CLASSIFICATION OF CHEATGRASS INVASION USING BIOPHYSICAL AND REMOTE SENSING DATA

Aaron Tuor & Kyle Larson
Pacific Northwest National Laboratory
{aaron.tuor, kyle.larson}@pnnl.gov

ABSTRACT

In this study we explore machine learning approaches which incorporate geophysical data and readings from multiple remote sensing platforms for predictive mapping of vegetation over very large, ecologically diverse regions. We focus specifically on mapping an invasive exotic annual grass, cheatgrass (*Bromus tectorum*), that has become a major land management issue in the Western United States. We explore the advantages of integrating imagery from both the LandSat7 and MODIS platforms which have complementary spectral bandwidth, spatial resolution, and temporal frequency. Our experiments demonstrate that machine learning methods integrating biophysical variables and imagery from multiple platforms can be beneficial to the predictive mapping of invasive vegetation. Our best performing model is used to create a map of coverage predictions to aid in land management over 288 million acres of the Western United States.

1 INTRODUCTION

Accurate and up to date maps of vegetation coverage are critical assets for land management agencies tasked with mitigating damage wrought by invasive species. Current methods for mapping invasive annual grasses from satellite and aerial imagery usually involve choosing one platform. However multiple platforms may potentially be useful, e.g., using finer spatial resolution imagery typically improves prediction, but platforms with more frequent temporal resolution are also important for detecting invasive species which maintain growth cycles distinct from native vegetation. In this work we demonstrate the effectiveness of machine learning models which integrate biophysical data and multiple satellite platfroms.

Introduced in the late 19th-century, cheatgrass is now found in every state in the contiguous U.S. Most prevalent in western states, it has become a dominant component in many shrubland and grassland ecosystems. Cheatgrass invasion poses a variety of threats to ecosystem function, rangeland health, and human safety. A central thread to many of these threats is the increase in fine fuels, which can lead to increased fire frequency and irreversible loss of native vegetation and wildlife habitat. Following a fire, cheatgrass is able to more effectively compete with native vegetation, giving rise to a positive feedback cycle of further invasion and fire. [Refs: Mack 1981; Pellant 1996; Bradley & Mustard 2005; USDA Plant Database]

While cheatgrass is considered ubiquitous throughout much of the western U.S., detailed spatial information about its presence and abundance are still lacking across this extent. Our study area encompasses the U.S. portion of the sage-grouse range, which is approximately 288 million acres. Figure 1 shows the region considered and location of field samples we use as ground truth for training predictive models.

2 DATASET

In this section we explain the data used in this study: field observations of cheatgrass cover, time series of MODIS and Landsat-7 data, and spatial data of various biophysical parameters.

We obtained 6650 field observations from multiple sources (Figure 2) that collected vegetation measurements between 2001 and 2014. All field data were collected along transects ranging from 25-m

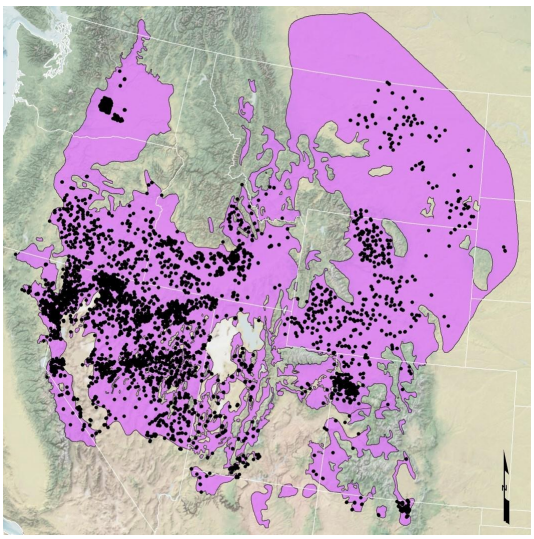

Figure 1: Study Region

Table 2.1. Data Source and Number of Field Measurements Evaluated.

Description	Number of Field Measurements Reviewed	Number of Field Measurements Accepted
BLM Assessment, Inventory, and Monitory Program	2043	1927
BLM Landscape Monitoring Framework	2335	2222
Global Invasive Species Information Network	4292	0
Joint Base Lewis-McCord Yakima Training Center	382	375
PNNL – Birds of Prey field campaign	92	81
PNNL – Hanford Vegetation	39	39
PNNL – Owyhee field campaign	30	29
PNNL – Shoshone field campaign	95	89
USGS SAGEMAP GIS database	820	818
Sagebrush Steppe Treatment Evaluation Project	1086	1070
U.S. Forest Service Forest Inventory Analysis	13170	0
Total	24384	6650

Figure 2: Field Data

Variable	Description	Source
C_1	Soil temperature and moisture regime classes	Chambers et al. 2014; SSURGO; STATSGO
C_2	Generalized vegetation cover type	LANDFIRE
C_4	EPA Level III Ecoregions.	U.S. EPA 2013
$X_{1:2}$	30-m gridded latitude & longitude	Google EarthEngine
X_3	Elevation above mean sea level.	USGS 2018
X_4	Potential Relative Radiation.	USGS 2018; Pierce et al. 2005
X_5	Median winter precipitation.	PRISM Climate Group
X_6	Median growing degree days.	Thornton et al. 2014
$X_{7:50}$	30-year normal climatic conditions.	PRISM Climate Group
$X_{51:323}$	Landsat-7 bands 1-10 (annual, spring, summer).	Roy et al. 2010
$X_{324:664}$	MODIS bands 1-8 (annual, spring, summer)	Didan 2015.

Table 1: Continuous Variables. Shaded variables were used in making the original Linear Discriminant Analysis model.

to 100-m in length. A strong break was observed in the distribution of cheatgrass canopy cover values at approximately 2 percent canopy cover (Figure 2) which coincidentally represented approximately equal percentages of the data (48 % and 52%); thus, we chose to map two classes of cheatgrass above and below this natural break. By selecting these classes we avoid the assumption of true absence for sites where cheatgrass was not observed.

The study area spans 23 EPA Level III ecoregions which we use as a categorical variable. Two other categorical biophysical variables soil moisture-temperature regimes and existing vegetation type were selected because they are useful indicators of biophysical conditions that affect plant community dynamics. We also include an additional 50 continuous variables related to climate, elevation, and location. Table 1 contains brief descriptions of all variables used in modeling cheatgrass coverage.

Time series of both annual- and seasonal-composite MODIS (Terra, 2001-2016) and Landsat-7 (2003-2012) data were used in this study. Both sets of imagery were composited to reduce residual cloud and aerosol contamination and reflect the time of peak vegetation vigor as determined by maximum NDVI within the composite period (Didan 2015; Roy et al. 2010). We used seasonal composite images for spring and summer periods which correspond to distinguishing periods in the life cycle of cheatgrass.

We selected four subsets of predictor variables that represent two generic approaches to mapping cheatgrass: one approach founded on ecological niche modeling, and the other founded on combining ecological-spectral modeling approaches. The purpose of using these subsets was twofold: to evaluate potential gains and losses in classification accuracy by combining ecological and spectral data; and to evaluate gains and losses in classification accuracy by combining data from multiple sensors. For the purpose of defining variable subsets

$$
\begin{aligned}
D_1(\mathbf{x}) &= \mathbf{x}_{1:50} && \text{Biophysical variables} && (1)\\
D_2(\mathbf{x}) &= \mathbf{x}_{1:50,326:664} && \text{Biophysical varibles + MODIS} && (2)\\
D_3(\mathbf{x}) &= \mathbf{x}_{1:325} && \text{Biophysical varibles + LandSat} && (3)\\
D_4(\mathbf{x}) &= \mathbf{x} && \text{All variables} && (4)\\
& && && (5)
\end{aligned}
$$

3 METHODS

We assessed four machine learning methods for predictive mapping of cheatgrass occurrence: Random Forest (RF), Logistic Regression (LR), a Deep Neural Network (DNN), and a joint Recurrent Neural Network model (JRNN). The classification objective for all models was to map two classes of cheatgrass. All continuous predictor variables were standardized by subtracting their respective mean and dividing by the standard deviation. The input to our machine learning models contains three values from categorical variables. We employ two simple strategies to map categorical values to vectors: one-hot vectors for RF, and embedding vectors for LR, DNN, and RNN. The LR, DNN, and RNN models were implemented in Tensorflow and RF models were implemented in scikit-learn.

4 RANDOM FOREST

We selected RF for comparison to DNN and RNN methods because it has been shown to perform well for predictive vegetation mapping (citations), be resilient to overfitting (citations), and provide competitive results compared to deep learning models in low resource data regimes (citations). Key parameters we tuned (and their final values) were as follows: sampling method (with and without replacement), criterion for splitting nodes (GINI index), maximum number of features (square root of the total number of features), minimum number of samples in a leaf node (1, 2, 4), minimum number of samples in a split node (2, 5, 10), maximum depth of a tree ([10, 110]), and the number of decision trees in the forest (10-200).

5 LOGISTIC REGRESSION

To provide perspective on the value of deep learning for deriving highly predictive feature representations for ecological modeling, we explore as a baseline linear LR models based directly on the standardized geospatial and biophysical predictor variables.

Since our data contains categorical variables derived from ecological analysis which we wish to incorporate into the model we define three learned embedding matrices for the categorical variables, $\mathbf{W}_{\texttt{cover}} \in \mathbb{R}^{15\times k}$, $\mathbf{W}_{\texttt{eco}} \in \mathbb{R}^{23\times k}$, and $\mathbf{W}_{\texttt{soil}} \in \mathbb{R}^{9\times k}$, corresponding to the 15 classes of derived generalized LANDFIRE cover classes, the 23 level III ecoregions, and the 9 soil temperature and moisture regime class values exhibited in the full study region. The rows of these matrices correspond to classes for their respective categories.

5.1 DEEP NEURAL NETWORK

We define the inputs to the neural network as a real valued vector $\widehat{\mathbf{x}}^{(i)}$, it's predictions $\widehat{\mathbf{y}}^{(i)} \in \mathbb{R}^2$, a probability distribution over 2 classes, and the ground truth for a field location as $\mathbf{y}^{(i)}$. Since we are interested in an ablative analysis of the contributions in modeling performance from using data from satellite platforms jointly and independently we define subsets on the full set of continuous variables defined as $\mathbf{x}$ in Table 1:

Particular deep neural network (DNN) configurations for classification and regresssion are depicted in Figure 3. The model consists of L hidden layers $\mathbf{h}_l$ which are recursively defined as:

$$\mathbf{h}_l = \texttt{RELU}(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{b}_l) \tag{6}$$

$$\mathbf{h}_0 = \widehat{\mathbf{x}} \tag{7}$$

where the nonlinear $\texttt{RELU}$ operation is defined as the elementwise vector valued operation:

$$\texttt{RELU}(\mathbf{x})_i = \texttt{max}(\mathbf{x}_i, 0) \tag{8}$$

The output of the DNN maps the final hidden representation $\mathbf{h}_L$ to a probability distribution over cheatgrass occurence classes of the form:

$$\widehat{\mathbf{y}} = \texttt{softmax}(\mathbf{W}_{L+1}\mathbf{h}_L + \mathbf{b}_{L+1}) \tag{9}$$

where the $\texttt{softmax}$ operation is defined as the elementwise vector valued operation:

$$\texttt{softmax}(\mathbf{x})_i = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}} \tag{10}$$

The parameters $\mathbf{W}$, $\mathbf{b}$ are optimized using stochastic gradient descent with the objective function being the cross-entropy loss between the predicted class distribution $\widehat{\mathbf{y}}$ and true class distribution $\mathbf{y}$. Beyond this basic formulation we also employ dropout for normalization to aviod overfitting and batch normalization for better conditioned gradient updates.

5.1.1 RECURRENT NEURAL NETWORK

A deep neural network as described in 5.1 has the potential as a universal function approximator to model cheatgrass distribution given the field data with an optimal Bayes error. In practice however, since the optimal function may be difficult to find during the optimization we developed a joint recurrent neural network (RNN) model intended to take greater advantage of the time series implicit in the satellite imagery data. In this section we detail a model for fusing time series of data from different satellite platforms into a latent representation that can be used for linear classification. The model consists of two separate bidirectional RNNs which model time series of the MODIS and LandSat derived imagery products respectively. A particular instantiation of this modeling approach is depicted in Figure 4. More formally we define two bidirectional Long Short Term Memory networks which are operations on sequences of pixel vectors from the MODIS and LandSat platforms respectively, $\texttt{LSTM}_{\texttt{MOD}}(\mathcal{X})$ and $\texttt{LSTM}_{\texttt{LAN}}(\mathcal{X})$. Given a seqence of vectors $\mathcal{X}$ each LSTM

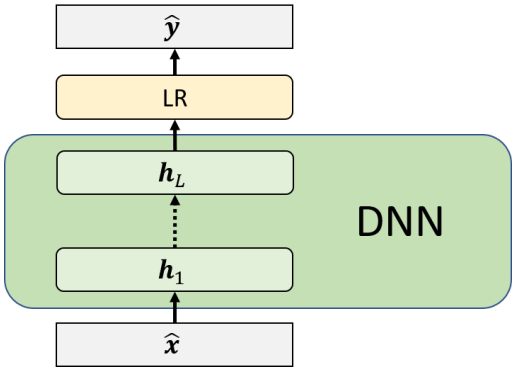

Figure 3: Deep Neural Network for Classification

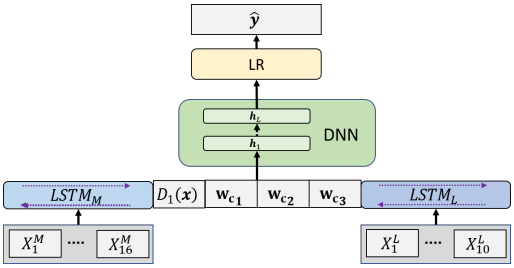

Figure 4: Recurrent Neural Network for Classification

outputs a condensed vector representation of the time series. The time series for the MODIS and LandSat data consist of sequences of 16, and 10 vectors respectively, one for each year that pixel values were obtained. From the notation developed in **??**, the t-th vector in the LandSat series is:

$$\mathcal{X}_t^{\texttt{LAN}} = \left[\mathbf{L}_{t,1}^{\texttt{annual}} \quad .... \quad \mathbf{L}_{t,10}^{\texttt{annual}} \quad \mathbf{L}_{t,1}^{\texttt{spring}} \quad .... \quad \mathbf{L}_{t,10}^{\texttt{spring}} \quad \mathbf{L}_{t,1}^{\texttt{summer}} \quad ... \quad \mathbf{L}_{t,10}^{\texttt{summer}}\right] \tag{11}$$

whereas the t-th vector in the MODIS time series is:

$$\mathcal{X}_t^{\texttt{MOD}} = \left[\mathbf{M}_{t,1}^{\texttt{annual}} \quad .... \quad \mathbf{M}_{t,10}^{\texttt{annual}} \quad \mathbf{M}_{t,1}^{\texttt{spring}} \quad .... \quad \mathbf{M}_{t,10}^{\texttt{spring}} \quad \mathbf{M}_{t,1}^{\texttt{summer}} \quad ... \quad \mathbf{M}_{t,10}^{\texttt{summer}}\right] \tag{12}$$

These resulting representation vectors are concatenated with the categorical embeddings and the continuous biophysical variables for the input to a deep neural network as described in 5.1 where:

$$\hat{\mathbf{x}} = \left[\mathbf{c}_1\mathbf{W}_{\texttt{soil}} \quad \mathbf{c}_2\mathbf{W}_{\texttt{cover}} \quad \mathbf{c}_3\mathbf{W}_{\texttt{eco}} \quad D_1(\mathbf{x}) \quad \texttt{LSTM}_{\texttt{MOD}}(\mathcal{X}^{\texttt{MOD}}) \quad \texttt{LSTM}_{\texttt{LAN}}(\mathcal{X}^{\texttt{LAN}})\right] \tag{13}$$

6 Results and Analysis

In this section we discuss the results of experiments and analyze the model predictions.

6.1 Performance on Field measurement locations

In this section we report statistics on field measurement locations.

6.1.1 Quantitative assessment

Quantitative results are reported for the average performance across cross-validation folds for each of the 4 subsets of input variables. Table 3 shows the outcomes of experiments in terms of accuracy for classification prediction. The best performing subset of input variables is bolded in each row of the tables, while the best overall performing model, the DNN using all available input data is highlighted in red. We note a a few trends. Incorporating satellite data of either form leads to better results. In addition, Random forest and LDA do not gain any advantage from incorporating both MODIS and LandSat imagery whereas the neural network models improve in accuracy from incorporating imagery from both platforms. Finally, we note that the recurrent neural network did not perform as well as the DNN, possibly due to the larger number of parameters to be fit on the same number of data points.

We chose the best performing model in terms of accuracy (DNN with all input variables) to create the final mapping product, and for this model we report several classification statistics for each of the four input variable subsets in Table 6. Here we the same trend for recall, F-score, and area under the ROC curve in that adding imagery from either sattellite platform gives the most significant improvement, with the LandSat imagery demonstrating more improvement. The best results for all these metrics are once again from the inclusion of imagery from both satellite platforms.

	Prec.	Rec.	F-scr	AUC	Acc.	Std.
D_1	0.755	0.820	0.768	0.825	0.767	0.021
D_2	0.795	0.820	0.808	0.870	0.796	0.021
D_3	0.818	0.832	0.822	0.887	0.812	0.019
D_4	0.819	0.827	0.822	0.889	0.814	0.012

Table 2: Performance across a range of metrics for most accurate LR models for each subset of input variables. The last column reports the standard deviation of the accuracy across different cross-validation folds.

	D_1	D_2	D_3	D_4
LR	76.722	79.620	81.194	81.443
LDA	73.816	78.279	**80.653**	79.826
RF	80.165	81.677	82.502	**82.798**
DNN	79.604	82.004	82.456	**83.203**
RNN	—	82.050	82.580	**83.184**

Table 3: Classification accuracy of models for subsets of data

	Prec.	Rec.	F-scr	AUC	Acc.	Std.
D_1	0.795	0.820	0.808	0.866	0.796	0.025
D_2	0.844	0.803	0.823	0.893	0.820	0.014
D_3	0.831	0.833	0.832	0.900	0.825	0.015
D_4	0.837	0.842	0.839	0.902	0.832	0.012

Table 4: Performance across a range of metrics for most accurate DNN models for each subset of input variables. The last column reports the standard deviation of the accuracy across different cross-validation folds.

	Prec.	Rec.	F-scr	AUC	Acc.
D_1	–	–	–	–	–
D_2	0.827	0.829	0.828	0.897	0.820
D_3	0.829	0.841	0.835	0.901	0.827
D_4	0.835	0.844	0.840	0.906	0.832

Table 5: Performance across a range of metrics for most accurate RNN models for each subset of input variables. The last column reports the standard deviation of the accuracy across different cross-validation folds.

	Std.	Prec.	Rec.	F-scr	AUC	Acc.	Std.
D_1	–	0.807	0.815	0.811	0.883	0.802	0.024
D_2	0.021	0.825	0.823	0.824	0.891	0.816	0.024
D_3	0.012	0.827	0.838	0.833	0.901	0.825	0.021
D_4	0.012	0.835	0.835	0.835	0.904	0.828	0.025

Table 6: Performance across a range of metrics for most accurate RF models for each subset of input variables. The last column reports the standard deviation of the accuracy across different cross-validation folds.

	D_1	D_2	D_3	D_4
Dropout Probability	0.275	0.350	0.30	.5
Layer 1	256	256	256	256
Layer 2	256	256	64	32
Layer 3	64	128	256	256

Table 7: Hyperparameters for most accurate DNN models for each subset of input variables.

6.1.2 TRAINING STATISTICS

In this section we analyse training statistics for the best performing model. We first look at plots of train versus test performance for the best performing cross-validation fold for each of the input variable subsets. Looking at several different metrics we see training is noisier and has a greater tendency to overfit with inclusion of the modis data. Conversely we see that incorporating the LandSat data into the model has a regularizing effect which limits the amount of overfitting so that test performance converges closer to train performance. We also consider the dropout probability for each data subset (for each model run a random dropout probability was selected to be between 0.5 and 1.0. Table **??** shows the dropout probabilities and hidden layer sizes for the different data subsets. The columns for using MODIS only or LandSat show that these models have comparable dropout probabilities and network size. The MODIS only input model does have the largest network which greatens the chance of overfitting but a large dropout probability which should have more of a regularizing effect. A more careful analysis with identical sized networks and dropout probabilities would confirm what the present analysis suggests (that the MODIS data has some properties which encourage overfitting). With careful use of early stopping we can see in figure 53 that the best loss can in fact be achieved by incorporating all the data.

6.1.3 QUALITATIVE ASSESMENT

In this section we characterize the overall performance of the best performing DNN models. Figure 16 shows a confusion matrix for classification for the best performing DNN model using all the input variables. The types of mistakes the model makes in classification are evenly balanced in terms of false positives and false negatives, however since a smaller proportion of the data contains negative examples it in fact slightly favors false positives. We can see this more clearly in figure 63 which plots accuracy in terms of actual cheat grass coverage values. The worst performance is in the 0-2 percent range which contains only negative examples, so the only mistake that can be made is a false positive.

Looking closer at prediction performance we next analyze the model in terms of accuracy across individual ranges of cheatgrass coverage, ecoregions, land cover classes, and soil and climactic conditions. Figures 60-61 show that the models incorporating the LandSat data attain 100 % accuracy for actual coverage ranging from 70-100 % cheat grass coverage and accuracy above 90 % up down to 40 % coverage. This is despite the fact that there are fewer examples in the data set for these high coverage locations. Since the overwhelming majority of false positives and false negatives fall within the 0-40 % range, this suggests that changing the class boundary from 2% to 40% may lead to more accurate classification at the expense of the overall utility of a final mapping product. We pay particular interest in judging if the model performs better across different ecoregions due to an unbalanced representation in the data set or something intrinsic in the environment itself. Across the other ways of splitting the data into ecoregion, soil moisture temperature regime, and land cover class we see that accuracy is not correlated with frequency of occurrence. Although there does seem to be a minimum threshold for number of examples for an ecoregion to attain acceptable accuracy, considering the low performance of ecoregions 25 and 16 and corresponding small number of examples.

- Map of field data symbolized by classification error

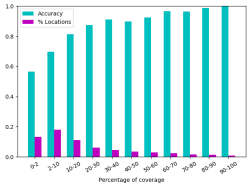

Figure 5: D_1

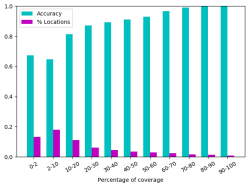

Figure 6: D_2

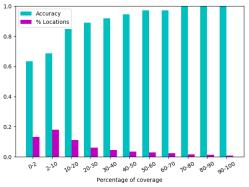

Figure 7: D_3

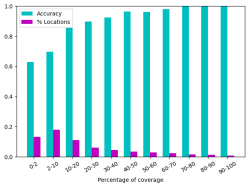

Figure 8: D_4

Figure 9: Accuracy for ranges of cheat grass coverage.

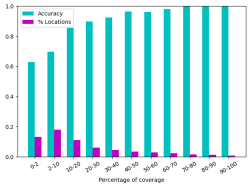

Figure 10: D_1

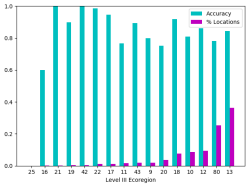

Figure 11: D_2

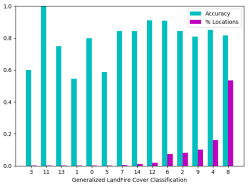

Figure 12: D_3

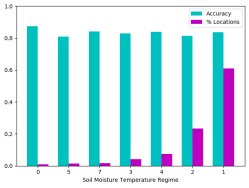

Figure 13: D_4

Figure 14: Accuracy for different subsets of the field locations.

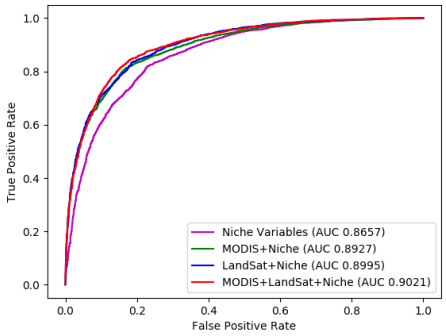

Figure 15: ROC curves for best performing DNN models on different data subsets

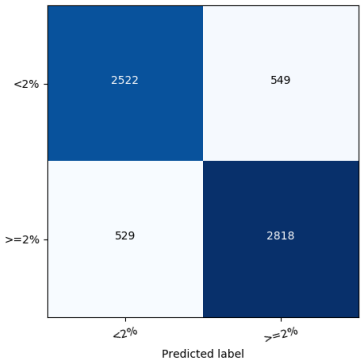

Figure 16: Confusion matrix for best performing DNN model.

6.2 Performance on Full Study Region

In this section we analyze the geographic variability in cross-validation model predictions among other things.

6.2.1 Pixel-wise count of cheatgrass class across model predictions

6.2.2 Block statistics (count) of cheatgrass class (analogous to block area), then look at range of block counts

7 Conclusion and Future Work

Acknowlegements

This work was funded by the Deep Learning for Scientific Discovery Laboratory Directed Research and Development investment at the Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.