# MAPPING INVASIVE PLANT SPECIES USING MACHINE LEARNING

A Technical Paper submitted to the Department of Computer Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Surbhi Singh

April 28, 2020

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Madhav Marathe, Department of Computer Science

## INTRODUCTION

The spread of invasive plant species is currently one of the greatest epidemics facing the agricultural industry, producing a loss of $137 billion per year in the United States alone. Due to the growth of human factors such as global travel and foreign imports, the prevalence of invasive species has risen substantially. The research proposed here aims to map invasive plants using remote sensing satellite imagery and machine learning algorithms. The first step to mitigating the presence of invasive plant species is understanding what factors are contributing to the spread. This can be studied using convolutional neural networks, which take an image as input and determine the importance of certain features. Invasive plant species are a global epidemic, but they are especially prevalent in biodiversity hotspots, such as the Chitwan Annapurna Landscape (CHAL) of Nepal. CHAL is the region of interest in this research for that very reason. A biodiversity hotspot is a region which is both rich in plant and animal species, but a region that is also threatened with destruction (Chepkemoi, 2017). With recent developments in high performance computing and machine learning, satellite imagery has become a viable tool in mapping plant species distributions. Three invasive plant species in this region of Nepal were studied using multiple types of imagery and convolutional neural network based architecture- *Lantana camara*, *Chromolaena odorata*, and *Parthenium hysterophorus*.

The main focus of the research was to experiment with different types of satellite images used to map species distribution. Many types of satellite imagery have been used to develop species distribution maps. Remote sensing using satellite imagery is the most viable option for tracking plant species with unique phenology or growth form over time. The use of multispectral remote sensing allows for more accurate identification because each layer of the image provides unique information.

The use of pan sharpened imagery to increase the resolution of the satellite imagery and improve prediction accuracy was explored. Pansharpening, short for panchromatic sharpening, is the merging of high-resolution panchromatic and lower resolution multispectral imagery for the creation of a single-high resolution image. This process was applied to all of the Nepal satellite imagery and the models were retrained to determine if these images provided better predictions for the spread of invasive plant species.

## FRAMEWORK

The goal of this research is to create species distributions maps to understand the spread of invasive plant species in the CHAL region of Nepal. These maps could provide insight on the introduction of spread and help mitigation efforts. This work attempts to develop a novel framework for using machine learning techniques and various types of satellite imagery to study the spread of invasive plant species. Convolutional Neural Networks (CNNs) have recently been explored for training models on hyperspectral and multispectral imagery. This type of neural network is a deep learning algorithm which takes an image as input and determines the importance of certain features (Saha, 2018). They are popularly used for mapping invasive species because of the ability of the algorithm to capture the spatial and temporal dependencies of the image. CNNs pinpoint which features are important in the image, and can be used to reduce the presence of invasive species by analyzing and identifying the environmental predictors. With these advanced techniques and high-resolution satellite imagery, modeling the spread of invasive plant species from remote sensed data has become a viable option (He et al., 2015). Due to the restricted size of the dataset, the number of layers used in the CNNs were kept small. LeNET-5 was used as a basis and the architecture was deepened to 4 convolutional layers.

The overview of the framework is shown below.

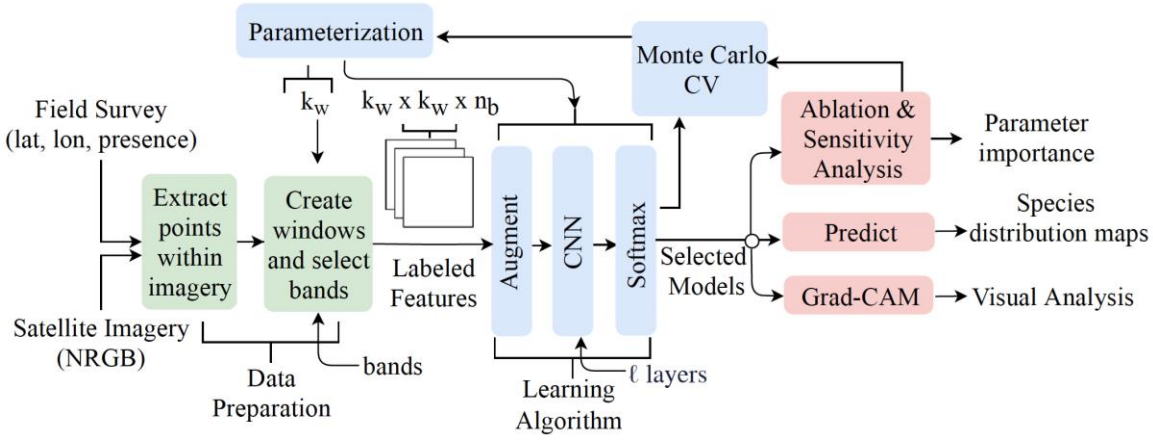


Figure 1: Framework Overview: Field reports were generated by a team in Nepal and used to create feature sets. This data was augmented and run through various CNN architectures to see which would create the most accurate predictions.

There are several challenges faced when collecting plant data including season, terrain, and biology of the plant. The CHAL area is a mountainous region, making it difficult to access some areas. Therefore, most of the data is collected by the roadside leading to roadside bias and less data points. Additionally, multiple plants can occupy the same space and are hard to identify. The season of data collection can affect the cloud coverage, fog spread, and color of the plants being studied. Because of the mentioned reasons and many more, free satellite images with sufficient resolution are not available. However, high-resolution images are not free and too expensive to obtain for the whole CHAL area. High resolution DigitalGlobe Worldview-2 satellite imagery was acquired for the region of interest. Because this imagery is expensive, 5 sub regions of the CHAL with diverse vegetation, elevation, and human activity were

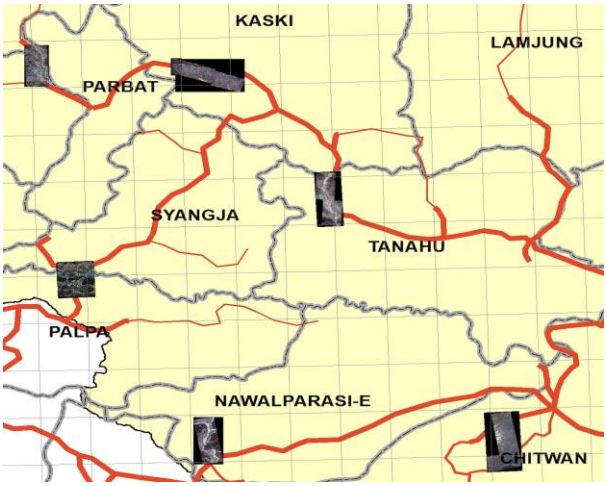


Figure 2: The CHAL region in Nepal with 6 specified regions for which DigitalGlobe imagery was acquired (excluding Palpa).

selected and images were acquired for those regions for the years of 2017 and 2018 (Adiga, Singh, & Venkatramanan, 2020). CNN models usually require large amounts of data to achieve good performance and tend to overfit if not given enough data. Therefore, plant expert's analysis will be needed to confirm the results of this limited data model.

## MODEL TRAINING AND SELECTION

To artificially increase the size of the limited dataset, data augmentation was performed by rotating and flipping feature vectors. Several popular techniques were then used to train and select the best models. Models were trained using binary cross-entropy loss and Monte Carlo cross validation was applied to see how the model would perform for a given parameter set. For each round, 20% of the data points were used as the test dataset and the rest 3:1 split between training and validation. Hyperparameters included batch size, epochs, dropout, and mad pooling. To identify the best models, the full factorial was initially used and then high performing values based on accuracy and F1-score were identified for further experiments. These experiments were run on University of Virginia's high performance computing system Rivanna which allowed 1000 tasks to be run simultaneously. Thus, allowing fast exploration of model parameters to select the best model.

## RESULTS

Initially, experiments were run on all of the non-pan sharpened imagery for all three species. M* was identified as the top performing set of parameters which was defined as window size =32, bands=nrgb, augmentation, dropout, batch normalization, max pooling, #epochs =100, and learning rate = 0.001. These values were used for training and evaluation with the Monte Carlo CV tests. The results are shown below.

| Spec. | Model | Accuracy | Std. dev. accuracy | F1 score | Std. dev. F1 score |
|---|---|---|---|---|---|
| *P. hysterophorus* | $\mathcal{M}^*(4,16)$ | 0.905 | 0.0334 | 0.895 | 0.0458 |
| | $\mathcal{M}^*(4,4)$ | 0.894 | 0.0548 | 0.888 | 0.0606 |
| | $\mathcal{M}^*(2,16)$ | 0.887 | 0.0440 | 0.885 | 0.0443 |
| | $\mathcal{M}^*(2,4)$ | 0.886 | 0.0541 | 0.881 | 0.0574 |
| | LeNET_avg | 0.894 | 0.0433 | 0.891 | 0.0454 |
| | LeNET_max | 0.889 | 0.0479 | 0.881 | 0.0596 |
| *C. odorata* | $\mathcal{M}^*(4,16)$ | 0.890 | 0.0594 | 0.879 | 0.0701 |
| | $\mathcal{M}^*(4,4)$ | 0.871 | 0.0817 | 0.849 | 0.1148 |
| | $\mathcal{M}^*(2,16)$ | 0.868 | 0.0868 | 0.840 | 0.1282 |
| | $\mathcal{M}^*(2,4)$ | 0.866 | 0.0773 | 0.842 | 0.1046 |
| | LeNET_avg | 0.858 | 0.0790 | 0.837 | 0.1219 |
| | LeNET_max | 0.850 | 0.0903 | 0.818 | 0.1305 |
| *L. camara* | $\mathcal{M}^*(4,16)$ | 0.709 | 0.0395 | 0.723 | 0.0410 |
| | $\mathcal{M}^*(4,4)$ | 0.683 | 0.0476 | 0.725 | 0.0353 |
| | $\mathcal{M}^*(2,16)$ | 0.710 | 0.0527 | 0.733 | 0.0401 |
| | $\mathcal{M}^*(2,4)$ | 0.653 | 0.0535 | 0.708 | 0.0404 |
| | LeNET_avg | 0.686 | 0.0617 | 0.725 | 0.0457 |
| | LeNET_max | 0.674 | 0.0418 | 0.710 | 0.0370 |

Table 1: Performance of best models for non-pan sharpened imagery under M* specifications.

**PANSHARPENING**

The DigitalGlobe imagery acquired contained both 4 and 8-band multispectral imagery with 1.84 m resolution and panchromatic imagery of 0.46m resolution. The 4 bands correspond to nrgb (near infrared, red, green, and blue). Generally, a multispectral image has a high spectral resolution while a panchromatic image has a high spatial resolution. Pansharpening can be used to increase the resolution of the multispectral imagery by merging the single band panchromatic image with the multispectral bands. Thus, when the two images are merged, it creates both high spectral and spatial resolution. This technique is used by most mapping software such as google maps. The two photos below are an example of a normal and pan sharpened image used from the CHAL region.

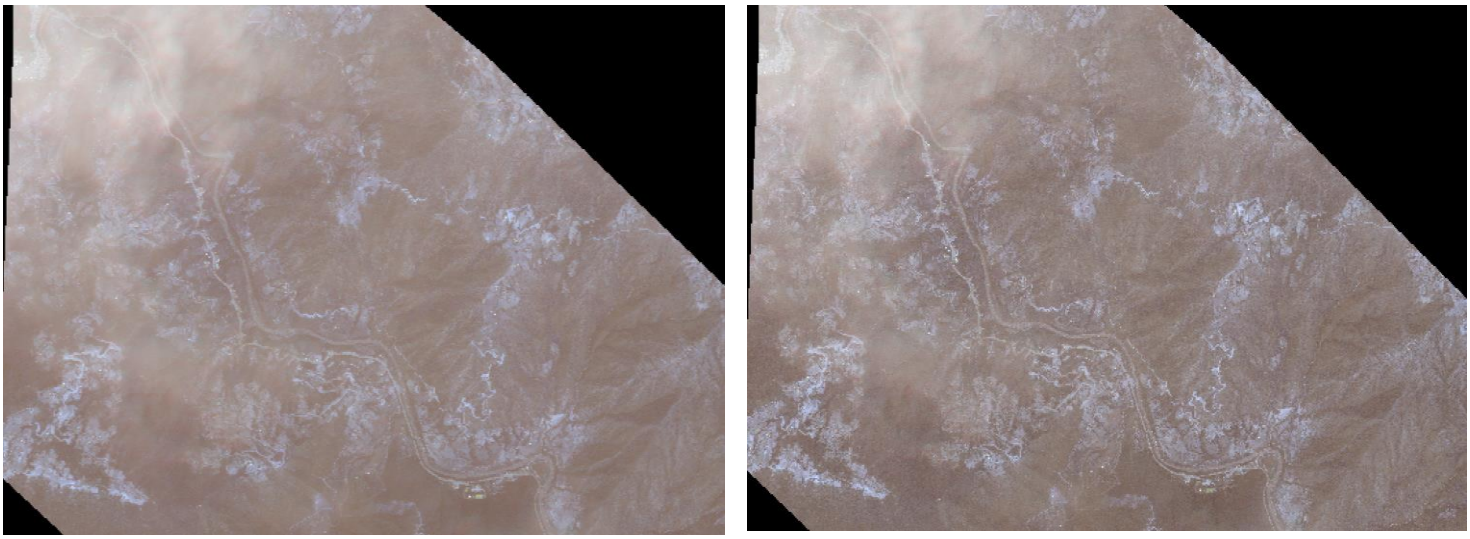

Figure 3: Example of regular DigitalGlobe satellite image of a part of CHAL (left) versus the more detailed image of same are after pansharpening (right).

The experiments performed on the original satellite imagery were repeated on the pan sharpened imagery using the M* specifications. The only parameter that was changed was the window size, which was updated to 128 due to the increased size of the pan sharpened images.

| Spec. | Model | Accuracy | Std. dev. accuracy | F1 score | Std. dev. F1 score |
|---|---|---|---|---|---|
| *P. hysterophorus* | $\mathcal{M}^*(4,16)$ | 0.900 | 0.0202 | 0.899 | 0.0238 |
| | $\mathcal{M}^*(4,4)$ | 0.868 | 0.0303 | 0.867 | 0.0288 |
| | $\mathcal{M}^*(2,16)$ | 0.876 | 0.0191 | 0.874 | 0.0209 |
| | $\mathcal{M}^*(2,4)$ | 0.863 | 0.0480 | 0.866 | 0.0438 |
| | LeNET_avg | 0.888 | 0.0356 | 0.889 | 0.0355 |
| *C. odorata* | $\mathcal{M}^*(4,16)$ | 0.865 | 0.0714 | 0.842 | 0.0902 |
| | $\mathcal{M}^*(4,4)$ | 0.867 | 0.0657 | 0.845 | 0.0882 |
| | $\mathcal{M}^*(2,16)$ | 0.869 | 0.0693 | 0.847 | 0.0894 |
| | $\mathcal{M}^*(2,4)$ | 0.840 | 0.0454 | 0.821 | 0.0584 |
| | LeNET_avg | 0.827 | 0.0433 | 0.799 | 0.0605 |
| | LeNET_max | | | | |
| *L. camara* | $\mathcal{M}^*(4,16)$ | 0.755 | 0.0481 | 0.774 | 0.0444 |
| | $\mathcal{M}^*(4,4)$ | 0.732 | 0.0302 | 0.754 | 0.0254 |
| | $\mathcal{M}^*(2,16)$ | 0.715 | 0.0489 | 0.737 | 0.0445 |
| | $\mathcal{M}^*(2,4)$ | 0.696 | 0.0516 | 0.710 | 0.0925 |
| | LeNET_avg | 0.735 | 0.0500 | 0.743 | 0.0626 |
| | LeNET_max | | | | |

Table 2: Performance of best models for pan sharpened imagery under M* specifications (with window size = 32).

While improvements were expected for all species, *P. hysterophor*us and *C. Odorata* performed comparably to the normal imagery and only *L.Camara* saw significant improvement. More improvement could

possibly be seen if another parameter set beyond M* was explored for the pan sharpened imagery.

**Species Distribution Maps**

Due to the roadside bias in the field survey data, species distribution maps were constructed along points which were near major roads and present in the satellite images. Using predictions from the best performing CNN models, the probability of presence was plotted for each species with Mapbox studio. Plant experts were consulted to see if the predictions match the actual spread of the plant species. While the *L. camara* and *P. hysterophorus* maps were mostly accurate, especially in the red areas which are heavily infested, the *C. odorata* map seemed to differ from the actual spread.

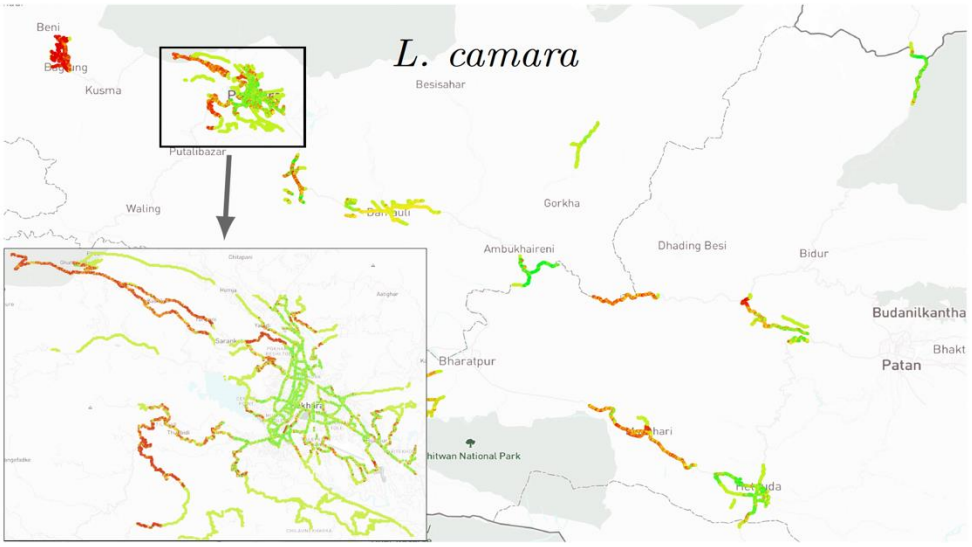


Figure 4: Species distribution map for Lantana Camara. The color represents the probability of presence: green (0) to yellow (0.5) to red (1).

## ANALYSIS

### ABLATION STUDIES

Ablation studies were used to understand the importance of band and hyperparameter choice. After identifying the consistently top-performing model M*(4,16), feature ablation was performed by seeing how the test F1 score changes for bands = nrgb. As we can see in Table 3,

| Species | r | g | b | n |
|---|---|---|---|---|
| *C. odorata* | -0.027 | -0.031 | -0.022 | 0.033 |
| *L. camara* | -0.018 | -0.002 | -0.003 | 0.004 |
| *P. hysterophorus* | 0.009 | 0.014 | -0.006 | 0.004 |

Table 3: Input Feature Ablation

the values are near 0 indicating that the ablation had minimal effects. For *C. odorota*, bands r, g, and b are significant and the F1 score is lowered when removed.

**GRADCAM FEATURE MAPS**

GradCAM analysis was performed to see which features were being highlighted as important from the satellite imagery (Selvaraju et al., 2017). GradCAM output for true positive (TP) and true negative (TN) predictions were generated to see whether factors such as roads or buildings influenced the model's decisions. Because roads and buildings are highlighted in both TP and TN cases, they are not the deciding factors for the CNN to classify the presence or absence of different plant species. Other exploratory analysis included seeing how models for different species differed in GradCAM output. Areas in which two or more species were identified as presence were used to generate GradCAM outputs as shown in Figure 5. While the highlighted regions looked similar in some instances, it was very different for others. Further GradCAM type analysis could be performed by exploring methods such as randomized input sampling or integrated gradients.

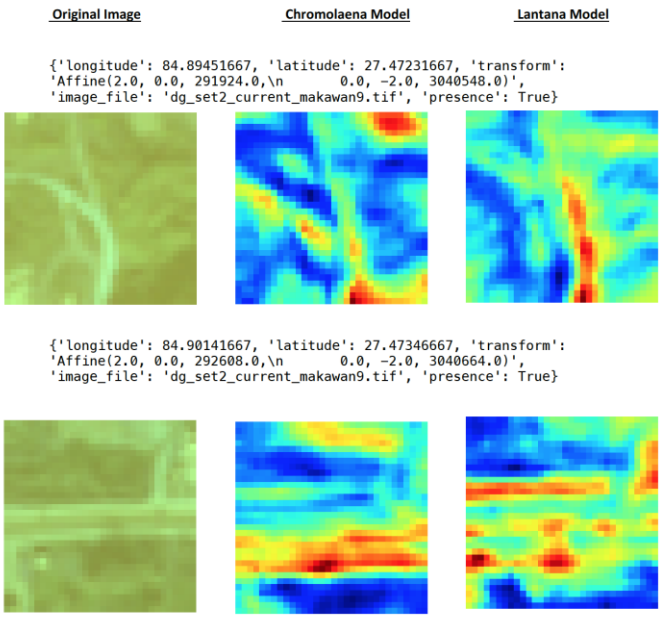


Figure 5: GradCAM outputs for regions where two species were present to visualize difference in important features for different species.

## CONCLUSION

The use of DigitalGlobe satellite imagery was studied to combat invasive plant species in Nepal using CNNs. The research shows that predictions of high accuracy can be achieved despite several complications with data collection and having a small dataset. The use of pansharpening to increase the accuracy of the CNN predictions shows potential but must be further explored for concrete results. In addition, several framework modifications can be investigated to make the process more efficient and accurate. LandSat8 imagery, which is low resolution (30m), is freely available and could be obtained for the whole region of interest. If the Landsat8 data produces comparable to the DigitalGlobe data, the need to purchase expensive images is eliminated or reduced. Understanding how invasive plant species originate and spread is crucial to mitigating their presence. As novel Machine learning methods are applied to larger amounts of detailed data, effective methods will be developed to study and monitor invasive plant species across the globe.

## WORKS CITED


Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (pp. 9505-9515).

Adiga, A., Singh, S., & Venkatramanan, S. (2020). Mapping invasive plants in a biodiversity hotspot using remote sensing data, under preparation.

Chepkemoi, J. (2017, March 28). What is a biodiversity hotspot? Retrieved from https://www.worldatlas.com/articles/what-is-a-biodiversity-hotspot.html.

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M. N., Schmidtlein, S., & Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, *1*(1), 4-1

Kobilinsky, D. (2016). Invasive species bigger threat in developing countries. *The Wildlife Society,* Retrieved from https://wildlife.org/invasive-species-bigger-threat-in-developing-countries/

Saha, S. (2018). A comprehensive guide to convolutional neural networks-the ELI5 way. Retrieved from https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*. doi:10.1109/iccv.2017.74

Stegelmeier, B. L., Field, R., Panter, K. E., Hall, J. O., Welch, K. D., Pfister, J. A., ... & Green, B. T. (2013). Selected poisonous plants affecting animal and human health. *Haschek and Rousseaux's Handbook of Toxicologic Pathology* (pp. 1259-1314). https://doi.org/10.1016/B978-0-12-415759-0.00040-6

Wang, L. (2008). Invasive species spread mapping using multi-resolution remote sensing data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *37*, 135-142.

Wu, Y., Yang, Y., Nishiura, H., & Saitoh, M. (2018). Deep learning for epidemiological predictions. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. doi:10.1145/3209978.3210077