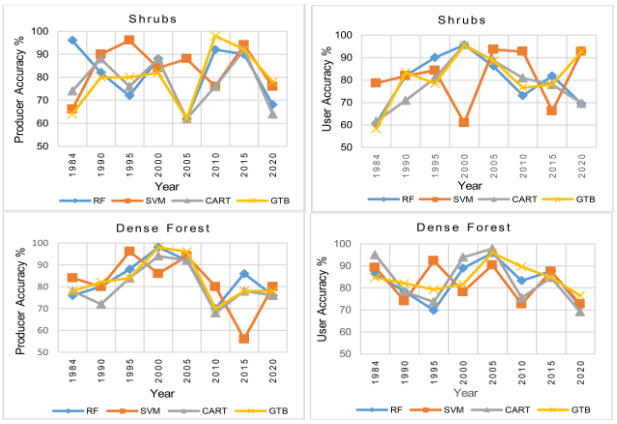


Figure 10. PA and UA results for vegetation classes.

For the six classes, urban built-up was mapped with highest average PA (96.8%), water (91.3%), grass (88%), bare-soil (84.4%), forest (82.1%) and shrubs (80.5%). However, measured using the average UA the results were: water (97.9%), bare-soil (97.0%), grass (92.3%), dense forest (83.2%), built-up (82.4), and shrubs (80.3%). The lowest average PA (77.5%) and lowest average UA (78.5%) were obtained using CART for the classification of shrubs. The average highest UA of 99.6% was achieved using SVM in the classification of water, whereas the highest average PA (98.5%) was from RF in the classification of urban built-up areas.

The results from the PA and UA analysis shows that the mapping of urban LULC classes and the performance of the classifiers are influenced by the sensor spectral resolution and time of image acquisition as characterized by the atmospheric conditions. The PA and UA metrices are further analyzed in terms of overall accuracy (OA), F1-score, TPR, FPR and the AUC.

4.2 Class classification metrics results

Table 1 presents the average metrics in terms of OA, F1-score, TPR, FPR and AUC measures for each class. For classification of built-up area, RF had the highest OA (96.4%) which was comparable to SVM at 96.2%, and the same trend was observed for the F1-score, TPR and the AUC values. However, CART, performing at equal OA as GTB, had the least FPR compared to all the classifiers in mapping the built-up areas (Table 1). RF mapped water bodies with the highest OA, F1-score, TPR and AUC scores for the study period. However, GTB outperformed all the classifiers.

The vegetation classes were mapped with relatively lower accuracy as compared to the other urban LULC classes. The results show that RF is the best classifier for mapping grass, SVM being the most suitable for mapping shrubland and GTB the most suitable for mapping forest cover. RF is recorded to be best classifier for detecting bare-soil at 97.5%, which is 1.5% more than the least accuracy from CART. For the overall average mapping of the LULC classes in Table 1, RF achieved the highest performance with OA of 95.9%, SVM (95.8%), GTB (95.6%) and CART (95.1%). The same performance pattern was observed from the overall average F1-score, TPR and AUC except for the FPR where CART tended to have lower FPR values compared to the better performing classifiers per LULC class.

Built-up	OA (%)	F1-Score	TPR	FPR	AUC
RF	96.4	0.907	0.884	0.003	0.918
SVM	96.2	0.900	0.843	0.008	0.917
CART	94.6	0.864	0.786	0.001	0.888
GTB	94.7	0.866	0.787	0.009	0.889
Bare-soil	**OA**	**F1-Score**	**TPR**	**FPR**	**AUC**
RF	97.5	0.927	0.983	0.026	0.979
SVM	97.0	0.911	0.966	0.029	0.969
CART	96.0	0.876	0.975	0.043	0.966
GTB	96.3	0.892	0.947	0.033	0.957
Water	**OA**	**F1-Score**	**TPR**	**FPR**	**AUC**
RF	99.0	0.951	0.970	0.008	0.981
SVM	98.8	0.938	0.995	0.013	0.991
CART	98.4	0.924	0.960	0.013	0.973
GTB	99.0	0.950	0.987	0.001	0.988
Grass	**OA**	**F1-Score**	**TPR**	**FPR**	**AUC**
RF	96.3	0.892	0.942	0.032	0.955
SVM	96.3	0.896	0.912	0.026	0.943
CART	96.1	0.893	0.877	0.002	0.929
GTB	96.7	0.905	0.929	0.025	0.952
Shrubs	**OA**	**F1-Score**	**TPR**	**FPR**	**AUC**
RF	92.5	0.795	0.778	0.041	0.868
SVM	93.2	0.814	0.816	0.036	0.878
CART	91.9	0.773	0.771	0.049	0.861
GTB	92.8	0.798	0.801	0.044	0.878
Forest	**OA**	**F1-Score**	**TPR**	**FPR**	**AUC**
RF	93.8	0.827	0.822	0.037	0.893
SVM	93.5	0.818	0.816	0.039	0.888
CART	93.5	0.816	0.829	0.043	0.893
GTB	94.2	0.835	0.841	0.037	0.902

Table 1. All-years average LULC class classification accuracy metrics.

4.3 Overall urban LULC mapping accuracy

Computed from the confusion matrix, the overall accuracy results are presented in Figure 11, indicating that for all the years and classes, RF performed better than all the classifiers with average OA of 87.8%. It is however observed that only SVM marginally outperformed RF in 1990,1995 and 2000. Despite the average OA for SVM being close to RF at 87.5%, the performance of RF is considered better than SVM, first based on the class accuracy metrices presented above, and the fact that the performance of RF is more stable across the classes and the different Landsat sensors as compared to SVM which exhibited non-uniform performances in the different years. With the most consistent performance across the years and data, GTB was the third best classifier with average OA of 86.4% and the least performing classifier is CART with 85.3%. Apart from the marginally lower performance of CART as compared to GTB in 1990 and 1995, both GTB and CART classifiers exhibited fairly stable performance across the years and the sensors as compared to SVM.

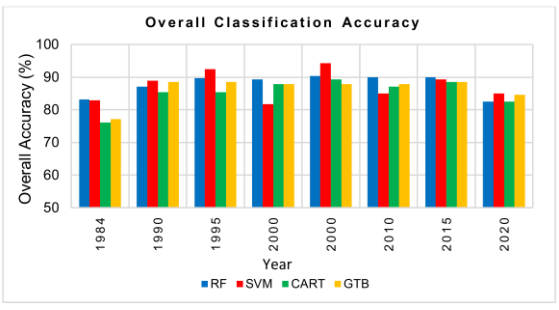


Figure 11. Average overall accuracy performance of the classifiers.

The results for the Kappa coefficient values describing the combined patterns of the yearly PA and UA are presented in Table 2. The least Kappa coefficient was observed in 1984 same as in the OA results in Figure 11 and this is attributed to the low spectral resolution in the MSS sensor with fewer spectral bands leading to spectral overlaps among the spectral classes. The 2020 results were slightly better than those of 1984, however lower than the rest of the years for Landsat TM and ETM+. This can be attributed to the narrower bandwidths in the L8-OLI as compared to Landsat TM and ETM+. Overall RF and SVM achieved equal average Kappa index accuracy of 0.85, while GTB and CART respectively achieved average Kappa indices of 0.84 and 0.82. The OA, Kappa coefficient and F1-score analyses show that the RF has the highest accuracy of all classifiers applied in this study.

	Kappa index			
Year	CART	RF	GTB	SVM
1984	0.711	0.798	0.724	0.793
1990	0.823	0.845	0.862	0.866
1995	0.823	0.875	0.862	0.909
2000	0.854	0.871	0.853	0.780
2005	0.871	0.884	0.853	0.931
2010	0.845	0.879	0.853	0.819
2015	0.862	0.879	0.862	0.871
2020	0.789	0.789	0.815	0.812

Table 2. Average Kappa coefficients for the classifiers per year.

4.4 *z*-score comparison of the classifiers

The results for the inter-comparison of the significance of the machine learning classifiers using the pairwise *z*-score test, such that *z*-score >1.96 is considered statistically different at 5% level of significance is presented in Table 3. The results shows that there is no significant difference between the classifiers in terms of the overall accuracy of performance. The notable significant difference is between CART and RF with a *z*-score >1 and *p*-value=0.093. The least observed difference is between RF and SVM at *p*-value=0.437.

Comparative classifier pairs	*z*-score	*p*-value	Significance Level
CART-RF	1.320	0.093	No
CART-GTB	-0.540	0.294	No
CART-SVM	0.992	0.161	No
RF-GTB	0.771	0.221	No
RF-SVM	0.159	0.437	No
GTB-SVM	0.505	0.307	No

Table 3. *z*-scores and *p*-values for classifier model pairs.

4.5 ROC for model performance evaluation

The area under ROC for all the prediction models is summarized in Table 1 and Figure 12. The ROC also indirectly indicates the differences in the optimal hyperparameters as set in the respective machine learning models. On average for all the classes, the RF model had the highest area under ROC curve among all prediction models. In Figure 12, RF depicts higher average AUC of 0.910, which is 0.057, 0.064 and 0.071 respectively higher than SVM, GTB and CART classifiers in mapping the urban LULC classes. The results in Figure 12 are the average results after tuning the classifier parameters to yield the best results for a given year and for the urban LULC classes.

5. DISCUSSIONS

As adopted in previous studies, for example in Sun et al. (2020), the current study utilized User's Accuracy, Producer's Accuracy, OA, Kappa coefficient and F1-score to evaluate the accuracy of the classifiers. In addition, AUC, TPR and FPR were also used to compare the LULC mapping results. The results from Landsat MSS and OLI sensors are observed to be lower than from TM and ETM+ by at least 5% in overall

accuracy. This, for the mapping of urban LULC features, are related to the low spectral resolution in Landsat-MSS sensor and to the narrower spectral bandwidths in the Landsat-OLI sensor as compared to the Landsat-TM and -ETM+ sensors. The overall results show that the maximum accuracy measured by the accuracy metrices is in the case of the RF classifier (OA=87.8% and Kappa=0.852) which is comparable to the SVM classifier results (OA=87.5% and Kappa=0.849). The noted difference between RF and SVM is that the performance of RF is more stable for all the years and not significantly influenced by the spectral and radiometric differences in the Landsat sensors. The stability of the RF classifier has been reported to be based on its increased number of trees, as well as the bagging and random concepts resulting into its efficiency and precision (Talukdar et al., 2020).

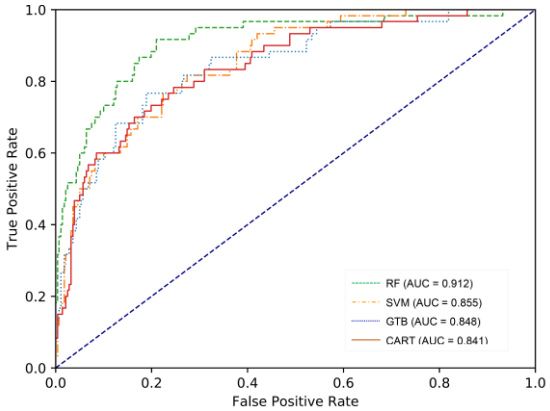


Figure 12. Average ROC curves for the classifiers.

GTB with OA=86.4% and Kappa=0.836, and CART with OA=85.3% and Kappa=0.822 presented the least performances compared to RF and SVM. However, both classifiers were stable in performance with minimal variability in the multitemporal classification accuracy. The lower performance by CART has been attributed to the decision trees being too sensitive to small changes in the training datasets and tends to overfit the model (Prasad et al., 2006). The results of the current study also point to the fact that CART is sensitive to variable input data which is attributed to the classifier having a single threshold for defining a node for splitting data into subsets. Based on its extended feature set (Georganos et al., 2018), GTB results were observed to be more stable with sensor and time, and exhibited the same performance trend as RF.

The superior performances of RF and SVM have also been attributed to the fact that the classifiers tend to be tolerant to noise (Breiman et al., 2001) and are significantly more robust towards both random and systematic noise of training data (Pelletier et al., 2017; Camargo et al., 2019). SVM performance was most effected by the radiometric and spectral resolution differences of the sensors as it recorded both the least and highest accuracy metrics results respectively in 2000 and 2005. This implies that the SVM requires continuous readjustment of the kernel function to minimize the classification errors (Pelletier et al., 2017). The performance of the classifiers are influenced by not only the data and case study, but also by the functionalities of the machine learning model. For example, simple decision-trees based classifiers like CART are not only sensitive to changes in the training datasets but also tend to overfit the model (Prasad et al., 2006). SVM on the other hand is effective in high dimensional spaces and performs adequately in situations where a clear margin of separation exists between classes and is computationally efficient. However, it requires a long training time for large datasets and is not intuitive, easy to understand or fine-tune (Huang et al., 2002).

Figure 13 presents a visual comparison of the results of the classifiers for the different classes in 2020 in relation to the ground-truth reference imagery from Google Earth. It is observed that all the classifiers mapped the correct shape and structural patterns for the different classes. SVM tended to underestimate the built-up areas, while GTB classified some urban areas as bare soils. RF and CART mapped the urban area with nearly the same degree of compactness. In mapping water bodies, RF and SVM were able to differentiate the land-water interfaces better than the other classifiers which tended to map the bare soils around the water bodies as built-up areas. RF detected the shape of the dam water body more accurately. Finally for the vegetation cover in 2020, it is observed in Figure 13 that RF and GTB mapped the forest and bare soil areas with the same degree of compactness, while SVM and CART tended to map the forest area as mixed with shrubs. The visual inspection results shows that the classifiers tend to have closely related results in terms of shape, however with different areal coverages.

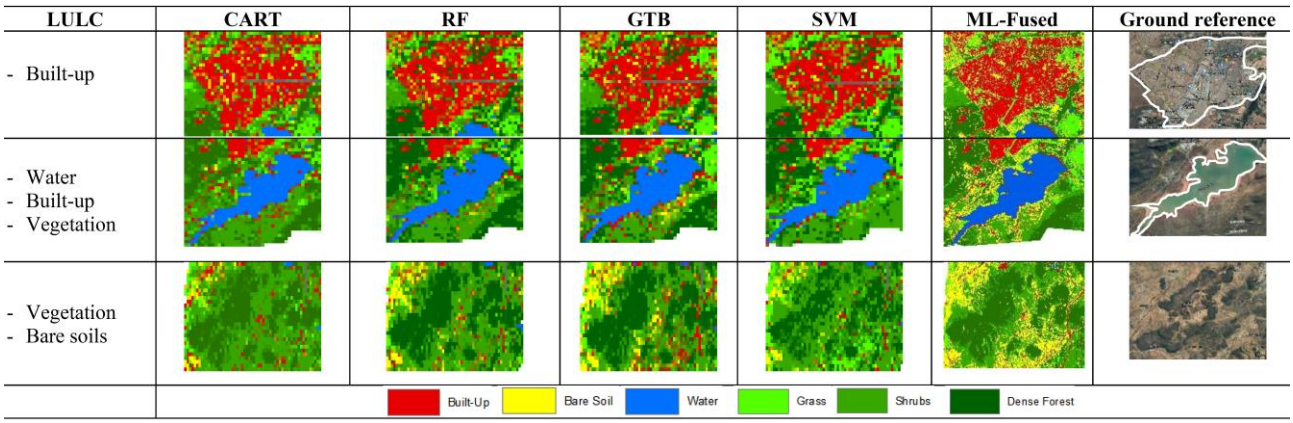


Figure 13. Image based ground-truth comparison of classification results for different LULC classes and at different locations within the study area for 2020.

Several studies have reported that for the same data and year of study, the performance of different classifiers in mapping of different LULC classes are not the same (Abadi et al., 2016). In the current study, this variation is also observed in the results

for the compared classifiers and the six urban LULC classes. Though not presented in this evaluation, it is also observed that the surface areas under each LULC class from a given classifier do not match each other. This is despite the fact that the classifier hyperparameters are tuned to yield the best results for the same study year. Deng et al. (2008) further reported that the LULC class areas also varied in the different Landsat sensor satellite data, and attributed this to the atmospheric, illumination and geometric variations. The observed differences are therefore attributed to the differences in the individual model parameter settings and classifier model functional approach (Talukdar et al., 2020). The differences in the accuracies have also been attributed to differences in the methods, time and space (Rodriguez-Galiano and Chica-Rivas, 2014).

Despite the observed marginal differences in the classification results, the study showed that the accuracies of the classifiers were similar at 5% level of significance. Hackman et al. (2017) argued that the advanced classification machine learning algorithms may not always have advantages when they were applied to process multispectral image data, and therefore focus should be on the abilities of the classifiers to extract specific LULC classes. With overall accuracies at less than 90% for all the classifiers, the study found that the high degrees of urban LULC interactions and mixing influences the classification results as the training classes consists of fractions of noise or impure pixels (Su et al., 2020).

To increase the overall accuracy of urban LULC mapping using the ML methods, focus should be on extraction of individual features and their post-classification feature fusion with the proposed multi-feature fusion concept depicted in Figure 14 and the ML-fused results compared in Figure 13. The multi-feature fusion is based on the concept of feature in-feature out (FEI-FEO) fusion where the extracted features are combined under mutual exclusivity boundary condition (Durrant-Whyte, 1988). The FEI-FEO is such that the most optimal outputs (features) from the classifiers represents different parts of the scene and are combined to obtain the complete scene global features. The fusion strategy results are found to be more accurate with compact LULC classes.

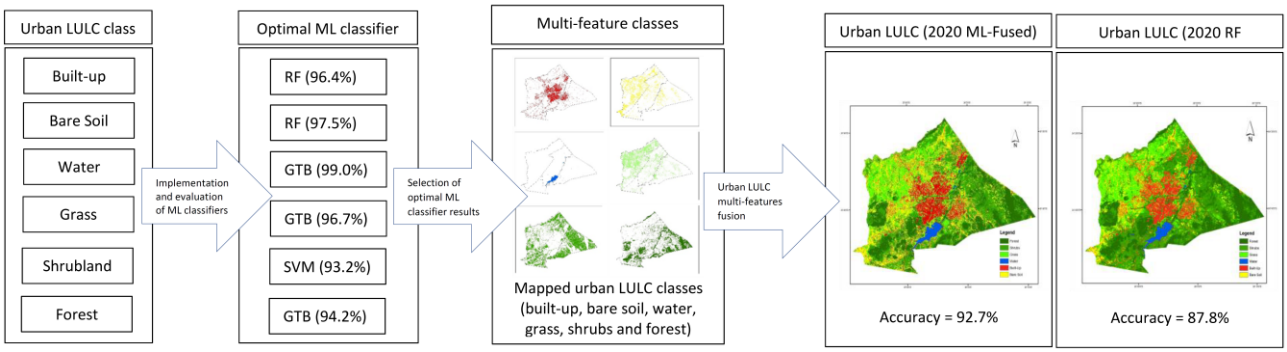


Figure 14. Multi-feature classification and fusion strategy.

6. CONCLUSIONS

This study was carried out, first to examine the accuracy of CART, RF and GTB and SVM machine classifiers for urban LULC mapping from multitemporal and multisensor Landsat data from 1984 to 2020 at half-decadal intervals. The results showed that for mapping built-up areas, RF and SVM presented the highest overall accuracy at above 85%. Bare-soil is best mapped using RF and CART with accuracy of 98%. SVM and GTB were most suitable for mapping water bodies, while the optimal classifiers for extracting the vegetation classes were grass (RF at 94.5%), shrubland (SVM at 81.5%) and forest (GTB at 84.3%). RF achieved the highest class-based performance with average overall accuracy (OA) of 95.9%, followed by SVM (95.8%), GTB (95.6%) and finally CART (95.1%). The same performance pattern was observed from the overall F1-score, True Positive Rate (TPR), Area under ROC curve (AUC) and False Positive Rate (FPR) metrices for the class specific classification accuracies. In terms of the combined average overall accuracy for the eight-epoch years, RF and SVM performed at the same level yielding the highest overall accuracy OA of 87.8% and 87.5% respectively. GTB and CART overall accuracy results were respectively at 86.4% and 85.3%. The z-score statistic revealed that in terms of the overall performance of the classifiers, the results were not statistically different. The results of the study shows that the accuracy of urban features cannot be generalized and depends on the sensor spectral resolutions, and is influenced by the temporal, atmospheric, illumination and geometric variations. The proposed post-classification feature fusion will increase the accuracy of urban LULC mapping especially for multisensor and multitemporal data. Further comparisons of the classifiers with neural network based models, and their applications in different case studies with varied number of training and validation data is recommended for future research.

ACKNOWLEDGEMENTS

This research project was funded by the Office of Research and Development (ORD) of the University of Botswana and the USAID Partnerships for Enhanced Engagement in Research (PEER) under the PEER program cooperative agreement number: AID-OAA-A-11-00012.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016: Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on Operating Systems Design and Implementation (OSDI'16), Savannah, GA, USA. pp. 265-283.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., et al., 2014: Geographic object-based image analysis–Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing* 87, 180–19.

Breiman, L., 2001: Random Forests. *Machine Learning* 45 (1), 5–32.

Camargo, F.F., Sano, E.E., Almeida, C.M., Mura, J.C., Almeida, T., 2019: A comparative assessment of machine-learning techniques for land use and land cover classification of the Brazilian tropical savanna using ALOS-2/PALSAR-2 polarimetric images. *Remote Sensing* 11, 1600.

Carranza-García, M., García-Gutiérrez, J., Riquelme, J.C., 2019: A framework for evaluating land use and land cover classification using convolutional neural networks. *Remote Sensing* 11, 274.

Deng, J.S., Wang, K., Deng, Y.H., Qi, G.J., 2008: PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing* 29, 4823–4838.

Durrant-Whyte, H.F., 1988. Sensor models and multisensor integration. *International Journal of Robotics Research* 7(6), 97–113.

Dutta, D., Rahman, A., Paul, S.K., Kundu, A., 2019: Changing pattern of urban landscape and its effect on land surface temperature in and around Delhi. *Environmental Monitoring and Assessment* 191, 551.

Fan, F., Weng, Q., Wang, Y., 2007: Land use land cover change in Guangzhou, China, from 1998 to 2003, based on Landsat TM/ETM+ imagery. *Sensors* 7, 1323–1342.

Friedman, J.H., 2002: Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38(4), 367–378.

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E., 2018: Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. *IEEE Geoscience and Remote Sensing Letters* 15(4), 607–611.

Ghosh, A., Joshi, P.K., 2014: A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery. *The International Journal of Applied Earth Observation and Geoinformation* 26, 298–311.

Hackman, K.O., Gong, P., Wang, J., 2017: New land-cover maps of Ghana for 2015 using Landsat 8 and three popular classifiers for biodiversity assessment. *International Journal of Remote Sensing*, 38, 4008–4021.

Heydari, S.S., Mountrakis, G., 2018: Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment* 204, 648–658.

Huang, C., Davis, L.S., Townshend, J.R.G., 2002: An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* 23, 725–749.

Johnson, B., Xie, Z., 2013: Classifying a high resolution image of an urban area using super-object information. *ISPRS Journal of Photogrammetry and Remote Sensing* 83, 40–49.

Khatami, R., Mountrakis, G., Stehman, S.V., 2016: A meta-analysis of remote sensing research on supervised pixel-based land cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment* 177, 89–100.

Maxwell, A.E., Warner, T.A., Fang, F., 2018: Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing* 39, 2784–2817.

Nery, T., Sadler, R., Solis-Aulestia, M., White, B., Polyakov, M., Chalak, M., 2016: Comparing supervised algorithms in Land Use and Land Cover classification of a Landsat time-series. In *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, China, 3 November 2016.

Nichols, J.A., Chan, H.W.H, BakerY, M., 2019: Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews* 11(1), 111–118.

Orieschnig, C.A., Belaud, G., Venot, J-P., Massuel, S., Ogilvie, A., 2021: Input imagery, classifiers, and cloud computing: Insights from multi-temporal LULC mapping in the Cambodian Mekong Delta. *European Journal of Remote Sensing* 54(1), 398-416.

Ouma, Y.O., Tateishi, R., 2008: Urban-trees extraction from Quickbird imagery using multiscale spectex-filtering and non-parametric classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 63(3), 333-351.

Pal, S., Talukdar, S., 2018: Assessing the role of hydrological modifications on land use/land cover dynamics in Punarbhaba river basin of Indo-Bangladesh. *Environment, Development and Sustainability* 22, 363–382.

Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C., Dedieu, G., 2017: Effect of training class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing*, 9(2), 173.

Pouteau, R., Jean-Yves Meyer, J-Y., Stoll, B., 2011: A SVM-based model for predicting distribution of the invasive tree Miconia calvescens in tropical rainforests. *Ecological Modelling* 222(15), 2631-2641.

Prasad, A., Iverson, L., Liaw, A., 2006: Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems* 9, 181–199.

Rodriguez-Galiano, V.F., Chica-Rivas, M., 2014: Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. *International Journal of Digital Earth* 7, 492–509.

Shih, H.C., Stow, D.A., Tsai, Y.H., 2019: Guidance on and comparison of machine learning classifiers for Landsat-based land cover and land use mapping. *International Journal of Remote Sensing* 40, 1248–1274.

Su, M., Guo, R., Chen, B., Hong, W., Wang, J., Feng, Y., Xu, B. 2020: Sampling Strategy for Detailed Urban Land Use Classification: A Systematic Analysis in Shenzhen. *Remote Sensing* 12, 1497.

Sun, J., Wang, H., Song, Z., Lu, J., Meng, P., Qin, S., 2020: Mapping Essential Urban Land Use Categories in Nanjing by Integrating Multi-Source Big Data. *Remote Sensing* 12, 2386.

Talukdar, S., Singha, P., Mahato, S., Praveen, B., Rahman, A., 2020: Dynamics of ecosystem services (ESs) in response to land use land cover (LU/LC) changes in the lower Gangetic plain of India. *Ecological Indicators* 112, 106121.