Supplemental Tables

Table S1. Variables that were initially selected according to their number of votes with a maximum of six votes. A variable received a vote if it fell within the top ten important variables for a given decision tree-based model.

Variable	Number of Votes
EVI (mean)	6
ARVI (mean)	6
SAVI (median)	5
Canopy Height (mean)	5
ARVI (minimum)	4
ARVI (maximum)	3
NDVI (minimum)	3
NDVI (median)	3
EVI (maximum)	3
Canopy Cover (range)	2

Table S2. Optimal hyperparameters for each multiclass model after variable selection as determined by grid search.

Hyperparameter	DT	BAG	RF	GB	XGB	LGB	ADA
Number of Trees	1	1000	1000	100	100	10	1000
Max Tree Depth	-	5	5	5	5	10	-
Max Number of Leaf Nodes	-	None	None	None	None	-	-
CCP Alpha	-	0.0	0.0	0.0	0.0	-	-
Learning Rate	-	-	-	0.01	0.01	0.1	0.1
Subsample	-	-	-	0.5	0.5	0.5	-
Loss Function	-	-	-	deviance	-	-	-
Split Function	-	-	-	Squared error	-	-	-
Number of Covariates Considered per Split	-	-	auto	log2	auto	-	-
Number of Leaves	-	-	-	-	-	10	-
Minimum Child Samples	-	-	-	-	-	10	-

Table S3. Multiclass model performance metrics between each refinement step.

	Overall Accuracy			Standard Error			Sensitivity			Specificity		
Model	Initial	Variable Selection	Final	Initial	Variable Selection	Final	Initial	Variable Selection	Final	Initial	Variable Selection	Final
DT	59.8%	57.9%	57.9%	1.537%	1.087%	1.057%	63.3%	57.9%	57.8%	70.6%	72.9%	72.9%
BAG	59.2%	58.1%	60.1%	0.883%	0.914%	1.682%	62.0%	61.3%	68.2%	70.2%	68.8%	67.9%
RF	59.8%	58.7%	60.1%	1.236%	1.655%	1.790%	62.5%	61.5%	67.5%	71.0%	70.5%	68.4%
GB	61.1%	60.8%	60.5%	1.262%	1.273%	1.103%	65.6%	67.1%	62.4%	71.3%	69.5%	73.1%
XGB	60.7%	60.2%	60.3%	1.723%	1.782%	1.451%	66.4%	67.1%	67.5%	70.1%	68.8%	68.3%
LGB	60.8%	60.3%	59.6%	1.377%	1.535%	1.255%	66.0%	67.0%	67.0%	70.3%	68.7%	70.8%
ADA	59.4%	58.6%	59.1%	1.314%	1.093%	1.760%	57.2%	57.6%	68.0%	75.8%	75.5%	71.8%

Supplemental Figures

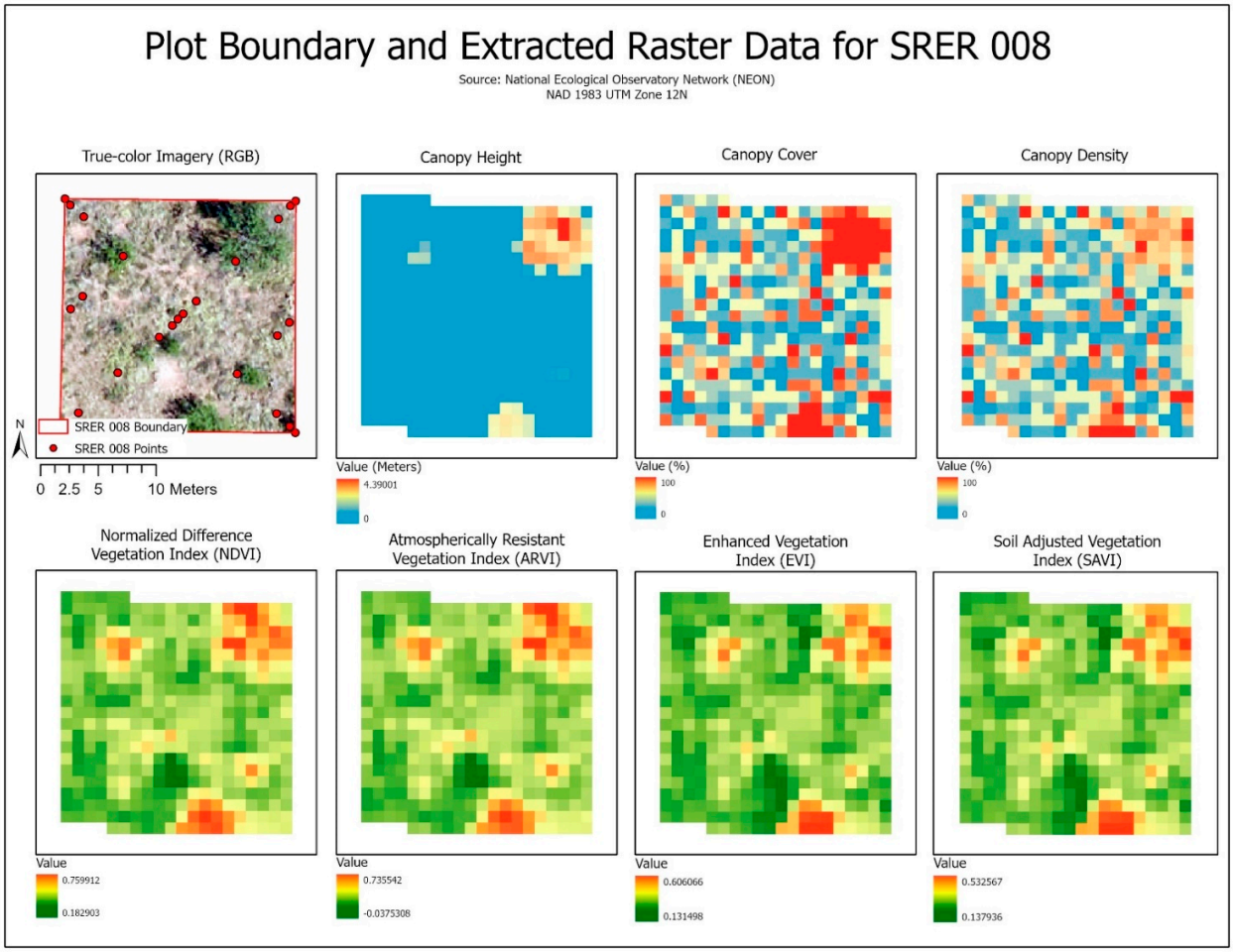

Figure S1. Rasterized versions of each variable used for this study. The extent shows a single distributed plot (SRER 008) and its associated variables [27,31–34].

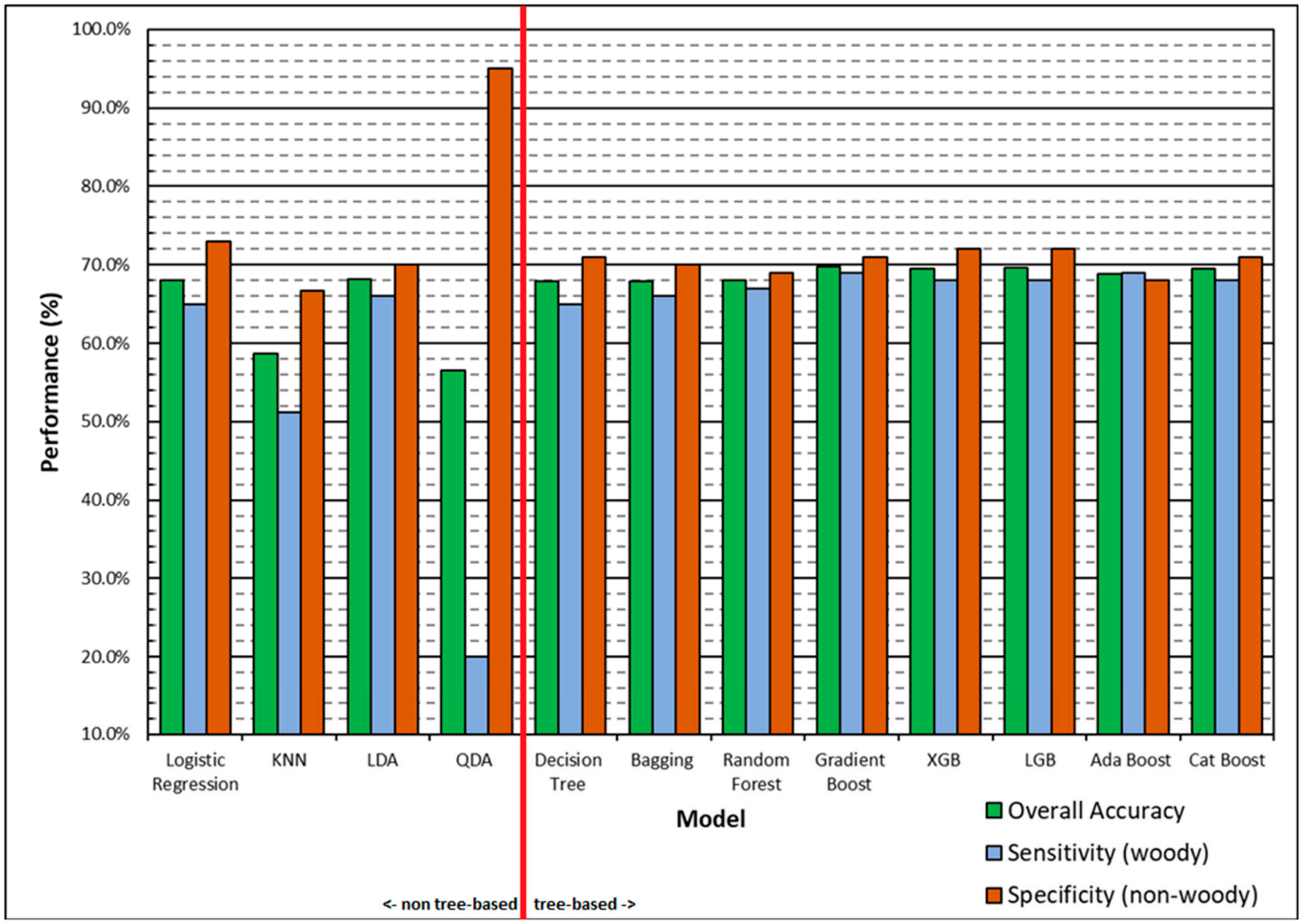

Figure S2. Initial binary classification performance showed consistently higher overall accuracy for all decision tree-based models (67.8–69.8%) relative to non-decision tree-based models (56.5%–68.0%). Decision-tree based models had consistently higher sensitivity (65.0–69.0%) and specificity (68.0–72.0%) compared to non-decision tree-based models (sensitivity = 20–66%, specificity = 66.7–95.0%). Gradient Boost had the highest overall performance (accuracy = 69.8%. sensitivity = 69.0%, specificity = 71.0%).

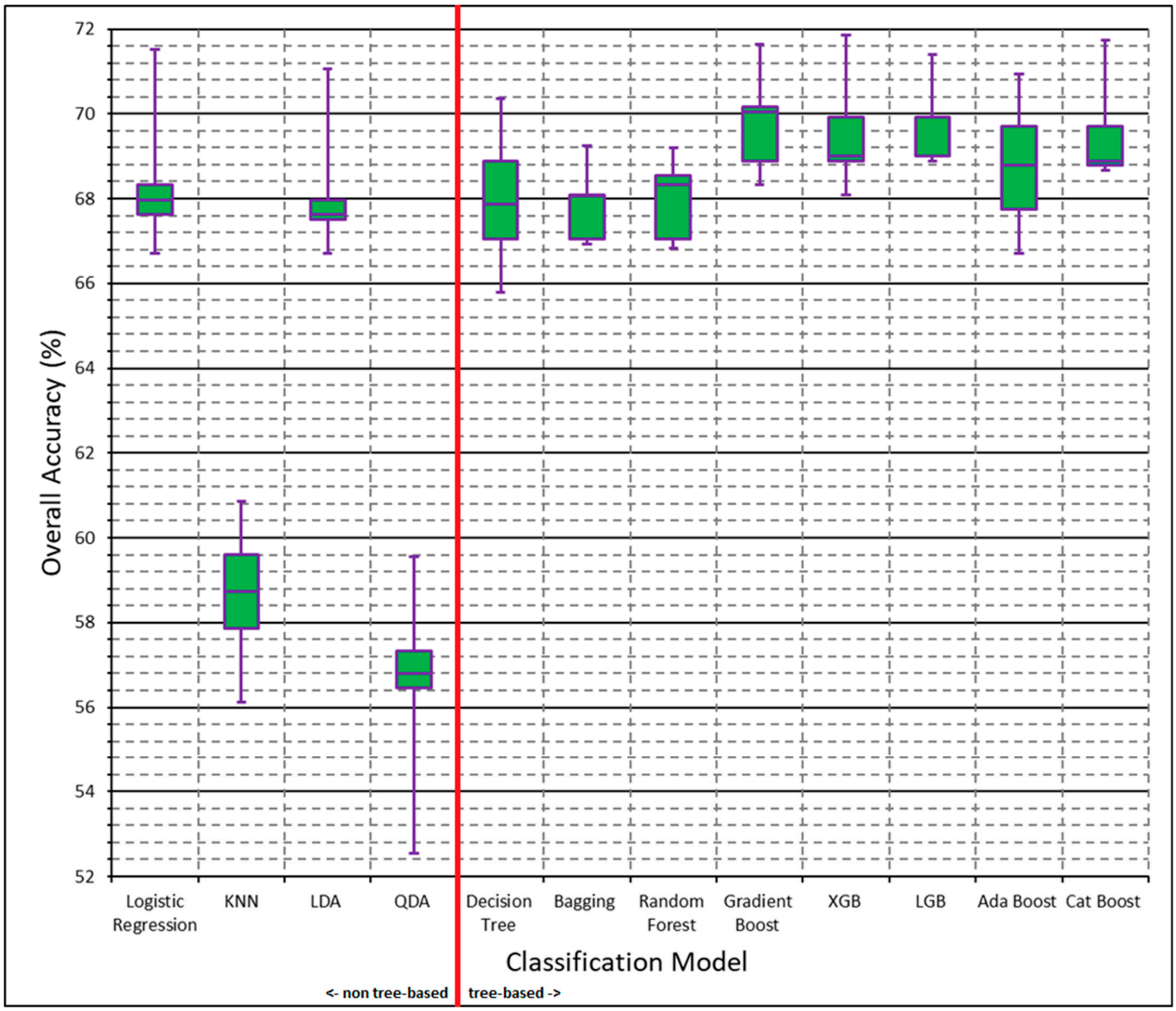

Figure S3. Initial binary model stability across the Five-Fold Cross Validation. Decision tree-based models were generally more stable than non-decision tree-based models as observed by smaller ranges of overall accuracy. Random Forest had the smallest range (2.4%), and Logistic Regression had the largest range (4.8%)

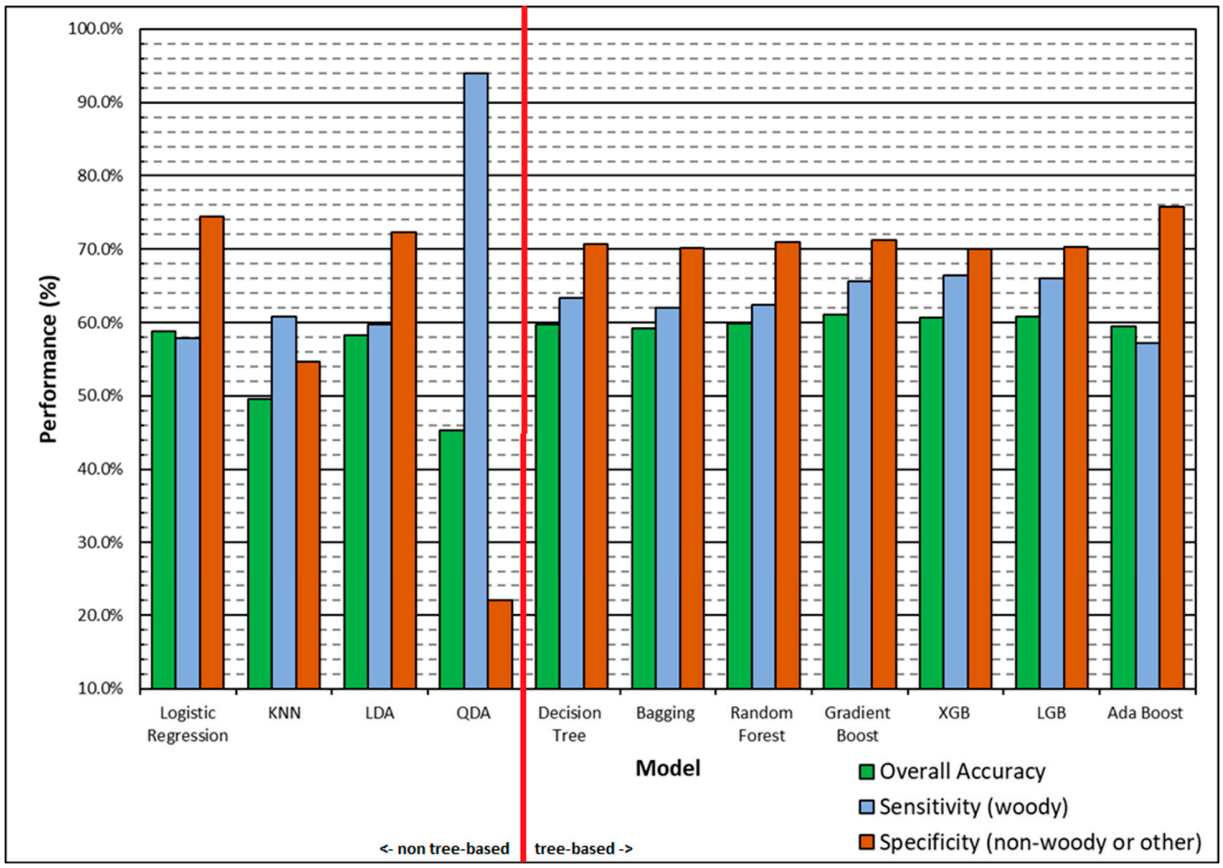

Figure S4. Initial multiclass classification performance showed consistently higher overall accuracy for all decision tree-based models (59.2–61.1%) relative to non-decision tree-based models (45.3–58.8%). Decision-tree based models had consistently higher sensitivity (57.2–66.4%) and specificity (70.1–75.8%) compared to non-decision tree-based models (sensitivity = 57.9–93.9%, specificity = 22.0–74.4%). Gradient Boost had the highest overall performance (accuracy = 61.1%. sensitivity = 65.6%, specificity = 71.3%).

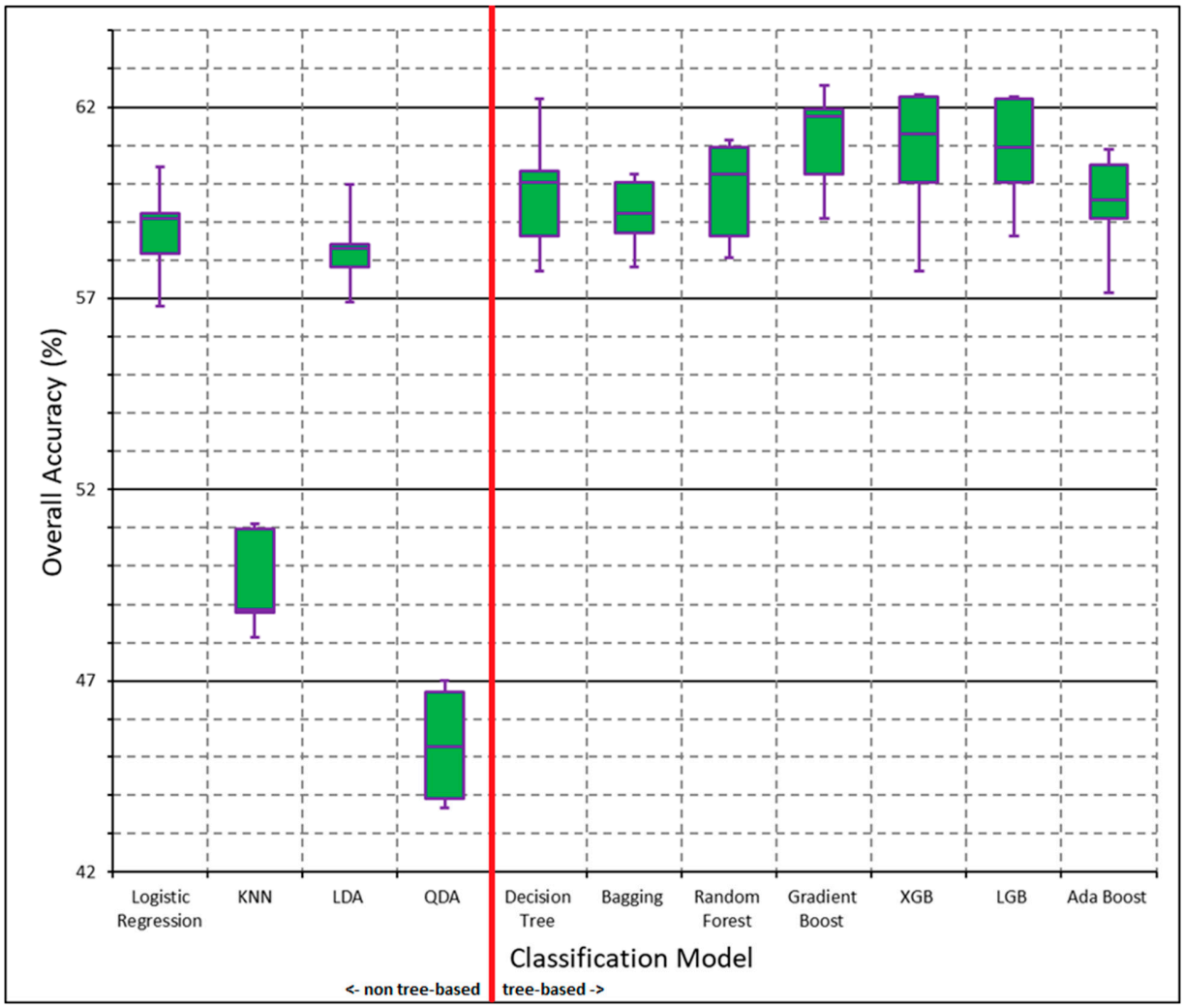

Figure S5. Initial multiclass model stability across the Five-Fold Cross Validation. Decision tree-based models were generally more stable than non-decision tree-based models as observed by smaller ranges of overall accuracy. Bagging had the smallest range (2.4%), and eXtreme Gradient Boost had the highest range (4.6%).

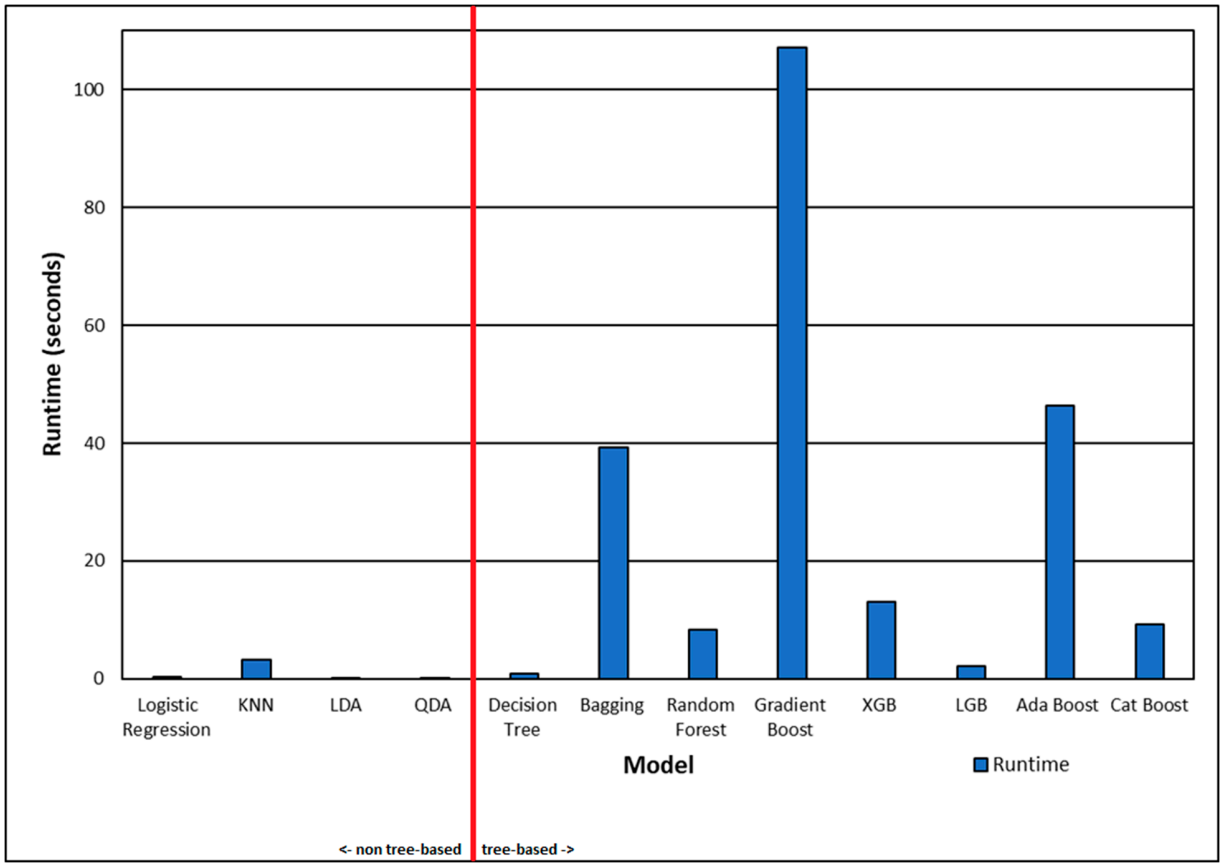

Figure S6. Initial binary model runtimes. Decision tree-based models took longer to run overall, with the fastest model being Quadratic Discriminant Analysis (0.05097 s) and the slowest model being Gradient Boost (107.2 s)

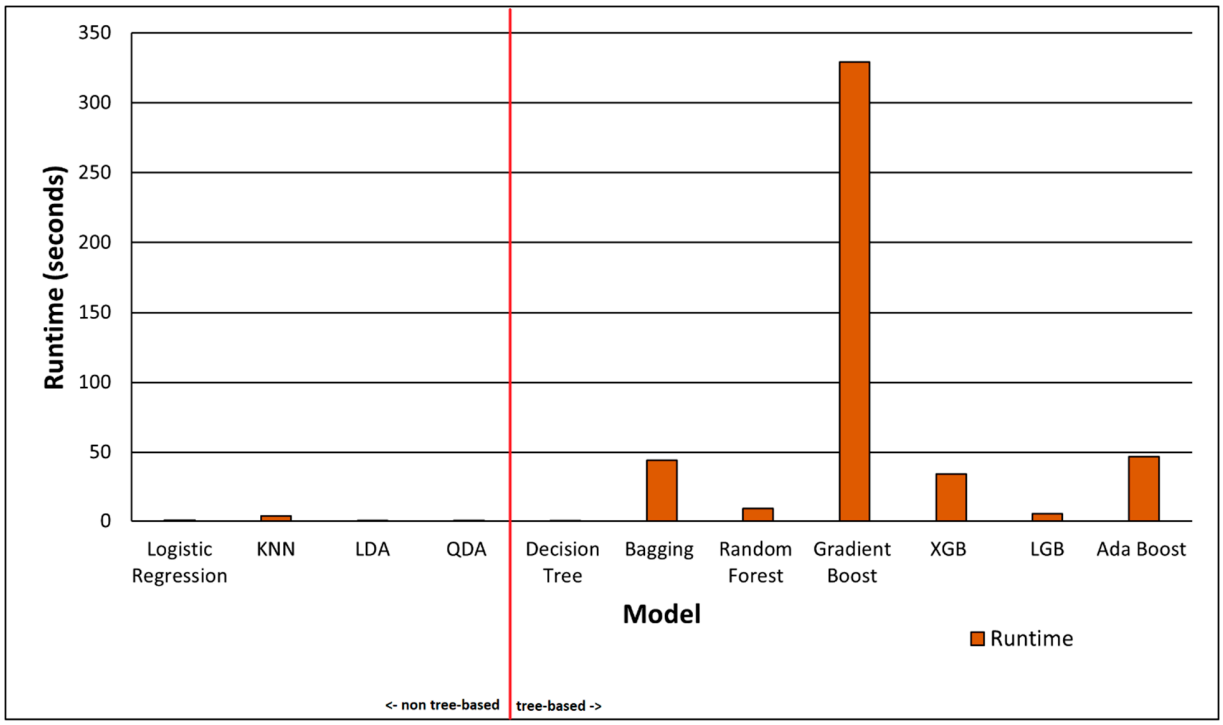

Figure S7. Initial multiclass model runtimes. Decision tree-based models took longer to run overall, with the fastest model being Quadratic Discriminant Analysis (0.0527 s) and the slowest model being Gradient Boost (329.1 s).

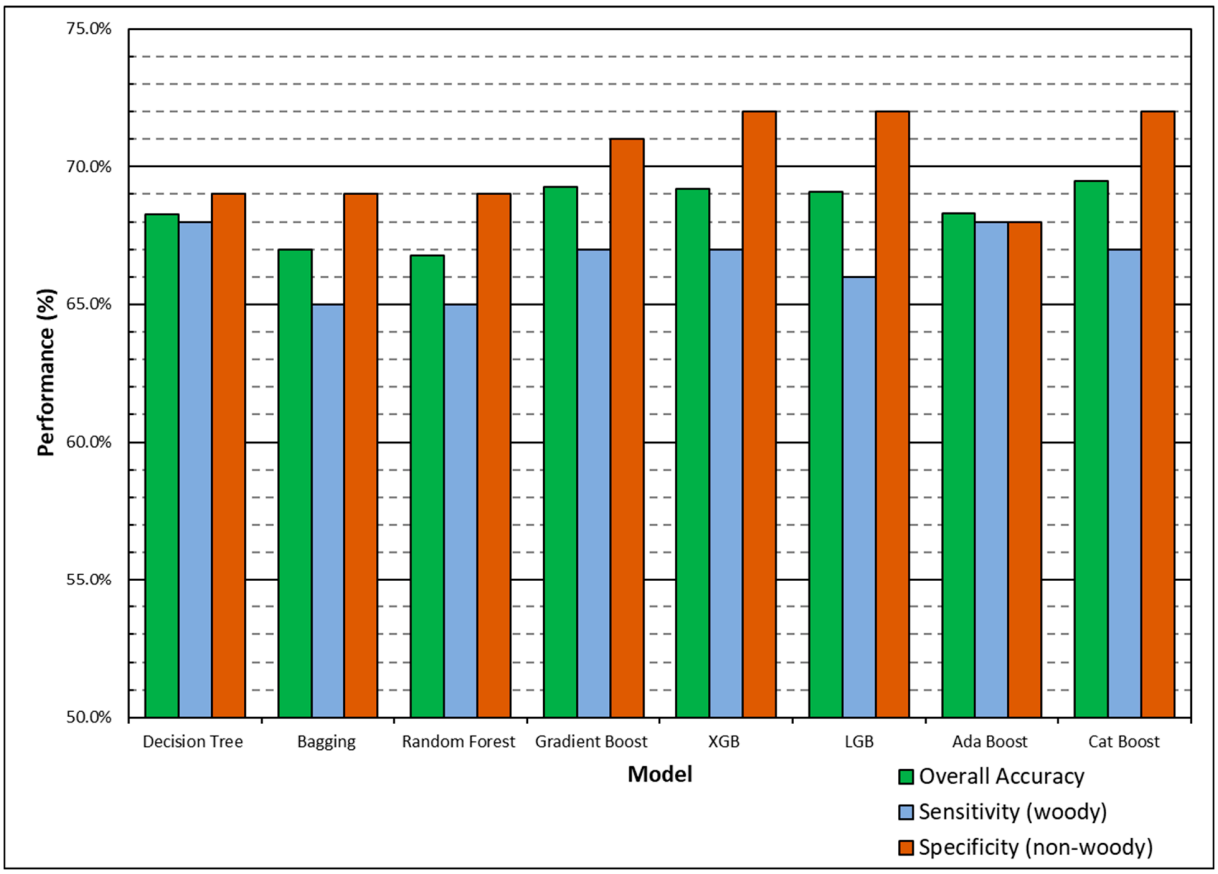

Figure S8. Binary model performance after final variable selection showed lower overall accuracy for all models (67.0–69.5%) compared to initial overall accuracy (67.8–69.8%). Cat Boost had the highest overall accuracy with 69.5%, Decision Tree and Ada Boost had the highest sensitivity with 68.0%, and eXtreme Gradient Boost, Light Gradient Boost, and Cat Boost had the highest specificity with 72.0%. Sensitivity remained lower than specificity for all models except Ada Boost.

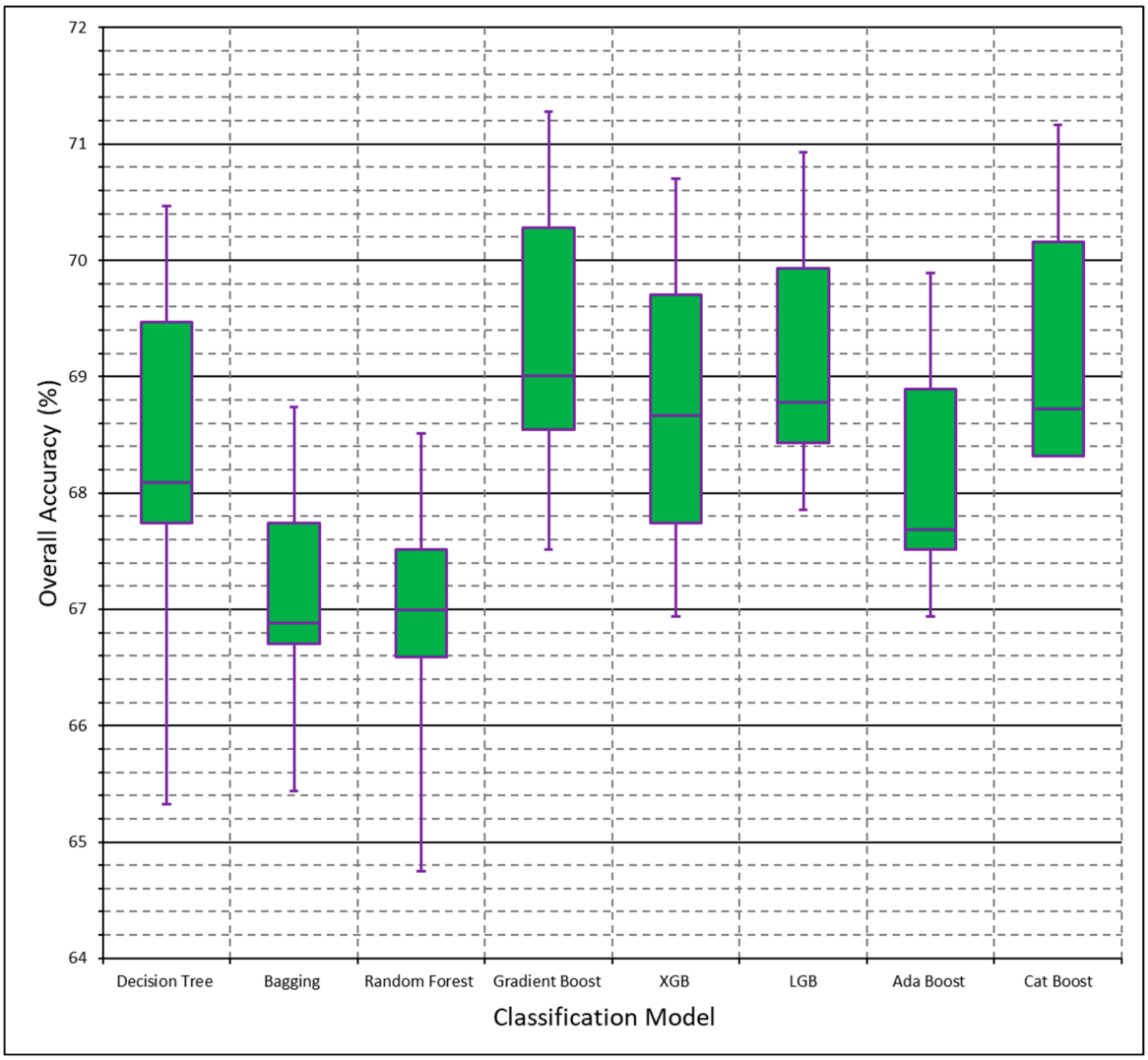

Figure S9. Binary model stability after variable selection. All models except single decision tree had similar ranges for overall accuracy. Decision Tree had the largest range with 5.0% and Light Gradient Boost had the smallest range with 2.3%.

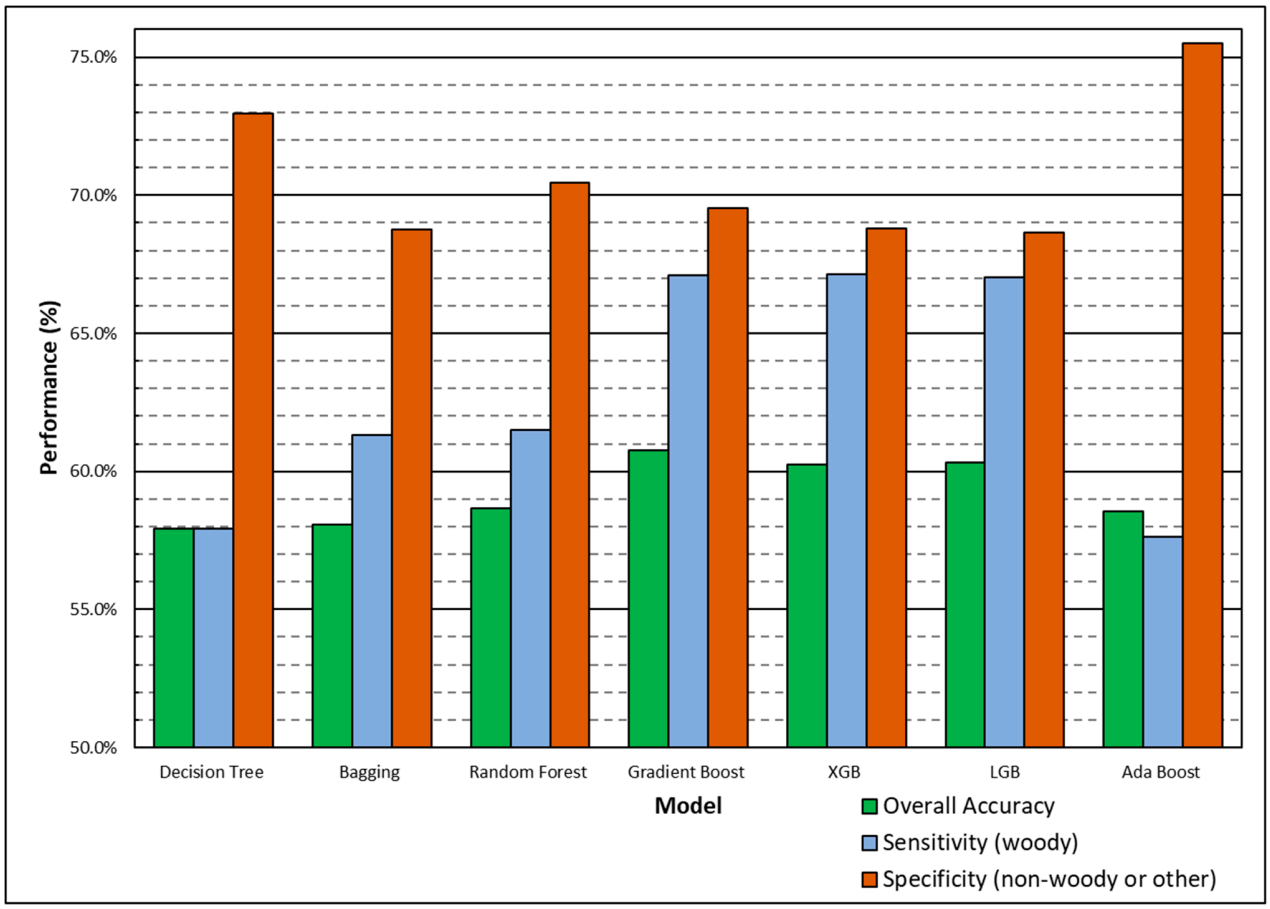

Figure S10. Multiclass model performance after final variable selection showed lower overall accuracy for all models (57.9–60.8%) compared to initial overall accuracy (59.2–61.1%). Gradient Boost had the highest overall accuracy with 60.8%, Gradient Boost and eXtreme Gradient Boost had the highest sensitivity with 67.1%, and Ada Boost had the highest specificity with 75.5%. Sensitivity remained lower than specificity for all models.

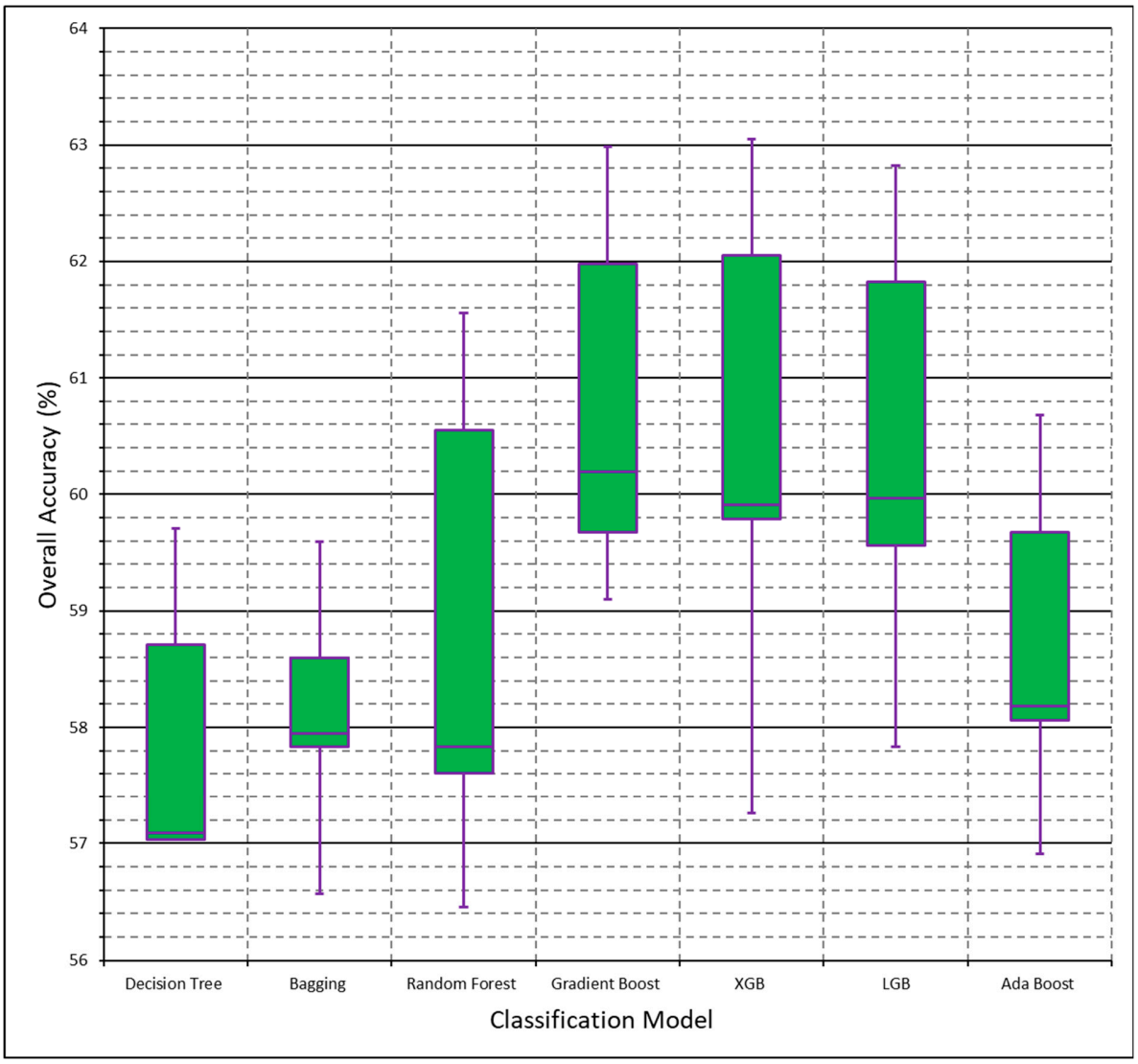

Figure S11. Multiclass model stability after variable selection. eXtreme Gradient Boost had the largest range with 4.8% and Decision Tree had the smallest range with 2.6%.

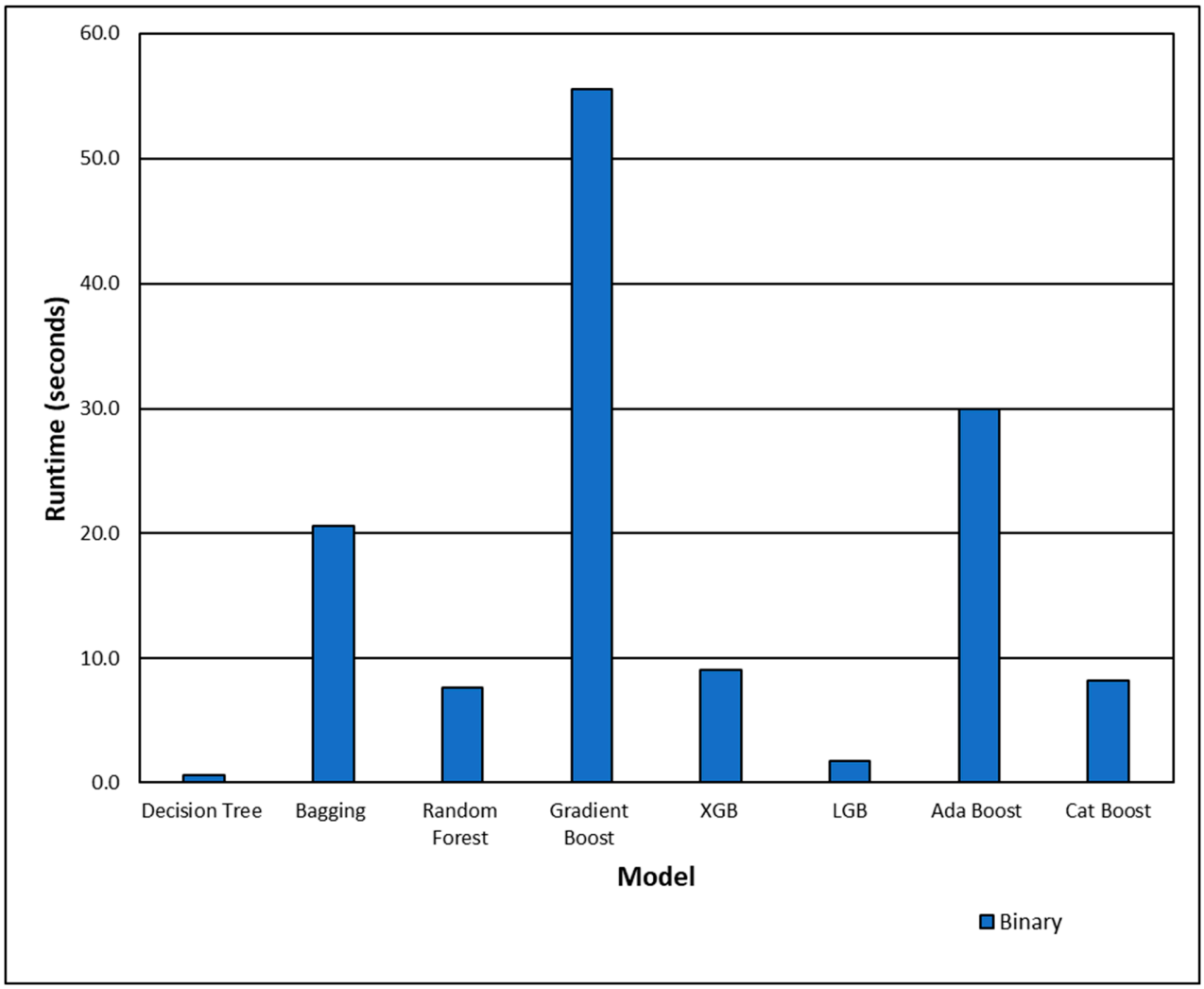

Figure S12. Runtimes for each binary model after variable selection. Gradient Boost remains with the longest runtime (55.54 s), and single decision tree with the shortest runtime (0.6075 s).

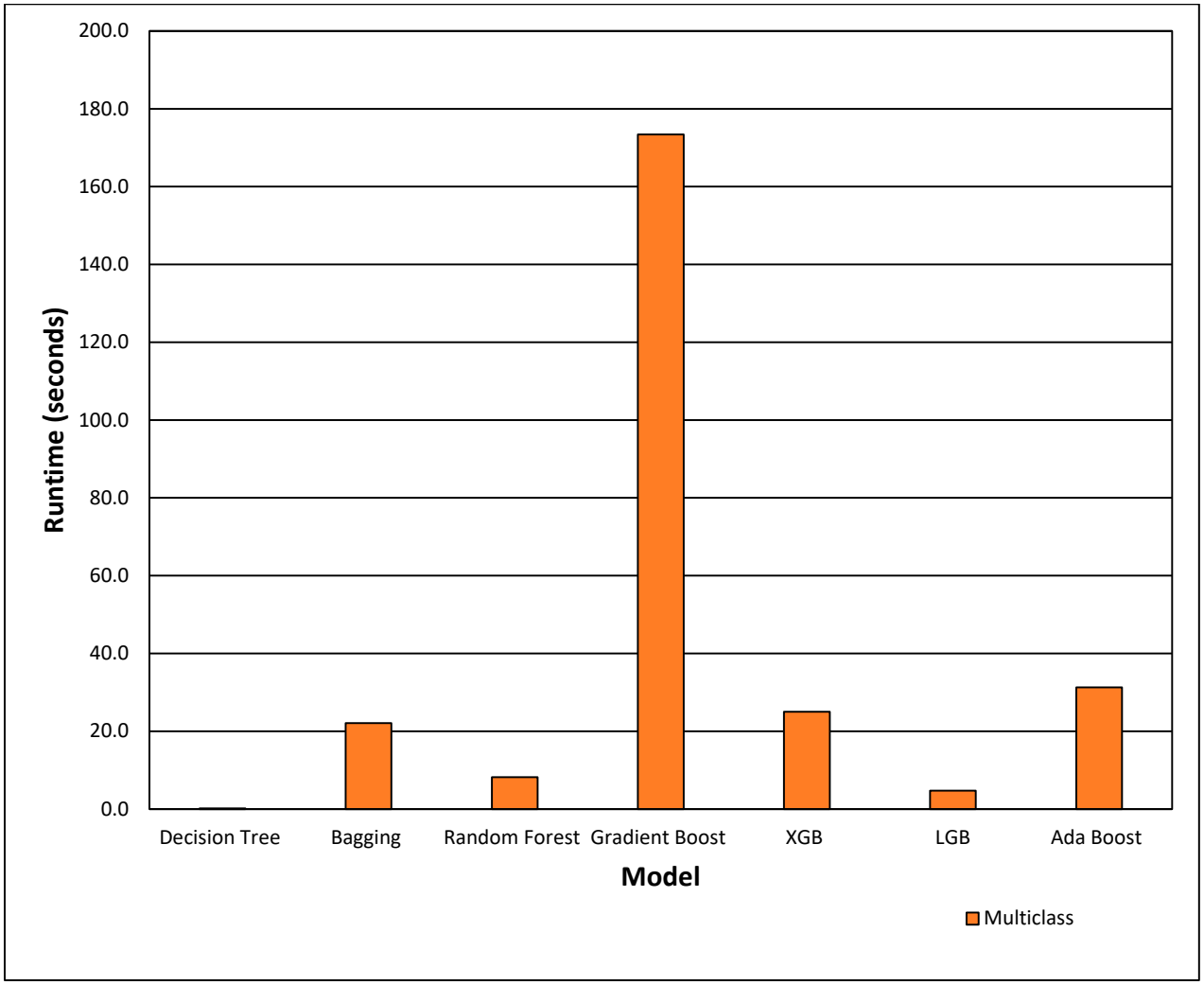

Figure S13. Runtimes for each multiclass model after variable selection. Gradient Boost remains with the longest runtime (173.4 s) and single decision tree with the shortest runtime (0.1687 s).

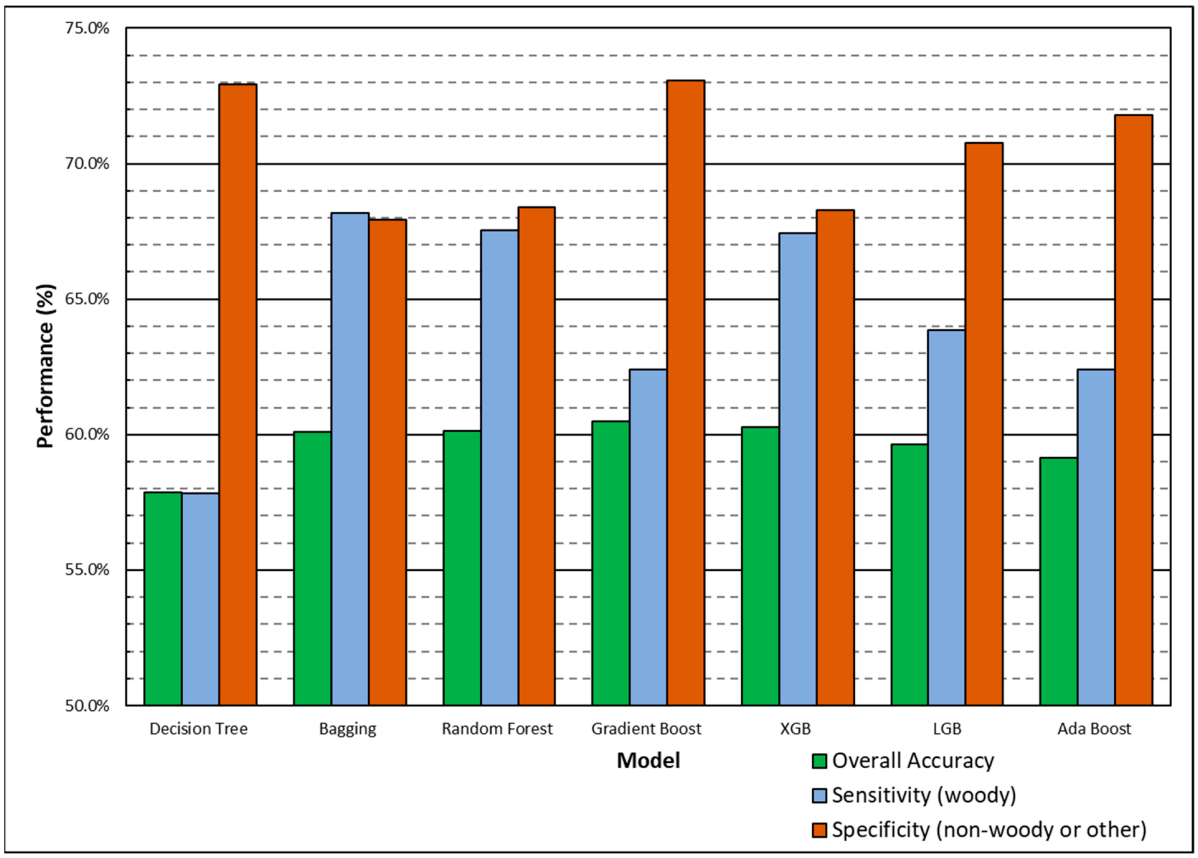

Figure S14. Finalized multiclass model performance showed a slight decline at the lower limit of overall accuracy compared to the initial run (57.9% from 59.2%) and slight decline at the upper limit of overall accuracy (60.5% from 61.1%). Specificity remained higher than sensitivity for all models except Bagging, with improvement from initial models seen in the lower and upper limits of sensitivity (57.8% from 57.2% and 68.2% from 66.4%, respectively). Gradient Boost performed with highest overall accuracy with 60.5% but has a larger difference between sensitivity and specificity compared to eXtreme Gradient Boost, which has an overall accuracy of 60.3%. Lowest sensitivity was observed from the Decision Tree model (57.8%) and highest sensitivity was observed from the Bagging model (68.2%). Lowest specificity was observed from the Bagging model (67.9%) and highest specificity was observed from the Gradient Boost model (73.1%).

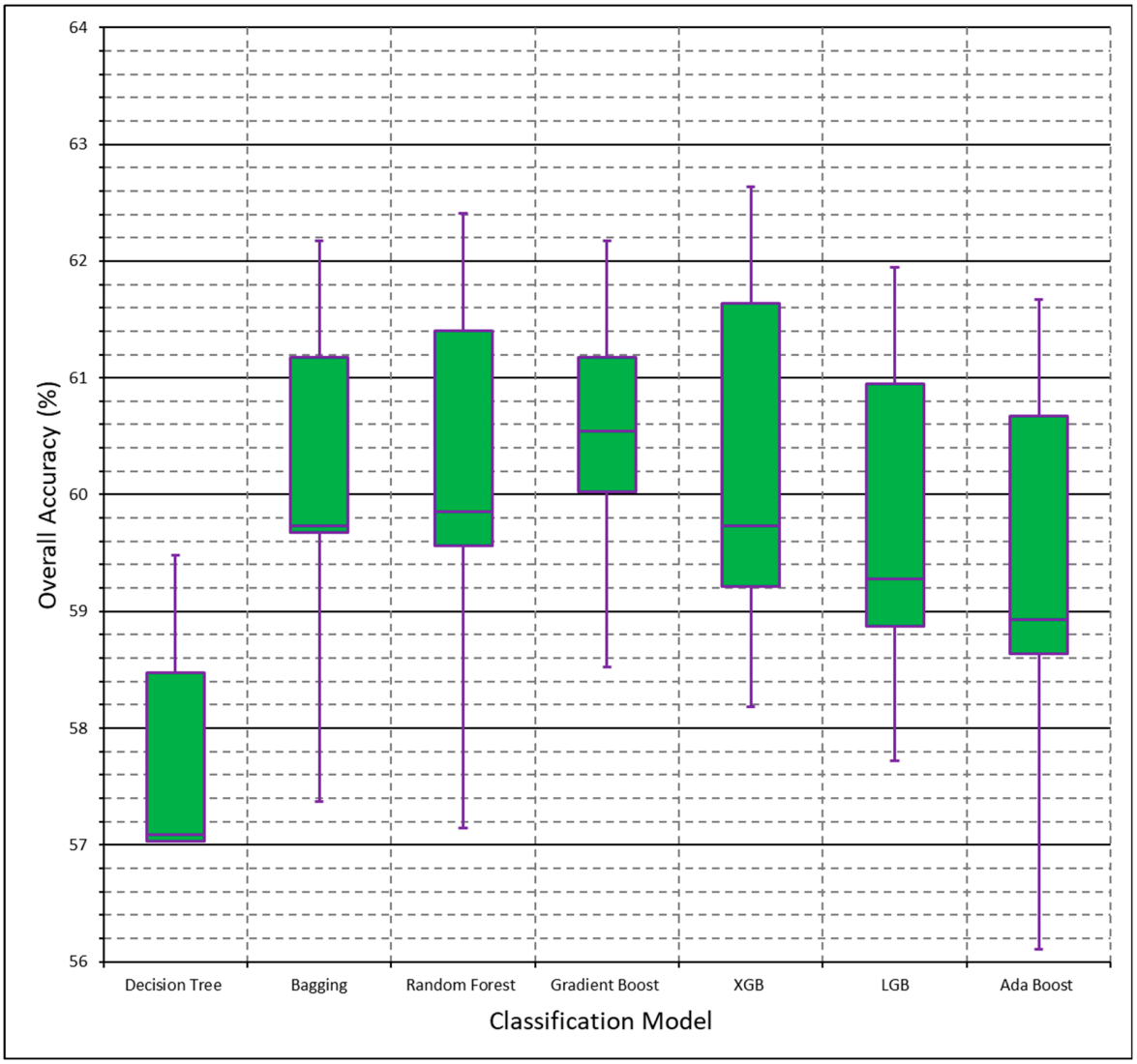

Figure S15. Finalized multiclass model stability showed that Bagging had the lowest range (2.6%), and Random Forest had the highest range (5.3%). This is a slight improvement to the lower limit compared to initial models (2.4%) and slight decline to the upper limit compared to initial models (4.6%)

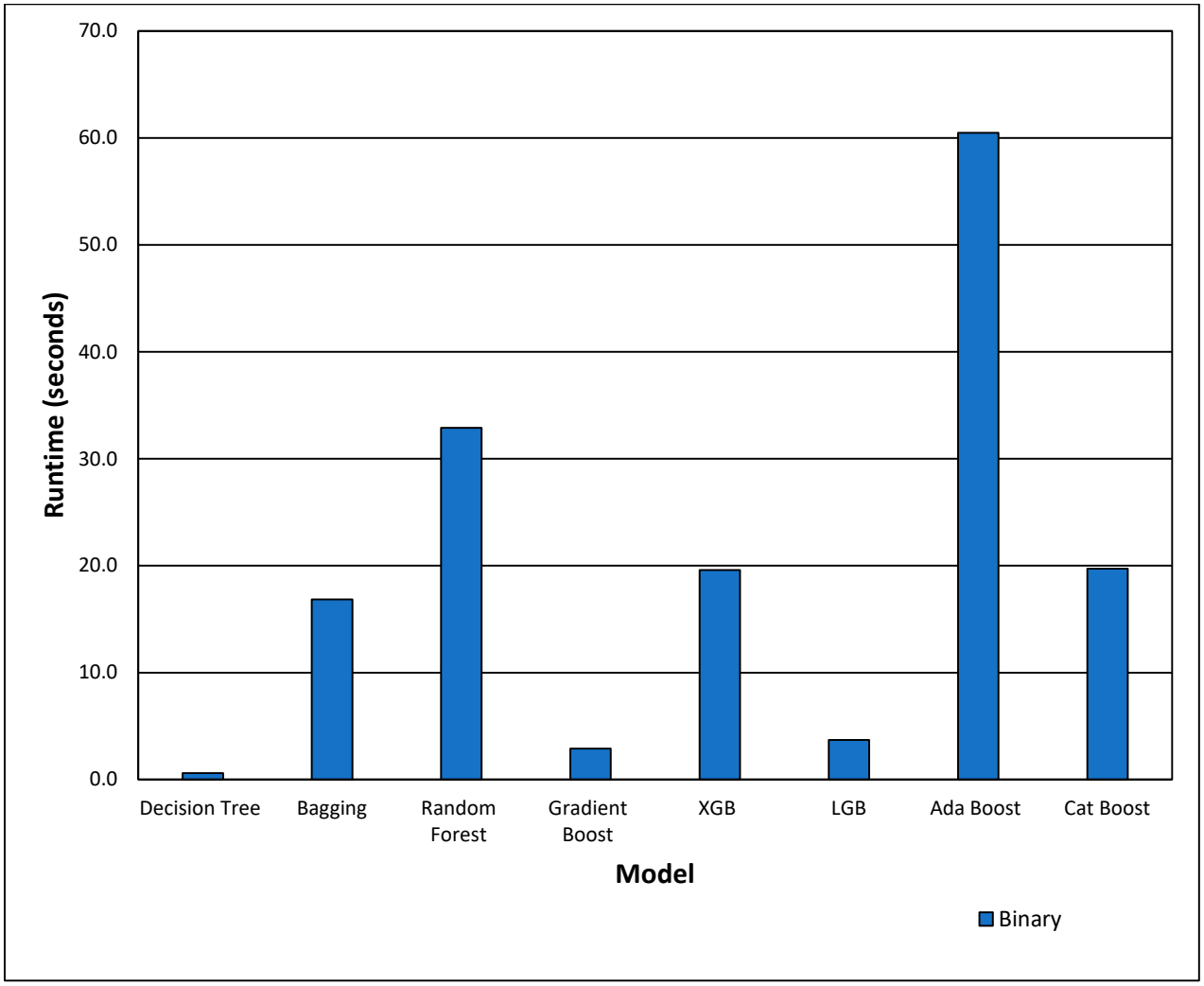

Figure S16. Finalized binary model runtimes. Decision tree had the shortest runtime (0.6054 s), and Ada Boost had the longest runtime (60.47 s).

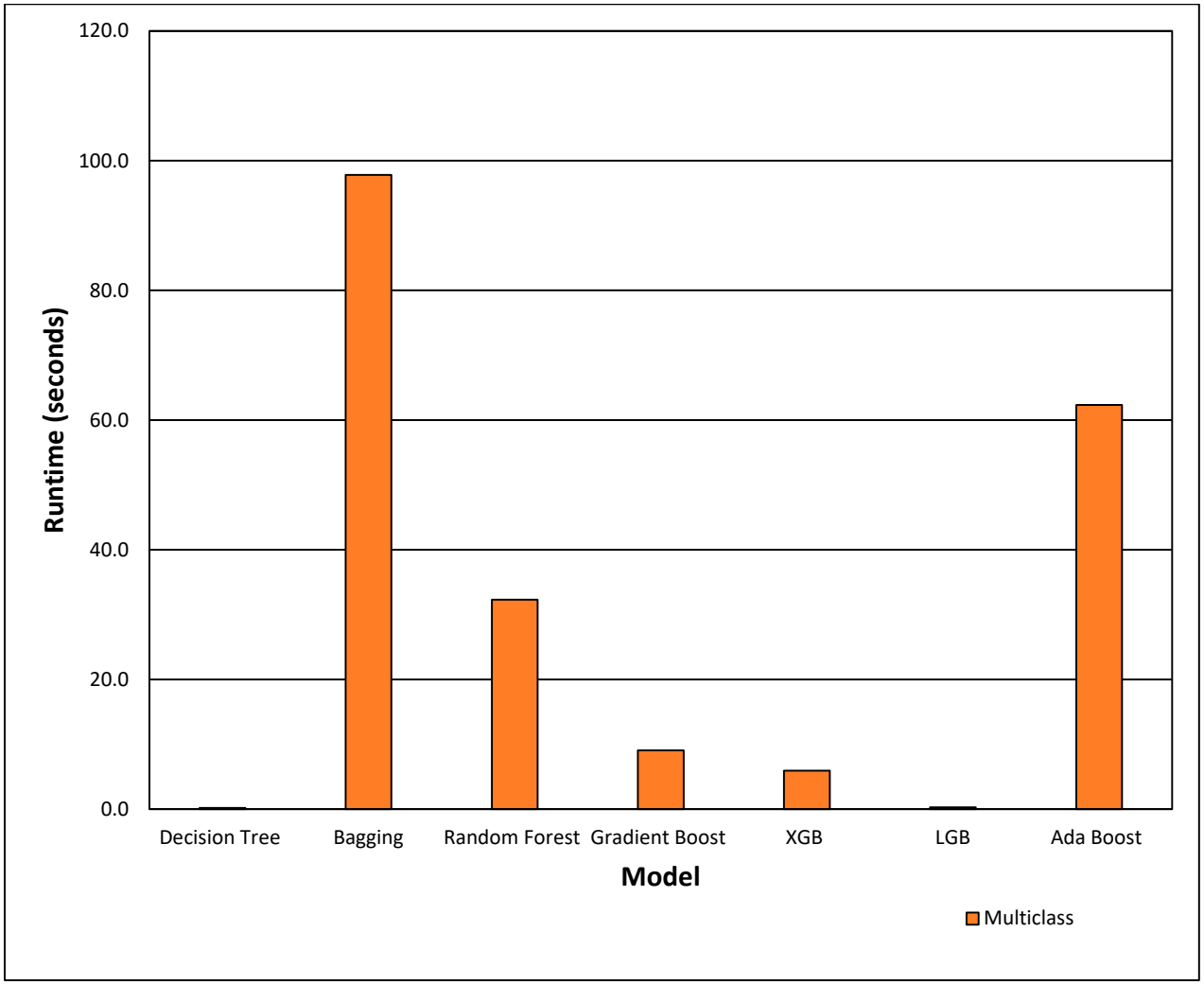

Figure S17. Finalized multiclass model runtimes. Decision tree had the shortest runtime (0.1946 s), and Bagging had the longest runtime (97.81 s).

SRER Plot 14: Low FWC

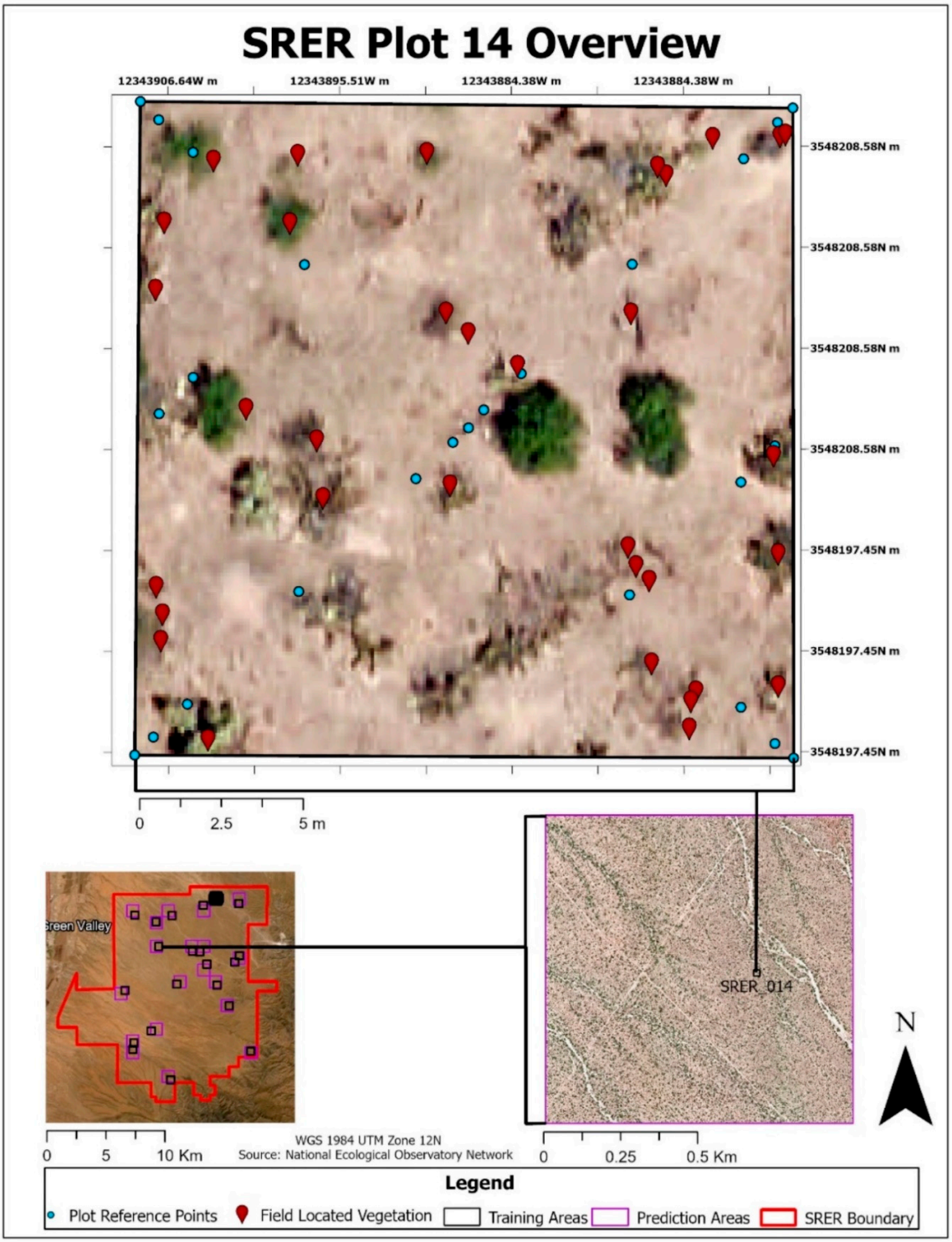

Figure S18. Overview of Plot 14 (low FWC). Field located vegetation is spread throughout the plot with some inaccuracies observed due to variations in GPS accuracy. Average FWC between binary and multiclass training data was 7.56%. This plot features a handful of live woody shrubs and cacti along with dead woody cover. The remainder is covered by bare ground or non-woody vegetation.

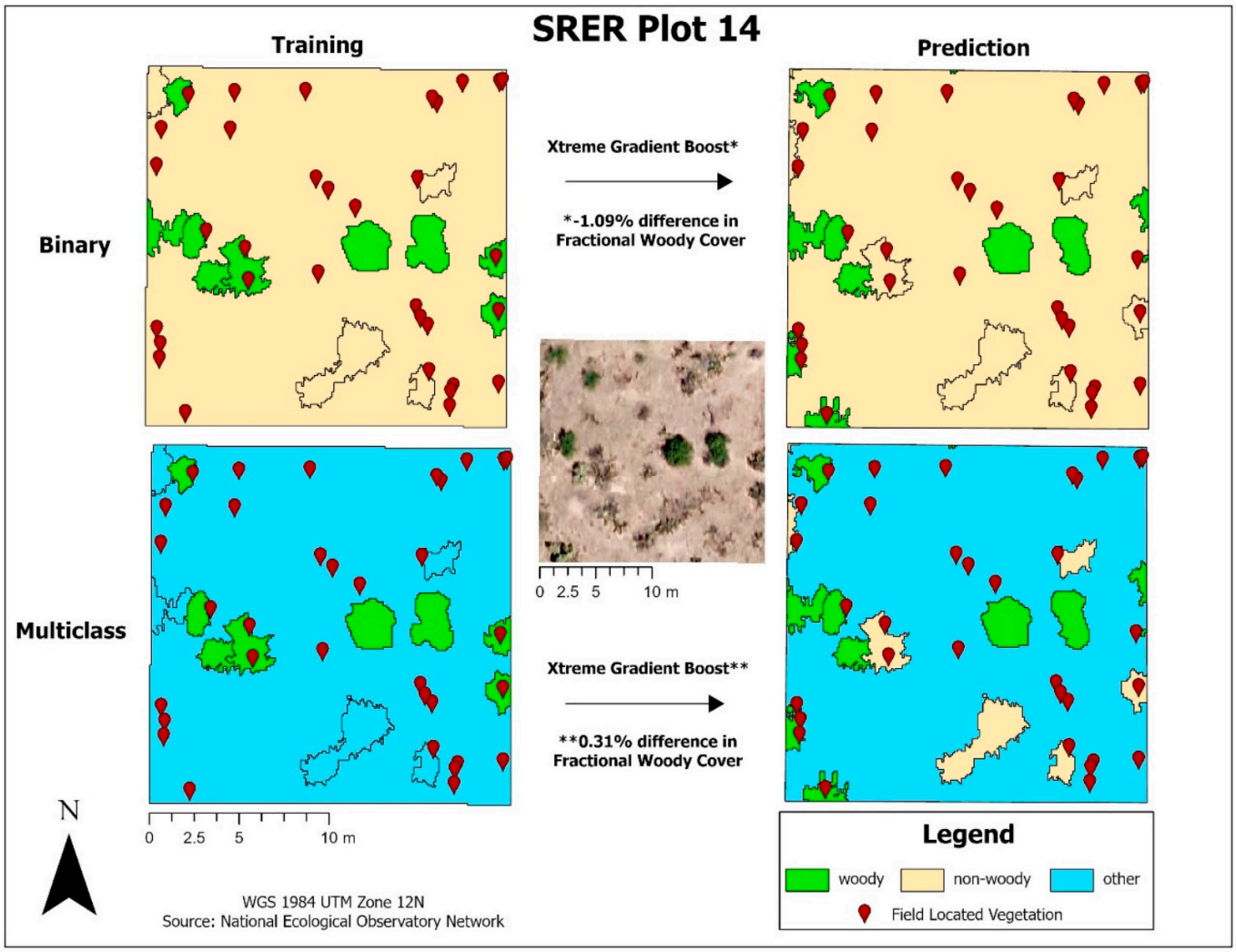

Figure S19. XGB predictions for Plot 14 under both binary and multiclass classification schemes. Despite slight differences in image segmentation, overall estimations of FWC are very similar between training and prediction methods. Most differences in classification occur in polygons that appear less green in imagery, but field verification indicates live vegetation (polygon just west of center for example). This presents a possible limitation in AOP or model ability to determine live plant status with current sensor/data resolutions. The multiclass scheme appears to be classifying the less green, but still live vegetation as non-woody and the surrounding bare ground as other.

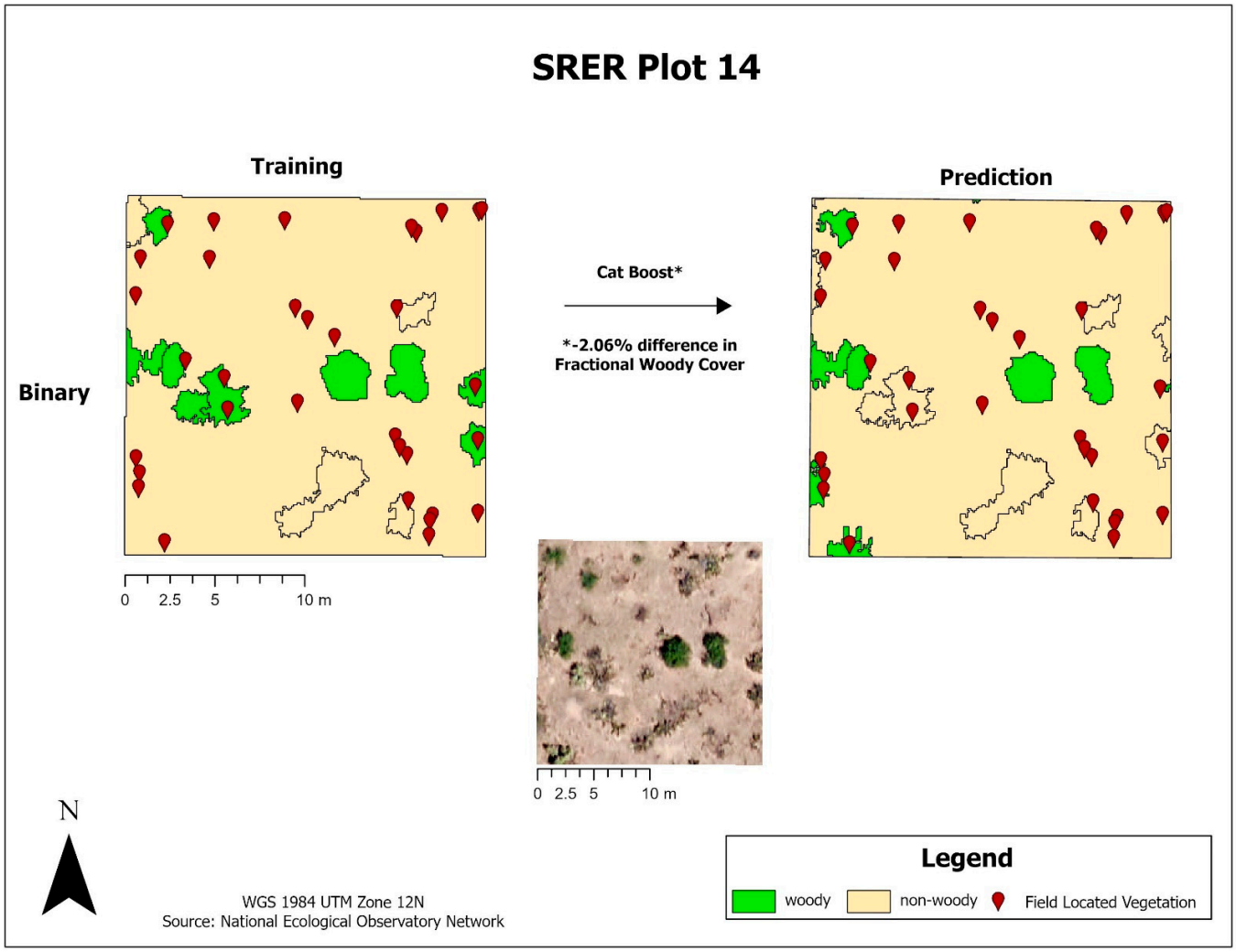

Figure S20. Cat Boost predictions for Plot 14. Multiclass is not shown, as it was not supported by the Cat Boost model. As observed in the XGB model, less-green live vegetation is being classified as non-woody.

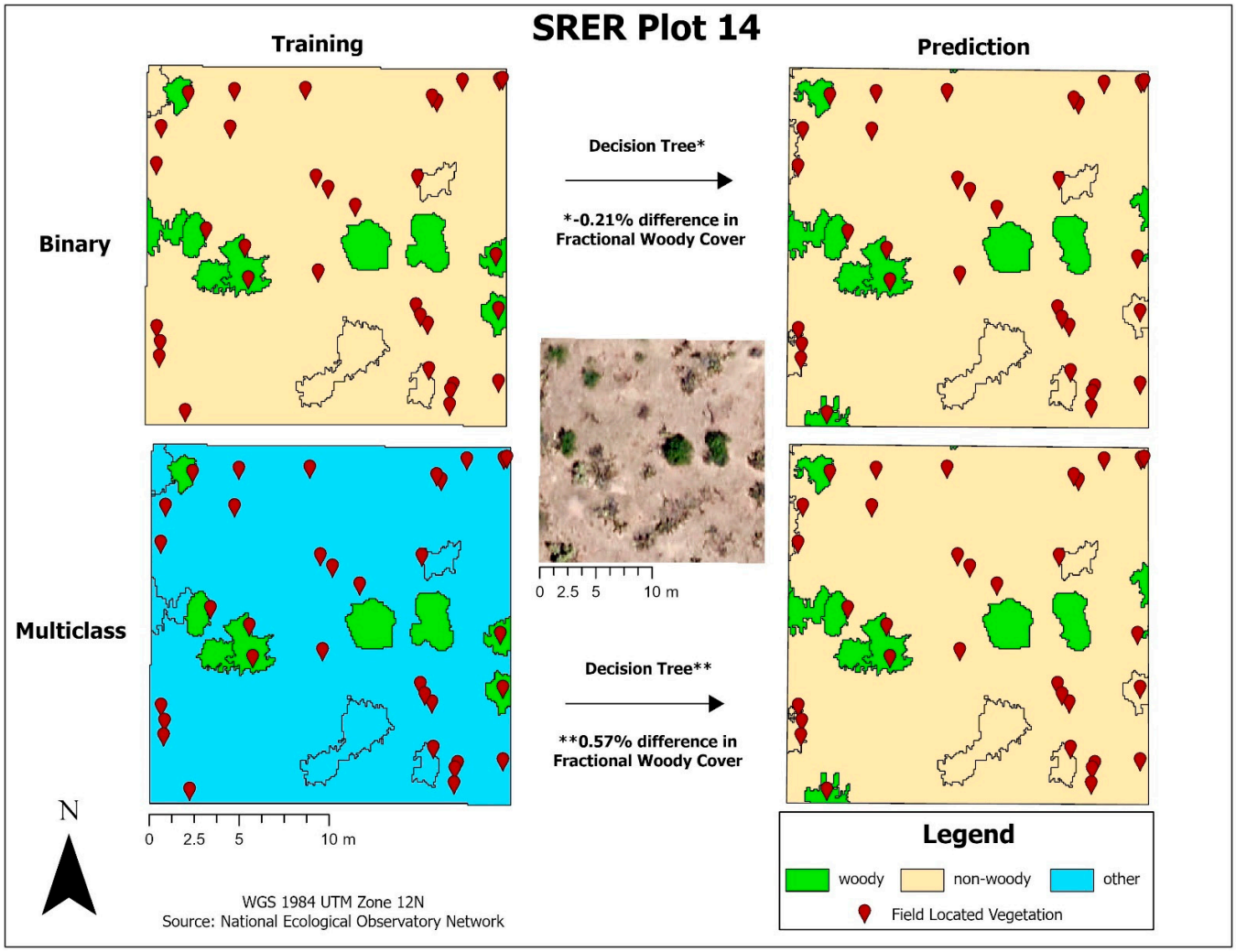

Figure S21. Decision Tree predictions for Plot 14 for the binary and multiclass schemes. Despite having the worst performance overall, the Decision Tree model performs very well at this low FWC plot and was able to accurately predict the less-green vegetation as woody a little better than XGB and Cat boost. Decision Tree also differed from eXtreme Gradient Boost for the multiclass prediction in that it did not predict any "other" cover.

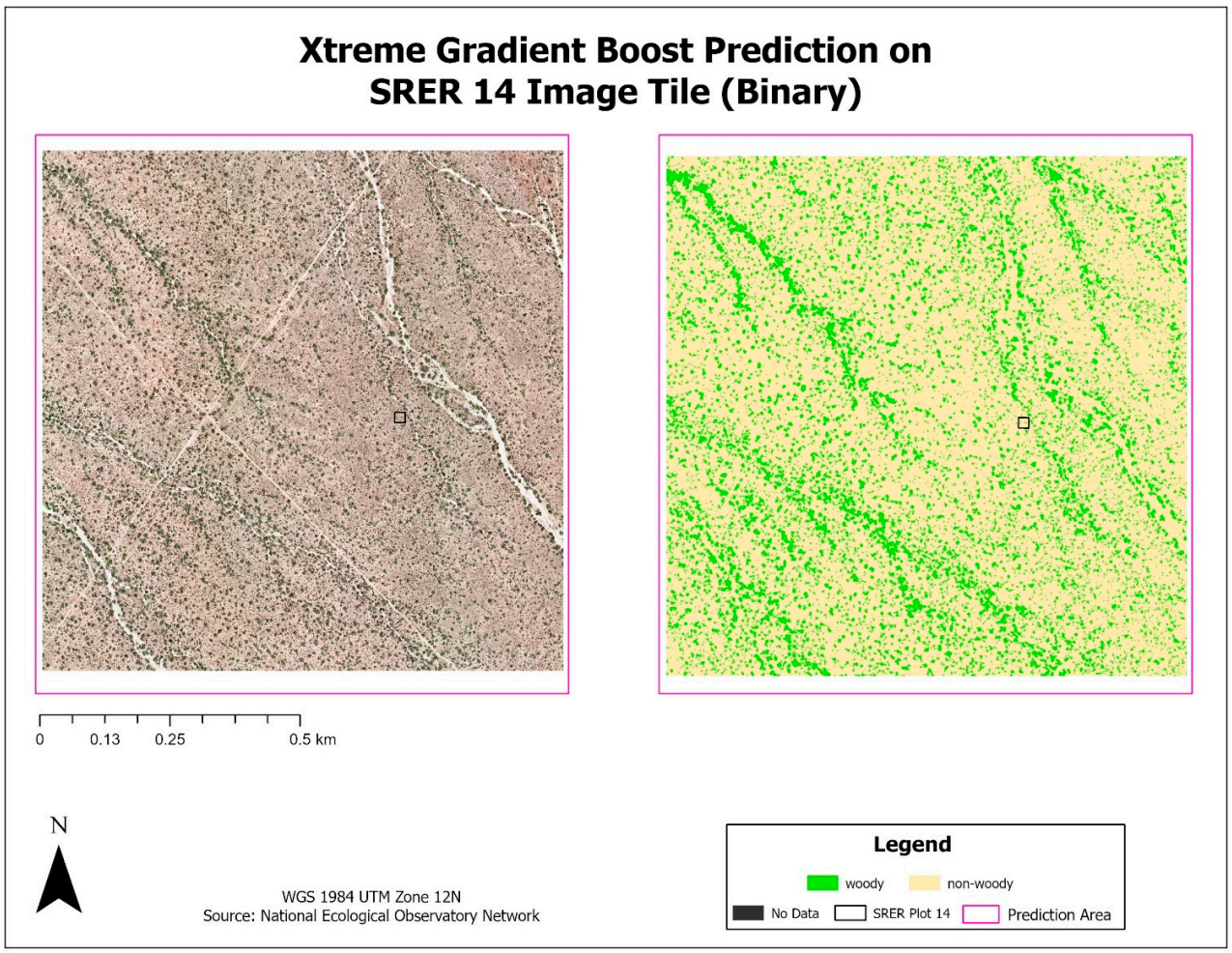

Figure S22. XGB predictions for the Plot 14 image tile using the binary classification scheme. FWC for the entire prediction area is 20.8%. Areas with higher densities of woody plant cover are clearly highlighted in green in both the RGB and prediction tile, along with some delineation of water and road/trail features.

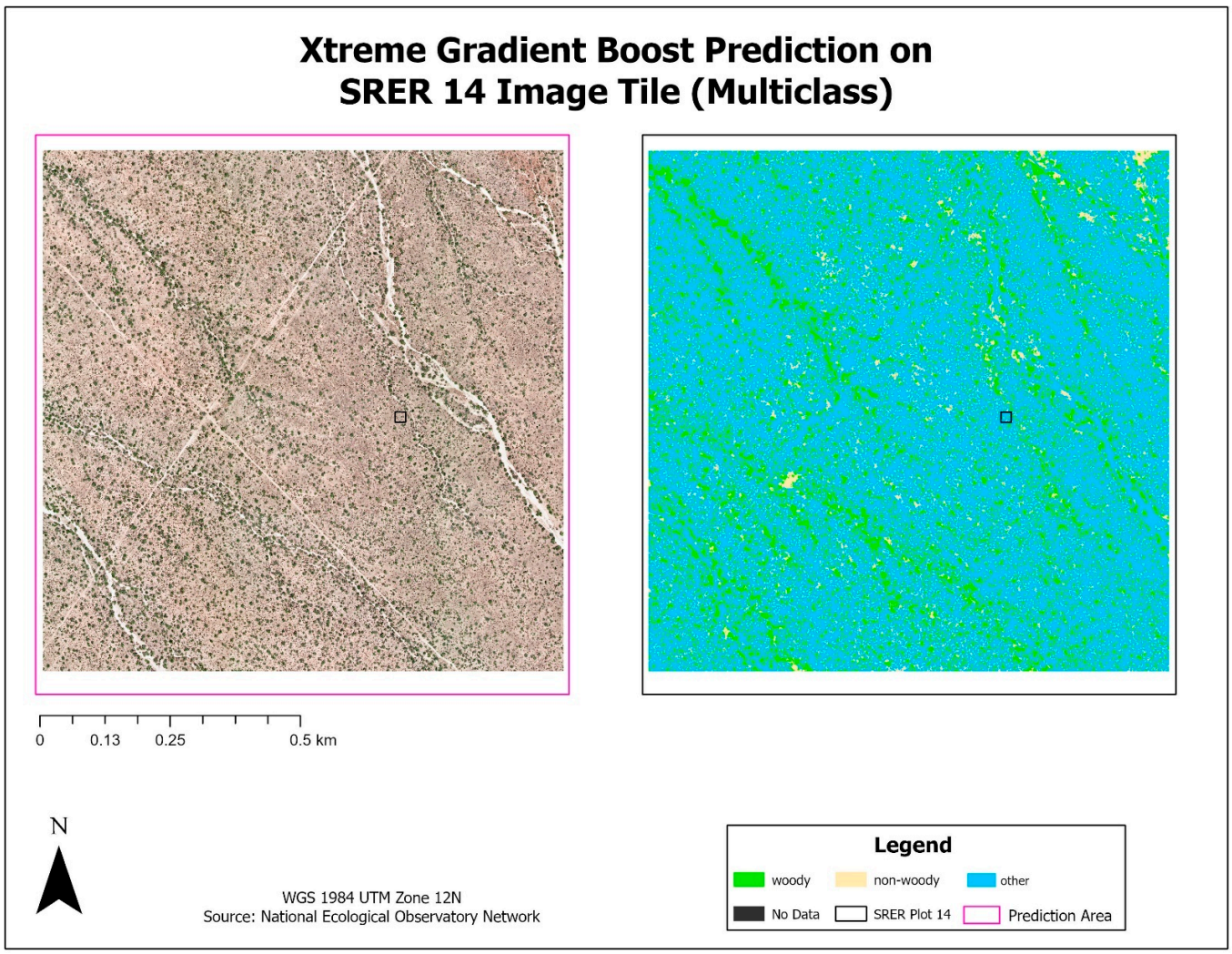

Figure S23. XGB model predictions for the Plot 14 image tile using a multiclass classification scheme. FWC for the entire prediction area is 20.5%. Major woody vegetation patterns appear in green for both the RGB and prediction tile, with areas of bare ground, roads, and water features being classified in blue as "other". Areas being classified as non-woody appear to have a reddish soil color or have areas of less-green vegetation.

Overall, Plot 14 had high agreeance in FWC between the training and prediction classification methods (mean dFWC for XGB = -0.70%). The XGB model slightly under predicted woody vegetation overall, as indicated by the negative dFWC value and visual inspection showing some woody polygons being classified as non-woody. This underprediction may be occurring due to differences in the ability of field versus remote sensing methods being able to discern live and dead vegetation. The prediction tiles for this plot highlights areas of woody vegetation consistently between classification schemes (dFWC between schemes = 0.3%). The multiclass scheme shows additional detail with the non-woody class potentially highlighting areas of interest for further field data collection and verification.

SRER Plot 19: High FWC

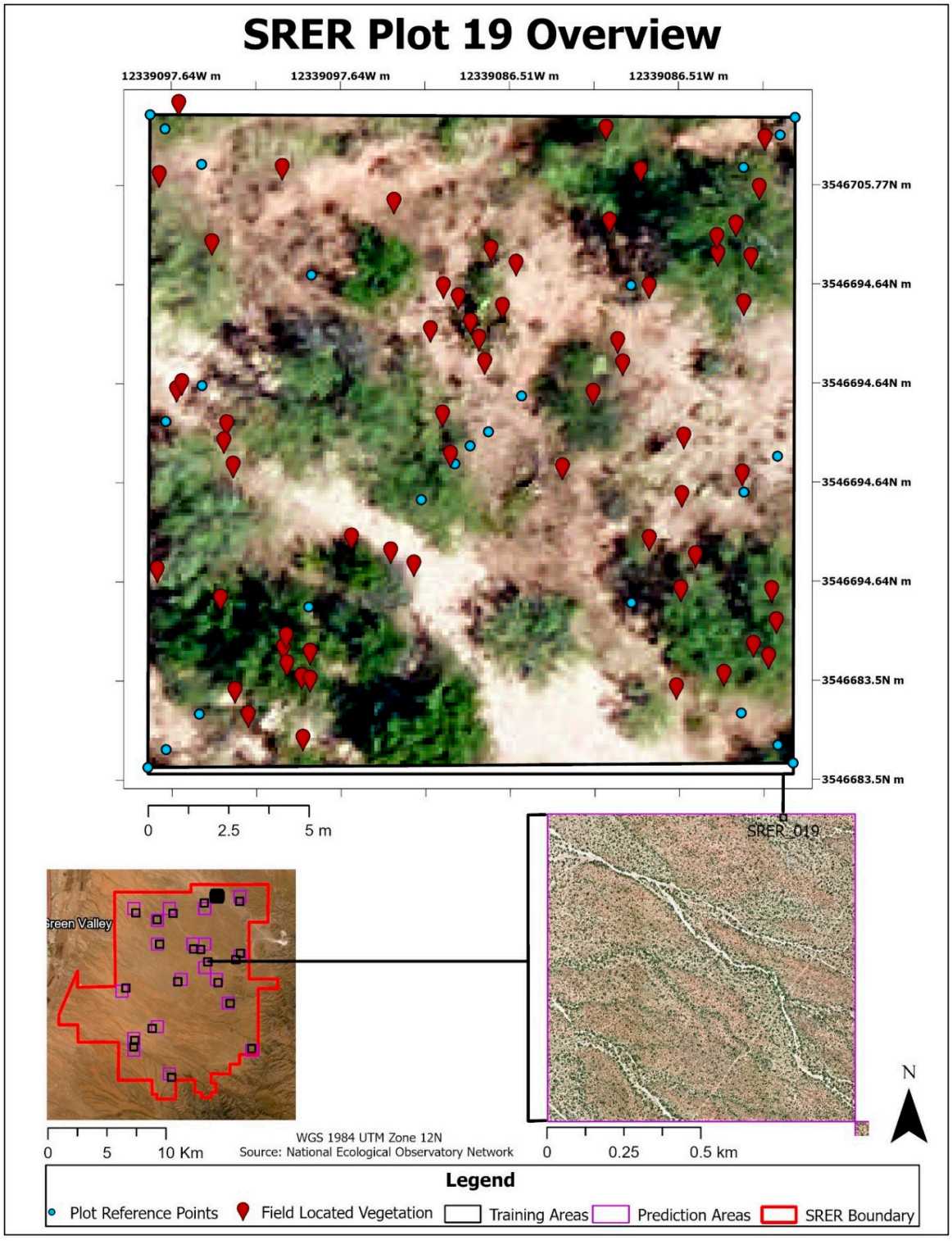

Figure S24. Overview of Plot 19 (high FWC). Average FWC between binary and multiclass training data was 66.16%. Vegetation for this plot is much denser and complex than Plots 14 and 46. Vegetation was sampled throughout the plot with some points being misaligned with imagery due to GPS inaccuracies.

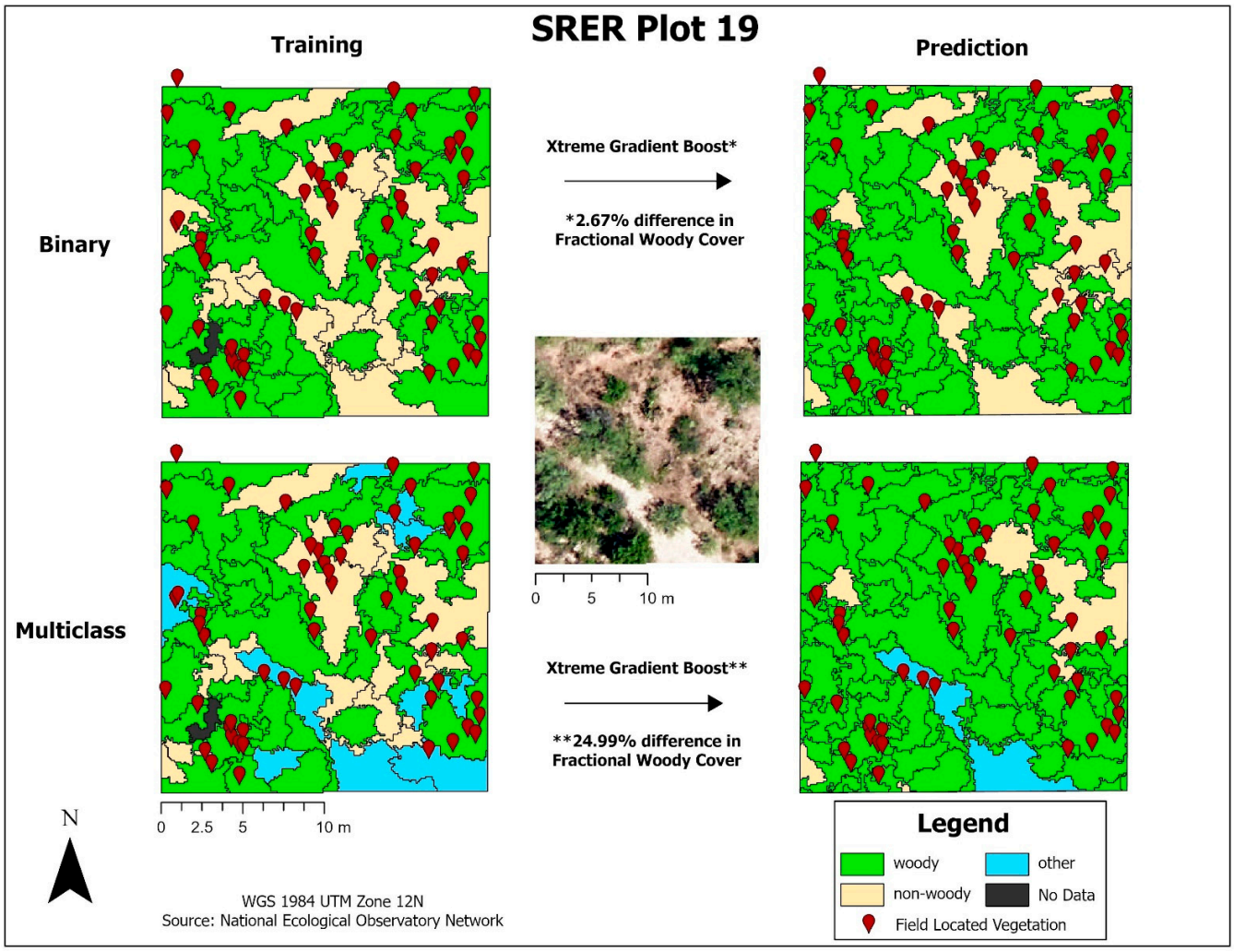

Figure S25. XGB predictions for Plot 19 under both binary and multiclass classification schemes. Overall estimations of FWC are similar between training and prediction methods for the binary scheme, but significantly larger (~25% overprediction) for the multiclass scheme. Most differences in classification occur in polygons that appear less green in imagery. Some of these areas are gaps between the woody canopy and may have high vegetation index values from herbaceous vegetation, leading to their classification as woody by the model (polygons just southeast of center for example). This presents a possible limitation in the model, as it may be relying more heavily on vegetation reflectance than vegetation structure to classify records as woody. Field verification of herbaceous vegetation locations would help determine if this is true or not, but herbaceous locations are not mapped by the TOS. This misclassification issue is more prevalent in the multiclass predictions, as seen in the map and high dFWC value. Model predictions also seem to predict woody vegetation in more clustered groups compared to the training data which is more sporadic. This highlights human ability to discern potential in vegetation relative to the model. Plots with high FWC are clearly the biggest source of error for this model.

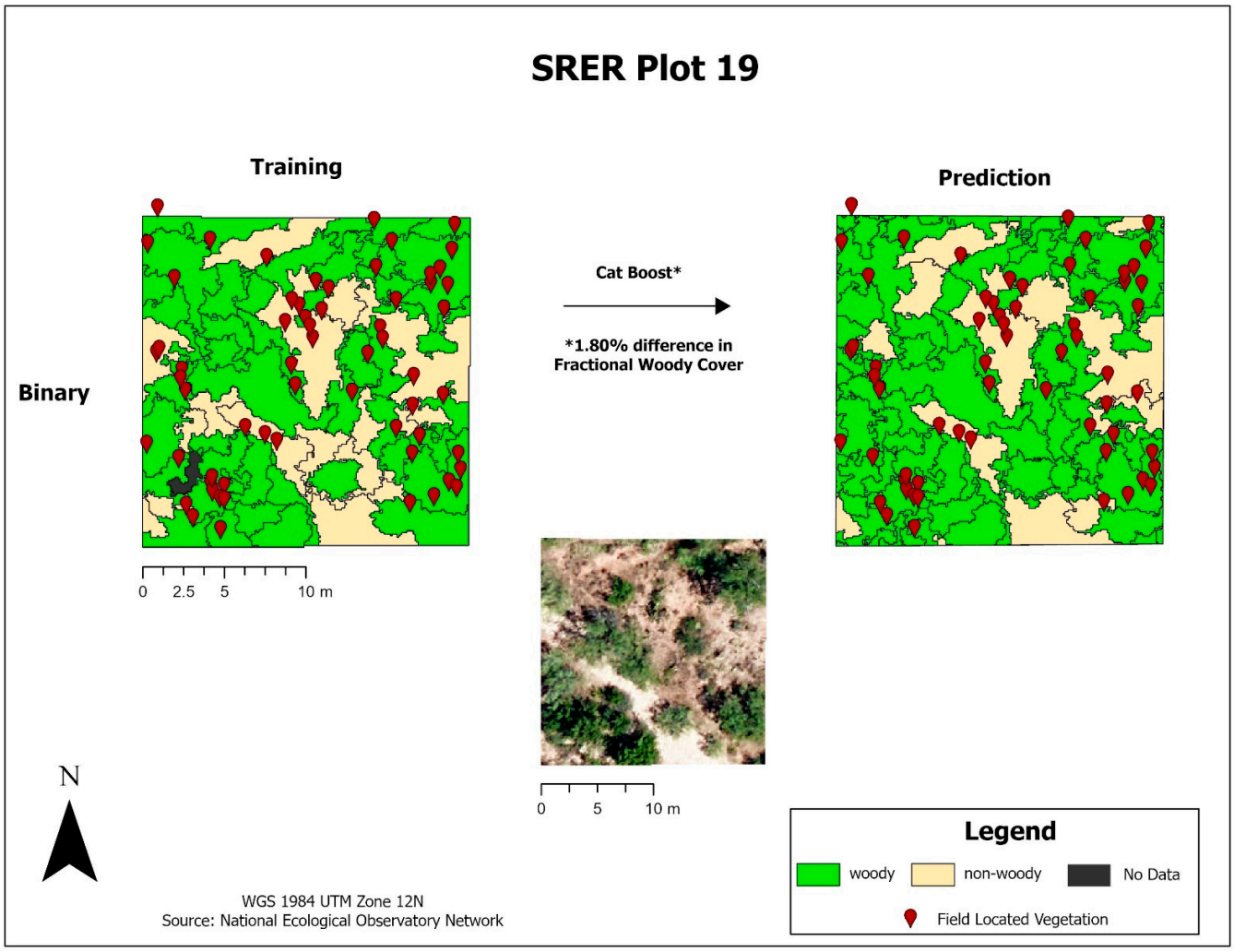

Figure S26. Cat Boost predictions for Plot 19. Multiclass is not shown, as it was not supported by the Cat Boost model. As observed in the XGB model, less-green gaps between woody vegetation are being classified as woody. Despite this misclassification, dFWC is relatively low with the model overpredicting woody vegetation by 1.8%.

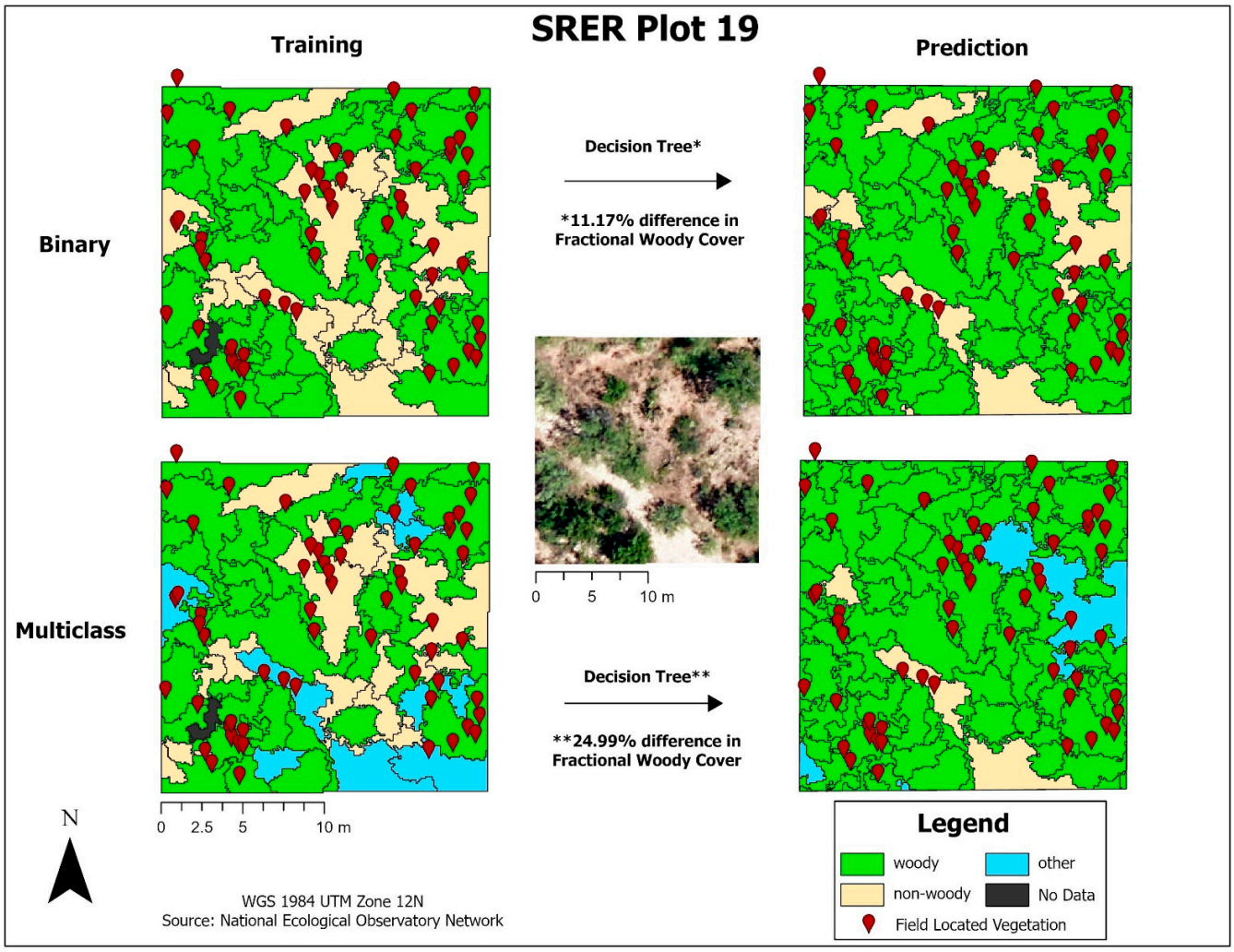

Figure S27. Decision Tree predictions for Plot 19 for the binary and multiclass schemes. The Decision Tree model performs relatively poorly in this high FWC plot under both classification schemes (dFWC = 11.17% and 24.99%). Decision Tree also predicted more "other" cover at this plot relative to plots 14 and 46. Decision Tree overpredicted FWC in the multiclass scheme with a relatively high dFWC value of 4.65%. Like the previous models, woody vegetation was predicted in gaps between the canopy with less-green vegetation.

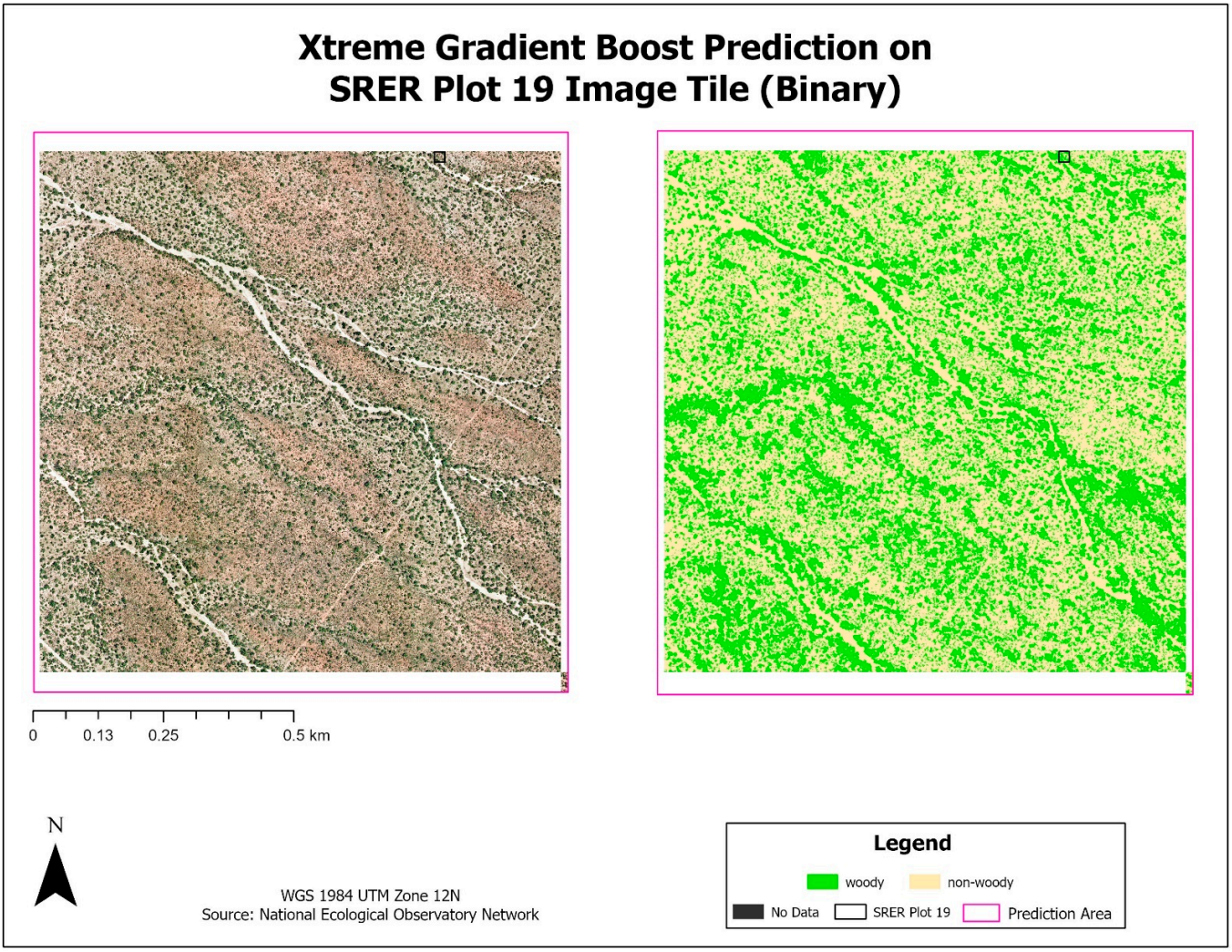

Figure S28. XGB predictions for the Plot 19 image tile using the binary classification scheme. FWC for the entire prediction area is 43.3%. Areas with higher densities of woody plant cover are clearly highlighted in green in both the RGB and prediction tile, along with some delineation of water and road/trail features.

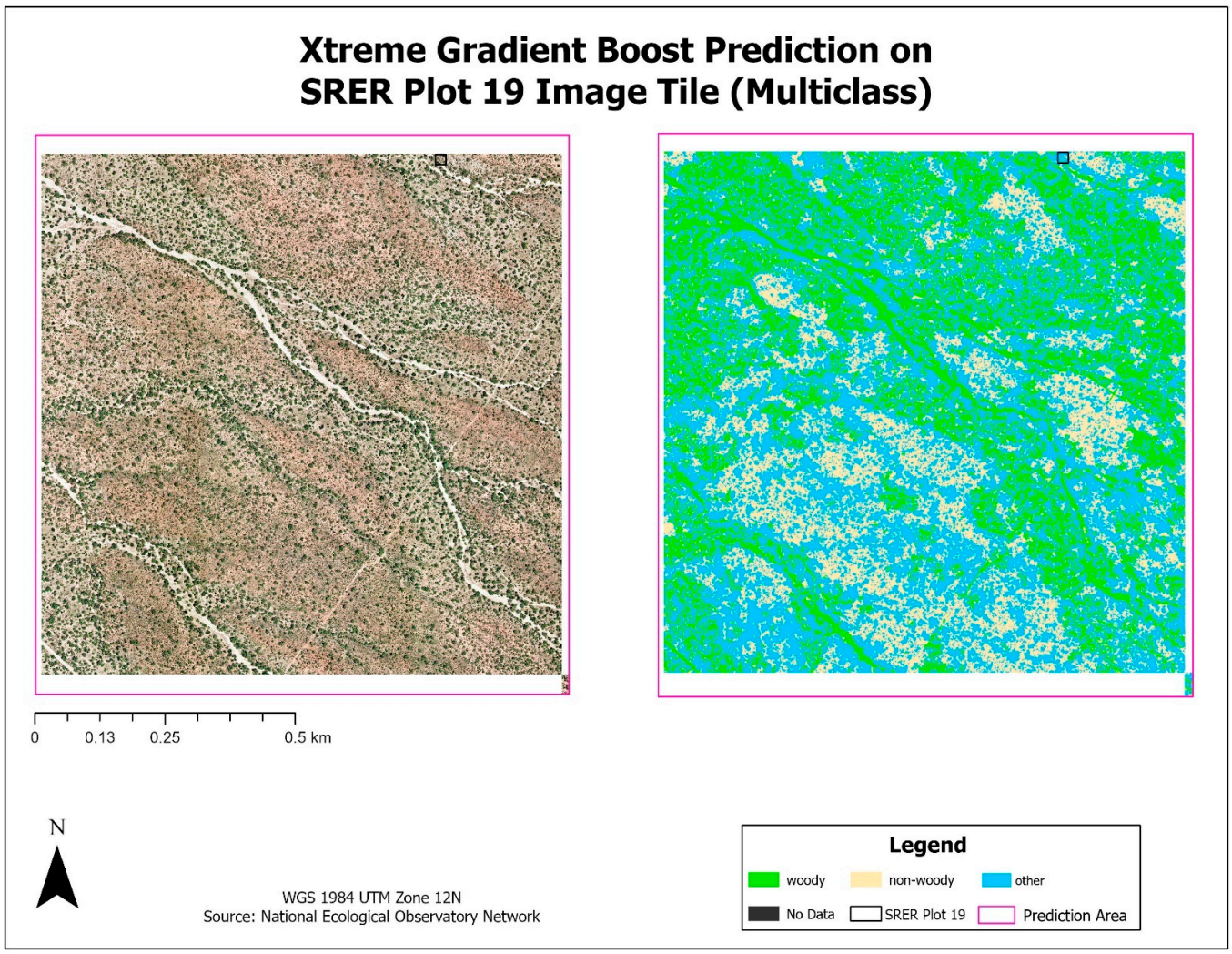

Figure S29. XGB model predictions for the Plot 19 image tile using a multiclass classification scheme. FWC for the entire prediction area is 48.2%. Major woody vegetation patterns appear in green for both the RGB and prediction tile but appear to be overpredicted when looking at more continuous distributions of woody plants in the prediction compared to the imagery. Some waterways also appear to be classified as woody vegetation. Some areas of bare ground, roads, and water features are being classified in blue as "other". Areas being classified as non-woody appear to have a reddish soil color.

Overall, Plot 19 had poor agreeance in FWC between the training and prediction classification methods (mean dFWC for XGB = 13.83%). The XGB model overpredicted woody vegetation overall, as indicated by the positive dFWC value and visual inspection showing clustered predictions in areas of woody vegetation and areas of potential non-woody vegetation and some waterways being classified as woody. This misclassification of higher complexity sites is consistent with suggestions by [23]. The prediction tiles for this plot highlights areas of woody vegetation with lower consistency between classification schemes compared to the less complex areas (dFWC between schemes = 4.9%). The multiclass scheme shows additional detail with the non-woody class potentially highlighting areas of interest for further field data collection and verification