Image Segmenters Final Report

Henry Dikeman, Ben DeNio, Reyhaneh Rahimi Nahouji

05/04/21

Abstract

Over the next decades, climate change combined with demographic and environmental pressures are expected to have significant disruptive effects on food systems, particularly in developing countries where less-adaptable, rainfed agricultural systems are dominant. Therefore, accurate and real-time survey information of agricultural fields is important for providing actionable information to farmers, community leaders, and policymakers by enabling them to evaluate aggregate crop statistics and generate mitigation plans to ensure food security. This project focuses on a limited agricultural area delineation task conducted with Iranian farmland image data captured with Sentinel 2 satellite imagery. Satellite pixel data was converted to record data and processed using several sets of classifier models. 12 classification algorithms were fit to this dataset and their performance characterized using accuracy, F1, precision, recall, and TPR/FPR scores. Classifiers were selected from a diverse range of classification modelling algorithms to cast a wide net and determine whether any models not yet pursued for this classification task show robust classification performance. Initial modelling suggested that random forest and multilayer perceptron models were the best general-purpose model for this task, with other models identified as optimal for high precision or high recall modelling. Further optimization of these classifiers was pursued, with the final optimal random forest model yielding a classification accuracy of **89.2%**, F1-score of **0.701**, and TPR/FPR score of **14.28**. For further analysis, we recommend extracting greater information from spatial and temporal attributes of Sentinel 2 data and determining to what extent classification performance improves.

Motivation and Literature Review

Land area classification from satellite imagery is a rapidly growing area of study. Land area classification seeks to reduce the various impacts of climate change and population growth. Data collected from agricultural maps can inform food system management, from local level administration to national policy making. This oversight includes a range of government policies, from water rights to subsidy allocation (Akbari et al., 2007; .Garcia-Pedrero et al., 2018).

Duro et al. (2012) used pixel-based and objective-based image analysis approaches for classifying land cover. They compared three supervised machine learning algorithms: decision tree (DT), random forest (RF), and the support vector machine (SVM), and found they performed equally in farmland classification tasks. Also, Kussul et al. (2017) compared RF and MLP with convolutional NNs which are widely used in remote sensing. They found convolutional NNs outperform the others in detecting agricultural fields, especially in summer.

Data Collection

Data was sourced from Sentinel 2 satellite imagery over Iran. ARC GIS was used to analyze the imagery and split imagery into manageable sections: six 1000 x 1000 pixel squares. The raw image data from these 6 images are stitched together and visualized below, first for each channel independently then for the original three channel RGB image:

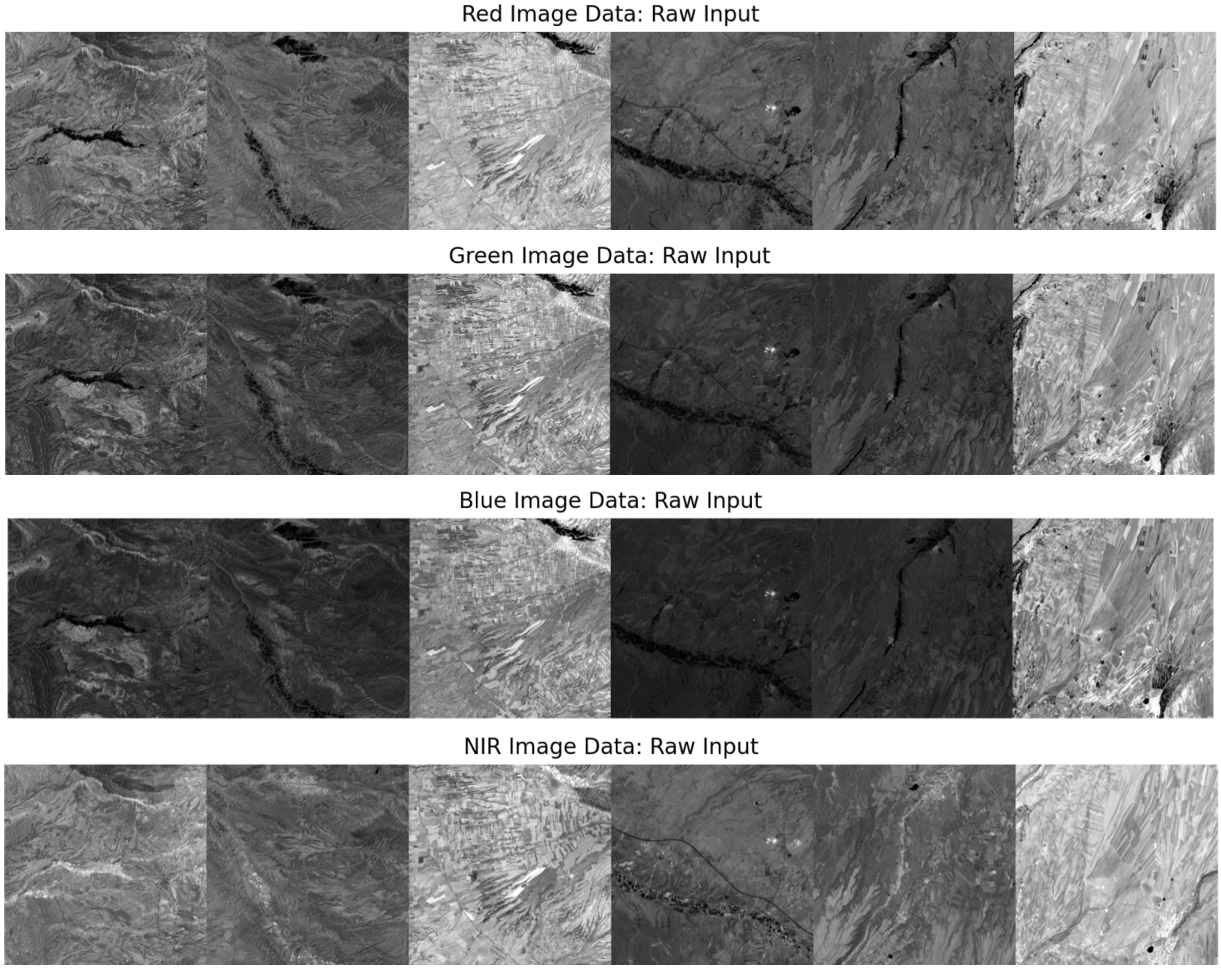


Figure 1 | Original Sentinel 2 image channels for all 6 images. From top to bottom: (1) red (2) green (3) blue and (4) NIR input channels

The original, raw satellite image was taken in Iran from a Sentinel 2 satellite, where the raw image is represented here:

Figure 2 | Raw image data, displayed as an RGB image for all 6 satellite images arrayed side-by-side.

Originally this research was planned to evaluate training images for a range of months. However, the process of labeling the data by hand was quite considerable even for this six 1000 x 1000 training image dataset. Instead of dedicating additional effort to growing the dataset, resources were instead directed towards effectively modelling this small set of images, treating it as a representative sample. For each 1000 x 1000 image, a binary label map was added to the imagery. This map was then used as labeled training data, shown below:

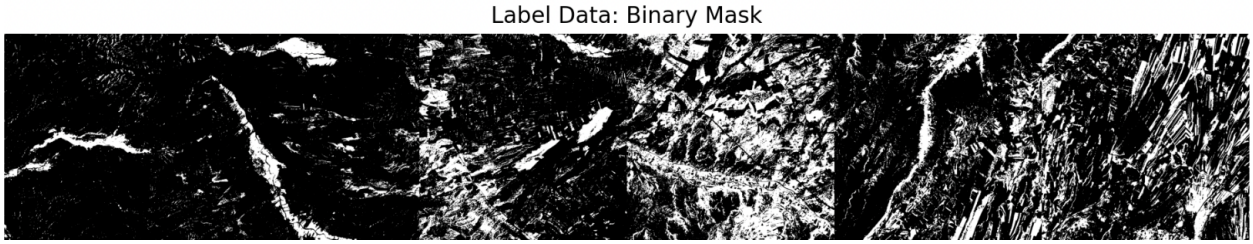
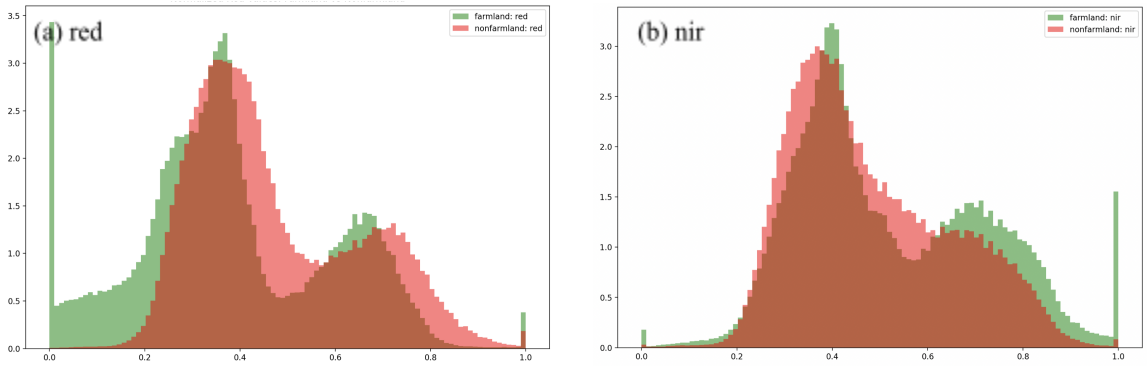


Figure 3 | A binary mask over Sentinel 2 satellite imagery, where white is farmland pixels.

This data was originally stored in raster format, which was then converted to CSV record format. Although this approach did not retain the spatial relationship between pixels, record format is preferred for most classifier models which rely upon a feature → class mapping.

Data Exploration

Before developing classifier models for the farmland classification task, this set of 6 labeled images was investigated to find identifiable patterns in the strength of certain wavelength bands with respect to class. First, the initial set of 4 channels (red, green, blue, nir) was investigated. The average channel intensity for farmland and nonfarmland areas was plotted below:



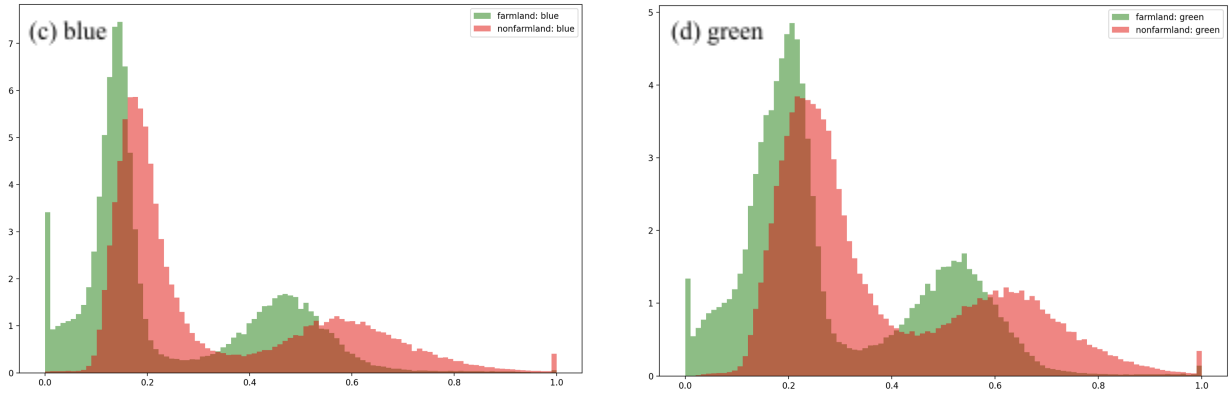


Figure 4 | Normalized histograms of pixel intensities stratified by class, where red indicates nonfarmland and green indicates farmland pixels. (a) red (b) near infrared (c) blue (d) green

Several conclusions can be drawn from this simple visualization: (1) there is no single distinguishing feature for this dataset (2) channel intensities are bimodal due to different overall brightness levels for different images (3) since these intensities are normalized in the range [0,1], it also appears that images captured in Sentinel 2 imagery run the risk of saturating certain channels given the presence of peaks at both 0 and 1. This implies that images captured at too dark or too bright of an overall light level diminish the ability to evaluate these pixels. These three factors increase the difficulty included in an already difficult classification task.

While (1) and (3) are problems inherent in the dataset and cannot be significantly rectified, the bimodal nature of the channel intensities can be somewhat overcome by developing synthetic features which are inherently normalized. The feature that was selected for this work is often incorporated into farmland classification, known as normalized difference vegetation index (NDVI). This feature is defined by the equation below:

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

This index attempts to capture a normalized representation of the presence of live, green vegetation on the pixel of interest. The physics behind this measure are relatively complex, but depend on the light absorption properties of chlorophyll compounds contained by living plants. The rationale behind the use of this feature is that farmland areas are more likely to contain living green vegetation than nonfarmland areas. The fact that this feature is, by nature, normalized across images taken at different overall brightness levels is an added benefit. When this feature is generated for all pixels across the sample space and visualized similarly to the original pixel channels above, the following image data is produced:

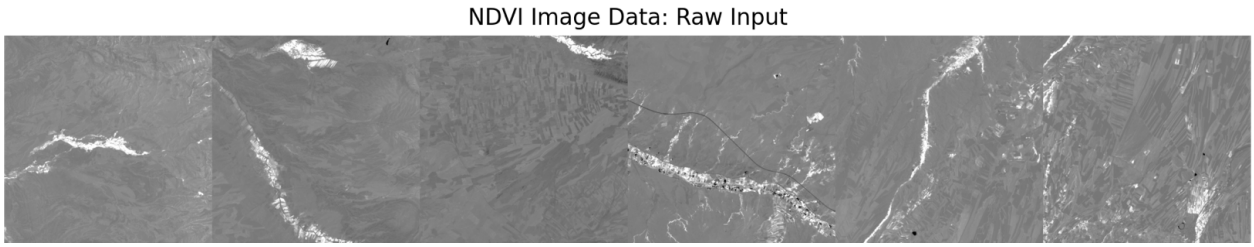


Figure 5 | NDVI synthetic feature for all six images, where white are areas of high NDVI

This NDVI image data is intuitive once you consider the region of study: Iran, as an arid region, is mostly devoid of vegetation and farmland barring riverbank regions. Thus, NDVI values reach their maximum primarily in these riverbank regions, as is clear above for the narrow, long strips of high NDVI values. Summary statistics of this NDVI data generated the intensity frequency graph shown below:

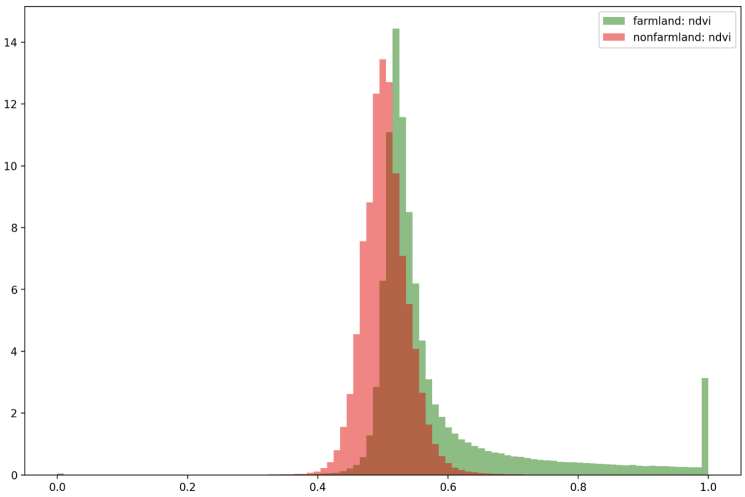


Figure 6 | Normalized NDVI values for all image data, with farmland areas indicated in green and nonfarmland areas indicated in red.

This NDVI synthetic feature is more independently distinguishing than the original set of image channels, with highly green areas readily distinguishable as farmland, but with most pixel areas still difficult to distinguish as farmland or nonfarmland (the peaks centered on 0.5)

All features were standardized using the StandardScaler method from sklearn before fitting classifiers. This was done in order to prevent overweighting certain channels according to differences in magnitudes. With this set of original features and synthetic features prepared and processed, the dataset was now ready to evaluate with classifier models.

Data Analysis

Initial Modelling

After image data was collected and transferred to a csv record format, a broad range of classifier models were applied to the dataset of 6,000,000 pixels to assess the relative performance of different classifiers on this dataset. Classifiers were selected from a diverse range of classification algorithms to evaluate the most successful classification scheme for this data. The families of classifiers tested on this dataset are listed below:

1. **Decision tree methods:** single decision tree and forest methods
2. **Bayesian methods:** Gaussian and Bernoulli naive Bayesian models
3. **Artificial neural network methods:** simple feedforward neural network model
4. **Ensemble methods:** voting classifiers, random forests, bagged and boosted models
5. **Similarity methods:** k-nearest neighbor
6. **Regression methods:** ridge classifier, support vector machine

These models were applied with default parameters using the sklearn library to assess the relative performance of each modelling approach. While there is some argument to be made for the fact

that parameter tuning can drastically change the outcome of a given modelling approach, it was determined for this set of models that the default sklearn parameters were reasonable and would serve as a rough estimate of modelling performance. The effect of tuning was considered when comparing models, and classifiers which performed with accuracies ±2% of each other were considered to have nearly equal performance.

Before conducting this initial analysis, hypotheses were formed as to which models would perform the farmland classification most robustly for this dataset. Given that no features are individually distinguishing and feature values are highly dependent on the overall brightness of each satellite image, this classification problem is considered to be quite difficult. It was expected that artificial neural network methods and ensemble methods would perform with the highest accuracy on this dataset given the conclusions of literature research on this topic.

It was also expected that naive Bayesian methods would perform poorly given the lack of conditional independence for bimodal features, which makes distinguishing between bright and dark brightness-level farmland areas difficult. Single decision tree methods were also expected to perform poorly on this dataset given their tendency to overfit, which is an even larger problem when grown to full depth on a dataset where performance is highly limited by nondistinguishing features.

After applying several classifiers from each group listed above, the following performance graph was generated, graphed using accuracy as a metric:

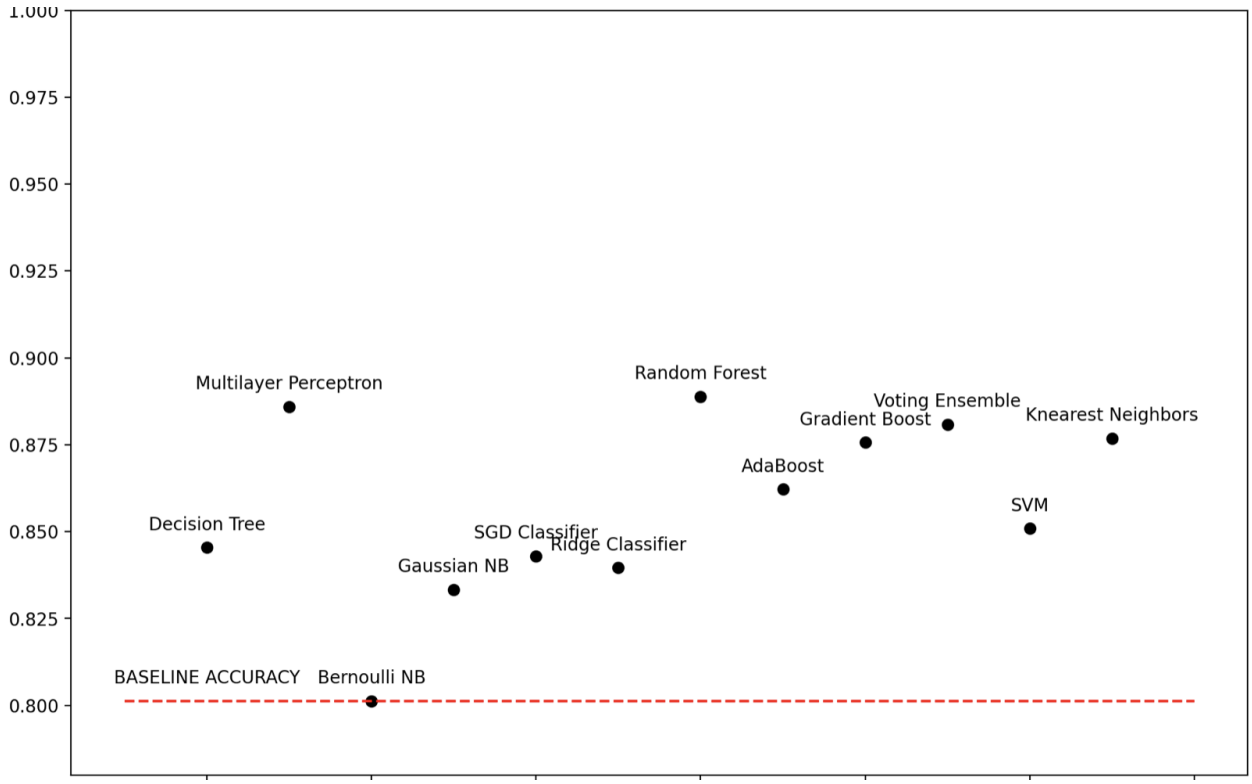


Figure 7 | Accuracy results for OOTB classifier models from sklearn, with baseline accuracy for a model predicting all areas as nonfarmland shown in red.

These performances were also tabulated below along with other measures of performance, including F1, precision, and recall:

Table 1 | Untuned classifier performance, including accuracy, F1, precision, recall, and TPR/FPR metrics of performance.

Classifier	Accuracy	F1	Precision	Recall	TPR/FPR
Decision tree	0.846	0.616	0.609	0.622	6.27
Multilayer perceptron	0.885	0.673	0.782	0.591	14.5
Bernoulli naive Bayes	0.801	0.0	N/A	0	N/A
Gaussian naive Bayes	0.833	0.395	0.710	0.273	9.85
Stochastic Gradient Descent	0.843	0.370	0.913	0.232	42.3
Ridge Classifier	0.840	0.340	0.933	0.208	56.0
Random forest	0.889	0.694	0.766	0.634	13.3
AdaBoost	0.862	0.568	0.753	0.457	12.3
Gradient Classifier	0.876	0.629	0.773	0.530	13.7
Ensemble Voting Classifier	0.881	0.657	0.766	0.576	13.22
Support vector machine	0.851	0.463	0.323	0.815	17.6
K-nearest neighbors	0.877	0.667	0.622	0.720	10.4

As expected, the neural network and random forest methods performed most robustly on this dataset, with high accuracy over 88% and F1 scores both above 0.67. These models also strike a balance between recall and precision, with similar TPR/FPR values. This makes sense, since these models have sufficient degrees of freedom to extract the maximum amount of distinguishing information from the set of overlapping original and synthetic features.

As hypothesized above, both Bayesian methods that were fit to this dataset (Bayesian and Gaussian methodologies), while simple in structure, were not able to fit the dataset with sufficient accuracy. This can largely be attributed to the lack of conditional independence in channels taken with different brightness levels and difficulty with an unbalanced dataset.

Surprisingly, k-nearest neighbors also performed quite well on this dataset, given the simplicity of this model. This makes sense, since often the pixels most similar to a sample pixel are the areas directly nearby, which are likely to also share its class label due to the spatial contiguity of farmland. As such, there are a few issues when attempting the generalize the performance of a nearest neighbors model on this classification task: (1) in order to replicate the performance of this model on different geographic regions, vast amounts of image data would need to be stored in memory rather than only the fitted model parameters, which may not be feasible for large image data (2) as the size of datasets grow, the small amount of spatial

information retained by nearby pixel similarity (as explained above) may diminish, causing nearest neighbors to regress to the mean of model accuracy.

Another surprising result was the performance of the ridge classifier model, which performed with near perfect precision, with 93% of samples estimated to be farmland truly reflected as ground truth farmland area. However, this high precision is paid for in the low recall for this model, with only 21% of farmland pixels recalled.

It is difficult to determine which measure of performance best encapsulates the purpose such a classifier might have when actually applied:

- If the purpose were to identify with absolute confidence the locations of some fraction of farmland areas from satellite imagery (in other words, if the price for a false positive were far higher than for a false negative) a high precision model such as the ridge classifier or stochastic gradient descent classifier may be preferred.
- If the purpose were to collect the maximum amount of farmland area from the classification model, minimizing the amount of true farmland that was missed (in other worse, a strong preference for false positives over false negatives) then a model with high recall and low precision, such as a support vector machine (SVM) or k-nearest neighbor model would be preferred.
- If the purpose of such a model is to estimate the proportion of area in a given area that contains farmland (in other words, no preference between false negatives and false positives) a more generally effective model such as the random forest model may be preferred.

For the purposes of this study, models which were generally effective in their evaluation of farmland vs nonfarmland pixels, with no preference for false negatives or false positives, were carried forward for further study: these were (1) **random forest model** (2) **multilayer perceptron**, matching reviewed scientific literature.

Stratified Sampling Data

As another attempt to analyze the data we collected data subsets in which the amount of farmland and non-farmland pixels were equal. There were 1,192,478 pixels classified as farmland and 4,807,522 classified as non-farmland in the original dataset. The same untuned classifiers from sklearn were run instead this time with the stratified data at 1,000,000 samples of both farmland and non-farmland.

Table 2 | Stratified Data Set OOTB classifiers using 1,000,000 samples of each class.

Classifier	Accuracy	F1	Precision	Recall	TPR/FPR
Decision tree	0.787	0.786	0.78	0.79	3.67
Multilayer perceptron	0.826	0.822	0.84	0.81	5.22
Bernoulli naive Bayes	0.688	0.745	0.63	0.91	1.70
Gaussian naive Bayes	0.648	0.529	0.79	0.40	3.85
Stochastic Gradient	0.751	0.743	0.77	0.72	3.29

Descent					
Ridge Classifier	0.772	0.770	0.77	0.77	3.46
Random forest	0.844	0.843	0.84	0.84	5.38
AdaBoost	0.802	0.805	0.79	0.82	3.87
Gradient Classifier	0.822	0.822	0.82	0.82	4.57
Ensemble Voting Classifier	0.825	0.824	0.82	0.82	4.74
Support vector machine	0.773	0.773	0.77	0.78	3.35
K-nearest neighbor	0.828	0.828	0.82	0.83	4.72

Interestingly, the accuracy of the stratified sample decreased by approximately 10% for most modelling schemes. Conversely, the F1 scores of the data set rose by 0.1 on average. The rationale behind this result is that since the data set is no longer skewed towards nonfarmland areas, weak learning by overprediction of nonfarmland is deterred. Other stratified sample sizes from 250,000 - 1,000,000 pixels were also tested and yielded similar results. Modelling with the original dataset was pursued since the highest performing models (random forest, multilayer perceptron) were relatively invariant to the balance of the dataset, implying this effect is exaggerated for weak learning models.

Hyperparameter tuning

Hyperparameter tuning was conducted for the random forest model from sklearn, since this model showed the highest performance during initial modelling. The out-of-the-box classifier yielded an accuracy of 88.9% and an F1 score of 0.694. The goal of hyperparameter tuning was to boost this performance by optimizing the parameters used for model training. In this way, hyperparameters can be roughly understood as a control on the ability of a model to underfit or overfit the sample data. In the case of random forest, decreasing or increasing the size of random forests (depth, number of decision trees, pixels required to split a node) is designed to decrease or increase the ability of the model to overfit trends in data features.

Tuning was conducted on a representative subset of the original dataset, composed of 60,000 pixels (1% of the original dataset). This subset of the original dataset was used in an effort to minimize the extensive training time needed for models on the full dataset. While accuracy was diminished on this subset of the dataset, it was assumed that any performance differences on this limited dataset would be replicated on the overall dataset when retrained on the full set of 6,000,000 sample pixels from all six images.

The random forest model was preferred over the MultilayerPerceptron from sklearn since training time was significantly lower for the random forest model, which was important during hyperparameter tuning given the limited computing resources available for this project. Hyperparameter tuning for this model was conducted in two stages:

1. Random grid search over a wide range of model parameters

2. Random grid search over a smaller space of parameter sets centered on the most successful parameter set from the random grid search

These two hyperparameter tuning steps were conducted using the RandomizedSearchCV method available from sklearn. The parameters used for the first step of hyperparameter tuning are listed below:

Table 3 | Hyperparameter set for first round of hyperparameter tuning of random forest

Hyperparameter	Definition	Range	# of Options
Number of estimators	Number of weak learners to use for random forest model	50-500	10
Maximum features	Number of features to evaluate when searching for optimal split	N/A	2
Maximum depth	Maximum depth to grow decision tree before stopping	10-110	12
Minimum samples for split	Minimum number of training samples to split a leaf	2-10	3
Minimum samples per leaf	Minimum number of training samples per leaf of estimator	1-4	3
Bootstrap samples	Whether or not to use a 'bootstrapped' set of samples to quickly train tree	N/A	2

This yields a parameter sample space of 4,320 sets of hyperparameters. This large set of hyperparameters makes direct evaluation of all sets of hyperparameters computationally prohibitive, hence the need for a randomized search during tuning. The randomized search was conducted for 100 iterations using 4-fold cross validation, and the best-performing model with respect to accuracy was stored for future use. The sampled sets of hyperparameters comprise 2.3% of the hyperparameter sample space. Somewhat inherent in this search for optimal hyperparameters is that, even if only a small fraction of the sample space is surveyed for the optimal parameter values, a reasonable and near-optimal set of hyperparameters can be derived from this sampling procedure without strong conditional effects from multiple changes simultaneously. The optimal set of parameters for the first hyperparameter tuning iteration was determined to be the following:

{'n_estimators': 450**, 'min_samples_split':** 10**, 'min_samples_leaf':** 4**, 'max_features':** 'sqrt'**, 'max_depth':** 100**, 'bootstrap':** True**}**

This optimized random forest model from the first round of hyperparameter tuning achieved an accuracy of 88.4% and an F1 score of 0.676. While this performance is a slight decrease from the

performance of the 'out-of-the-box' model, this tuning was conducted with only 1% of the sample data, a set of 60,000 pixels. It is expected that a model trained on the full dataset would reach a higher performance. Regardless, this model is only the first optimum from a broad search of the hyperparameter sample space.

Following this round of hyperparameter tuning, a second iteration of hyperparameter tuning was again conducted with 100 iterations using 4-fold cross-validation. The set of parameters used for the second round of hyperparameter tuning was as follows:

Table 4 | Hyperparameter set for second round of hyperparameter tuning of random forest

Hyperparameter	Definition	Range	# of Options
Number of estimators	Number of weak learners to use for random forest model	400-500	5
Maximum features	Number of features to evaluate when searching for optimal split	N/A	1
Maximum depth	Maximum depth to grow decision tree before stopping	90-110	5
Minimum samples for split	Minimum number of training samples to split a leaf	8-12	3
Minimum samples per leaf	Minimum number of training samples per leaf of estimator	3-5	3
Bootstrap samples	Whether or not to use a 'bootstrapped' set of samples to quickly train tree	N/A	1

This represents a sample space of 225 sets of parameters, for which 100 iterations of the RandomizedSearchCV with 4-fold cross validation represents 44% of the sample space. The optimal set of parameters for this second round of hyperparameter tuning was as follows:

{'n_estimators': 450**, 'min_samples_split':** 12**, 'min_samples_leaf':** 5**, 'max_features':** 'sqrt'**, 'max_depth':** 95**, 'bootstrap':** True**}**

Insofar as hyperparameters are considered a control on the ability of models to underfit or overfit the dataset, it appears that increasing the amount of samples required to split a node and increasing the number of samples required at each leaf node increased the performance of the model, most likely by preventing the random forest model and its weak learners from overfitting the training dataset.

The final performance of this optimized set of hyperparameters was an accuracy of 89.2% and an F1 score of 0.701, with a TPR/FPR score of 14.28. This represents an overall accuracy gain of 0.3%, which is quite small given the extensive effort expended in tuning hyperparameters. Put simply, it may be that further optimization of hyperparameters was not sufficient to improve the accuracy of the classification, and that the initial round of classification

using out-of-the-box classification models approached the maximum classification performance on this image dataset for this classifier.

Results

The generally effective classification models, with a reasonable balance between precision and recall for this dataset were determined to be (1) random forest models and (2) multilayer perceptron models. A more complete cataloguing of models evaluated in this work is shown below:

1. **High precision models:**
    a. Ridge classifier
    b. Stochastic gradient descent classifier
2. **High recall models:**
    a. Support vector machine
    b. K-nearest neighbor classifier
3. **General use models:**
    a. Random forest classifier
    b. Multilayer perceptron

While k-nearest neighbors also performed well on this dataset, it was omitted from the list of optimal classifiers due to expected difficulties in model size and generalizability. Further, k-nearest neighbors is also not reflected as an accepted model for farmland image classification in the relevant research literature.

Many other models were tested and their performance characterized, with high precision models such as ridge classifiers and gradient descent models, as well as high recall models such as support vector machines and k-nearest neighbors also explored in this work. According to the purpose of any given farmland research, all of these models may have some utility in performing farmland classification tasks.

Additional tuning of hyperparameters for the leading model, the random forest model, did not yield significant performance benefit on the classification task, with the performance of this final, optimal classifier reaching an accuracy of 89.2% and an F1 score of 0.701.

Future research should focus on efficient methods of encoding spatial information in image datasets, since this approach is expected to yield more distinguishing feature data for farmland classification, and also to reduce the effect of natural land variability on classification results. Given the limitations of a single set of images for farmland classification, it was also expected that data taken from multiple timepoints could increase the accuracy of classifier models, albeit with a sharp increase in the complexity of generated classifier models.

Conclusion

Farmland classification from satellite imagery is an important area of research for future climate change mitigation work, especially in vulnerable regions with limited on-the-ground support. Future data mining efforts in this area will be essential when allocating local resources to affected residents and estimating long-term trends in food production networks. This task of farmland classification is challenging due to the small size of cropped fields, low vegetative cover within and outside cropped areas, and high correlation between the seasonal trajectories in the signatures of cropped and uncropped surfaces. The main challenges in this project were the limited number of labeled data for training, small size of cropped fields in the images (producing an asymmetric dataset), and the failure of any image channel or combination of channels being

fully distinguishing for classification. This project used a small set of 6 images from Sentinel 2 image data in Iran to evaluate the performance of different farmland classifiers. Results showed that the random forest and multilayer perceptron have the best performance, in agreement with previous work. These models have sufficient degrees of freedom to extract the maximum amount of distinguishing information from the set of overlapping original and synthetic features which results in the high performance of them in classification. Additional tuning of hyperparameters did not yield significant performance benefits, suggesting that generic versions of these classifiers approached the maximum classification performance on this dataset. For future work, it is suggested to use higher resolution imagery that would enable the derivation of texture features, and the use of more than a year of data in the time series to ameliorate the effects of seasonal conditions. Also, applying algorithms with a concept of spatial proximity between pixels such as CNNs or graph clustering methods show promise in extracting additional information from image data.

Group Members Contribution

Ben: Created label data set in ARCGIS, contributed to writing/editing of reports, presentation, worked on stratified data collection and analysis.
Reyhaneh: Worked on performance analysis of the classifiers, calculating contingency tables, and scores. Contributed to writing/editing of reports, presentation.
Henry: Programmed data IO and data processing/cleaning functions for CSV data. Programmed classifier training code, performed hyperparameter tuning of random forest model, and implemented residual neural network model. Performed data exploration of channel features (histogram stuff) and implemented an algorithm to compute NDVI.

References

Duro, D.C., Franklin, S.E. and Dubé, M.G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote sensing of environment, 118, pp.259-272.

Akbari, M., Mamanpoush, A.R., Gieske, A., Miranzadeh, M., Torabi, M. and Salemi, H.R., 2006. Crop and land cover classification in Iran using Landsat 7 imagery. International Journal of Remote Sensing, 27(19), pp.4117-4135.

Kussul, N., Lavreniuk, M., Skakun, S. and Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), pp.778-782.

Garcia-Pedrero, A., Gonzalo-Martin, C. and Lillo-Saavedra, M., 2017. A machine learning approach for agricultural parcel delineation through agglomerative segmentation. International journal of remote sensing, 38(7), pp.1809-1819.

Github Link

https://github.com/hankdikeman/DataMining5523