# NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY (NGA)
# RESEARCH PERFORMANCE PROGRESS REPORT (RPPR)

*For instructions, please visit*
*https://www.nga.mil/Partners/ResearchandGrants/Pages/AcademicResearchProgram.aspx*

| AWARD INFORMATION | |
|---|---|
| 1. Federal Agency: **National Geospatial-Intelligence Agency** | 2. NGA Grant Award Number: HM04762010005 |
| 3. Project Title: Spatial Dynamics and Analysis of Crops using Super Multispectral Image Resolution and Radar Fusion | |
| 4. Award Period of Performance Start Date: 04/01/2020 | 5. Award Period of Performance End Date: 03/31/2022 |

| PRINCIPAL INVESTIGATOR/PROJECT DIRECTOR | |
|---|---|
| 6. Last Name and Suffix: Beck | 7. First and Middle Name: John Malone |
| 8. Title: Principal Research Scientist IV | |
| 9. Email: john.beck@uah.edu | 10. Phone Number: 256-824-5147 or 256-289-1036 |

| AUTHORIZING OFFICIAL | |
|---|---|
| 11. Last Name and Suffix: Greene | 12. First and Middle Name: Gloria W. |
| 13. Title: Assistant Vice President Contracts and Grants | |
| 14. Email: osp@uah.edu | 15. Phone Number: 256-824-2657 |

| REPORTING INFORMATION | |
|---|---|
| Signature of Submitting Official: *Gloria W. Greene* | Digitally signed by Gloria W. Greene, MA, CRA DN: cn=Gloria W. Greene, MA, CRA, o=The University of Alabama in Huntsville, ou=Office of Sponsored Programs, email=greeneg@uah.edu, c=US Date: 2022.04.11 14:24:32 -05'00' |
| 16. Submission Date and Time Stamp: 04/10/2021 | 17. Reporting Period End Date: 04/11/2021 |
| 18. Reporting Frequency: ☒ Annual ☐ Semi-Annual ☐ Quarterly | 19. Report Type: ☐ Not Final ☒ Final |

| RECIPIENT ORGANIZATION | |
|---|---|
| 20. Recipient Name: The University of Alabama in Huntsville | |
| 21. Recipient Address: 301 Sparkman Drive, Huntsville, AL 35899 | |
| 22. Recipient DUNS: 949687123 | 23. Recipient EIN: 63-0520830 |

| ACCOMPLISHMENTS |
|---|
| 24. What were the major goals and objectives of this project? The major goals of this project were to use deep learning with convolution neural networks (CNN) for improving the spatial resolution of multi spectral imagery (MSI) using a corresponding high-resolution panchromatic image and then fusing the downscaled high-resolution MSI image with a synthetic aperture radar (SAR) image to make a new GEOINT product for assisting in the analyzing of agricultural crops. **Please see Appendix A.** |
| 25. What was accomplished under these goals?  A detailed literature review was conducted on using deep earning technologies for conducting super-resolution and fusing SAR data. Four convolution neural networks (CNN) were developed, tested, and evaluated (a traditional supervised CNN, a generative adversarial network (GAN), a PanColorGAN, and an unsupervised CNN with an advanced specialized loss function) against well-known and published pan-sharpening methods. Results were evaluated for usefulness in analysis of agricultural crops. **Please see Appendix A for more information**. |
| 26. What opportunities for training and professional development has the project provided? <br><br> Principal Investigator Beck attended and presented at the the 2020 and 2021 Intelligence Community Academic Research Symposium (ICARS).  Several of the presentations at ICARS provided the team with new insights into advanced deep learning technologies and cutting edge research in the areas of remote sensing and geospatial technologies. |
| 27. How were the results disseminated to communities of interest?  Principal Investigator Beck gave presentations at the 2021 and 2022 ICAR meetings. He also gave a presentation where he demonstrated the research in this project to NASA's Advanced Information Systems Technology Office (AIST). NASA's AIST leadership are very interested in the super-resolution research in this project.  NASA is interested in this project due to its vast holdings of moderate resolution imagery. |
| 28. What do you plan to do during the next reporting period to accomplish the goals and objectives? <br><br> Not applicable. This is the final report. |

| PRODUCTS |
|---|
| 29. Publications, conference papers, and presentations. Principal Investigator Beck presented the project at the 2020 and 2021 Intelligence Community Academic Research Symposiums (ICARS). He also presented a poster at the 2019 AGU Annual Conference. **See Appendix A for references.** |
| 30. Technologies or techniques <br><br>  Machine learning and deep learning architectures for pan / multi-spectral sharpening with synthetic aperture radar (SAR ) fusion. Analysis of agriculture crops. |
| 31. Inventions, patent applications, and/or licenses <br><br>    N/A |
| 32. Other products <br><br>  N/A |

*Attach a separate document if more space is needed for #6-10, or #24-50.*

| PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS |
|---|

**33. What individuals have worked on this project?**
John M. Beck, Ph.D. (Principal Investigator)
Charles Collins (Developer/Computer Scientist)
Evans Criswell (Computer Scientist)
Xiang Xi. Ph.D. (Remote Sensing Scientist)

**34. Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?**

No.

**35. What other organizations have been involved as partners?**

The NASA Advanced Information Systems Technology (AIST) program under its Earth Science Technology Office (ESTO) is interested in this project and its technology.

**36. Have other collaborators or contacts been involved?**

Yes, I have contacted Dr. David W. Messinger.  He has a similar funded project with NGA. I attended his research webinar with NGA and I contacted him so we could collaborate.

| IMPACT |
|---|

**37. What was the impact on the development of the principal discipline(s) of the project?**
An innovative approach for conducting pan-sharpening and fusing SAR data using deep learning and advanced loss functions provided the remote sensing community with new techniques and results that will expand the use of super-resolution methods for improving lower resolution imagery.

**38. What was the impact on other disciplines?**
Super-resolution imagery combined with SAR provides improved GEOINT products that can impact imagery analysis in all civilian and military domain such as agriculture, forestry, hydrology, and object detection such as vehicles.

**39. What was the impact on the development of human resources?**

N/A

**40. What was the impact on teaching and educational experiences?**

This research provided students the opportunity to learn more about GEOINT product development and imagery analysis. Results of this research was exposed to 50 students involved in UAH's Intelligence Community Center of Academic Excellence (IC CAE) in critical technologies. **Please see Appendix A.**

**41. What was the impact on physical, institutional, and information resources that form infrastructure?**

Not applicable.

**42. What was the impact on technology transfer?**

**Please see Appendix A.**

*Attach a separate document if more space is needed for #6-10, or #24-50.*

| |
|---|
| 43. What was the impact on society beyond science and technology? |
| New data sources for analyzing crops, predicting yields, and estimating food sources during the early growing stages. |
| 44. What percentage of the award's budget was spent in foreign country(ies)? |
| None. |

| CHANGES/PROBLEMS |
|---|
| 45. Changes in approach and reasons for change |
| **Please see Appendix A.** |
| 46. Actual or anticipated problems or delays and actions or plans to resolve them |
| COVID-19 delayed some of our initial testing of the CNN architectures due to remote login issues into our computer systems that utilize Graphic Processing Units (GPUs).  Additionally, because of COVID the team was not able to travel to conferences to present the project findings. |
| 47. Changes that had a significant impact on expenditures |
| Dr. Susan Bridges, an original Co-I of the project, unexpectedly retired. |
| 48. Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents |
| None. |
| 49. Change of primary performance site location from that originally proposed UAH began telework from home due to the COVID-19 outbreak from March 16, 2020 to May 31, 2021. On June 1, 2021 UAH returned to normal working operations on campus. |

| PROJECT OUTCOMES |
|---|
| 50. What were the outcomes of the award? |
| We have designed, evaluated, and tested four different CNN models for pan-sharpening Multi-spectral imagery (MSI) with panchromatic imagery with varying results. We also evaluated several strategies for fusing the Synthesis Aperture Radar (SAR) data with the pan-sharpened imagery so we could create a new GEOINT product for agricultural crop analysis. We developed statistical metrics to evaluate our results compared to other known methods and each other. The SAR data fusion portion of this project had mixed results that did effect the success of the crop analysis negatively. We conducted a series of classification techniques using all of our new products compared to using only MSI imagery. Results showed that using a pan-sharpened image did not improve crop classification of row crops (corn, cotton, wheat, and soybeans). Introducing texture from the SAR image did improve classification. Results showed that our PanColorGAN produced the best results followed by the basic GAN, and then the supervised CNN. Specialized loss functions were not successfull. **Please see Appendix A for expanded details on the outcomes of the award including lesson learned**. |

*Attach a separate document if more space is needed for #6-10, or #24-50.*

| DEMOGRAPHIC INFORMATION FOR SIGNIFICANT CONTRIBUTORS (VOLUNTARY) | |
|---|---|
| **Gender:**<br><br>☒Male<br><br>☐Female<br><br>☐Do not wish to provide | **Ethnicity:**<br><br>☐Hispanic or Latina/o<br><br>☒Not Hispanic or Latina/o<br><br>☐Do not wish to provide |
| **Race:**<br><br>☐American Indian or Alaska Native<br><br>☐Asian<br><br>☐Black or African American<br><br>☐Native Hawaiian or other Pacific Islander<br><br>☒White<br><br>☐Do not wish to provide | **Disability Status:**<br><br>☐Yes<br><br>   ☐Deaf or serious difficulty hearing<br><br>   ☐Blind or serious difficulty seeing even when wearing glasses<br><br>   ☐Serious difficulty walking or climbing stairs<br><br>   ☐Other serious disability related to a physical, mental, or emotional condition<br>☒No<br><br>☐Do not wish to provide |

*Attach a separate document if more space is needed for #6-10, or #24-50.*

## *Appendix A. NGA Final Report Grant# HM04762010005*

**24. What were the major goals and objectives of this project?**

The major goals of this project were to use deep learning with convolution neural networks (CNN) for improving the spatial resolution of multispectral imagery (MSI) using a corresponding high-resolution panchromatic image and then fusing the downscaled high-resolution MSI image with a synthetic aperture radar (SAR) image to make a new GEOINT product to assist in the analyzing of agricultural row crops. These goals were achieved by accomplishing the following objectives: (1) To adapt UAH's deep learning CNN image super-resolution method to enhance the spatial resolution of high-resolution (2.0 m/pixel) MSI imagery using a corresponding fine-resolution (0.5 m/pixe) panchromatic image,(2) fuse downscaled fine-resolution MSI images from Objective 1 with fine-resolution (< 10.0 m/pixel) SAR data**,** and (3) evaluate how agricultural crop analysis can be enhanced using the new GEOINT product.

**25. What was accomplished under these goals?**

*Literature Review:*

We conducted a detailed literature review to examine the latest technologies in deep learning for conducting pan sharpening and radar data fusion. Research has shown that Convolutional Neural Networks (CNNs) can be an effective method for improving the spatial resolution of multispectral imagery (MSI) using a corresponding high-resolution panchromatic image [1-7]. In fact, several studies have demonstrated the ability to combine MSI data with panchromatic [1], microwave [3], and SAR data [4] with good success. Many different algorithms and techniques have been tested including Intensity-Hue-Saturation (IHS) [5,8], Principal Component Analysis (PCA) [4,6,9], Discrete Wavelet Transformation (DWT) [7], Support Vector Regression (SVR) [10,11], and various combinations of these methods [12]. Several of these methods use a higher-spatial resolution panchromatic image to improve the spatial resolution of the lower-spatial resolution MSI data. Although successful, these techniques often display high spectral distortions because the panchromatic and MSI bands of data do not overlap entirely in the electromagnetic spectrum. Recent studies have successfully used deep learning to address this issue [13,14,15,16]. One of the most promising directions of deep learning is data fusion of SAR and optical data as these data modalities are completely different from each other both in terms of geometric and radiometric appearance [17].

Deep learning has gained popularity over the past decade due to its capacity to learn data representations using both supervised and unsupervised classifiers. Because of

this ability, many researchers are now exploring the use of deep learning technologies to perform image super-resolution and data fusion with varying degrees of success [19]. Deep learning develops computational models consisting of multiple processing layers that learn hierarchical features at multiple levels of abstraction [20]. There are two major types of deep learning neural networks, CNNs, and recurrent neural networks (RNN). Both approaches have consistently outperformed traditional machine learning methods when evaluated with benchmark datasets [18,19,20].

Principal Investigator (PI) Beck demonstrated that deep learning CNNs can be successfully trained to downscale (increase) the resolution of low-resolution satellite-based imagery by learning a mapping of features to corresponding higher-resolution imagery [21]. A trained CNN architecture can then be used to increase the resolution of low-resolution satellite imagery when no corresponding high-resolution imagery is available. Results indicated that the CNN automatically adjusted the dynamic range of the low-resolution image to match the range of the target high-resolution image and the CNN approach outperformed other standard image interpolation techniques (Table 1). Images produced by this study were notably more detailed than those produced by common interpolation methods (Figure 1).

*Table 1. Comparison of Image Quality Produced by CNN and Other Standard Methods.*

| Enhancement Method | Peak Signal to Noise Ratio | Structural Similarity Index |
|---|---|---|
| CNN | 46.99 | 0.9862 |
| Nearest Neighbor | 32.39 | 0.9550 |
| Bilinear | 32.43 | 0.9576 |
| Cubic | 32.45 | 0.9593 |
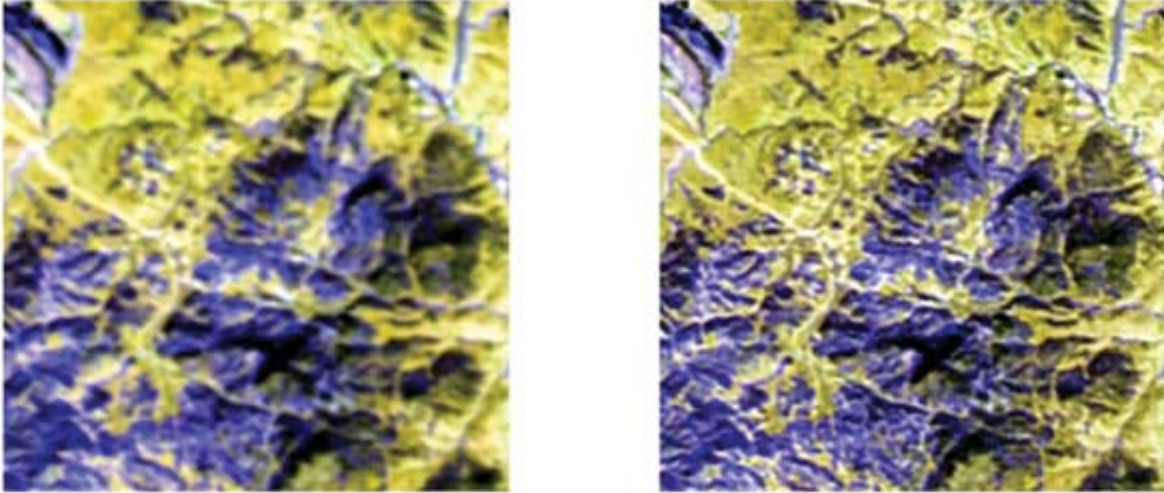
*Figure 1. Comparison of a low-resolution satellite image (left) and CNN super-resolution image (right) for a region where no corresponding high-resolution image is available.*

In Beck's study, the initial architecture for the panchromatic sharpening task was similar to that used for standard image super-resolution [16, 18-22]. In contrast to CNNs developed for classification tasks where the output is a single class, this CNN was used for super-resolution where learned hierarchical features are used to produce an improved resolution dataset. For super-resolution of MSI data, the network maintained the spatial dimensions of the panchromatic image. The task of generating each grid point in the output was a regression task rather than a classification task. The network learned a set of features that made use of information embedded in neighboring data points to generate an interpolated value for each grid point simultaneously. Unlike classification networks, super-resolution networks are purely convolutional and contain no pooling or fully connected layers. For standard image super-resolution, the output values are integers, but for image super-resolution problems in which SAR data will be fused, outputs are real-valued grid points.

For this project, we conducted several experiments with alternative network parameters, architectures, and data handling options. Recent CNN literature reports improved performance with more complex architectures such as deep residual networks, generative adversarial networks (GANs), and capsule networks [23-28]. Many of these new techniques were initially designed for image classification and needed to be truncated or modified for the purpose of the regression problem central to a resolution improvement using a panchromatic image and SAR data fusion task. PI Beck conducted initial experiments with generative adversarial networks for super-resolution and radar fusion of satellite-based precipitation data that showed great promise [22].

Generative Adversarial Networks (GANs) are an emerging deep learning architecture that can include other forms of deep learning including CNNs and autoencoders, a type of neural network which compresses an input into smaller feature information. The GAN's purpose is to generate a target image indistinguishable from the real image. GANs accomplish this by employing a two-branch architecture consisting of a generator branch and a discriminator branch. The generator branch generates new data derived from sample data and the discriminator branch acts as the adversary calculating the probability of the generated image being real or fake. This training process repeats until the discriminator believes the generated MSI is a high-quality pan-sharpened image. Accurate hierarchical features are learned and preserved using this deep learning method while they would be lost or inaccurate if a traditional pan-sharpening algorithm was applied.

We also investigated using a super-resolution generative adversarial network (SRGAN) which consisted of three blocks of layers, each with a specific function: generation, feature extraction, and discrimination [23, 24]. The generation block ingested the input MSI samples and generated a synthetic sample in the same feature space as the target domain. The synthetic sample was then passed to both the feature extraction block and the discriminator block. The feature extraction block compared the loss between the features of the synthetic sample and the features of the target sample. The discrimination block attempts to classify the generated synthetic sample as real or synthetic/fake; the goal was to generate samples such that it is difficult to discriminate between real target samples and generated samples. The discrimination and feature extraction blocks back-propagate the loss to train the generation block. During testing and inference, only the generation block was used to generate enhanced MSI samples and fuse the SAR data.

We also explored and expanded our testing of the GAN by fine-tuning the CNN parameters according to the following research [25-35]. Finally, we explored the development and use of an unsupervised deep-learning CNN architecture using advanced loss functions.

### *Convolutional Neural Network (CNN) Architectures:*

For this investigation, we developed four different CNN architectures or models for pan-sharpening multispectral data with panchromatic data: a traditional supervised CNN, an unsupervised CNN using a specialized loss function, a super-resolution generative adversarial network (SRGAN) [24,25], and a novel idea using a PanColorGAN [36]. All four of our models were implemented using TensorFlow. The following is a detailed account of each model and the data we used throughout the project.

**Study Area:**

The study area for this project consisted of approximately 15,000 hectares of Tennessee River Valley land, just west of Huntsville, Alabama. This area was chosen because of its close proximity to UAH and its vast amount of row crop agriculture (cotton, corn, and soybeans). Recent trends in cropping practices have shifted away from traditional cotton farming towards planting more corn and soybeans.

**Data Used:**

For this project, we acquired imagery from the Airbus DS Pleiades 1B satellite which consisted of a fine-resolution panchromatic image with a 0.5 meters/pixel spatial resolution and a four-band multispectral image with a spatial resolution of 2.0 meters/pixel. This scale factor of 4 between the panchromatic and multispectral bands is frequently encountered in super-resolution and pansharpening research. Our Pleiades product was delivered as a 16-bit GeoTIFF and was collected on May 18, 2019. This time period was chosen because most of the agricultural row crops had already been planted and were in their early stages of growth.

**Data Pre-processing**:

Supervised machine learning models require a ground truth image against which to compare the model output. Our supervised models did not have natural truth or target images because coincident multispectral imagery with the same spatial resolution of the panchromatic imagery was unavailable. To circumvent this, we used a process called the Wald protocol to create input and truth pairs. The Wald Protocol uses the patches from the original multispectral imagery as truth samples [37].

A set of degraded panchromatic and multispectral patches are generated to serve as the input samples. The original scale factor is retained for the input and target patches. This way the models can be applied to the native resolution multispectral and panchromatic samples during testing. The original scale factor is retained for the input and target patches so the models can be applied to the native resolution of the multispectral and panchromatic samples during testing.

For all models, the imagery was divided into smaller patches. Patches were 256 x 256 pixels at the resolution of the panchromatic imagery and 64 x 64 pixels at the resolution of the multispectral imagery. The output of our models had "banding" around the edges of the image. This is an artifact of convolution and padding operations. Since our patches overlapped spatially, we were able to remove the banding effect during image reconstruction after model inference. The pixel dimensions of the necessary overlap correlated to the dimensions of the convolutional filters used by our model layers.

**1.) Traditional Supervised CNN**

At the beginning of this project we decided to try a very simple supervised CNN architecture consisting of three convolution layers, the first two being followed by a Rectified Linear Unit (ReLU) (Figure 2.) Since the dimensions of the image must be maintained throughout the model, pixel padding was applied to the output of each convolution layer. The parameters for the CNN consisted of the numbers of filters in the first and second convolution layers (the third layer's size was determined by the number of bands in the output image) as well as the convolution kernel sizes for the first, second, and third layers. A learning rate was also specified throughout the model.
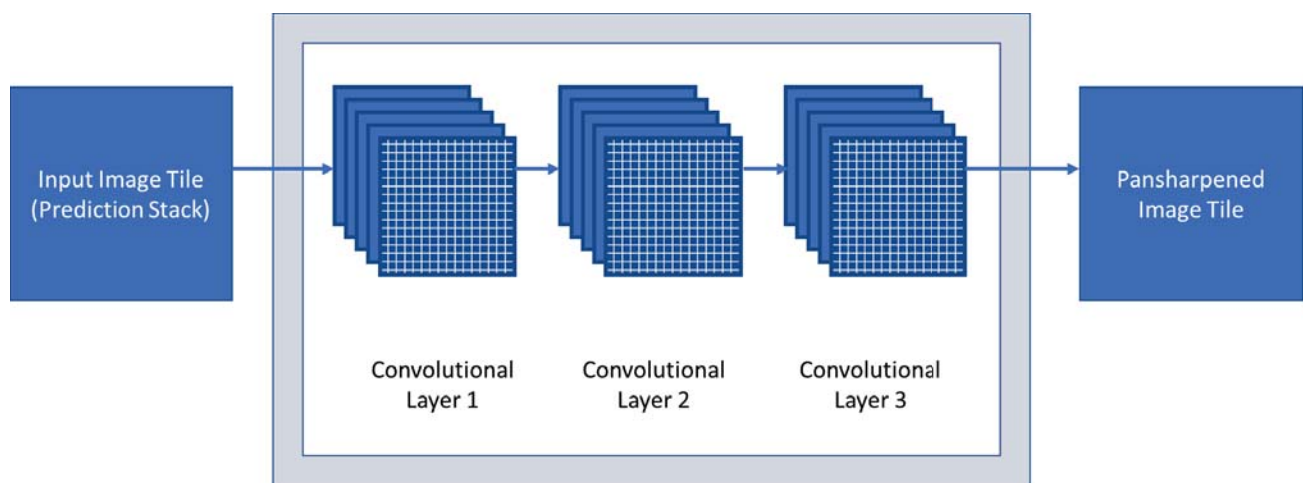


*Figure 2. Traditional Supervised CNN Architecture for this project.*

This CNN did not contain pooling layers since the purpose was not to recognize objects in images but to perform pansharpening. Pooling layers described here for completeness, subsample the input from the previous layer using an aggregation function such as mean or maximum. Using even a small pooling kernel of 2x2 and a stride of 2, data will be reduced by 75 percent from the previous layer, with only the mean or maximum of each kernel making it to the next layer. With the goal of pansharpening, keeping all of the information from the previous layer is necessary, and reducing the data by aggregating would be counterproductive. Figure 3 represents the basic CNN data workflow.
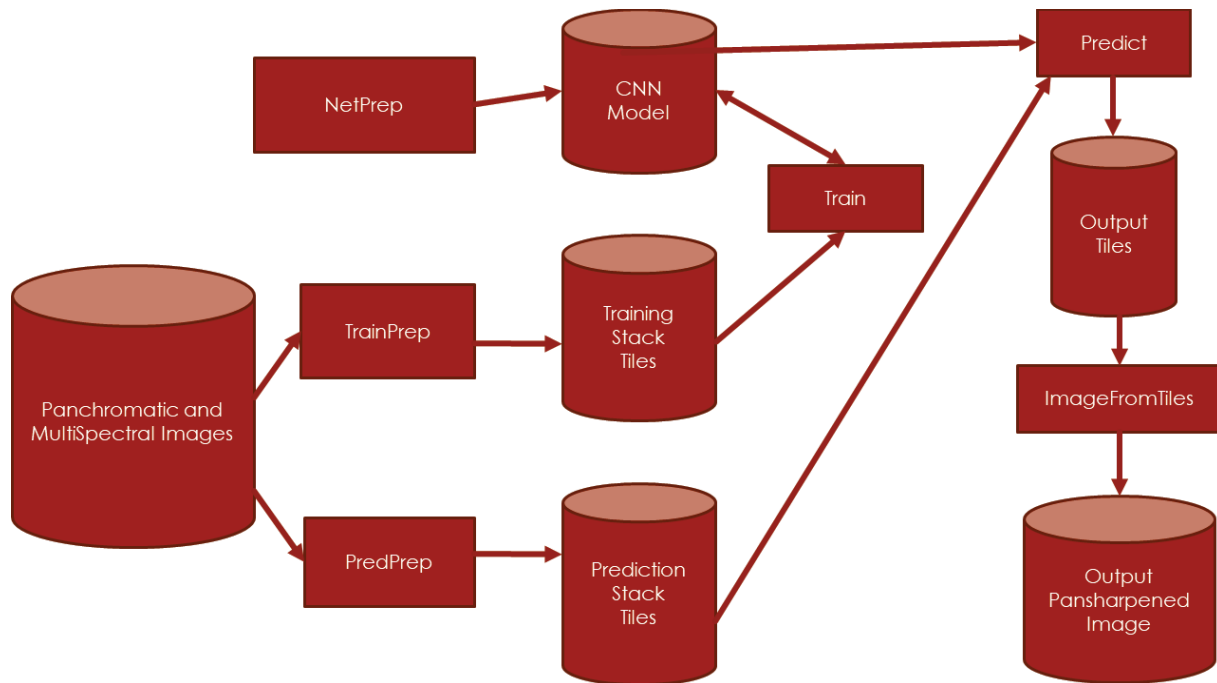
*Figure 3. Example workflow from our traditional supervised CNN.*

Panchromatic input patches were 256 x 256 pixels, and multispectral input patches were 64x64 pixels. All data were normalized by the dynamic range prior to training. The number of filters used was 256, 64, 4, and filter dimensions of 13, 9, 9 respectively. The Adam optimizer with a learning rate of 0.0001 was used with this model. We found that the Adam optimizer improved learning performance (as opposed to Stochastic Gradient Descent) and that the Adam optimizer provided convergence much more quickly than the standard stochastic gradient descent with a momentum approach. The training set was generated using the Wald protocol and the multispectral and panchromatic image patches are stacked into 5-band images which are used as model input.

For the Wald protocol processing, we separated each individual multispectral band data into separate arrays, instead of combining them into a multichannel image. We also calculated the average and standard deviation of the values in each channel and saved them to a file stored in the output model directory so the prediction script could read it and adjust the output levels to have the same average. This helped the problem somewhat, but often, the result was an unsaturated image. Considerable difficulty was encountered in getting the color of the output channels in predicted images to match the color in the input images. We discovered that if we passed the input tile through the network multiple times it would increase the quality of the output with respect to the color matching. We also noticed that the error statistic greatly decreased the first few passes, with diminishing returns later on. We found that eight epochs were sufficient to produce good results.

Utilizing the MSI of the Tennessee Valley and its corresponding panchromatic image we were able to generate a high-quality pan-sharpened image. Although our results from this basic supervised CNN were promising we are having an issue with our model where a small grid-like pattern is produced where the patches were mosaicked together for the scene. To avoid this artifact, input patches were generated such that there was a 128 pixel (or 50%) overlap. See Figure4 for an example of the output.



*Figure 4. Example output from our traditional supervised CNN.*

## 2.) Unsupervised CNN

For our second experiment, we developed an unsupervised CNN model based on recent work by Zhou and others where they used a specialized loss function to improve the pan-sharpening results by developing a function in which the relationships between the input multispectral and panchromatic images and the fused multispectral image were used to design the spatial constraints and spectral consistency, respectively [38]. This CNN did not require the use of the Wald protocol to construct a surrogate training set. Instead, this model only used the panchromatic and multispectral images at their original resolutions as the base training set. A compound loss function was then used which evaluated the pan-sharpened image output against the original imagery and intermediate imagery derived from the original data. The following steps and equations were implemented based on the journal article by Zhou [37].

A few preprocessing steps were performed prior to training:

- We resampled the original multispectral imagery to the same spatial resolution as the panchromatic imagery. We also resampled the panchromatic imagery to the spatial resolution of the multispectral imagery. These sets will be referred to as high resolution multispectral and low resolution panchromatic respectively.
- All four sets were normalized by the dynamic range of the sensor, 12 bits.
- An operation during the loss function requires a learned hyperparameter vector named $\alpha$. $\alpha$ is used to approximate the panchromatic image as a linear combination of the multispectral bands. This was formally defined as

$$I = \sum_{c=1}^{C} \alpha_c M_c + e_1, \ c = 1, 2,..., C$$

where M is the multispectral image, C is the number of bands in M, and $e_1$ is additive random noise. To solve for $\alpha$, we first resampled the panchromatic image to the spatial resolution of the multispectral image. We then used TensorFlow to solve the least-squares estimation. While $\alpha$ is solved for the relationship between the low resolution panchromatic and native multispectral imagery, it can also be applied to native panchromatic and high-resolution multispectral imagery.

Our unsupervised model used a novel loss function consisting of three main components: spatial loss, spectral loss, and quality with no reference (QNR).

- *Spatial loss:* The spatial loss calculation used three inputs: an original panchromatic patch $P$, a patch generated by the model $\widehat{M}$, and the vector $\alpha$.
- First an approximation of P is calculated using the combination of bands from $\hat{M}$

$$I = \sum_{c=1}^{C} \alpha_c \widehat{M}_c + e_1, \ c = 1, 2,..., C$$

- Then two metrics were computed.

$$L_{d1} = MSE(P, \ I)$$
$$L_{d2} = 1 \ - \ SSIM(P, \ I)$$

where MSE was the mean-squared error and SSIM is a structural similarity index.

- The final spatial loss was then defined as

$$L_d = L_{d1} + L_{d2}$$

- *Spectral loss:* Spectral loss calculation uses high-resolution multispectral patches $\tilde{M}$ and model output $\widehat{M}$. A Gaussian blur $K$ is applied to $\widehat{M}$. $K * \widehat{M}$ can then be considered an approximation of resampled multispectral image $\tilde{M}$. Spectral loss calculation is otherwise similar to that of spatial loss:

$$L_{s1} = MSE(\tilde{M}, \widehat{M})$$

$$L_{s2} = 1 - SSIM(\tilde{M}, \widehat{M})$$

$$L_s = L_{s1} + L_{s2}$$

- *QNR:* QNR calculation uses an original panchromatic patch $P$, a low-resolution panchromatic patch $p$, an original multispectral patch $m$, and a patch generated by the model $\widehat{M}$. QNR consists of a spectral distortion index $D_\lambda$ and a spatial distortion index $D_s$, both of which use the universal image quality index Q. The Q index is defined as follows:

$$Q(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \times \frac{2\bar{x}\bar{y}}{\bar{x}^2 + \bar{y}^2} \times \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

The spectral distortion is the loss between the Q index of band pairs in $\widehat{M}$ and the corresponding band pairs in $m$.

$$D_\lambda = \sqrt{\frac{1}{C(C-1)} \sum_{c=1}^{C} \sum_{r=1, r \neq c}^{C} |Q(\widehat{M}_c, \widehat{M}_r) - Q(m_c, m_r)|^2}$$

The spectral distortion is the loss between the Q index each band in $\widehat{M}$ and $P$ and the corresponding band pairs in $m$ and $p$.

$$D_\lambda = \sqrt{\frac{1}{C} \sum_{c=1}^{C} |Q(\widehat{M}_c, P) - Q(m_c, p)|^2}$$

QNR was then defined as

$$QNR = (1 - D_\lambda)^i \cdot (1 - D_s)^j$$

where *i* and *j* are used as weights for each component. In our case, we then set

$$i = j = 1$$

The final loss returned by this component was calculated using

$$L_{QNR} = 1 - QNR.$$

- The final compound loss was defined as

$$L_{SSQ} = L_d + L_s + L_{QNR}$$

In this architecture, the model takes two main inputs: the high-resolution multispectral imagery and the original panchromatic imagery. Additionally, the original multispectral imagery and low-resolution panchromatic imagery are used in the model's loss function as described above. Features are extracted from the panchromatic imagery using a convolution layer and a leaky ReLU. The HRMS is passed through a series of layer blocks referred to as fusion units. Each fusion unit consists of two convolution layers, a leaky ReLU, an elementwise addition layer that adds the HRMS to the ReLU output, and a concatenation of the panchromatic features. Our implementation used three fusion units. The output of the fusion unit is passed through a final convolution layer and leaky ReLU. The HRMS is added to this output to form the pan-sharpened multispectral output.

We used TensorFlow to implement this model and used the following hyperparameters:

- Convolution filter dimension: 3x3
- Convolution filters: 64 interior, 4 terminal layers
- 1-pixel padding after convolution layers
- Adam optimizer with weight decay: 1e-3 learning rate, 1e-4 decay
- Batch size: 16

The unsupervised model was trained for 1000 epochs with checkpoints save every 100 epochs. The model converged such that spectral loss was favored, resulting in output nearly identical to the resampled high-resolution multispectral input. The spatial loss was lowest at the 200 and 300 epoch checkpoints while still maintaining a low spectral loss. Results from the 200 epoch checkpoint were selected as the preferred output. Unfortunately, we encountered many difficulties with this model due to mode collapse.

## 3.) Generative Adversarial Network (GAN)

Generative Adversarial Networks (GAN) are an emerging deep learning architecture that can include other forms of deep learning including CNNs. The GAN's purpose is to generate data indistinguishable from the real data from a target distribution. GANs accomplish this by employing a two-branch architecture consisting of a generator branch and a discriminator branch. The generator branch generates new data derived from sample data and the discriminator branch acts as the adversary calculating the probability of the generated image being real or fake. This adversarial game trains both the process of identifying generated items as well as the generating of said items very effectively. Accurate hierarchical features are learned and preserved using this deep learning method while they would be lost or inaccurate if a traditional algorithm was applied. See Figure 5 for an example of the GAN two-branch architecture.
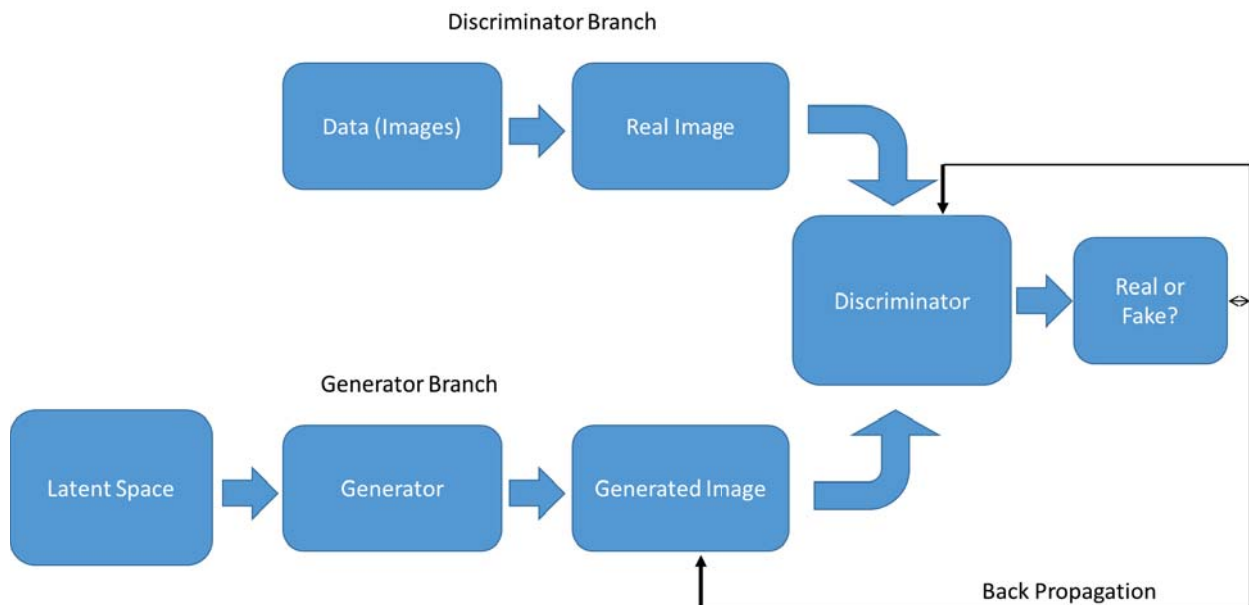


*Figure 5. Example of a two-branch architecture for a GAN.*

Utilizing the same imagery from the Tennessee Valley and a GAN deep learning architecture approach we generated high-quality pan-sharpened imagery. The panchromatic and multispectral imagery was used as a training set where one image was the test image and the rest of the imagery was used by the GAN model to train the CNN. The GAN model's architecture consisted of pixel and patch discriminator branches fed from the output of the two-stream generator branches. The generator branch attempted to generate a higher quality image than the low-resolution MSI using the high-resolution pixel details of the panchromatic imagery, each attempt was passed through the discriminator which determined if parameters needed to be adjusted to generate a suitable image. The discriminator and generator were made up of

convolution layers, batch normalization, and LeakyReLU activations. On a high level, these branches have layers that are extracting the features of the imagery in the training dataset. Considering our "noisy" problem set, we use an Adam optimizer to handle gradients.

For the training process, we utilized the deep learning computation power of a GPU to train the model for 80 epochs. We then applied the trained model to every patch in the test imagery dataset, (this is the imagery the model has not trained on). The post-processing processes include reconstructing these high-resolution MSI patches into a final image that can be used for observation and analysis. Considerable difficulty was encountered in getting the color of the output channels in predicted images to match the color in the input images.  See Figure 6 for an example of the output.



*Figure 6. Output example from the pan-sharpening generative adversarial network (PSGAN).*

## 4.) PanColorGAN

Our fourth CNN was based on a totally different strategy in which the high-level idea was to use the bands of the multispectral image to colorize corresponding features in the panchromatic image. This CNN was based on the work of F. Ozcelik and others in a paper titled "Rethinking CNN-Based Pansharpening: Guided Colorization of Panchromatic Images via GANs" [36].  Since this model is also a GAN, there are two constituent models: a generator and a discriminator. The training set was constructed by downsampling and then upsampling the multispectral image. This degraded multispectral image was then used as input to the generator network, and the original

MS was used as ground truth for training the discriminator. This resampling process traditionally uses the same scale factor, but the PanColorGAN authors opted to randomize the scale factor to have a more diverse training set. This in theory improves the general quality of the generator output during testing.

The generator takes the degraded four-band multispectral image and a single-band grayscale image as input. The latter is the degraded multispectral image converted to grayscale during training, but the panchromatic image is used instead during testing. The generator extracts spatial features from the grayscale image while mapping the color from the multispectral image to the learned spatial features. The output is a high-quality pan-sharpened image.

The discriminator is trained on alternative batches of the degraded multispectral image, the degraded multispectral image converted to grayscale, and generator output (the "fake" batch) and the degraded multispectral image, the degraded multispectral image converted to grayscale, and the original multispectral image (the "real" batch). Visual and statistical results indicated this method overall was superior to our other three deep learning architectures.

**Synthetic Aperture Radar (SAR) Fusion**

**SAR Data**

We acquired Sentinel-1A Level 1 SAR product over the same area of the Tennessee River Valley just west of Huntsville, Alabama just a few days after the multispectral image was acquired (May 25, 2019). The product was delivered with single-polarization (VV or HH) for Wave mode and dual-polarization (VV+VH or HH+HV) or single polarization (HH or VV) for SM, IW, and EW modes. The spatial resolution of the image we used was approximately 10 x 10 meters.

**SAR Data Processing**

The SAR data were preprocessed using the Sentinel Application Platform (SNAP), a common architecture for all Sentinel Toolboxes is being jointly developed by Brockmann Consult, SkyWatch, and C-S. These were the steps we used for pre-processing within the SNAP software:
- Radar/Apply Orbit File
- Radar/Radiometric/S-1 Thermal Noise Removal
- Radar/Radiometric/Calibrate (Sigma0)
- Radar/Sentinel-1 TOPS/S-1 TOPS Deburst
- Radar/Geometric/Terrain Correction/Range-Doppler Terrain Correction
- Open panchromatic imagery in SNAP

- Raster/Geometric/Collocation
  - Master: panchromatic imagery
  - Slave: SAR data
  - Cubic resampling

## 1.) Traditional Fusion Method

A well-known process for combining low-resolution features from one image with high-resolution features from another image is to convert the data to the frequency domain and combine the two using a crossover frequency then convert back to the spatial domain. The SAR data is of a very low resolution compared to the multispectral imagery so a natural approach is to upsample it to the resolution of the multispectral image then use a frequency-based transform (Fourier or Wavelet) and produce a new transform array with the low frequencies of the SAR combined with the higher frequencies of the multispectral channel, then do an inverse transform to produce the fused image. An important manipulation of the input images before attempting the fusion is to make both images have the same mean and standard deviation before performing the Fourier or wavelet transforms. Otherwise, one image or the other will likely dominate the blended result in an undesirable way.

After we preprocessed the Sentinel SAR image, tested a wavelet transform-based fusion technique to combine the SAR data, multispectral imagery, and panchromatic imagery based on current literature [41]. We first tested a fusion method using intensity-hue-saturation (IHS) transforms and à trous wavelet transforms (ATWT). This method used the panchromatic, multispectral, and SAR images to generate a four-band fused product. The multispectral and SAR data were resampled to the spatial resolution of the panchromatic imagery before applying this method. First, an IHS transform was applied to the multispectral image.

The generalized intensity from this transform was used in the two following steps. An ATWT was applied to the generalized intensity. Additionally, the generalized intensity was histogram matched with the panchromatic image. An ATWT was then applied to the matched data (result). Separately, texture information was extracted from the SAR data. We used homogeneity via a gray level co-occurrence matrix as per the source paper [41]. The high pass details of the transformed histogram were then modulated by the SAR texture. An inverse ATWT was then applied to the modulated result and then to the low pass details of the transformed general intensity of the multispectral data. Finally, an inverse IHS transform was applied to the preceding output, and the spectral bands obtained from the IHS transform were used to produce the final fusion product.

We also developed and tested a modified version of the above method which produced a one-band fused product from the panchromatic and SAR data. An ATWT was applied

to the panchromatic image, and the homogeneity matrix was then obtained from the SAR data. The SAR texture was then used to modulate the high-pass details as in the previous method. The final product was then produced by applying an inverse ATWT to the modulated high-pass details and original low-pass details. As a third test, we also tried this same method; however, we bypassed the SAR texture extraction step and simply used the original SAR data. Results from this method were acceptable.

## 2.) Modified Unsupervised Deep Learning CNN for Fusion

As one of the objectives of this study, we wanted to evaluate using deep learning and CNNs to automatically fuse all of the panchromatic, multispectral, and SAR data together. To test this theory, we modified the traditional supervised CNN by adding the SAR data into the input and adding an extra band into the final product. We were hoping by adding the SAR as an extra input to the supervised CNN model we would be able to perform both pan-sharpening and data fusion simultaneously. However, when we introduced the SAR image to the model it destroyed the network's ability to pan-sharpen the multispectral image. To have a "reference" for the Wald protocol, we needed to downsample the SAR by the resolution factor difference in the multispectral and pansharpening, which was 4 in each direction. The multispectral and panchromatic channels are similar enough with respect to the spatial resolution to accomplish this task. However, the order of magnitude of the SAR vs. the panchromatic data would require such downsampling that extremely little resolution would remain for the Wald protocol. Unfortunately, the large-scale difference (1/20) between the SAR data and panchromatic data was too great for the deep learning model to overcome and the results were poor.

## 3.) Modified Supervised Deep Learning CNN for Fusion

Since the supervised learning deep learning method was unsuccessful, we then developed a deep learning-based fusion method by modifying the unsupervised CNN. The multispectral imagery was replaced by the SAR data and the loss function was adjusted to accommodate single-band output.

The SAR data were resampled to the spatial resolution of panchromatic imagery and this new high-resolution SAR was used as input, while the original SAR data was used in the loss function. The panchromatic imagery was resampled to the spatial resolution of the SAR for use in the loss function. A scalar value $\beta$ was learned via least squares to satisfy the equation:

$$p = \beta S + e_1$$

where S was the SAR data and p was the low-resolution panchromatic image.

- The loss function was modified as follows.:
  - *Spatial loss:*

    $I = \beta\hat{S} + e_1$ where $\hat{S}$ is the model output.

    $L_{d1} = MSE(P, I)$

    $L_{d2} = 1 - SSIM(P, I)$

    $L_d = L_{d1} + L_{d2}$
  - *Spectral loss:*

    $\tilde{S} = K * \hat{S}$

    $L_{s1} = MSE(\tilde{S}, \hat{S})$

    $L_{s2} = 1 - SSIM(\tilde{S}, \hat{S})$

    $L_s = L_{s1} + L_{s2}$
  - $D_s$:

    Because applying $D_\lambda$ to the SAR data does not make sense, only the $D_s$ component of QNR was used.

    $$D_s = \sqrt{|Q(\hat{S}, P) - Q(s, p)|^2}$$
  - The final loss is defined as $L_{SSQ} = L_d + L_s + D_s$

Because of the large disparity in spatial resolutions between the SAR and panchromatic data, this model had difficulty converging and did not yield acceptable results.

**Agricultural Row Crop Evaluation**

Our next step in the project was to conduct four experiments to examine crop type classification performances. The four experiments included 1) using the multispectral band image only, 2) using multispectral band + vegetation index derived from the multispectral image, 3) using spectral band + vegetation index + Gray-Level Co-Occurance Matrix (GLCM) textures, and 4) using spectral band + vegetation index + GLCM textures and features from SAR image.

For each experiment, we evaluated classification cases by using the original multispectral band images and by using our pan-sharpened image. Results were compared to analyze how much the pansharpening improved crop type classification.

**Feature Generation**

***Multispectral features:***

Four spectral bands from our Pleiades satellite image are used in the study: 1) Blue band, 2) Green band, 3) Red band, and 4) Near-infrared band. As a result, we extract four spectral features from these bands of images.

***Vegetation index features:***

Vegetation index plays an important role in crop type classification. In the reviewed paper [42], a number of vegetation indexes derived from the optical images were found in the literature for crop type classification, including the well-known Normalized Difference Vegetation Index (NDVI). Considering the four spectral bands available, these 6 vegetation related indexes are selected for our experiments [43][1]: 1) NDVI, 2) the Atmospherically Resistant Vegetation Index (ARVI), 3) the Soil Adjusted Vegetation Index (SAVI), 4) the Green Chlorophyll Index (GCI), 5) the Structure Insensitive Pigment Index (SIPI), and 6) the Simple Ratio.

***Texture features from Gray-Level Co-occurrence Matrix (GLCM):***

The Gray-Level Co-occurrence Matrix (GLCM) texture features developed by Haralick et al [44] is a well-known texture feature set used in image analysis and remote sensing applications. In this study, we generate the GLCM texture features using the SeNtinel Application Platform (SNAP) software [45]. Ten texture features from the GLCM are calculated from the SNAP: 1) Contrast, 2) Dissimilarity, 3) Homogeneity, 4) Angular Second Moment (ASM), 5) Energy, 6) Maximum Probability, 7) Entropy, 8) GLCM Mean, 9) GLCM Variance, and 10) GLCM Correlation. The GLCM is calculated using a 5x5 window size with a displacement of 1 and all angles (0, 45, 90, 135).

***Features from SAR Data:***

The SAR data in this study is from the Sentinel-1 SAR imagery. The SAR images are preprocessed by using SNAP, including thermal noise removal, application of orbital file, radiometric calibration, and terrain correction. The SAR images are then collocated with the optical MSI images so that features are ready to be extracted in alignment with the MSI images. There are two SAR bands, VH and VV. Each band contains the values of Sigma0, the calibrated backscatter coefficients. We constructed a total of 22 features for each pixel of interest, including 4 MSI spectral features, 6 vegetation indexes from MSI

images, 10 GLCM texture features from MSI, and backscatter coefficients from the 2 SAR polarizations.

### *Crop type ground truth data and crop regions of interest:*

Our study was focused on the classification of four agricultural row crops and on grassland grown in the Tennessee River Valley in 2019: Corn, Cotton, Soybean, Double Crops (Winter Wheat/Soybean), and Grassland. The ground truth data come from the map of crop types from USDA's National Agricultural Classification Survey.

To reduce the number of pixels in the learning process, selected regions of interest were manually extracted from the crop map, with each region representing a unique surface type. A total of 33 regions were covered by the MSI images. Of the 33 regions, 7 of them were classified as corn, 6 cotton, 11 soybeans, 6 double-crop, and 3 of them were grassland fields. In the MSI image, there was a total of 861,820 corn pixels, 501,681 cotton pixels, 717,501 soybean pixels, 328,658 double-crop pixels, and 41,874 grassland pixels. The percentages of the five crop types were 35.15%, 20.46%, 29.27%, 13.41%, and 1.71% for corn, cotton, soybean, double-crops, and grassland, respectively. The grassland pixels were significantly less than the four-row crops. The total number of MSI pixels in the regions of interest was 2,451,534.

### 26. What opportunities for training and professional development has the project provided?

This project provided the foundation for UAH staff and student professional development in the areas of remote sensing and artificial intelligence. During this project's period of performance, UAH established an Intelligence Community Center of Academic Excellence (IC CAE) in critical technologies. With this program, UAH selected a number of outstanding students to become IC CAE Scholars. Each scholar was assigned a mentor to develop specialized and focus short research projects. Two of Dr. Beck's students were able to conduct deep learning and remote sensing research based on techniques from this project. One of the student's research was entered into the 7th annual Research Horizons Day at UAH, a showcase of collaborative research projects under the mentorship of faculty and staff. The scholar was awarded one of the top awards for the College of Science in the Research Horizons Virtual Poster Sessions. In his acceptance letter, he noted that his research project provided him with a valuable learning experience in using open-source deep learning software, state-of-the-art computer hardware systems, and advanced mathematical techniques. In addition, Dr. Beck developed a short deep learning and remote sensing training program that he gave each year to the IC CAE scholars. This grant provided Dr. Beck

and his team with invaluable insight into new deep learning technologies that would not have happened if it were not for this grant opportunity.

## 29. Publications, conference papers, and presentations.

**Conference presentations:**

Beck, J.M., Spatial Dynamics and Analysis of Crops using Super Multispectral Image Resolution and Radar Fusion. 7th Annual Intelligence Community-Academic Research Symposium (ICARS). Virtual Presentation. 22 September 2021.

Beck, J.M., Spatial Dynamics and Analysis of Crops using Super Multispectral Image Resolution and Radar Fusion. 7th Annual Intelligence Community-Academic Research Symposium (ICARS). Virtual Presentation. 30 September 2020.

Daigle, M. and J. Beck. Use of Autoendcoders in a Convolutional Neural Network for Image Fusion. The University of Alabama in Huntsville Intelligence Community Center of Academic Excellence (IC CAE) Summer Colloquium. July 2020.

Pagani, S. and J. Beck. Applications of Deep Learning: Pan Sharpening and Radar Fusion. The University of Alabama in Huntsville Intelligence Community Center of Academic Excellence (IC CAE) Summer Colloquium. July 2020.

## 45. Changes in approach and reasons for change.

The main challenge of the radar data fusion aspect of this project is the large-scale factor between the SAR data and the pan-sharpened image. The SAR is about 5 times the size of the MSI in pixel resolution and 20 times larger than the panchromatic. This large variation in scale factors was unreasonable for an all-in-one super-resolution deep learning model solution. From the initial results of the study from year 1, we have learned that we will need to fuse the SAR with the MSI imagery first, and then use the pan-sharpening CNN model to further enhance the fused product.

## 50. What were the outcomes of the award?

**Other outcomes (results) of the award are as follows:**

As with all neural networks, a primary challenge is the tweaking of hyper-parameters in order to produce quality results consistently. Starting with the unsupervised CNN, the PSGAN, and the PSColorGAN, we have a reliable baseline method in terms of performing pan sharpening. A major challenge in designing a neural network for fusing the three images will be the second step of incorporating the SAR. While PSGAN may serve as a good starting platform for this step, it is unlikely that the same

hyper-parameters would produce a good result with different types of imagery. Additionally, it is possible that the pan-sharpening step must be tweaked to improve the end result after SAR fusion. SAR is a fundamentally different dataset than MSI, heavy modification of parameters was required to achieve reliable results. In the end, the best method of fusion of the SAR data with the multispectral data was through the use of traditional mathematical algorithms such as IHS transforms and ATWT transformations.

**Pan-sharpen Image Results / Metrics**

To evaluate our results we conducted a series of five statistical tests comparing our unsupervised CNN, the basic GAN, and the PanColorGAN pan-sharpening model with 18 traditional pan-sharpening algorithms based on examples from recent publications [39, 40] for a total of 21 methods. *We did not use the results from our supervised CNN with the loss function because visual inspection showed deficiencies with our results due to mode collapse*. In this document, we denote the pan-sharpened image as the multispectral image sharpened through the panchromatic image using a pansharpening method, it is also referred to as a fused image.

In [39], traditional pansharpening methods are classified as one of two approaches: component substitution (CS) and multiresolution analysis (MRA). The CS-based methods are based on the substitution of components obtained through a spectral transformation with a high-resolution panchromatic image. The MRA-based methods are based on the injection into upsampled MS bands of spatial details obtained from multiple resolution decomposition of the panchromatic image [39]. In [39], intensity hue saturation (IHS), Brovey Transform (BT), principal component analysis (PCA), gram-schmidt (GS), GSA, BDSD, and PRACS are considered as CS-based methods, while high-pass filter (HPF), SFIM, Indusion, MTF-GLP, MTF-GLP-CBd, additive A Trous wavelet transform (ATWT), ATWT-M2, MTf-GLP-HPM, MTF-GLP-HPM-PP, and AWLP are considered as MRA-based methods. In this study, the EXP method does not involve any sharpening process using the panchromatic image. Instead, multispectral image bands are generated by upsampling multispectral bands through bi-cubic extrapolation. The three deep learning methods are referred to as PanColorGAN, Unsupervised, and a basic GAN.

The full resolution assessment approach as described in [39] was taken. For the full resolution approach, a comparison is made among the original panchromatic image, the multispectral image that was upsampled to the panchromatic image resolution (and used for pansharpening), and our three pan-sharpened images (result images). The qualities of the pan-sharpening methods were evaluated by both visual and quantitative analyses with respect to spatial and spectral fidelity. The five metrics were Spectral

Angle Mapper (SAM), Spatial Correlation Coefficient (SCC), D-lambda, D-S, and Quality with No Reference (QNR) [39].

The 21 pan-sharpen methods were assessed for 16 sub-areas using a full resolution validation approach and metrics were calculated on the five indices: SAM, SCC, QNR, D-lamda, and D-s. The Tennesse River Valley image was divided into 16 sub-areas so that the image size of each sub-area would fit into the memory of our computers to pan sharpen the image and evaluate the performance efficiently. The QNR metric is the combination of the two metrics, D-lamda and D-s. For this assessment, both of them contribute to the QNR index equally (i.e. coefficients are set as 0.5 for both of them). For each of the 16 sub-areas, the five indices were calculated based on the 512 x 512 image blocks. Each of the 16 sub-areas consists of more than 120 such blocks. As a result, the mean and standard deviation of each index was obtained for each sub-areas. The 21 methods were ranked for each of the five indices and for each of the 16 sub-areas. Therefore for the full TN valley image, we have 16 ranking lists for each index, with each one influenced by the image content in the sub-area. By comparing the 16 ranking lists for each index, we can to some extent understand how much each index is affected by the image content. The following are the results of our evaluation:

**SAM Index:**

For the SAM index, we observed that the best method was EXP, as it ranks first for 14 of 16 sub-areas. This was expected as the EXP method in our assessment uses images upsampled from original multispectral images by interpolation. No panchromatic image information is incorporated into the fused (pan-sharpened) image. It is then downsampled to the original multispectral size and compared with the original multispectral image for the SAM index. As a result, we expect the minimum spectral distortion for the EXP method. It is interesting that for a couple of sub-areas, the Indusion and Brovey methods are on top of the ranking list. The Brovey method is almost consistently the second-best on the SAM index.

The SAM index calculation in this assessment is different from that used in [39]. In [39], SAM is calculated using a reduced resolution approach. In this approach, the panchromatic image and original multispectral image are downsampled (reducing resolution) so that downsampled panchromatic image has the same resolution/dimension as that of the original multispectral image. The downsampled multispectral image is then 'pan-sharpened' by the downsampled panchromatic image to the resolution of the original multispectral image. As a result, the fused multispectral image has the same spatial resolution as that of the original multispectral image. The original multispectral image then serves as a reference image to assess the pan-sharpened image and pansharpening method.

In general, the SAM index ranking calculated using the full resolution approach showed similar characteristics as those in [39] which uses a reduced resolution approach. In

[39], the CS-based approach often shows a higher spectral distortion (i.e. poor SAM index) but a better visual appearance of fused images. From the SAM ranking spreadsheet, the CS-based methods such as PCA, GS, GSA, IHS are generally in lower ranking as compared to MRA-based methods such as Indusion, HPF, SFIM. The exception is the Brovey method which is consistently on top of the ranking. Our three DL-based methods are among the CS-based methods in spectral distortion using the SAM index. The PanColorGAN ranked higher than other methods in this study (**see attachment 1**).

## SCC Index:

The SCC index is a measure for spatial enhancement of pan-sharpened images as it calculates the correlation between a sharpened image and a pan-sharpened image that contains spatial details with higher spatial resolution. As observed in [39], the pan-sharpened images using CS-based methods are often superior in visual appearance. This was consistent with the SCC index ranking (**see attachment 2**). In most of the sub-areas, the EXP method ranks last in the SCC index. The CS-based methods such as IHS, Brovey, GS, and GSA ranks on top and the MRA-based methods are generally ranked lower. The EXP method ranks the last for most of the sub-areas, as expected (no information in the panchromatic image is included in upsampled MS bands). Of our three DL-based methods the PanColorGAN method ranks the highest and ranks in the middle of the 21 methods while other our other models rank consistently lower.

## D-Lambda Index:

Results from the D-lambda ranking (**see attachment 3**) indicated the best method was EXP, as expected as little spectral distortion is expected without pansharpening involvement. The second best was the BDSD method, consistent with the results found in [39]. The Brovey transform method ranks in the lower part of the list, contrary to that of the SAM index ranking. Our three DL methods rank higher (5, 6, and 8) as compared to those in SAM ranking results. Of the three methods, the ranking order was Unsupervised, PanColorGAN, and GAN on average.

## D-S Index:

Results from the D-S ranking (**see attachment 4**) indicated the Unsupervised and GAN methods ranked at the top of the list. Compared to the SCC index where component substitution (CS) methods generally perform better (i.e. IHS, Brovey, GS, GSA), these methods rank on the lower part of the list. So spatial distortion from these two indices, i.e. SCC and D-S, seems not to agree well in general. Our PanColorGAN method ranks in the middle of the list.

**QNR Index:**

Finally we assess these methods based on the QNR index. As the name suggests, the QNR (quality w/no reference) index does not require a reference image to estimate the quality of a pan-sharpened image. The QNR is the combination of the two indices: D-lambda and D-S. The former measures the spectral distortion while the latter measures the spatial distortion. For any two multi-spectral bands of the pan-sharpened image, the Universal Image Quality Index (UIQI) or Q-index is calculated which measures the similarity of the two images through correlation and the differences in luminance and contrast. The Q-index for the two fused bands is then compared with the Q-index for the two corresponding original multispectral image bands. The smaller the difference between the two Q-index, the more consistent the fused image bands are corresponding to the original MS bands, the less the spectral distortion, and the better the fused image quality is considered.

The D-lambda is the average of the Q-index difference from all pairs of multispectral bands. On the other hand, the D-S is calculated based on the Q-index between the fused image bands and the panchromatic image. For each multispectral band, two Q-index values are calculated. The first Q-index is calculated using the fused image band and panchromatic image. The second Q-index is calculated using the original multispectral band and the downsampled panchromatic image. The difference of the two Q-index is a measure of the spatial consistency of the fused multispectral image corresponding to that of the original multispectral image band. The smaller the difference, the better the quality of the fused image. D-S is the average value from all the multispectral bands. In this method, the downsampled panchromatic image is an approximation of the one that is supposed to be observed at a lower resolution from the panchromatic sensor.

Results from the QNR ranking (**see attachment 5**) indicated that the EXP, BDSD, and our Unsupervised and GAN methods were among the best. Our PanColorGAN method ranks relatively in the middle of the list (9 out of 21 methods). In this current QNR calculation, the D-Lambda and D-S contribute equally. Changing the ratio of the combination is expected to change the ranking order of these methods. Figure 7 shows the images using the 19 methods along with the panchromatic image. It is from subset 1 of the Tennesse River Valley image. The 512 x 512 block is on row 2 and column 14 of the subset 1 image.

Figure 7. First row, left to right: 1) panchromatic image, 2) upsampled MSI images, 3) PCA, 4) IHS, 5) Brovey, Second row, left to right: 6) BDSD, 7) GS, 8) GSA, 9) PRACS, 10) HPF, Third row, left to right: 11) SFIM, 12) Indusion, 13) ATWT, 14) AWLP, 15) ATWTM2, Fourth row, left to right: 16) PanColorGAN, 17) MTF_GLP, 18) MTF_GLP_HPM_PP, 19) MTF_GLP_HPM, 20) MTF_GLP_CBD.

## Example of Pan-sharpening Results
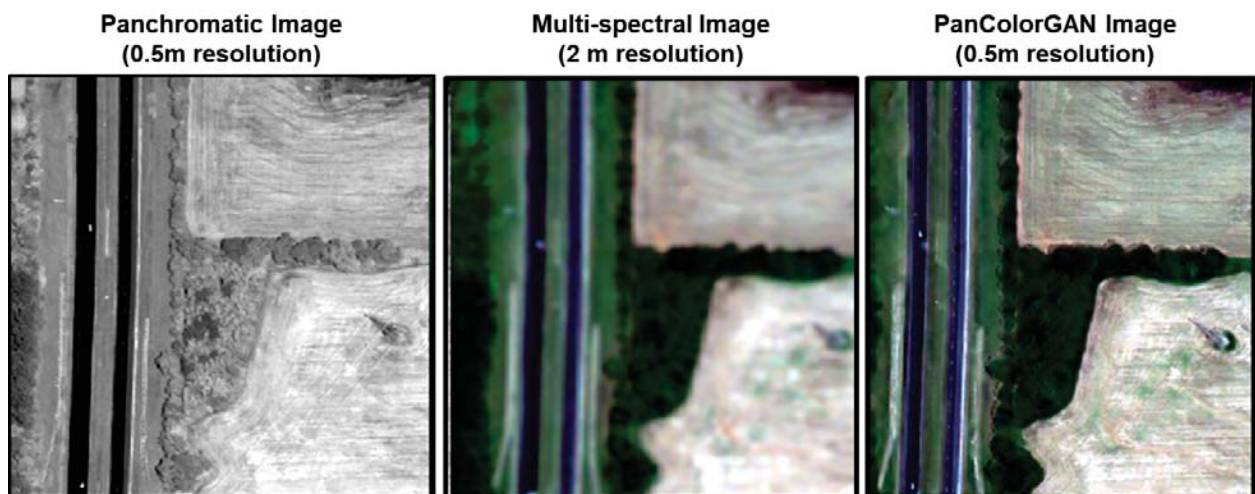


Figure 8. Results from our PanColorGAN pan-sharpening results.

**Basic CNN**
*(fuzzy)*

**PanColorGAN**
*(crisp detail)*
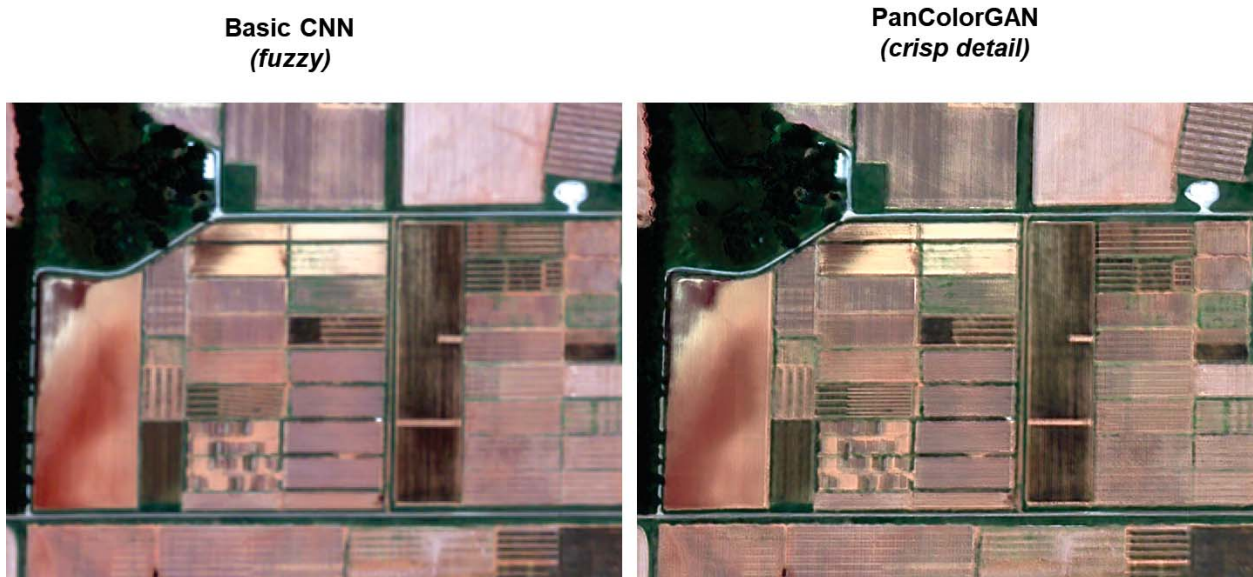
Figure 9. Comparison of the supervised CNN and the PanColorGAN pan-sharpening models.

## Synthetic Aperture Radar (SAR) Fusion Results

One of the biggest challenges encountered was the large gap in spatial resolution between the available SAR data and our imagery. This made it particularly difficult to extract useful features from the SAR and fuse them with the pan-sharpened image. We initially attempted introducing SAR data as an additional input to our deep learning pan-sharpening models with the goal of improving the quality of model output. We observed no qualitative or quantitative improvements in spatial details with this approach. Therefore, we decided to use IHS and  Wavelet Transform fusion techniques which produced good results. The resulting GEOINT product was a five-band pan-sharpened image consisting of a blue band, green band, red band, near-infrared band, and SAR band.

## Lessons Learned

### *Edge Effects:*

Just as in signal processing in general, running image tiles through a convolutional neural network for "prediction" will usually result in edge effects around the borders of the output tiles. The easiest way around the problem is to have redundancy in the tiling of the image, with the standard approach being to use half the tile width as the tile offset. If tiles are 256x256, then use a stride of 128 within the image, creating roughly twice the tiles horizontally and vertically. Since the 256x256 output tiles will usually have edge effects, a few pixels can be removed from each edge and still have plenty of redundancy to not have gaps in the resulting image except for a few pixels around the

edge. Removing eight pixels from the edges of tiles when building the output image suffices when convolution operators of sizes 9x9 and 5x5 were used in layers. Larger operators may require a larger number of pixels to be removed.

### *Sufficient Network Training:*

A convolutional neural network that has not been trained sufficiently can produce output that can mislead the developer into thinking that problems exist in other parts of the implementation, such as the data preparation and image synthesis portions of the code. Completely wrong relative color levels can result despite individual channels viewed as monochrome can look fine in isolation. Passing the training tile set through the CNN several epochs until diminishing returns are gained can greatly improve the output quality and made problems disappear. Just eight epochs of training made relative color levels improve from completely wrong to very closely matching the data.

### *Separation of Functionality and Reuse of Intermediate Steps:*

When working with large images that require amounts of RAM, instead of tiling the image and keeping the whole tile set in RAM during the network training, separating the Wald protocol and tiling process into a separate program and writing the tiles to a directory allows their reuse in training a previously created CNN multiple epochs, or even multiple CNNs with different parameters without having to regenerate the tiles each time. The process for creating the tiles for the prediction can be separated in the same manner as the prediction script to greatly reduce the memory requirements required in creating network output tiles. Making an image from the final output tiles into a separate process is a good idea as well. During development, these separations of duties into different programs make debugging significantly easier.

## Crop type classification and performance evaluation

### *Crop type classification using MSI spectral bands only*

*Generating feature set for model learning*

For the sake of model learning with balanced samples (i.e. model learning is not overwhelmed by a huge number of samples and each class type contributes a relatively equal number of samples in model learning), we decided to choose roughly 50000 samples from the total number of over 2 million samples for model learning, with each class contributing 10000 samples.

Two strategies were taken to select these samples from the sample pool to form a feature set. For the first one, we randomly selected five regions of interest, with each

one representing a crop type. And we extracted 10000 samples from each of the five regions of interest. We called this scheme the select-region method.

For the second approach, the total number of selected samples for each crop type was set as 10,000 for balanced learning. But the samples were extracted from all the regions of interest and there were no left-out regions. For each crop type, the number of pixels selected from a region was proportional to the region size of a given crop type. This way, the number of samples for crop types is balanced, and pixels from all the regions are involved in the model learning process. We call this approach the all-region method. We expect better performance from the all-region method as samples from all regions are involved in the model learning. We expect the worse classification performance for the select-region method but would like to know how much worse it is as compared to the all-region method. If the classification performance using the select-region method is not significantly worse than that of the all-region method, the crop type signatures seem well separated and intro-class dissimilarity would not be significant.

Two strategies for the all-region method are also evaluated. For the first strategy, the number of pixels selected from a region into the learning feature set for a crop type was proportional to the size of the region with the crop type. The larger the region, the more samples were selected into the feature set for learning. For the second strategy, the number of pixels selected from a region into the learning feature set for a crop type was equal for all the regions with the crop type. This way, samples in the learning feature set are not only balanced among crop types but also balanced with regard to the regions for a crop type. We call the first strategy the all-region ratio-sample method, and the second one the all-region equal-sample method.

*Model performance evaluation metrics*

The confusion matrix brings fundamental information about classifier performance. From the confusion matrix, we estimate the accuracy of model prediction for each class and how capable each class can be accurately identified. Besides the confusion matrix, other metrics such as overall accuracy, precision, recall, F1 score are often used to evaluate a classifier's performance.

In this study, we use all relevant metrics to evaluate the random forest (RF) model performance for crop type classification. The first metric evaluated was overall accuracy, which is the fraction of correct predictions over all samples. A classification report summarizes the precision, recall, and F1 score metrics for each class. The precision is the ratio of correct prediction to the overall prediction for a class. It shows the quality of a model's prediction for a class. The higher the value, the fewer the non-class instances were assigned to the class. The recall is the ratio of correct prediction to all the instances for a class. It shows the capability of a model in identifying the occurrences

for a class. The higher the recall value, the fewer the class instances are assigned to other classes (i.e. not identified by the model).  The F1 score is a harmonic mean of precision and recall metrics, which is a balanced score considering both false positive and false negative occurrences for a class. Through these metrics, we can assess how the RF model performs on each crop type. The macro average and weighted average of these metrics are also available, with the former the average of all classes for each metric and the latter the weighted average of these metrics by the percentage of the samples of each class. The weighted average of the F1 score is a good measure of a multi-class classifier performance generally.

The third metric was Cohen's kappa statistics. Cohen's kappa measures how much better a classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class and is considered good for multi-class and imbalanced class problems [6]. The maximum value of Cohen's kappa is 1 when a classifier predicts perfectly. A value of 0 suggests a classifier is no better than a random guess according to the sample distribution.

The fourth metric was Matthew's correlation coefficient (MCC) which is used as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 is an average random prediction, and -1 is an inverse prediction. The statistic is also known as the phi coefficient." The MCC ranges between -1 and 1, with 1 indicating the best agreement between predictions and actual class labels.

The fifth metric was the Area Under the ROC (AUROC) where ROC stands for Receiver Operating Characteristics and is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate." The AUROC is the area under the ROC curve. The maximum AUROC is 1. The larger the value, the excellent the model performance.

The sixth metric we used was the Log-loss score. A good interpretation of the log-loss metrics is given in [47]. The lower the log-loss score, the better a prediction model performs.

*Classification performance of Random Forest (RF) model using default parameters*

*Confusion matrix for testing set and independent set*

| | | Testing set | | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 2480 | 0 | 0 | 30 | 0 | | Corn | 366685 | 9570 | 2822 | 282847 | 189857 |
| | Cotton | 0 | 2372 | 158 | 0 | 0 | | Cotton | 1080 | 386341 | 96391 | 7161 | 708 |
| Actual | Soybean | 0 | 96 | 2370 | 0 | 0 | Actual Soybean | 39969 | 260216 | 185350 | 219111 | 2855 |
| | DblCrop | 22 | 0 | 0 | 2455 | 8 | | DblCrop | 1739 | 20 | 27 | 182170 | 134702 |
| | Grassland | 1 | 0 | 0 | 14 | 2494 | | Grassland | 6002 | 213 | 0 | 15863 | 9792 |

Testing set accuracy: 0.97368

*Classification report for independent set*

```
classification report
                                precision    recall   f1-score    support

                         Corn       0.88       0.43       0.58     851781
                       Cotton       0.59       0.79       0.67     491681
                      Soybean       0.65       0.26       0.37     707501
DoubleCrop/WinterWheat_Soybean     0.26       0.57       0.36     318658
                    Grassland       0.03       0.31       0.05      31870

                     accuracy                             0.47    2401491
                    macro avg       0.48       0.47       0.41    2401491
                 weighted avg       0.66       0.47       0.50    2401491
```

*Other metrics for independent set*

Accuracy 0.470681755625984
BalancedAccuracy 0.4714305720273916
AUROC 0.7820539978424056
Kappa 0.3439492443357288
Matthew Coef 0.3695712432333883
Logloss 8.906852477296693

*Results using training feature set from the all-region ratio-sample method*

*Confusion matrix for testing set and independent set*

| | | Testing set | | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 2408 | 1 | 39 | 17 | 49 | | Corn | 750679 | 5640 | 20661 | 24063 | 50743 |
| | Cotton | 22 | 2351 | 114 | 3 | 8 | | Cotton | 8852 | 428408 | 54387 | 5 | 32 |
| Actual | Soybean | 11 | 127 | 2303 | 9 | 4 | Actual Soybean | 95240 | 137790 | 451656 | 18025 | 4796 |
| | DblCrop | 13 | 0 | 10 | 2469 | 10 | | DblCrop | 21528 | 567 | 4593 | 270561 | 21412 |
| | Grassland | 80 | 1 | 5 | 32 | 2410 | | Grassland | 1617 | 93 | 400 | 756 | 29005 |

Testing set accuracy: 0.95559

## Classification report for independent set

```
classification report
                                 precision    recall  f1-score    support

                         Corn       0.86      0.88      0.87     851786
                       Cotton       0.75      0.87      0.81     491684
                      Soybean       0.85      0.64      0.73     707507
DoubleCrop/WinterWheat_Soybean     0.86      0.85      0.86     318661
                    Grassland       0.27      0.91      0.42      31871

                     accuracy                           0.80    2401509
                    macro avg       0.72      0.83      0.74    2401509
                 weighted avg       0.82      0.80      0.81    2401509
```

## Other metrics for independent set

Accuracy 0.8037900336829885
BalancedAccuracy 0.8300230383035071
AUROC 0.9459264596140856
Kappa 0.7342767302674598
Matthew Coef 0.7380914488476223
Logloss 1.5798480599850815

## Results using training feature set from the all-region equal-sample method

## Confusion matrix for testing set and independent set

| | | Testing set | | | | | | | Independent set | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | Predicted | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 2371 | 1 | 26 | 14 | 75 | | Corn | 697979 | 22709 | 24470 | 55373 | 51254 |
| | Cotton | 23 | 2317 | 129 | 0 | 20 | | Cotton | 2941 | 427026 | 61522 | 145 | 51 |
| Actual | Soybean | 11 | 163 | 2316 | 13 | 5 | Actual | Soybean | 33590 | 168465 | 438822 | 61238 | 5387 |
| | DblCrop | 11 | 2 | 9 | 2467 | 21 | | DblCrop | 23015 | 1689 | 5066 | 252772 | 36120 |
| | Grassland | 84 | 3 | 14 | 40 | 2362 | | Grassland | 1323 | 20 | 222 | 426 | 29880 |

Testing set accuracy: 0.9468672481395535

## Classification report for independent set

```
classification report
                                 precision    recall  f1-score    support

                         Corn       0.92      0.82      0.87     851785
                       Cotton       0.69      0.87      0.77     491685
                      Soybean       0.83      0.62      0.71     707502
DoubleCrop/WinterWheat_Soybean     0.68      0.79      0.73     318662
                    Grassland       0.24      0.94      0.39      31871

                     accuracy                           0.77    2401505
                    macro avg       0.67      0.81      0.69    2401505
                 weighted avg       0.81      0.77      0.78    2401505
```

## Other metrics for independent set

Accuracy 0.7688840955983852
BalancedAccuracy 0.8077851987937287
AUROC 0.9454816098495885
Kappa 0.6914018676234226
MatthewCoef 0.6969904202666065
Logloss 1.346309664914619

*Comparing results from using the two feature sampling methods*

The overall accuracy of RF model on testing data for the select-region method is over 97.3%. This result suggests that the 5 regions of distinct crops are highly distinguishable using the RF model. The high accuracy for the testing set also implies that within-field pixels share high similarity in spectral features. However, the accuracy of the RF model drops significantly to 47.1% when applied to other crop field samples in the independent set. This result is most likely caused by the dis-similarity of spectral features among samples of different fields sharing the same crop type. The RF model cannot perform well with samples from only selected crop fields for learning. The results from the all-region method further confirm this conclusion.

For the RF model using the all-region ratio-sample method, the overall accuracy for the testing set and independent set were 95.5% and 80.3%, respectively. The accuracy of the testing set using all-region ratio-sample samples is about 2% less than that using select-region samples. For the all-region method, the RF model is trained using samples from all regions which are expected to present more diverse spectral signatures. As a result, the model performance on the test set slightly decreases. However, the over accuracy for the independent sample set reaches 80%, significantly higher than using the select-region method.

For the two all-region methods, the ratio-sample method performs slightly better than of the equal-sample method. The overall accuracy for the equal-sample method is 76.9%, about 3.4% lower than that of the ratio-sample method. This result also implies a somewhat disparity of spectral signatures of individual regions within a crop type. The ratio-sample method allows more training samples in the feature set for larger regions, therefore the RF model is learned to favor larger regions. As a result, large regions perform better which contributes more to the overall model performance.

These results above seem to confirm our speculation earlier that a classifier needs to be learned with a well-sampled feature set (i.e. samples extracted from all regions to be representative). In this case, the samples need to be from all crop fields as spectral signatures from individual fields of a crop type may present a wide dynamic range. Spectral signatures from individual fields may not be representative.

Comparing other metrics, including AUROC, Kappa, MatthewCoefficient, and Logloss, the overall RF model performance is significantly better using the all-region method than that using the select-region method.

*Feature importance ranking*

The RF model was ranked by the importance of features in the classification process. This feature importance is available as a class attribute. For spectral only RF learning, we have 4 features, i.e. the 4 spectral bands. The following table shows the model feature importance outputs from feature sets using select-region, all-region ratio-sample, and all-region equal-sample methods.

| Feature importance | | | | | |
|---|---|---|---|---|---|
| select-region | | all-region ratio-sample | | all-region equal-sample | |
| blue | 0.33657225 | blue | 0.36599277 | blue | 0.31027418 |
| nir | 0.30558308 | nir | 0.27322675 | nir | 0.28707258 |
| red | 0.27168806 | red | 0.1824947 | red | 0.2052926 |
| green | 0.08615661 | green | 0.17828578 | green | 0.19736063 |

From the table, the feature importance rankings are the same for all three methods, though the contributing factors are different for each method. The Blue band is the most important feature and the green band is the least important feature. For the select-region method, the Blue, NIR, and Red bands are key features for the RF model while for all-region methods, all the features are important and needed for good classifier performance.

*Parameter optimization for RF models*

For the above-mentioned experiments, the RF models are trained with default model parameters. An effort is made for optimal parameter selection for RF models. We use the grid-search method for optimal parameter selection. In the grid-search method, each parameter is given a list of parameters to search for. The optimal parameter set is the one on the "multi-dimensional" parameter grids that the RF model delivers the best performance, in our case overall accuracy.

For both the select-region method and all-region methods, the setting for model parameter grid search is the same. Four model parameter sets are searched for optimal values. The search options for a number of tree estimators are set as [25, 50, 75, 100, 200, 500]. The max_features are set as ['auto', 'sqrt', 'log2']. The max_depth are set as [4, 7, 10, 20, 50, 100, 200]. The criterion parameter are from options of ['gini', 'entropy']. The cross-validation (CV) fold is set as 5.

The optimal parameter set of RF models for each sample generation method is shown in the following table.

| | Optimal model parameters | | |
|---|---|---|---|
| | Select-region | all-region ratio-sample | all-region equal-sample |
| n_estimators | 200 | 200 | 500 |
| max_features | auto | auto | auto |
| max_depth | 20 | 10 | 50 |
| criterion | entropy | entropy | entropy |

From the table above, the entropy method with which tree splitting is determined by information gain is selected. The "auto" method, which set max_features as the square root of a number of features when looking for the best feature split is selected. The optimal max_depth ranges from 10 to 50 and the number of trees in the forest range from 200 to 500. Following are the model performances for each of the three methods.

RF model performance using the select-region method:

| | | Independent set | | | | |
|---|---|---|---|---|---|---|
| | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 372769 | 8819 | 2870 | 280913 | 186410 |
| | Cotton | 244 | 386638 | 96018 | 8047 | 734 |
| Actual | Soybean | 41737 | 267818 | 172068 | 222386 | 3492 |
| | DblCrop | 1538 | 18 | 23 | 182887 | 134192 |
| | Grassland | 5641 | 193 | 0 | 15933 | 10103 |

Training+testing set accuracy:  0.99876

*Classification report for independent set*

```
classification report
                                   precision    recall   f1-score    support

                           Corn        0.88      0.44       0.59     851781
                         Cotton        0.58      0.79       0.67     491681
                        Soybean        0.63      0.24       0.35     707501
DoubleCrop/WinterWheat_Soybean        0.26      0.57       0.36     318658
                      Grassland        0.03      0.32       0.06      31870

                       accuracy                             0.47    2401491
                      macro avg        0.48      0.47       0.40    2401491
                   weighted avg        0.65      0.47       0.50    2401491
```

*Other metrics for independent set*

Accuracy 0.46823619159930224
BalancedAccuracy 0.4716269855381919
AUROC 0.7856715404853365
Kappa 0.34088462227677996
MatthewCoef 0.36665017228982394
Logloss 8.72056788263012

The RF model performance using optimal parameter set is slightly worse than that using default parameter setting. The overall accuracies are 46.82% and 47.07%, respectively from models with optimal parameters and from default parameters. Keep in mind that the result from the optimal parameter search is the average of 5 cross-validation results.

*RF model performance using all-region ratio-sample method*

| | | Training set + testing set | | | | | | | | Independent set | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 9563 | 0 | 105 | 118 | 209 | | Corn | 741056 | 1546 | 23222 | 36339 | 49623 |
| | Cotton | 81 | 9294 | 580 | 19 | 23 | | Cotton | 7865 | 421907 | 61888 | 17 | 7 |
| Actual | Soybean | 30 | 477 | 9376 | 106 | 5 | Actual | Soybean | 96101 | 124610 | 459831 | 22263 | 4702 |
| | DblCrop | 68 | 0 | 17 | 9893 | 19 | | DblCrop | 22604 | 2 | 3690 | 276062 | 16303 |
| | Grassland | 240 | 0 | 18 | 207 | 9534 | | Grassland | 1570 | 1 | 467 | 1026 | 28807 |

Training+testing set accuracy: 0.9535432755792085

*Classification report for independent set*

```
classification report
                                        precision     recall    f1-score    support

                                 Corn        0.85       0.87        0.86     851786
                               Cotton        0.77       0.86        0.81     491684
                              Soybean        0.84       0.65        0.73     707507
      DoubleCrop/WinterWheat_Soybean        0.82       0.87        0.84     318661
                            Grassland        0.29       0.90        0.44      31871

                             accuracy                               0.80    2401509
                            macro avg        0.71       0.83        0.74    2401509
                         weighted avg        0.82       0.80        0.81    2401509
```

*Other metrics for independent set*

Accuracy 0.8026882264442898
BalancedAccuracy 0.8296401634319486
AUROC 0.9507652624266579
Kappa 0.7327991634385813
MatthewCoef 0.7357285895510765
Logloss 0.6142519898828813

Again the performance from optimal parameter set and from default set is trivial. The overall accuracies are 80.27% and 80.38%, respectively from models with optimal parameters and from default parameters.

*RF model performance using all-region equal-sample method*

| | | | Training set + testing set | | | | | | | | Independent set | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Predicted | | | | | | | | Predicted | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 9995 | 0 | 0 | 0 | 1 | | Corn | 698672 | 25310 | 22581 | 54397 | 50825 |
| | Cotton | 0 | 9995 | 1 | 0 | 0 | | Cotton | 3221 | 426561 | 61645 | 135 | 123 |
| Actual | Soybean | 0 | 0 | 9999 | 0 | 0 | Actual | Soybean | 33629 | 172296 | 435760 | 60231 | 5586 |
| | DblCrop | 0 | 0 | 0 | 9996 | 0 | | DblCrop | 22044 | 1600 | 4853 | 250553 | 39612 |
| | Grassland | 0 | 0 | 0 | 0 | 9999 | | Grassland | 1304 | 27 | 213 | 420 | 29907 |

Training+testing set accuracy: 0.9999599887968631

*Classification report for independent set*

```
classification report
                                  precision    recall    f1-score    support

                          Corn       0.92        0.82       0.87       851785
                        Cotton       0.68        0.87       0.76       491685
                       Soybean       0.83        0.62       0.71       707502
 DoubleCrop/WinterWheat_Soybean      0.69        0.79       0.73       318662
                     Grassland       0.24        0.94       0.38        31871

                      accuracy                               0.77      2401505
                     macro avg       0.67        0.81       0.69      2401505
                  weighted avg       0.80        0.77       0.77      2401505
```

*Other metrics for independent set*

Accuracy 0.7667912413257519
BalancedAccuracy 0.8056699246372823
AUROC 0.9469168705218404
Kappa 0.6887441239914291
MatthewCoef 0.6945978663964455
Logloss 1.14857048980024

The performance from optimal parameter set and from default set is trivial. The overall accuracies are 76.68% and 76.89%, respectively from models with optimal parameters and from default parameters.

The default parameters of the RF model are as follows. The n_estimators (number of trees in the forest) is 100, the criteria were the gini method, the max_depth (max depth of the tree) is no limit, that is, the tree nodes are expanded until all leaves are pure or until all leaves contain less than the minimum sample splitting requirement. The max_features is auto. Comparing the results from RF models with grid-search optimal parameters with those using default parameters, it seems that parameter selection for the RF model has little impact on the classification performance. That means that improving crop type classification may most likely be achieved by selecting more impactful features than the model setting. We are going to include the vegetation indexes and texture features to seek better performances.

These spectral-only feature experiments give us the bottom-line accuracy that the RF model can deliver for the five crop type classification, around 75% to 80% using the four spectral bands of worldview-2 MSI images. Next, we followed the same procedure to examine the spectral-only classification performances using the pan-sharpened MSI images from our three deep-learning methods.

## *Crop type classification using Pan-sharpened MSI spectral bands only*

## **Results from PanColor GAN deep learning method**

## **Results using training feature set from the select-region method**

| | | Testing set | | | | |
|---|---|---|---|---|---|---|
| | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 38955 | 19 | 3 | 746 | 267 |
| | Cotton | 18 | 36003 | 3897 | 0 | 0 |
| Actual | Soybean | 3 | 2935 | 37116 | 0 | 0 |
| | DblCrop | 1146 | 2 | 0 | 38514 | 412 |
| | Grassland | 341 | 0 | 0 | 590 | 39033 |

| | | Independent set | | | | |
|---|---|---|---|---|---|---|
| | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 11159332 | 26990 | 32139 | 1450090 | 968953 |
| | Cotton | 264680 | 5700261 | 1884682 | 7745 | 8316 |
| Actual | Soybean | 4616311 | 2654396 | 1410815 | 2622588 | 24163 |
| | DblCrop | 381953 | 130 | 340 | 2891631 | 1830001 |
| | Grassland | 229917 | 206 | 0 | 58742 | 219565 |

Testing set accuracy:  0.948105

*Classification report for independent set*

```
classification report
                                   precision    recall   f1-score    support

                          Corn          0.67      0.82       0.74   13637504
                        Cotton          0.68      0.72       0.70    7865684
                       Soybean          0.42      0.12       0.19   11328273
DoubleCrop/WinterWheat_Soybean          0.41      0.57       0.48    5104055
                     Grassland          0.07      0.43       0.12     508430

                      accuracy                               0.56   38443946
                     macro avg          0.45      0.53       0.45   38443946
                  weighted avg          0.56      0.56       0.53   38443946
```

*Other metrics for independent set*

Accuracy 0.5561761011733811
BalancedAccuracy 0.5331813771849921
AUROC 0.7960605384771109
Kappa 0.4089401890016554
MatthewCoef 0.42494359840217977
Logloss 7.805613264560757

## Results using training feature set from the all-region ratio-sample method

| | | Testing set | | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 38445 | 63 | 405 | 223 | 825 | | Corn | 11557458 | 296301 | 371326 | 372732 | 1189698 |
| | Cotton | 236 | 37580 | 2219 | 76 | 143 | | Cotton | 51072 | 6777162 | 984106 | 1381 | 8001 |
| Actual | Soybean | 235 | 3039 | 36373 | 80 | 41 | Actual | Soybean | 1527594 | 2078875 | 7293756 | 391239 | 186820 |
| | DblCrop | 230 | 22 | 107 | 39050 | 499 | | DblCrop | 493685 | 73825 | 98540 | 4056019 | 532000 |
| | Grassland | 849 | 13 | 117 | 583 | 38544 | | Grassland | 32597 | 6780 | 5290 | 18643 | 595100 |

Testing set accuracy:  0.9499742496137442

*Classification report for independent set*

```
classification report
                                  precision    recall   f1-score    support

                           Corn       0.85       0.84       0.84   13787515
                         Cotton       0.73       0.87       0.79    7821722
                        Soybean       0.83       0.64       0.72   11478284
DoubleCrop/WinterWheat_Soybean       0.84       0.77       0.80    5254069
                      Grassland       0.24       0.90       0.38     658410

                       accuracy                              0.78   39000000
                      macro avg       0.70       0.80       0.71   39000000
                   weighted avg       0.81       0.78       0.78   39000000
```

*Other metrics for independent set*

**Accuracy** 0.7763973076923077
BalancedAccuracy 0.8031939431651584
AUROC 0.9399522649133555
Kappa 0.6997863299203768
MatthewCoef 0.7039118419894466
Logloss 1.5061117058961349


## Results using training feature set from the all-region equal-sample method

| | | Testing set | | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 37885 | 61 | 331 | 316 | 1449 | | Corn | 10031789 | 568729 | 335708 | 379522 | 2321757 |
| | Cotton | 344 | 36771 | 2737 | 95 | 190 | | Cotton | 90689 | 6910693 | 834758 | 14647 | 14901 |
| Actual | Soybean | 169 | 4073 | 35464 | 110 | 102 | Actual | Soybean | 1246854 | 2434792 | 6367283 | 1169580 | 109769 |
| | DblCrop | 377 | 26 | 194 | 38562 | 772 | | DblCrop | 292275 | 168292 | 77519 | 2591722 | 1974251 |
| | Grassland | 1338 | 37 | 223 | 828 | 37543 | | Grassland | 20717 | 3197 | 4498 | 7285 | 472734 |

Testing set accuracy:  0.9311389670845063

*Classification report for independent set*

```
classification report
                                       precision    recall  f1-score    support

                              Corn         0.86       0.74      0.79   13637505
                            Cotton         0.69       0.88      0.77    7865688
                           Soybean         0.84       0.56      0.67   11328278
     DoubleCrop/WinterWheat_Soybean       0.62       0.51      0.56    5104059
                         Grassland         0.10       0.93      0.18     508431

                          accuracy                              0.69   38443961
                         macro avg         0.62       0.72      0.59   38443961
                      weighted avg         0.77       0.69      0.71   38443961
```

*Other metrics for independent set*

**Accuracy 0.6860432773823696**
BalancedAccuracy 0.7227653047790219
AUROC 0.899346650661117
Kappa 0.5890949654764603
MatthewCoef 0.6006834180962203
Logloss 3.537414087289885

## Performance comparison between using original MSI images and using PanColorGAN images

Using select-region method

```
classification report                                            classification report
                  precision  recall  f1-score  support                             precision  recall  f1-score   support

            Corn     0.88     0.43     0.58     851781                      Corn       0.67     0.82      0.74   13637504
          Cotton     0.59     0.79     0.67     491681                    Cotton       0.68     0.72      0.70    7865684
         Soybean     0.65     0.26     0.37     707501                   Soybean       0.42     0.12      0.19   11328273
DoubleCrop/WinterWheat_Soybean  0.26  0.57  0.36  318658  DoubleCrop/WinterWheat_Soybean  0.41   0.57   0.48    5104055
       Grassland     0.03     0.31     0.05      31870                 Grassland       0.07     0.43      0.12     508430

        accuracy                      0.47    2401491                   accuracy                          0.56   38443946
       macro avg     0.48     0.47     0.41    2401491                  macro avg       0.45     0.53      0.45   38443946
    weighted avg     0.66     0.47     0.50    2401491               weighted avg       0.56     0.56      0.53   38443946
```

Above are the classification reports from using original MSI images on the left and from using PanColorGAN images on the right.

This classification report presents the main classification metrics on individual classes (crop types) and overall performance, including precision, recall, and f1-score metrics and the number of samples that support these metrics for each class. It also provides overall model accuracy, unweighted average of precision, recall, and f1-score from individual classes (macro avg), and weighted average of precision, recall, and f1-score values. Through the classification report, we not only get an understanding of how the model performs overall but also of individual classes.

Comparing the two classification reports from the select-region method, model crop type classification performances improve using the pan-sharpened MSI images over than using original MSI images. The weighted average of the F1 score increases by 3%, from 0.50 to 0.53. Individually, F1 scores for 4 of the 5 crop types improves except

Soybean. The weighted average of precision decreases by 10% while the recall increases by 9%. The metrics improvements are shown for all of the five crop types as well as the overall performances.

B. Using the all-region ratio-sample method

```
classification report
                            precision    recall  f1-score   support

                   Corn         0.86      0.88      0.87    851786
                 Cotton         0.75      0.87      0.81    491684
                Soybean         0.85      0.64      0.73    707507
DoubleCrop/WinterWheat_Soybean  0.86      0.85      0.86    318661
              Grassland         0.27      0.91      0.42     31871

               accuracy                             0.80   2401509
              macro avg         0.72      0.83      0.74   2401509
           weighted avg         0.82      0.80      0.81   2401509

classification report
                            precision    recall  f1-score   support

                   Corn         0.85      0.84      0.84  13787515
                 Cotton         0.73      0.87      0.79   7821722
                Soybean         0.83      0.64      0.72  11478284
DoubleCrop/WinterWheat_Soybean  0.84      0.77      0.80   5254069
              Grassland         0.24      0.90      0.38    658410

               accuracy                             0.78  39000000
              macro avg         0.70      0.80      0.71  39000000
           weighted avg         0.81      0.78      0.78  39000000
```

Again, above are the classification reports from using original MSI images on the left and from using PanColorGAN images on the right.

As expected, overall classifier performance using the all-region ratio-sample method is significantly better than that using the select-region method. The weighted F1 score increases from 0.53 to 0.78. However, as compared with the performance of using original MSI images, the weighted average F1 score decreases from 0.81 to 0.78. The overall accuracy also decreases from 0.8 to 0.78.

C. Using the all-region equal-sample method

```
classification report
                            precision    recall  f1-score   support

                   Corn         0.92      0.82      0.87    851785
                 Cotton         0.69      0.87      0.77    491685
                Soybean         0.83      0.62      0.71    707502
DoubleCrop/WinterWheat_Soybean  0.68      0.79      0.73    318662
              Grassland         0.24      0.94      0.39     31871

               accuracy                             0.77   2401505
              macro avg         0.67      0.81      0.69   2401505
           weighted avg         0.81      0.77      0.78   2401505

classification report
                            precision    recall  f1-score   support

                   Corn         0.86      0.74      0.79  13637505
                 Cotton         0.69      0.88      0.77   7865688
                Soybean         0.84      0.56      0.67  11328278
DoubleCrop/WinterWheat_Soybean  0.62      0.51      0.56   5104059
              Grassland         0.10      0.93      0.18    508431

               accuracy                             0.69  38443961
              macro avg         0.62      0.72      0.59  38443961
           weighted avg         0.77      0.69      0.71  38443961
```

Again, above are the classification reports from using original MSI images on the left and from using PanColorGAN images on the right.

The weighted average F1 score is 0.69 using the all-region equal-sample method using pan-sharpened images while the F1 score is 0.77 using original MSI images. Also, the F1 score using the all-region equal-sample method decreases from 0.78 to 0.69, about

a 9% decrease in absolute percentage as compared with that using the all-region ratio-sample method.

From the results above, the classifier performs the best using the all-region ratio-sample method. The overall accuracy is 78%.  The classification performance using pan-sharpened images is close to but not as good as that using the original low spatial resolution MSI images. The overall accuracies are 80% and 78%, respectively, from original MSI images and from pan-sharpened images.

**Results from supervised deep learning method**

In this section, we would like to compare side-by-side the results from using supervised images with that from using PanColorGAN images as we already compared results using PanColorGAN images to that using original MSI images.

A. Using the select-region method

```
classification report
                            precision   recall  f1-score   support
                  Corn         0.67       0.82     0.74    13637504
                Cotton         0.68       0.72     0.70     7865684
               Soybean         0.42       0.12     0.19    11328273
DoubleCrop/WinterWheat_Soybean 0.41       0.57     0.48     5104055
             Grassland         0.07       0.43     0.12      508430

              accuracy                             0.56    38443946
             macro avg         0.45       0.53     0.45    38443946
          weighted avg         0.56       0.56     0.53    38443946
```

```
classification report
                            precision   recall  f1-score   support
                  Corn         0.59       0.75     0.66    13636195
                Cotton         0.64       0.75     0.69     7865684
               Soybean         0.44       0.11     0.17    11328273
DoubleCrop/WinterWheat_Soybean 0.45       0.40     0.43     5104055
             Grassland         0.09       0.81     0.16      507790

              accuracy                             0.52    38441997
             macro avg         0.44       0.56     0.42    38441997
          weighted avg         0.53       0.52     0.49    38441997
```

The classification report on the left side above is from using PanColorGAN images and on the right side is from using supervised images.

Looking at the F1 scores for individual classes, supervised images perform better for the Grassland type but worse on the other four crop types. The overall weighted average of the F1 score from supervised images is 0.49, less than that from PanColorGAN images, which is 0.53. From these metrics, it seems CPanColorGAN images deliver a little better performance in distinguishing among the 5 crop types.

B. Using the all-region ratio-sample method

```
classification report
                            precision   recall  f1-score   support
                  Corn         0.85       0.84     0.84    13787515
                Cotton         0.73       0.87     0.79     7821722
               Soybean         0.83       0.64     0.72    11478284
DoubleCrop/WinterWheat_Soybean 0.84       0.77     0.80     5254069
             Grassland         0.24       0.90     0.38      658410

              accuracy                             0.78    39000000
             macro avg         0.70       0.80     0.71    39000000
          weighted avg         0.81       0.78     0.78    39000000
```

```
classification report
                            precision   recall  f1-score   support
                  Corn         0.81       0.75     0.78    13636198
                Cotton         0.62       0.71     0.66     7865686
               Soybean         0.72       0.60     0.66    11328278
DoubleCrop/WinterWheat_Soybean 0.78       0.39     0.52     5104058
             Grassland         0.09       0.91     0.17      507791

              accuracy                             0.65    38442011
             macro avg         0.61       0.67     0.56    38442011
          weighted avg         0.73       0.65     0.68    38442011
```

The classification report on the left side above is from using PanColorGAN images and on the right side is from using supervised images.

The weighted average F1 score from using PanColorGAN images is 0.78 while it is 0.68 from using supervised images. For all individual scores except the recall score for Grassland, the model performs better using PanColorGAN images than using supervised images.

## C. Using all-region equal-sample method

```
classification report
                             precision    recall  f1-score   support

                      Corn        0.86      0.74      0.79  13637505
                    Cotton        0.69      0.88      0.77   7865688
                   Soybean        0.84      0.56      0.67  11328278
DoubleCrop/WinterWheat_Soybean    0.62      0.51      0.56   5104059
                 Grassland        0.10      0.93      0.18    508431

                  accuracy                            0.69  38443961
                 macro avg        0.62      0.72      0.59  38443961
              weighted avg        0.77      0.69      0.71  38443961

classification report
                             precision    recall  f1-score   support

                      Corn        0.81      0.67      0.73  13636196
                    Cotton        0.55      0.74      0.63   7865688
                   Soybean        0.73      0.56      0.63  11328278
DoubleCrop/WinterWheat_Soybean    0.71      0.46      0.56   5104059
                 Grassland        0.10      0.92      0.18    507791

                  accuracy                            0.63  38442012
                 macro avg        0.58      0.67      0.55  38442012
              weighted avg        0.71      0.63      0.65  38442012
```

The classification report on the left side above is from using PanColorGAN images and on the right side is from using supervised images.

The weighted average F1 score from using PanColorGAN images is 0.71 while it is 0.65 from using supervised images. For all individual scores except precision score for double crops fields, the model performs better from using PanColorGAN images than from those using supervised images.

Based on these results, we can get the conclusion that using spectral bands only, PanColorGAN images perform better in crop type classification than that using supervised pan-sharpened images. But the performance using pan-sharpened images is not as good as that of using original MSI images.

**Results from GAN deep learning method**

In this section, we would like to compare side-by-side the results from using GAN images with that from using PanColorGAN images.

## A. Using the select-region method

```
classification report                                    classification report
                             precision  recall f1-score  support                                  precision  recall f1-score  support

                      Corn        0.67    0.82     0.74  13637504                          Corn        0.60    0.73     0.66  13637769
                    Cotton        0.68    0.72     0.70   7865684                        Cotton        0.64    0.75     0.69   7865684
                   Soybean        0.42    0.12     0.19  11328273                       Soybean        0.49    0.14     0.22  11328273
DoubleCrop/WinterWheat_Soybean    0.41    0.57     0.48   5104055  DoubleCrop/WinterWheat_Soybean    0.42    0.38     0.39   5104055
                 Grassland        0.07    0.43     0.12    508430                     Grassland        0.09    0.81     0.16    508270

                  accuracy                          0.56  38443946                      accuracy                        0.52  38444051
                 macro avg        0.45    0.53     0.45  38443946                     macro avg        0.45    0.56     0.42  38444051
              weighted avg        0.56    0.56     0.53  38443946                  weighted avg        0.55    0.52     0.49  38444051
```

The classification report on the left side above is from using PanColorGAN images and on the right side is from using GAN images.

Looking at the F1 scores for individual classes, GAN images perform better for Grassland and Soybean types but worse on the other three crop types. The overall weighted average of the F1 score from GAN images is 0.49, the same as that from supervised images, and is less than that from PanColorGAN images, which is 0.53.

## B. Using the all-region ratio-sample method

classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Corn | 0.85 | 0.84 | 0.84 | 13787515 |
| Cotton | 0.73 | 0.87 | 0.79 | 7821722 |
| Soybean | 0.83 | 0.64 | 0.72 | 11478284 |
| DoubleCrop/WinterWheat_Soybean | 0.84 | 0.77 | 0.80 | 5254069 |
| Grassland | 0.24 | 0.90 | 0.38 | 658410 |
| | | | | |
| accuracy | | | 0.78 | 39000000 |
| macro avg | 0.70 | 0.80 | 0.71 | 39000000 |
| weighted avg | 0.81 | 0.78 | 0.78 | 39000000 |

classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Corn | 0.84 | 0.85 | 0.85 | 13637772 |
| Cotton | 0.71 | 0.84 | 0.77 | 7865686 |
| Soybean | 0.81 | 0.61 | 0.70 | 11328278 |
| DoubleCrop/WinterWheat_Soybean | 0.83 | 0.72 | 0.77 | 5104058 |
| Grassland | 0.20 | 0.90 | 0.33 | 508271 |
| | | | | |
| accuracy | | | 0.76 | 38444065 |
| macro avg | 0.68 | 0.79 | 0.68 | 38444065 |
| weighted avg | 0.79 | 0.76 | 0.77 | 38444065 |

The classification report on the left side above is from using PanColorGAN images and on the right side is from using GAN images.

The weighted F1 score from using GAN images is 0.77, 0.01 less than that from using the PanColorGAN images, significantly better than that using supervised images, which is 0.68.

## C. Using all-region equal-sample method

classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Corn | 0.86 | 0.74 | 0.79 | 13637505 |
| Cotton | 0.69 | 0.88 | 0.77 | 7865688 |
| Soybean | 0.84 | 0.56 | 0.67 | 11328278 |
| DoubleCrop/WinterWheat_Soybean | 0.62 | 0.51 | 0.56 | 5104059 |
| Grassland | 0.10 | 0.93 | 0.18 | 508431 |
| | | | | |
| accuracy | | | 0.69 | 38443961 |
| macro avg | 0.62 | 0.72 | 0.59 | 38443961 |
| weighted avg | 0.77 | 0.69 | 0.71 | 38443961 |

classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Corn | 0.86 | 0.84 | 0.85 | 13637770 |
| Cotton | 0.69 | 0.84 | 0.76 | 7865688 |
| Soybean | 0.81 | 0.59 | 0.68 | 11328278 |
| DoubleCrop/WinterWheat_Soybean | 0.74 | 0.63 | 0.69 | 5104059 |
| Grassland | 0.17 | 0.92 | 0.28 | 508271 |
| | | | | |
| accuracy | | | 0.74 | 38444066 |
| macro avg | 0.65 | 0.77 | 0.65 | 38444066 |
| weighted avg | 0.78 | 0.74 | 0.75 | 38444066 |

The classification report on the left side above is from using PanColorGAN images and on the right side is from using GAN images.

The weighted F1 score from using GAN images is 0.75, better than that from using the PanColorGAN images, which is 0.71, and significantly better than that using supervised images, which is 0.65.

**So based on the classification results from the pan-sharpened images using three deep learning methods, the performances from using PanColorGAN images are comparable to those from using GAN images. They are both better than that using supervised images. The classifier performances are somewhat worse than that of using original MSI images, i.e. pan-sharpened images have not improved classifier performances for crop type classification**. These are the classifier results

using spectral band data only. We then added the vegetation index information to see if classification performance was improved.

## Crop type RF classification using a combination of features from original MSI images

The experiments in this section is to test model performance improvement when vegetation index features are included. There is a total of 6 vegetation index features, so the total number of features increases to 10. This section is about the experiment using features derived from original MSI images. Since we come to the conclusion from experiments in previous sections that the model performs best using the all-region ratio-sample method, in the following sections, we focus our experiments on this learning strategy.

## Results from using MSI spectral + vegetation index features, all-region ratio-sample method

| | | Testing set | | | | | | | Independent set | | | |
| | | | Predicted | | | | | | | Predicted | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 2392 | 1 | 40 | 22 | 59 | | Corn | 748843 | 4866 | 19170 | 23408 | 55499 |
| | Cotton | 22 | 2352 | 112 | 5 | 7 | | Cotton | 5474 | 427195 | 58951 | 36 | 28 |
| Actual | Soybean | 14 | 123 | 2303 | 10 | 4 | Actual | Soybean | 97431 | 125843 | 461122 | 17543 | 5568 |
| | DblCrop | 15 | 1 | 6 | 2470 | 10 | | DblCrop | 26474 | 384 | 4392 | 267718 | 19693 |
| | Grassland | 73 | 0 | 4 | 34 | 2417 | | Grassland | 1521 | 62 | 435 | 836 | 29017 |

Testing set accuracy:  0.9550256081946222

```
classification report
                                 precision    recall  f1-score   support

                          Corn       0.85      0.88      0.86    851786
                        Cotton       0.77      0.87      0.81    491684
                       Soybean       0.85      0.65      0.74    707507
DoubleCrop/WinterWheat_Soybean     0.86      0.84      0.85    318661
                     Grassland       0.26      0.91      0.41     31871

                      accuracy                           0.81   2401509
                     macro avg       0.72      0.83      0.74   2401509
                  weighted avg       0.83      0.81      0.81   2401509
```

Independent set metrics

Accuracy 0.8052832614826761
BalancedAccuracy 0.8300653745734792
AUROC 0.9470398859118007
Kappa 0.7361600371764991
MatthewCoef 0.7395028625501928
Logloss 1.4830595518599834

The overall accuracy of RF model using spectral features only is 0.8037900336829885, which is about the same as that from using spectral and vegetation index features,

which is 0.8052832614826761. The vegetation index features do not show improvement in the RF classifier performance using the original MSI images.

The feature importance ranking from the RF model is as follows.

feature importance
sipi 0.21622587492690096
blue 0.16672065657137938
gci 0.11619976404608513
savi 0.08463612407093617
sr 0.07903142731516716
red 0.07573081794186802
green 0.07524434362257881
ndvi 0.0702170520157898
nir 0.060910383482836276
arvi 0.05508355600645836

Several vegetation indexes are on top of the list such as sipi and gci while the rest of the features show relatively the same impact on the RF classifier.

### Results from using MSI vegetation index features, all-region ratio-sample method

| | | Test set | | | | | | | | Independent set | | | | |
| | | | Predicted | | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 2245 | 1 | 79 | 85 | 104 | | Corn | 735502 | 3825 | 30157 | 35847 | 46455 |
| | Cotton | 31 | 2267 | 183 | 8 | 9 | | Cotton | 1976 | 417115 | 72473 | 43 | 77 |
| Actual | Soybean | 43 | 374 | 1993 | 36 | 8 | Actual | Soybean | 77257 | 149984 | 437962 | 32871 | 9433 |
| | DblCrop | 45 | 1 | 8 | 2414 | 34 | | DblCrop | 36431 | 424 | 5484 | 255873 | 20449 |
| | Grassland | 108 | 0 | 22 | 111 | 2287 | | Grassland | 1916 | 19 | 350 | 1890 | 27696 |

Testing set accuracy: 0.8967669654289373

```
classification report
                                    precision    recall   f1-score    support

                             Corn        0.86      0.86       0.86     851786
                           Cotton        0.73      0.85       0.78     491684
                          Soybean        0.80      0.62       0.70     707507
     DoubleCrop/WinterWheat_Soybean      0.78      0.80       0.79     318661
                        Grassland        0.27      0.87       0.41      31871

                         accuracy                             0.78    2401509
                        macro avg        0.69      0.80       0.71    2401509
                     weighted avg        0.80      0.78       0.78    2401509
```

Independent set metrics

Accuracy 0.7804043207833075
BalancedAccuracy 0.8005618711717097
AUROC 0.9325982403224766
Kappa 0.7031057561194036
MatthewCoef 0.7062655602826173
Logloss 1.7226970773933241

Feature importance

sipi 0.320545582925083
gci 0.20457297053049986
ndvi 0.15675880583840524
savi 0.12773286044818036
sr 0.10771856179977221
arvi 0.08267121845805935

The overall accuracy of RF model using the vegetation features only is a couple of percents lower than that from using spectral band features only, which is 0.804. SIPI and GCI are still the top two features of importance, consistent with the results from using the spectral band + vegetation index feature scheme. The NDVI feature ranks higher though.

## Results from using MSI all features (spectral + vegetation index + texture), all-region ratio-sample method

| | | Test set | | | | | | | Independent set | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Predicted | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 2470 | 3 | 8 | 7 | 26 | | Corn | 736060 | 8131 | 12042 | 57244 | 38309 |
| | Cotton | 6 | 2459 | 23 | 1 | 9 | | Cotton | 4180 | 444340 | 43045 | 24 | 95 |
| Actual | Soybean | 3 | 43 | 2403 | 4 | 1 | Actual | Soybean | 49531 | 87713 | 511994 | 55390 | 2879 |
| | DblCrop | 5 | 1 | 2 | 2490 | 4 | | DblCrop | 41563 | 318 | 2025 | 270981 | 3774 |
| | Grassland | 11 | 0 | 0 | 4 | 2513 | | Grassland | 999 | 12 | 9 | 1378 | 29473 |

Testing set accuracy:  0.9871158770806658

```
classification report
                                   precision      recall    f1-score     support

                           Corn         0.88        0.86        0.87      851786
                         Cotton         0.82        0.90        0.86      491684
                        Soybean         0.90        0.72        0.80      707507
  DoubleCrop/WinterWheat_Soybean       0.70        0.85        0.77      318661
                      Grassland         0.40        0.92        0.55       31871

                       accuracy                                 0.83     2401509
                      macro avg         0.74        0.85        0.77     2401509
                   weighted avg         0.85        0.83        0.83     2401509
```

Independent set metrics

Accuracy 0.8298315767294647
BalancedAccuracy 0.8533280229218636
AUROC 0.9650328979910802
Kappa 0.7698827474169325
MatthewCoef 0.7724930063092298
Logloss 0.6246185565591569

feature importance

sipi 0.17153523376840848
blue 0.12592461972475263
gci 0.08260824610322996
ndvi 0.07036808914627055
savi 0.06872145158232129
green 0.060774958805092046
sr 0.05970585747263723
GLCM_Variance 0.05622869448432492
GLCM_Mean 0.056213628257675015
arvi 0.05337219856627159
red 0.05136384949399235
nir 0.046690003509700484
GLCM_Correlation 0.028114749351384652
GLCM_contrast 0.018152182878802856
GLCM_dissimilarity 0.012873237763203125
GLCM_homogeneity 0.01140896544327458
GLCM_entropy 0.01009662840006263
GLCM_ASM 0.006880370058932887
GLCM_energy 0.005619758322849141
GLCM_MAX 0.0033472768668133977

Using all features (spectra l+ vegetation index + textures) scheme, the RF model overall accuracy is 0.83, about 2.5% higher than that using spectral + vegetation index scheme, which is 0.805. So texture does contribute to the improvement of the RF classifier.

**Results from using MSI texture features, all-region ratio-sample method**

| | | Test set | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 1416 | 180 | 149 | 354 | 415 | | Corn | 352838 | 53205 | 22707 | 237601 | 185435 |
| | Cotton | 209 | 1609 | 595 | 36 | 49 | | Cotton | 33370 | 299877 | 134590 | 13258 | 10589 |
| Actual | Soybean | 130 | 557 | 1597 | 65 | 105 | Actual | Soybean | 82448 | 173046 | 316944 | 75958 | 59111 |
| | DblCrop | 300 | 9 | 16 | 1921 | 256 | | DblCrop | 82875 | 6803 | 7358 | 156467 | 65158 |
| | Grassland | 280 | 14 | 47 | 253 | 1934 | | Grassland | 7231 | 618 | 831 | 2537 | 20654 |

Testing set accuracy:  0.678377080665813

```
classification report
                                precision    recall   f1-score    support

                         Corn        0.63      0.41       0.50     851786
                       Cotton        0.56      0.61       0.58     491684
                      Soybean        0.66      0.45       0.53     707507
DoubleCrop/WinterWheat_Soybean      0.32      0.49       0.39     318661
                    Grassland        0.06      0.65       0.11      31871

                     accuracy                             0.48    2401509
                    macro avg        0.45      0.52       0.42    2401509
                 weighted avg        0.58      0.48       0.51    2401509
```

Independent set metrics

Accuracy 0.47752475630947044
BalancedAccuracy 0.522233567767892
AUROC 0.7925183875364817
Kappa 0.333642448917542
MatthewCoef 0.34392312403647307
Logloss 3.3436229880335193

Feature importance

GLCM_Mean 0.25070542750680685
GLCM_Variance 0.24079295650766627
GLCM_Correlation 0.12105231587381532
GLCM_contrast 0.06987984319606184
GLCM_entropy 0.06620396737608636
GLCM_homogeneity 0.05863947442767483
GLCM_dissimilarity 0.05159693959832221
GLCM_ASM 0.051000274051327815
GLCM_energy 0.05056395018383
GLCM_MAX 0.039564851278408406

The overall accuracy for the testing set and independent set are 0.678 and 0.478, respectively, significantly worse than those using spectral and/or vegetation index features. This result suggests that texture features alone are not good to distinguish different crop types. However, adding texture features to spectral and vegetation index features does improve classifier performance, while adding vegetation index features to spectral features does not show improvement. This may be because texture features add some information for the classifier to use while vegetation index features are highly correlated to spectral features so not much classifier improvement is observed.

### 3.3.5 Comparing results with different feature sample volumes from using MSI texture features, all-region ratio-sample method

Experiments in this section intend to examine the impact of feature volume size on classifier learning and performance. Three volume sizes are tested: 25000 samples, 100000 samples, and 200000 samples. Only the classification reports are shown which provide insight into the RF classifier performance on an independent set. All features (spectral + vegetation index + texture) are included in the model training.

A feature set with 25000 samples for learning

```
classification report
                                  precision    recall  f1-score   support

                          Corn         0.89      0.82      0.86    856785
                        Cotton         0.83      0.91      0.87    496685
                       Soybean         0.89      0.74      0.81    712507
 DoubleCrop/WinterWheat_Soybean         0.67      0.85      0.75    323661
                     Grassland         0.43      0.92      0.58     36871

                      accuracy                             0.82   2426509
                     macro avg         0.74      0.85      0.77   2426509
                  weighted avg         0.84      0.82      0.83   2426509
```

Feature set with 100000 samples for learning

```
classification report
                          precision    recall  f1-score   support

                   Corn       0.91      0.88      0.89    841785
                 Cotton       0.83      0.91      0.87    481685
                Soybean       0.91      0.76      0.83    697507
DoubleCrop/WinterWheat_Soybean   0.76      0.88      0.81    308661
              Grassland       0.34      0.93      0.49     21871

               accuracy                           0.85   2351509
              macro avg       0.75      0.87      0.78   2351509
           weighted avg       0.87      0.85      0.85   2351509
```

Feature set with 200000 samples for learning

```
classification report
                          precision    recall  f1-score   support

                   Corn       0.94      0.84      0.89    821785
                 Cotton       0.85      0.92      0.88    461684
                Soybean       0.91      0.80      0.85    677507
DoubleCrop/WinterWheat_Soybean   0.73      0.92      0.81    288661
              Grassland       0.03      0.99      0.06      1871

               accuracy                           0.86   2251508
              macro avg       0.69      0.89      0.70   2251508
           weighted avg       0.88      0.86      0.87   2251508
```

From these classification reports, we see that RF overall accuracy increases from 0.82 to 0.85 when the training sample size increases from 25000 to 100000, and from 0.85 to 0.86 when sampling size increases to 200000. Increasing sample size seems to improve performance as more representative samples might be available for the classifier. Further doubling the sample size to 200000 does not improve overall accuracy a lot. So 100000 samples might be a good size for RF classifier training. The overall accuracy of 0.86 is the highest.

**Performances of other supervised learning algorithms from using MSI all features, all-region ratio-sample method**

**KNN method with all features**

KNN classifier is examined with k from 1 to 20. The overall accuracies for these k values are within 75% - 77%. The highest score is 0.7705 when k = 2. All features are used in the classifier. Here is the classification report for k = 2

```
classification report
                                precision    recall  f1-score   support

                       Corn          0.84      0.79      0.81    851786
                     Cotton          0.73      0.90      0.80    491684
                    Soybean          0.85      0.65      0.74    707507
DoubleCrop/WinterWheat_Soybean       0.65      0.78      0.71    318661
                  Grassland          0.39      0.90      0.55     31871

                   accuracy                              0.77   2401509
                  macro avg          0.69      0.80      0.72   2401509
               weighted avg          0.79      0.77      0.77   2401509
```

Experiments were also run on KNN classifiers with spectral only features and spectral + vegetation index features.

An experiment was also done by forcing training samples taken proportional to the number of samples in a region for a crop type instead of randomly splitting samples into training and test sample sets. We have a little concern that the random splitting train sample set may leave some regions less representative by selecting few if no samples in the training set for model learning. Here is the classification report.

```
classification report
                                precision    recall  f1-score   support

                       Corn          0.84      0.79      0.81    851786
                     Cotton          0.72      0.89      0.80    491684
                    Soybean          0.85      0.65      0.73    707507
DoubleCrop/WinterWheat_Soybean       0.65      0.77      0.71    318661
                  Grassland          0.39      0.90      0.54     31871

                   accuracy                              0.77   2401509
                  macro avg          0.69      0.80      0.72   2401509
               weighted avg          0.79      0.77      0.77   2401509
```

The classifier with k=2 delivers the best overall accuracy, which is 0.766, slightly more than that of the random splitting scheme as shown above. So approach to split sample set might not be a concern at all.

In the experiments above, features are not normalized. For KNN classifiers, features are expected to be normalized. We also examine the impact of the normalization process on the KNN classifiers. The classification report is shown below.

```
classification report
                                precision    recall  f1-score   support

                       Corn          0.87      0.77      0.82    851786
                     Cotton          0.79      0.91      0.85    491684
                    Soybean          0.85      0.72      0.78    707507
DoubleCrop/WinterWheat_Soybean       0.61      0.79      0.69    318661
                  Grassland          0.40      0.88      0.55     31871

                   accuracy                              0.79   2401509
                  macro avg          0.70      0.81      0.74   2401509
               weighted avg          0.81      0.79      0.79   2401509
```

The classifier with k=2 delivers the best overall accuracy, which is 0.788, an improvement of about 2% as compared with the results without feature normalization.

**KNN with spectral features only**

```
classification report
                              precision    recall  f1-score   support

                   Corn           0.88      0.87      0.87    851786
                 Cotton           0.78      0.86      0.82    491684
                Soybean           0.83      0.68      0.75    707507
DoubleCrop/WinterWheat_Soybean    0.81      0.87      0.84    318661
              Grassland           0.29      0.90      0.44     31871

               accuracy                               0.81   2401509
              macro avg           0.72      0.83      0.74   2401509
           weighted avg           0.83      0.81      0.81   2401509
```

The best overall accuracy is achieved when k = 9, the score is 0.81

**KNN with spectral + vegetation index features**

```
classification report
                              precision    recall  f1-score   support

                   Corn           0.87      0.87      0.87    851786
                 Cotton           0.77      0.87      0.82    491684
                Soybean           0.84      0.68      0.75    707507
DoubleCrop/WinterWheat_Soybean    0.82      0.86      0.84    318661
              Grassland           0.30      0.89      0.45     31871

               accuracy                               0.81   2401509
              macro avg           0.72      0.83      0.75   2401509
           weighted avg           0.83      0.81      0.82   2401509
```

The best overall accuracy is achieved when k = 10, the score is 0.8.

So based on the experiment runs on three different feature sets, the one using spectral features and the one using both spectral and vegetation index features performs about the same with an overall accuracy of 0.81. However, the overall accuracy drops to 0.77 when all features (i.e. including texture features) are used.

**Crop type classification using a combination of features using pan-sharpened MSI images**

**Results from using PanColorGAN images**

**RF classifier using spectral + vegetation index features**

<table>
<tr><td colspan="2"></td><td colspan="5">Testing set</td></tr>
</table>

| | | Corn | Cotton | Soybean | DblCrop | Grassland |
|---|---|---|---|---|---|---|
| | | | | Predicted | | |
| | Corn | 38353 | 59 | 437 | 244 | 868 |
| | Cotton | 240 | 37645 | 2140 | 86 | 143 |
| Actual | Soybean | 251 | 3089 | 36311 | 76 | 41 |
| | DblCrop | 225 | 23 | 115 | 39014 | 531 |
| | Grassland | 858 | 8 | 125 | 598 | 38517 |

| | | Corn | Cotton | Soybean | DblCrop | Grassland |
|---|---|---|---|---|---|---|
| | | | | Independent set | | |
| | | | | Predicted | | |
| | Corn | 12023625 | 153645 | 343475 | 383272 | 733490 |
| | Cotton | 46151 | 6950171 | 863976 | 606 | 4782 |
| Actual | Soybean | 1215768 | 2006704 | 7449527 | 507649 | 148630 |
| | DblCrop | 358534 | 14710 | 78865 | 4308132 | 343817 |
| | Grassland | 22872 | 2160 | 5393 | 12009 | 465997 |

```
classification report
                                        precision    recall  f1-score   support

                                 Corn        0.88      0.88      0.88  13637507
                               Cotton        0.76      0.88      0.82   7865686
                              Soybean        0.85      0.66      0.74  11328278
        DoubleCrop/WinterWheat_Soybean      0.83      0.84      0.84   5104058
                            Grassland        0.27      0.92      0.42    508431

                             accuracy                            0.81  38443960
                            macro avg        0.72      0.84      0.74  38443960
                         weighted avg        0.83      0.81      0.82  38443960
```

With spectral + vegetation index features used, the overall accuracy from PanColorGAN data is 0.81, about the same as that from the original MSI images, which is also 0.81 (precision to two digits).

## RF classifier using all features (spectral + vegetation index + texture) features

| | | Corn | Cotton | Soybean | DblCrop | Grassland |
|---|---|---|---|---|---|---|
| | | | | Testing set | | |
| | | | | Predicted | | |
| | Corn | 39382 | 21 | 174 | 45 | 339 |
| | Cotton | 47 | 39713 | 348 | 11 | 135 |
| Actual | Soybean | 74 | 635 | 39034 | 2 | 23 |
| | DblCrop | 127 | 8 | 29 | 39585 | 159 |
| | Grassland | 140 | 7 | 13 | 46 | 39900 |

| | | Corn | Cotton | Soybean | DblCrop | Grassland |
|---|---|---|---|---|---|---|
| | | | | Independent set | | |
| | | | | Predicted | | |
| | Corn | 11884350 | 88178 | 391559 | 471711 | 801709 |
| | Cotton | 34307 | 6948766 | 876508 | 2891 | 3214 |
| Actual | Soybean | 916385 | 1762629 | 8062449 | 502574 | 84241 |
| | DblCrop | 526002 | 4084 | 48030 | 4480103 | 45839 |
| | Grassland | 14820 | 766 | 954 | 15725 | 476166 |

Classification report

```
classification report
                                        precision    recall  f1-score   support

                                 Corn        0.89      0.87      0.88  13637507
                               Cotton        0.79      0.88      0.83   7865686
                              Soybean        0.86      0.71      0.78  11328278
        DoubleCrop/WinterWheat_Soybean      0.82      0.88      0.85   5104058
                            Grassland        0.34      0.94      0.50    508431

                             accuracy                            0.83  38443960
                            macro avg        0.74      0.86      0.77  38443960
                         weighted avg        0.84      0.83      0.83  38443960
```

With all features being used, the overall accuracy from PanColorGAN data is 0.83, about the same as that from original MSI images, which is also 0.83 (precision to two digits).

**KNN classifier using spectral features only**

```
classification report
                                precision    recall   f1-score    support

                         Corn        0.89      0.88       0.88   13637507
                       Cotton        0.76      0.87       0.82    7865686
                      Soybean        0.84      0.68       0.75   11328278
DoubleCrop/WinterWheat_Soybean       0.83      0.87       0.85    5104058
                    Grassland        0.31      0.91       0.47     508431

                     accuracy                             0.82   38443960
                    macro avg        0.73      0.84       0.75   38443960
                 weighted avg        0.83      0.82       0.82   38443960
```

The optimal KNN classifier is when k = 9 (k is tested from 1 to 9), which is 0.818, about the same as that from using spectral features from original MSI images, which is 0.81.

**KNN classifier using spectral + vegetation index features**

```
classification report
                                precision    recall   f1-score    support

                         Corn        0.89      0.88       0.88   13637507
                       Cotton        0.76      0.87       0.82    7865686
                      Soybean        0.84      0.68       0.75   11328278
DoubleCrop/WinterWheat_Soybean       0.83      0.87       0.85    5104058
                    Grassland        0.31      0.91       0.47     508431

                     accuracy                             0.82   38443960
                    macro avg        0.73      0.84       0.75   38443960
                 weighted avg        0.83      0.82       0.82   38443960
```

The optimal KNN classifier is when k = 9 (k is tested from 1 to 9), which is 0.818, the same as that from using spectral only features of PanColorGAN images.

**Results from using supervised pan-sharpened MSI images**

**RF classifier using spectral + vegetation index features**

| | | Testing set | | | | | | | Independent set | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | Predicted | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 38473 | 40 | 204 | 417 | 827 | | Corn | 11605601 | 207922 | 305297 | 476909 | 1040469 |
| | Cotton | 313 | 38279 | 1498 | 22 | 142 | | Cotton | 49649 | 6637216 | 1178358 | 385 | 78 |
| Actual | Soybean | 72 | 1584 | 37848 | 173 | 91 | Actual | Soybean | 1566491 | 2301906 | 7069331 | 208567 | 181983 |
| | DblCrop | 352 | 6 | 145 | 38790 | 615 | | DblCrop | 602346 | 53552 | 100161 | 3754668 | 593331 |
| | Grassland | 1049 | 34 | 249 | 907 | 37867 | | Grassland | 27286 | 571 | 5027 | 19255 | 455652 |

```
classification report
                                   precision    recall   f1-score    support

                         Corn         0.84       0.85       0.84    13636198
                       Cotton         0.72       0.84       0.78     7865686
                      Soybean         0.82       0.62       0.71    11328278
DoubleCrop/WinterWheat_Soybean        0.84       0.74       0.79     5104058
                    Grassland         0.20       0.90       0.33      507791

                     accuracy                               0.77    38442011
                    macro avg         0.68       0.79       0.69    38442011
                 weighted avg         0.80       0.77       0.78    38442011
```

The overall accuracy of RF classifier using spectral + vegetation features from the supervised method is 0.77, less than that from using PanColorGAN images, which is 0.81.

## RF classifier using all features

```
classification report
                                   precision    recall   f1-score    support

                         Corn         0.86       0.90       0.88    13636198
                       Cotton         0.78       0.81       0.80     7865686
                      Soybean         0.81       0.70       0.75    11328278
DoubleCrop/WinterWheat_Soybean        0.83       0.84       0.83     5104058
                    Grassland         0.43       0.98       0.60      507791

                     accuracy                               0.81    38442011
                    macro avg         0.74       0.84       0.77    38442011
                 weighted avg         0.82       0.81       0.81    38442011
```

The overall accuracy of the RF classifier using all features from supervised images is 0.81, increased by 0.04 as compared with that without using texture features. The overall accuracy of using all features derived from original MSI images is 0.83.

Based on the overall accuracy measure using the RF classifier, the images using PanColorGAN images deliver a little better result.

## KNN classifier using spectral features only

```
classification report
                                   precision    recall   f1-score    support

                         Corn         0.84       0.85       0.85    13636198
                       Cotton         0.73       0.84       0.78     7865686
                      Soybean         0.81       0.64       0.71    11328278
DoubleCrop/WinterWheat_Soybean        0.83       0.78       0.80     5104058
                    Grassland         0.24       0.88       0.37      507791

                     accuracy                               0.78    38442011
                    macro avg         0.69       0.80       0.70    38442011
                 weighted avg         0.80       0.78       0.78    38442011
```

The overall accuracy of the KNN classifier from the supervised spectral feature method is 0.78, less than that from using PanColorGAN images, which is 0.81.

## Results from using GAN pan-sharpened MSI images

### RF classifier using spectral + vegetation index features

```
classification report
                               precision    recall   f1-score    support

                        Corn        0.84      0.85       0.85   13637772
                      Cotton        0.71      0.84       0.77    7865686
                     Soybean        0.81      0.61       0.70   11328278
DoubleCrop/WinterWheat_Soybean      0.83      0.72       0.77    5104058
                   Grassland        0.21      0.90       0.33     508271

                    accuracy                             0.76   38444065
                   macro avg        0.68      0.79       0.68   38444065
                weighted avg        0.79      0.76       0.77   38444065
```

The overall accuracy of RF classifier using spectral + vegetation features from GAN method is 0.76, less than that from using PanColorGAN images, which is 0.81, and from using the supervised method, which is 0.77.

### RF classifier using all features

```
classification report
                               precision    recall   f1-score    support

                        Corn        0.86      0.87       0.86   13637772
                      Cotton        0.84      0.82       0.83    7865686
                     Soybean        0.84      0.75       0.79   11328278
DoubleCrop/WinterWheat_Soybean      0.82      0.86       0.84    5104058
                   Grassland        0.32      0.92       0.47     508271

                    accuracy                             0.82   38444065
                   macro avg        0.73      0.84       0.76   38444065
                weighted avg        0.84      0.82       0.83   38444065
```

The overall accuracy of the RF classifier using all features from GAN images is 0.82, increased by 0.06 as compared with that without using texture features. The overall accuracy of using all features derived from original MSI images is 0.83. Using GAN all features, RF models deliver slightly better overall accuracy than that from using the supervised method, which is 0.81.

Based on the overall accuracy measure using the RF classifier, the images using PanColorGAN images deliver the best result, as compared to that using supervised or basic GAN models.

### KNN classifier using spectral features only

```
classification report
                              precision    recall  f1-score   support

                   Corn           0.85      0.86      0.85  13637772
                 Cotton           0.73      0.84      0.78   7865686
                Soybean           0.81      0.63      0.71  11328278
DoubleCrop/WinterWheat_Soybean    0.82      0.77      0.79   5104058
              Grassland           0.24      0.89      0.38    508271

               accuracy                               0.78  38444065
              macro avg           0.69      0.80      0.70  38444065
           weighted avg           0.80      0.78      0.78  38444065
```

The overall accuracy of the KNN classifier from the GAN spectral feature method is 0.78, the same as that using the supervised method, less than that from using PanColorGAN images, which is 0.81.

### Summary of experiments on multispectral imagery

Generally two groups of experiments have been done. The first group focused on the investigation of crop type classification using original multispectral images (MSI images). These results let us know what we can expect from classification performance using 'low resolution' MSI images. The second group focused on comparisons of classifier performances using original MSI images and pan-sharpened images from using PanColorGAN, supervised, and GAN methods. This tells us if pan-sharpened images can enhance crop type classification performances.

### Comparison of experiment results from using original MSI images

| Classifier | Method | Features | | | | |
|---|---|---|---|---|---|---|
| | | Spectral Only | Index only | Texture only | Spectral+Index | All features |
| | select region | (0.47, 0.50) | | | | |
| | select region grid-search | (0.47, 0.50) | | | | |
| | all region ratio sample | (0.80, 0.81) | (0.78, 0.78) | (0.48, 0.51) | (0.81, 0.81) | (0.83, 0.83) |
| | all region ratio sample grid-search | (0.80, 0.81) | | | | (0.83, 0.83) |
| RF | all region equal sample | (0.77, 0.78) | | | | |
| | all region equal sample grid-search | (0.77, 0.78) | | | | |
| | all region ratio sample 25K | | | | | (0.82, 0.83) |
| | all region ratio sample 100K | | | | | (0.85, 0.85) |
| | all region ratio sample 200K | | | | | (0.86, 0.87) |
| | all region ratio sample | (0.81, 0.81) | | | (0.81, 0.81) | (0.77, 0.77) |
| KNN | all region ratio sample ratio contribution | | | | | (0.77, 0.77) |
| | all region ratio sample ration contribution normalization | | | | | (0.79, 0.79) |
| | all region ratio sample 25K | (0.80, 0.80) | | | | |
| SVM | all region ratio sample linear kernel | | | | | (0.80, 0.80) |
| | all region ratio sample rbf kernel | | | | | (0.83, 0.83) |

Results of various experiments on crop type classification using MSI images

This figure shows the classifier performance for the experiments on MSI images. For each experiment, the first value is the overall accuracy, and the second one is the

weighted F1 score. Though the sample distribution is somewhat imbalanced (the grassland type counts significantly less than the other four crop types), the overall accuracy and F1 score are very consistent.

To summarize, three classifiers are examined in these experiments: Random Forest (RF), K nearest neighbor (KNN), and Support Vector Machine (SVM). The RF classifier is the major classifier in the experiments.

There are 20 features extracted from the optical images. Four features are directly from the four spectral bands. There are also six vegetation index features and 10 GLCM texture features calculated. In the spreadsheet, spectral only indicates that 4 spectral features are used for classification. Index only corresponds to 6 vegetation index features. Texture only corresponds to 10 texture features. Spectral+index indicates that a combination of 4 spectral features and 6 vegetation index features was used. All feature indicates that all the 20 features are used in the experiment.

There are a number of experiments conducted using different combinations of classifiers with training sample generation strategies. For the select region method, the training samples come from 5 manually selected crop regions, with each one representing one crop type. For all region equal sample methods, training samples come from all the crop regions identified over the Tennessee Valley, with each region contributing an equal number of samples to the dataset. For all region ratio sample methods, training samples come from all the crop regions, with samples from each region proportional to the region size. For all these three methods, the total number of samples in the training set is 50000. For most of the experiments, the classifiers (RF and SVM) are used with default parameter settings.

To test the impact of a number of training samples on classifier performance, three additional experiments were done for the RF classifier with a number of samples of 25K, 100K, and 200K, as shown in the spreadsheet. An experiment was done for the KNN classifier with a number of samples of 25K. For the SVM classifier, two experiments were done using two kernels: linear kernel and Radial Basis Function (RBF), with default model parameters.

Results of RF experiments using the three training set sample selection strategies show that the select region method delivers the worst performance with an overall accuracy of only 0.47. All region methods show significantly superior performances with an overall accuracy of 0.77 and 0.80, respectively for ratio-sample and equal-sample methods. These results suggest that in order to obtain high crop type classification results, samples from every crop field should be selected.

As for features used in the classifier, the RF classifier with texture features only has the worst performance with an accuracy of 0.48. The overall accuracy improves to 0.78 for using vegetation index features only. Using spectral features only and combinations of spectral, index, and texture features, overall accuracies range from 0.80 to 0.83 with the highest accuracy from using all features.

RF classifier accuracy increases from 0.82 to 0.86 when the number of training samples increases from 25K to 200K. It seems that the improvement in accuracy might be saturated with increasing training samples.

RF classifier optimization using grid-search with 5 fold cross-validation did not show improvement over using default model parameters.

The overall accuracy of the KNN classifier using spectral features only is the same as that of the RF classifier, which is 0.81. A combination of spectral and index features (a total of 10) delivers the same performance. However, KNN accuracy drops to 0.77 when all features are used. Is this due to the curse of dimensionality? KNN accuracy decreases slightly from 0.81 to 0.80 when the number of training samples decreases from 50K to 25K.

For the KNN classifier, two additional experiments were done to check the impact of alternative feature preparation processes. The first one, called 'all-region ratio sample ratio contribution', differs from the 'all-region ratio sample' in that in the former case training sample set is forced to include the number of samples for a region that is proportional to its size in the training set while in the latter case the feature set is randomly split into training sample set and testing sample set. The former guarantees the number of samples being used to train the classifier for a region. From the spreadsheet, the two strategies made no difference, most probably suggesting that the random splitting strategy did a good job in selecting the training sample set.

The second one, 'all-region ratio sample ratio contribution normalization', goes one further step in normalizing the features using the z-score method so that all features have zero means and 1 standard deviation (i.e. all features have the same dynamic range in value). For the RF classifier, this feature normalization may not be a big issue. For KNN and SVM classifiers, feature normalization is expected to have an impact. From the spreadsheet, with normalization, KNN accuracy using all features increases from 0.77 to 0.79.

For the SVM classifier, the overall accuracies are 0.80 and 0.83, respectively for linear kernel and RBF kernel. The performance of the SVM RBF model is the same as that of the RF classifier.

Through these experiments, we got the knowledge of what we can get for crop type classification using the original MSI images.

## Comparison of classifier performance from using original MSI images and pan-sharpened images

| Metrics | Method | Features | | |
|---|---|---|---|---|
| | | Spectral Only | Spectral+Index | All Features |
| Accuracy | MSI | 0.8 | 0.81 | 0.83 |
| | Charles | 0.78 | 0.81 | 0.83 |
| | Evans1 | 0.65 | 0.77 | 0.81 |
| | Evans2 | 0.76 | 0.76 | 0.82 |
| Weighted F1 | MSI | 0.81 | 0.81 | 0.83 |
| | Charles | 0.78 | 0.82 | 0.83 |
| | Evans1 | 0.68 | 0.78 | 0.81 |
| | Evans2 | 0.77 | 0.77 | 0.83 |
| Kappa | MSI | 0.73 | 0.74 | 0.77 |
| | Charles | 0.7 | 0.75 | 0.77 |
| | Evans1 | 0.54 | 0.69 | 0.74 |
| | Evans2 | 0.68 | 0.68 | 0.76 |
| MCC | MSI | 0.74 | 0.74 | 0.77 |
| | Charles | 0.7 | 0.75 | 0.77 |
| | Evans1 | 0.54 | 0.69 | 0.75 |
| | Evans2 | 0.68 | 0.68 | 0.76 |

RF classifier performances from using original MSI images and pan-sharpened images using PanColorGAN (Charles), supervised (Evans1) and basic GAN (Evans2) methods

RF classifier results from using original MSI images and pan-sharpened images using PanColorGAN, supervised and GAN methods are shown in Figure 4.2.1. Four performance metrics are shown for RF classifiers using each of the four image sets. They are overall accuracy, weighted F1 score, Kappa coefficient, and Matthew correlation coefficient.

Using the spectral features only, the best metrics are obtained from using original MSI images, and the second-best is from using PanColorGAN images. The metrics from using Evans methods are the lowest among these classifiers.

When vegetation index features are used, the metrics from using PanColorGAN images slightly outperform those from using original MSI images, while metrics from using supervised images improve the most. When using all features, the metrics from all four image sources are close to each other. And the models perform the best using all features. There is no advantage of classification from using the pan-sharpened images as compared to that using original MSI images though.

## Classification performance as shown by regions

In this section, we show the results from classifier performance based on crop region using RF classifier.

| RegionID | CropType | pixelCount | Select | Select-opt | All-ratio | All-ratio-opt | All-equal | All-equal-opt |
|---|---|---|---|---|---|---|---|---|
| | | | | | Features: Spectral only  Classifier: RF  Accuracy: Recall | | | |
| 1 | Corn | 124852 | 0.93858183 | 0.939522168 | 0.874801465 | 0.842979158 | 0.853075577 | 0.861307363 |
| 100 | Cotton | 5590 | 0.06940966 | 0.067084079 | 0.616535864 | 0.509582041 | 0.909785933 | 0.912589195 |
| 128 | Soybean | 36036 | 0.35228146 | 0.34774927 | 0.252546857 | 0.267405865 | 0.322657785 | 0.319640163 |
| 129 | Soybean | 40432 | 0.16449842 | 0.151117926 | 0.838144925 | 0.828413053 | 0.871163626 | 0.83758824 |
| 131 | Soybean | 96655 | 0.15492215 | 0.11910403 | 0.373966509 | 0.419608008 | 0.399995822 | 0.410962338 |
| 132 | Soybean | 74830 | 0.05259922 | 0.053735133 | 0.91958042 | 0.929554941 | 0.902057602 | 0.904709081 |
| 133 | Soybean | 158550 | 0.08020183 | 0.085115106 | 0.478294241 | 0.486411114 | 0.439987059 | 0.444262597 |
| 134 | Soybean | 51086 | 0 | 0 | 0.586898263 | 0.516744417 | 0.56717221 | 0.568726707 |
| 14 | Corn | 7280 | 0.41840659 | 0.39739011 | 0.747359644 | 0.763757643 | 0.831852358 | 0.838174983 |
| 141 | Soybean | 4060 | 0.02783251 | 0.039408867 | 0.884865135 | 0.869130869 | 0.890828308 | 0.91875595 |
| 142 | Soybean | 70057 | 0.69761765 | 0.679318269 | 0.643158032 | 0.699526643 | 0.615404639 | 0.582402962 |
| 143 | Soybean | 94809 | 0.62292609 | 0.518400152 | 0.878561954 | 0.886113726 | 0.811874334 | 0.802183174 |
| 144 | Soybean | 45822 | 0.00349177 | 0.004168303 | 0.926257082 | 0.921454497 | 0.915703694 | 0.901253535 |
| 157 | Soybean | 45164 | 0.65751483 | 0.68162696 | 0.661681823 | 0.681755922 | 0.622212179 | 0.625059315 |
| 2 | Corn | 360378 | 0.40695325 | 0.424246208 | 0.961647628 | 0.947927692 | 0.833781864 | 0.834698426 |
| 210 | DoubleCrop | 136554 | 0.06664763 | 0.069921057 | 0.796193353 | 0.827817221 | 0.630463792 | 0.618149872 |
| 215 | DoubleCrop | 19138 | 0.99365288 | 0.993324579 | 0.790148739 | 0.769454624 | 0.897378663 | 0.900068681 |
| 216 | DoubleCrop | 47432 | 0.97236043 | 0.975459605 | 0.896083846 | 0.890843463 | 0.929095835 | 0.929095835 |
| 233 | Grassland | 5362 | 0.25233122 | 0.238530399 | 0.545565899 | 0.542871142 | 0.633316905 | 0.638245441 |
| 235 | DoubleCrop | 75978 | 0.92035852 | 0.921687857 | 0.900552486 | 0.911656508 | 0.9133249 | 0.907080956 |
| 237 | DoubleCrop | 18148 | 0.91525237 | 0.916519727 | 0.796658331 | 0.838429188 | 0.787525786 | 0.779395704 |
| 24 | Corn | 83244 | 0.57365095 | 0.57726683 | 0.958944567 | 0.956513813 | 0.948249731 | 0.947565268 |
| 240 | DoubleCrop | 31408 | 0.9975484 | 0.997612073 | 0.949463107 | 0.962138377 | 0.964259297 | 0.964023939 |
| 249 | Grassland | 12488 | 0.99677939 | 0.997584541 | 0.96958535 | 0.959692696 | 0.962736313 | 0.964375478 |
| 254 | Grassland | 24024 | 0.24821012 | 0.264152514 | 0.960518401 | 0.955432821 | 0.956212846 | 0.956309507 |
| 3 | Corn | 108486 | 0.05941781 | 0.055334329 | 0.88999142 | 0.891698064 | 0.91658727 | 0.916185619 |
| 37 | Cotton | 188595 | 0.95744562 | 0.958778241 | 0.93178277 | 0.916509771 | 0.922253904 | 0.922398344 |
| 38 | Cotton | 132935 | 0.43722872 | 0.4432467 | 0.809334848 | 0.812458745 | 0.762396301 | 0.759455774 |
| 39 | Cotton | 58870 | 0.91319857 | 0.905622558 | 0.943463265 | 0.932578817 | 0.939269981 | 0.941245367 |
| 4 | Corn | 120887 | 0.24228412 | 0.237411798 | 0.680252751 | 0.683700883 | 0.613984714 | 0.612210047 |
| 41 | Cotton | 55258 | 0.87270621 | 0.87190995 | 0.756762745 | 0.745886958 | 0.866547246 | 0.848969996 |
| 43 | Cotton | 60433 | 0.90763325 | 0.903695001 | 0.876918401 | 0.838389978 | 0.864617898 | 0.876733541 |
| 5 | Corn | 56693 | 0.45356021 | 0.464645038 | 0.700001786 | 0.676393378 | 0.714862565 | 0.708017963 |
| | | 2451534 | | | | | | |

Selected regions: 1, 37, 128, 215, 249

The figure above shows the classifier results by region using the three training set strategies. The first one constructs a training set from samples of selected regions. The selected regions are #1, #37, #128, #215 and #249. These regions are shown in the row of gray shades with region ID highlighted in yellow. The second and the third ones are the 'all-region ratio sample' method and the 'all-region equal sample' method. For each method, two columns of results are shown, one corresponding to the default RF model, and the other one as 'opt', referring to RF model parameter optimization. The number of samples in each region is given in column 3. The number shown in the results are recall values for each region, that is, the ratio of the pixels correctly predicted by the RF classifier for a region to the total pixels in the region.

From the figure above, it is seen that recall values with the select region method are higher than that using the all-region method, as expected as these regions are well represented in the RF classifier through the selected samples. And for the rest of the regions, on the contrary, the recall values are lower using the select region method than those using all region methods. Using all-region methods, the recall values are significantly improved for the regions that show poor performance in the select region method. For region 134 (Soybean crop type), the recall values improve from 0 to more than 50%. This figure justifies the superior performance using the training samples from all regions.

For all of the experiment results shown at this point, the samples used to train the classifier to count a very small percentage of the total samples in these regions (a total of 33 regions). The total number of samples in these regions from original MSI images is over 2.45 million. Our default sample size for model training is 50K, about 2% of the total samples. For the experiment with 200K samples, it is still a little over 8%. The main consideration of using a small percentage of samples is for speeding up the model training process, especially for experiments using pan-sharpened images. A huge training set could require significant model training time. Besides, 50000 samples are not a small number for the model to learn, as long as they are representative. However, experiments do show improved classifier performance with enlarging the training sample set.

As extended efforts, experiments were conducted with all pixels from the crop regions taken into a training set for model learning using original MSI images. Again a number of experiments were done using a combination of feature selection schemes (spectral only, vegetation index only, texture only, and combinations) to examine the RF model performance with a full sample set. For these experiments, the training sample set is randomly split into two 50%-50% sections, one for model learning, the other for model testing. The following figure shows the results.

| RegionID | CropType | Features | | | | |
|---|---|---|---|---|---|---|
| | | Spectral | Index | Texture | Spectral+index | All |
| 1 | 1 | 98.90751 | 98.58152 | 92.84993 | 98.92833 | 99.66681 |
| 2 | 1 | 98.79238 | 98.39807 | 92.62275 | 98.80154 | 99.55186 |
| 3 | 1 | 97.99329 | 95.32013 | 95.53859 | 98.02832 | 99.57045 |
| 4 | 1 | 91.545 | 84.15297 | 86.62884 | 91.43911 | 98.42249 |
| 5 | 1 | 95.83613 | 93.63505 | 77.4632 | 95.9032 | 99.48636 |
| 14 | 1 | 98.15934 | 95.02747 | 85.53571 | 98.18681 | 98.75 |
| 24 | 1 | 99.17712 | 99.23718 | 77.82182 | 99.19394 | 99.65523 |
| 37 | 2 | 95.73265 | 94.67324 | 83.79384 | 95.86627 | 97.99305 |
| 38 | 2 | 90.545 | 93.90153 | 55.69865 | 90.8617 | 99.79388 |
| 39 | 2 | 96.82691 | 94.69849 | 67.2176 | 96.88976 | 98.75488 |
| 41 | 2 | 96.11459 | 93.69503 | 78.33617 | 96.03496 | 98.91418 |
| 43 | 2 | 97.57583 | 93.55981 | 59.80342 | 97.64036 | 98.32211 |
| 100 | 2 | 79.03399 | 83.1127 | 50.33989 | 79.69589 | 88.6941 |
| 128 | 5 | 69.65257 | 59.97336 | 77.11455 | 69.35564 | 94.08647 |
| 129 | 5 | 98.38494 | 92.49357 | 75.76672 | 98.47398 | 99.46082 |
| 131 | 5 | 83.22591 | 72.63463 | 96.63235 | 83.29729 | 97.65661 |
| 132 | 5 | 98.27743 | 97.2217 | 72.1168 | 98.24669 | 98.8975 |
| 133 | 5 | 97.52444 | 95.56985 | 76.43015 | 97.54273 | 98.99842 |
| 134 | 5 | 90.18322 | 76.75097 | 48.4575 | 90.20867 | 97.14012 |
| 141 | 5 | 99.03941 | 96.89655 | 51.00985 | 98.867 | 99.26108 |
| 142 | 5 | 99.78018 | 99.51325 | 75.08886 | 99.7759 | 99.96003 |
| 143 | 5 | 99.06443 | 98.79547 | 88.11083 | 99.07709 | 99.5127 |
| 144 | 5 | 97.48374 | 97.49684 | 72.23386 | 97.63214 | 99.48278 |
| 157 | 5 | 87.03171 | 69.84767 | 78.13745 | 87.10477 | 97.71721 |
| 210 | 26 | 97.64928 | 95.45894 | 58.7167 | 97.66612 | 99.51008 |
| 215 | 26 | 91.1067 | 91.38363 | 61.98662 | 91.34706 | 91.88003 |
| 216 | 26 | 96.59513 | 90.63923 | 61.11064 | 96.60778 | 99.04706 |
| 233 | 176 | 69.67549 | 61.67475 | 35.71429 | 69.3771 | 89.55614 |
| 235 | 26 | 97.55456 | 95.65006 | 59.11185 | 97.55456 | 99.10632 |
| 237 | 26 | 91.60238 | 91.56381 | 70.04629 | 91.81728 | 95.17853 |
| 240 | 26 | 98.37303 | 95.74631 | 56.13856 | 98.33164 | 99.09577 |
| 249 | 176 | 93.53573 | 89.04998 | 95.69048 | 93.41557 | 98.96668 |
| 254 | 176 | 95.64186 | 94.26823 | 46.74908 | 95.55861 | 96.48268 |
| Trainset Accuracy | | 100% | 100% | 91% | 100% | 100% |
| Testset Accuracy | | 91.20% | 86.10% | 64.70% | 91.30% | 97.70% |

CropType: 1- Corn; 2- Cotton; 5-Soybean; 26- DoubleCrop; 176: Grassland

It is surprising that with full sample set learning, the RF model using spectral features only delivers more than 90% of overall accuracy. Including index, features do not improve the accuracy. However, the accuracy rises up to 97% when texture features are included. For index only and texture only schemes, the overall accuracies are 86% and 64%, respectively. High accuracy of 97% is pretty satisfactory using the original MSI images, which leaves little room to improve with pan-sharpened images. This suggests that with the current pixel-based learning strategy, the MSI images with a spatial resolution of several meters perform well enough for crop type classification. Further thoughts may be taken for argument validity.

**Classification performances with SAR data included**

A number of experiments have been conducted for crop type classification performance with SAR data included. The SAR data was from the European Sentinel satellite. Both VV and VH polarization bands were used. Since the MSI data was the observation on May 18, 2019. Two days of SAR images over Tennessee Valley were obtained and used in the experiments: May 13, 2019, and May 25, 2019.

**Comparison of RF classifier performance on a combination of MSI and SAR features**

| Features | Accuracy | F1 score | Kappa | MCC |
|---|---|---|---|---|
| MSI spectral only | 0.912 | 0.91 | 0.879 | 0.879 |
| MSI vegetation index only | 0.861 | 0.86 | 0.811 | 0.812 |
| MSI textures only | 0.647 | 0.64 | 0.505 | 0.509 |
| MSI spectral + index | 0.913 | 0.91 | 0.88 | 0.88 |
| MSI spectral + index + texture | 0.977 | 0.98 | 0.968 | 0.968 |
| SAR sigma0 (VH + VV) (05/25/19) | 0.53 | 0.51 | 0.349 | 0.352 |
| SAR sigma0 (VH + VV) (05/13/19) | 0.555 | 0.55 | 0.384 | 0.385 |
| SAR sigma0 (VH + VV) (05/13/19 and 05/25/19) | 0.693 | 0.68 | 0.578 | 0.579 |
| SAR sigma0 + texture  (VH + VV) (05/25/19) | 0.822 | 0.82 | 0.754 | 0.756 |
| SAR sigma0 + texture  (VH + VV) (05/13/19) | 0.83 | 0.83 | 0.766 | 0.768 |
| SAR sigma0 + texture  (VH + VV) (05/13/19 and 05/25/19) | 0.915 | 0.91 | 0.883 | 0.883 |
| MSI spectral + SAR sigma0 (VH + VV) (05/25/2019) | 0.941 | 0.94 | 0.92 | 0.92 |
| MSI spectral + SAR sigma0 (VH + VV) (05/13/2019) | 0.944 | 0.94 | 0.923 | 0.923 |
| MSI spectral + SAR sigma0 (VH + VV) (5/13/19 and 5/25/19) | 0.959 | 0.96 | 0.943 | 0.943 |

The Figure above shows the results for RF classifier performance using a number of combinations of SAR and MSI images. The performance is shown in 4 metrics: overall accuracy, weighted F1 score, Kappa index, and Matthew Correlation Coefficients.

The first five rows in the figure show the RF results using a combination of optical features from MSI images and they have been discussed early and will not be discussed here.

The next two rows show the RF performances using two SAR polarization bands (VH + VV) only. These results are viewed as the baseline RF classifier performance that SAR data can deliver. From the figure, the overall accuracies are 53% to 55.5%, respectively for the two days. Compared with RF performance using MSI spectral features only, which is 91.2%, the performance of using SAR data only is significantly worse. Using

two days of SAR data together increases overall accuracy to 69.3%, a significant improvement, but still significantly worse than using MSI bands only. Using two days combination of SAR data can be considered a time series of SAR data, and published literature already indicates the effectiveness of time series of SAR data in improving classifier performance. This experiment also confirms this conclusion. Extending the time series of SAR data (i.e. adding more days of SAR data) is expected to further improve classifier performance.

Experiments were also done for SAR-only data, including the GLCM texture features calculated from both VH and VV bands, for a total of 22 features. Results are shown for both days. For 05/13/19, the overall accuracy increases from 55.5% to 83% with the addition of GLCM texture features. Similarly, on 05/25/19, the overall accuracy increases from 53% to 82.2%. From the two cases, SAR texture features supplement significant information for crop type classification.

We also tested the SAR-only data for crop type classification using VV and VH bands features and GLCM features calculated from the two bands for two days. The overall accuracy increases from 69.3% to 91.5% when texture features were added, about the same as that using MSI spectral bands features.

Further, we tested the classifier performance using combined MSI and SAR features. Using the 4 spectral MSI band and the SAR VH + VV bands features, the overall accuracy is about 94% for the two days test. And with consideration of two daytime series, the overall accuracy increases to 96%. These results show that a combination of MSI and SAR features slightly improves classifier performance.

**RF classifier performance on training sample size using all MSI and SAR features**

For this group of experiments, the focus aims on the dependency of RF classifier performance on the training sample size using all available features. A total of 64 features are available, including 4 MSI spectral bands, 2 SAR bands (VV+VH) for each day (total of 4 for two days), 6 vegetation indexes, 10 MSI texture features, 20 VV+VH SAR texture features for each day(total of 40). The following figure shows the results.

**All MSI+SAR features (total 64)**

| # of Sample | % | Accuracy |
|---|---|---|
| 1211 | 0.05 | 0.895 |
| 2431 | 0.1 | 0.914 |
| 12187 | 0.5 | 0.946 |
| 60961 | 2.5 | 0.968 |
| 121928 | 5 | 0.975 |
| 243862 | 10 | 0.982 |
| 487733 | 20 | 0.987 |
| 1219350 | 50 | 0.993 |
| 1463215 | 60 | 0.994 |
| 1707086 | 70 | 0.994 |

From the figure above, the RF performance has little improvement when the training sample size reaches 10% of the total samples.

| All MSI+SAR features (total 64) | |
|---|---|
| % train sample | Accuracy |
| 5 | 0.507 |
| 20 | 0.507 |
| 50 | 0.506 |
| 70 | 0.507 |

The learned RF models show great performances for the samples in the region of interest. Then experiments were done to apply these learned models to all the 5 crop samples in the image but not in the regions of interest. This will show how the learned RF models perform on the unseen samples. Four RF models were tested. The four models were learned using training samples of 5%, 20%, 50%, and 70% from the regions of interest. The overall accuracies for the unseen samples from these 4 models are basically the same, 50.7%, a huge decrease compared to over 97% for the samples in the regions of interest. These results are consistent with those from model learning using select-region methods in earlier experiments and indicate that the learned models cannot perform well for the unseen samples.

Experiments were also conducted to check the impact of feature reduction through the principal component analysis (PCA) on the RF classifier performance. The PCA was run on the 64 MSI and SAR features and first, we selected the top 20 components were selected as the features for RF classifier learning. The training sample size is still set as 70%. The overall accuracy is 0.993, just slightly less than that using all 64 original features. Then the trained RF model was used to classify all the unseen samples in the Tennessee valley image that are not included in the selected regions and are one of the 5 crop types of interest (Corn, Cotton, Soybean, double-crop, and Grassland), with the same strategy as in the previous paragraph. The accuracy drops significantly, to 0.478. This accuracy value is close to the performance of the RF model that used training samples from the selected regions, though for that experiment only 4 spectral bands of MSI features were used. This result further confirms that the RF model using MSI and SAR features can perform very well on the crop regions that it learned but cannot be generally applied to other regions that it does not see and learn. This suggests that the features extracted from different crop fields with the same crop type are quite diverse. The following is the confusion matrix for the experiment using 20 PCA components.

| | | Test set | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 257291 | 31 | 757 | 192 | 130 | | Corn | 2882199 | 173056 | 770480 | 473135 | 152596 |
| | Cotton | 108 | 147489 | 644 | 17 | 0 | | Cotton | 625415 | 1819979 | 2074786 | 118693 | 35529 |
| Actual | Soybean | 682 | 865 | 212394 | 176 | 2 | Actual | Soybean | 1805140 | 874471 | 3214270 | 519692 | 148518 |
| | DblCrop | 708 | 4 | 209 | 97459 | 14 | | DblCrop | 521988 | 28452 | 132624 | 987822 | 32887 |
| | Grassland | 655 | 0 | 34 | 16 | 11741 | | Grassland | 1103327 | 61034 | 157373 | 372135 | 419272 |

Overall accuracy: 0.993                    Overall accuracy: 0.478

Next we selected the top 5 PCA components for the RF model learning. Following are the confusion matrices.

| | | Test set | | | | | | | | Independent set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted | | | | | | | | Predicted | | | |
| | | Corn | Cotton | Soybean | DblCrop | Grassland | | | Corn | Cotton | Soybean | DblCrop | Grassland |
| | Corn | 227261 | 3390 | 12785 | 14223 | 742 | | Corn | 2539004 | 214313 | 909986 | 622011 | 166152 |
| | Cotton | 839 | 129148 | 18046 | 224 | 1 | | Cotton | 531227 | 2144623 | 1715016 | 252212 | 31324 |
| Actual | Soybean | 18994 | 27428 | 165154 | 2462 | 81 | Actual | Soybean | 1518393 | 1607551 | 2549683 | 747366 | 139098 |
| | DblCrop | 17311 | 53 | 1635 | 79305 | 90 | | DblCrop | 715039 | 50847 | 89207 | 811032 | 37648 |
| | Grassland | 4124 | 4 | 426 | 416 | 7476 | | Grassland | 1030862 | 64303 | 252008 | 415806 | 350162 |

Overall accuracy: 0.832                      Overall accuracy: 0.430

From this experiment, reducing the number of PCA components from 20 to 5 significantly decreases the overall accuracy from 0.993 to 0.832. Overall accuracy for independent samples decreases from 0.478 to 0.430. So it seems the optimal number of principal components for RF model learning is somewhere between 5 and 20.

Further notice that the overall accuracy of the RF model using 5 PCA components is worse than that trained using only 4 MSI spectral band features, which has the overall accuracy of 0.912. That seems to suggest that applying PCA for feature reduction may not deliver better model classification performance than using original features.

**Classifier performance with training samples from an entire image using all MSI and SAR features**

For all the experiments above, the feature set for model training are collected from the selected regions of interest representing the 5 crop types of interest. Since we manually identify and crop the regions out of the image, pixels are expected to be more homogeneous in spectral, index, and texture features within a region than without the region. And learned models are as a result to have higher overall accuracy for these pixels.

At last, experiments were conducted to test the model performance using training samples directly and randomly extracted from the image instead of from manually carefully selected regions.

```
classification report
                                      precision    recall   f1-score    support

                          Corn          0.63        0.63      0.63      5062370
                        Cotton          0.71        0.64      0.67      4984715
                       Soybean          0.60        0.63      0.61      6952021
 DoubleCrop/WinterWheat_Soybean        0.62        0.66      0.64      1931101
                     Grassland          0.49        0.49      0.49      2052275

                      accuracy                                0.62     20982482
                     macro avg          0.61        0.61      0.61     20982482
                  weighted avg          0.62        0.62      0.62     20982482
```

The classification report shown above was from the model run with all learning samples randomly selected from the MSI image for the 5 crop types. Since the pixel volume is too large, only 5% of the samples were selected for model learning and only 4 MSI spectral bands were used as features so the laptop computer can handle the learning

process. From the report, the overall accuracy is 62%, as compared to over 90% using the selected regions of interest. The total number of samples supporting this result is over 20 million. We further extended this experiment to evaluate the model performance for all surface scene types in the MSI image instead of the 5 crop types of interest using the same training sample selection strategy.

```
classification report
              precision    recall  f1-score   support

           1       0.55      0.60      0.57   5275268
           2       0.68      0.63      0.65   5195375
           4       0.00      0.00      0.00       223
           5       0.54      0.59      0.56   7244593
           6       0.03      0.00      0.01       207
          10       0.29      0.01      0.02       224
          24       0.13      0.01      0.03     31244
          26       0.54      0.64      0.59   2012304
          29       0.15      0.03      0.05      8826
          37       0.14      0.03      0.05    319315
          44       0.08      0.00      0.01       209
          46       0.30      0.06      0.10      1547
          59       0.28      0.05      0.09      4511
          61       0.58      0.51      0.55   1273193
         111       0.65      0.53      0.58    136108
         121       0.27      0.21      0.24   1578831
         122       0.24      0.20      0.22    757773
         123       0.30      0.26      0.28    440353
         124       0.41      0.30      0.35    113336
         131       0.21      0.08      0.11     25039
         141       0.28      0.19      0.23   2710717
         142       0.33      0.19      0.24    409146
         143       0.10      0.01      0.02    235596
         152       0.18      0.03      0.06     97322
         176       0.29      0.25      0.27   2138364
         190       0.57      0.75      0.65   4999008
         195       0.45      0.18      0.26    123615
         222       0.17      0.01      0.02       433
         225       0.13      0.02      0.03       196

    accuracy                           0.52  35132876
   macro avg       0.31      0.22      0.24  35132876
weighted avg       0.50      0.52      0.51  35132876
```

This time, 2% of all the samples from the image were randomly extracted for model learning. From the classification report, the overall accuracy is 52% with only 4 MSI spectral band features used. A total of 29 surface scene types are included. Some scene types have a very limited number of samples in the images. The five crop types that we are interested in are type 1 for Corn, type 2 for Cotton, type 5 for Soybean, type 26 for DoubleCrop and type 176 for Grassland.

The F1 score for Corn drops from 0.63 to 0.57 when all surface types are classified. For Cotton, the F1 score drops from 0.67 to 0.65. For soybeans, the F1 score drops from 0.6` to 0.56. For double crops, the F1 score drops from 0.64 to 0.59. And for Grassland, it drops from 0.49 to 0.27. So grassland type is a difficult surface type to be distinguished so as the number of surface types increases the performance of grassland type deteriorates significantly.

**Overall Summary**

Four convolution neural networks were developed, tested, and evaluated (a traditional supervised CNN, a generative adversarial network (GAN), a PanColorGAN, and an unsupervised CNN with an advanced specialized loss function against well-known and published pan-sharpening methods. Results were evaluated for usefulness in the identification of agricultural crops. Our results indicated that the PanColorGAN was the most successful followed by the basic GAN, and then the supervised CNN. Our attempt to use an unsupervised CNN with a specialized loss function was unsuccessful. The biggest issue we encountered was the large-scale difference between the panchromatic image, the multispectral image, and SAR data. Deep learning attempts to fuse the data together into a single product were problematic. We found that traditional mathematical algorithms such as IHS and various forms of wavelet transformations were very successful. With this knowledge, we were able to pan-sharpen and fuse together a new GEOINT product consisting of pan-sharpened blue, green, red, nir, and polarization bands.

We used this new product to identify and classify agricultural row crops (corn, wheat, cotton, and soybeans) in the Tennessee River Valley. To evaluate we conducted a variety of experiments for crop type classification using a machine learning approach, mainly the Random Forest classifier. The experiments include the performance comparison between using pan-sharpened images and using original MSI images. According to our results, there are no obvious benefits to using the pan-sharpened images for crop type classification, at least with our new dataset.

From these experiments, we have seen that using training samples from carefully selected regions of interest, the RF (and other models) can perform very well in distinguishing among the 5 crop types if the models are learned with representative samples. Even using MSI spectral band features, the overall accuracy can reach over 90%. The accuracy improves with including texture features, and with including SAR polarization band features and their texture features. The overall accuracy can be as high as 99%. However, applying these well-trained models to other 'unseen' pixels (pixels in other areas with the same surface types) the performance decreases significantly. If we randomly select the sample for model learning, the overall accuracy is only over 60% for the 5 crop types and drops further to 50% for all surface type classification. This suggests the broad spectrum of feature characteristics even for the same crop type over the Tennessee River Valley, making the high-quality classification a challenge. However, model high performance over the 'selected regions' with high homogeneity/uniformity of feature characteristics is encouraging.

The most encouraging results from our new product were when we included the GLCM texture features calculated from both VH and VV bands. The overall accuracy increased from 53% to 83% with the addition of GLCM texture features. Supplementing SAR texture features information for crop type classification was significant. We are encouraged to continue research in this field and hope to acquire higher-resolution SAR data to improve our results.

**References**

1. Saraf, A. IRS-1C-LISS-III and PAN data fusion: an approach to improve remote sensing-based mapping techniques. International Journal of Remote Sensing, 1999. 20(10): p. 1929-1934.

2. Saroglu, E., et al. Fusion of multisensor remote sensing data: assessing the quality of resulting images. Int. Arch. Photogram. Rem. Sens. Spatial. Inform. Sci, 2004. 35: p. 575-579.

3. Venkataraman, G., et al. Fusion of Optical and Microwave Remote Sensing data for snow cover mapping. in Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International. 2004. Ieee.

4. Pal, S., T. Majumdar, and A.K. Bhattacharya. ERS-2 SAR and IRS-1C LISS III data fusion: A PCA approach to improve remote sensing based geological interpretation. ISPRS journal of photogrammetry and remote sensing, 2007. 61(5): p. 281-297.

5. Harris, J.R., R. Murray, and T. Hirose. IHS transform for the integration of radar imagery with other remotely sensed data. Photogrammetric Engineering and Remote Sensing, 1990. 56(12): p. 1631-1641.

6. Yonghong, J. Fusion of Landsat TM and SAR image based on principal component analysis. Remote Sensing Technology and Application, 1998. 13(1): p. 46-49.

7. Zhou, J., D. Civco, and J. Silander. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. International Journal of Remote Sensing, 1998. 19(4): p. 743-757.

8. Huang, D. and M. Yang. SAR and Multispectral images fusion based on Contourlet and HIS Transform. in Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on. 2010. IEEE.

9. Bannari, A., et al. PALSAR-FBS L-HH Mode and Landsat-TM Data Fusion for Geological Mapping. Advances in Remote Sensing, 2016. 5(04): p. 246.

10. Rao, C., et al. Fast spatiotemporal data fusion: merging LISS III with AWiFS sensor data. International Journal of Remote Sensing, 2014. 35(24): p. 8323-8344.

11. Radhika, K., C. Rao, and V.K. Prasad. Enhancement of AWiFS Spatial Resolution with SVM Learning. in Advanced Computing (IACC), 2016 IEEE 6th International Conference on. 2016. IEEE.

12. Li, S., H. Yin, and L. Fang. Remote sensing image fusion via sparse representations over learned dictionaries. IEEE Transactions on Geoscience and Remote Sensing, 2013. 51(9): p. 4779-4789.

13. Masi, G., et al. Pansharpening by convolutional neural networks. Remote Sensing, 2016. 8(7): p. 594.

14. Zhong, J., et al. Remote sensing image fusion with convolutional neural network. Sensing and Imaging, 2016. 17(1): p. 1-16.

15. J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley. "PanNet: A Deep Network Architecture for Pan-Sharpening," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 1753-1761.doi: 10.1109/ICCV.2017.193

16. Wei Yao, Zhigang Zeng, Cheng Lian, Huiming Tang. Pixel-wise regression using U-Net and its application on pansharpening, Neurocomputing, Volume 312, 2018, Pages 364-371, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.05.103.

17. Schmitt, M., Hughes, L. H., and Zhu, X. X.: THE SEN1-2 DATASET FOR DEEP LEARNING IN SAR-OPTICAL DATA FUSION, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-1, 141-146, https://doi.org/10.5194/isprs-annals-IV-1-141-2018, 2018.

18. Basu, S., et al. Deepsat: a learning framework for satellite imagery. 2015. ACM.

19. LeCun, Y., Y. Bengio, and G. Hinton. Deep learning. Nature, 2015. 521(7553): p. 436-444.

20. Dong, C., et al. Learning a deep convolutional network for image super-resolution. in European Conference on Computer Vision. 2014. Springer.

21. C.B. Collins, J.M. Beck, S.M. Bridges, J.A. Rushing, S.J. Graves, 2017. In GeoAI'17: Deep Learning for Multisensor Image Resolution Enhancement 1st ACM SIGSPATIAL Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery, November 7-10, Los Angeles Area, CA, USA, 8 pages. https://doi.org/10.1145/3149808.3149815.

22. J.M. Beck, C.B. Collins, S.M. Bridges, S.J. Graves, S. Wingo, 2018. "Machine Learning for Global Precipitation Measurement (GPM) Dual-Frequency Precipitation Radar (DPR) Resolution Improvement", AGU Fall Meeting, Washington, D.C, USA.

23. Ledig, C., et al. Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint, 2017.

24. Nagano, Yudai, and Yohei Kikuta. "SRGAN for super-resolving low-resolution food images." Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. 2018.

25. Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

26. Lim, B., et al. Enhanced deep residual networks for single image super-resolution. 2017.

27. Y. Wei, Q. Yuan, H. Shen, and L. Zhang. "Boosting the Accuracy of Multispectral Image Pansharpening by Learning a Deep Residual Network," in IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1795-1799, Oct. 2017. doi: 10.1109/LGRS.2017.2736020

28. Kim, J., J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. 2016.

29. Hinton, G., N. Frosst, and S. Sabour. Matrix capsules with EM routing. 2018.

30. Sabour, S., N. Frosst, and G.E. Hinton. Dynamic routing between capsules. in Advances in Neural Information Processing Systems. 2017.

31. Uijlings, J.R., et al. Selective search for object recognition. International journal of computer vision, 2013. 104(2): p. 154-171.

32. Girshick, R., et al. Rich feature hierarchies for accurate object detection and semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

33. He, K., et al. Mask r-CNN. in Computer Vision (ICCV), 2017 IEEE International Conference on. 2017. IEEE.

34. X. Liu, Y. Wang and Q. Liu. "Psgan: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 873-877. doi: 10.1109/ICIP.2018.8451049

35. Wei Yao, Zhigang Zeng, Cheng Lian, Huiming Tang. Pixel-wise regression using U-Net and its application on pansharpening, Neurocomputing, Volume 312, 2018, Pages 364-371, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.05.103.

36. Ozcelik, F., Alganci, U., Sertel, E., & Unal, G. (2020). Rethinking CNN-Based Pansharpening: Guided Colorization of Panchromatic Images via GANs. arXiv preprint arXiv:2006.16644.

37. Wald, T. Ranchin, M. Mangolini Fusion of satellite images of different spatial resolution: assessing the quality of resulting images Photogramm. Eng. Remote Sens., 63 (1997), pp. 691-699.

38. S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via Unsupervised Convolutional Neural Networks," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 4295-4310, 2020, doi: 10.1109/JSTARS.2020.3008047.

39. Gemine Vivone et al. "A Critical Comparison Among Pansharpening Algorithms," IEEE Trans. On Geoscience and Remote Sensing, vol. 53, No. 5, May 2015

40. Pushparaj, Jagalingam, and Arkal Vittal Hegde. "Evaluation of pan-sharpening methods for spatial and spectral quality." Applied Geomatics 9.1 (2017): 1-12.

41. P. Iervolino, R. Guida, D. Riccio, and R. Rea, "A Novel Multispectral, Panchromatic and SAR Data Fusion for Land Classification," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 10, pp. 3966-3979, Oct. 2019, doi: 10.1109/JSTARS.2019.2945188.

42. Orynbaikyzy, Aiym, Ursula Gessner, and Christopher Conrad. "Crop type classification using a combination of optical and radar remote sensing data: A review." international journal of remote sensing 40.17 (2019): 6553-6595.

43. Haralick, Robert M., Karthikeyan Shanmugam, and Its' Hak Dinstein. "Textural features for image classification." IEEE Transactions on systems, man, and cybernetics 6 (1973): 610-621.

44. Sentinel Application Platform (SNAP), https://step.esa.int/main/toolboxes/snap/.

45. Performance Measures: Cohen's Kappa statistics, https://thedatascientist.com/performance-measures-cohens-kappa-statistic/.

46. Understanding AUC-ROC curve, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

47. Intuition behind Log-loss score,
https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a.