Towards Localizing Nomadic Pastoralist Settlements from Remote Sensing Data: Proposal of a Novel Dataset of Settlements in Nyangatom, Ethiopia and Exploration of Deep Learning Classification Methods

Benjamin Liu
Stanford University
bencliu@stanford.edu

Abstract

For several years, nomadic pastoralists (NP) have escaped the surveying duties of world-wide healthcare systems. By definition, NP are individuals who travel seasonally with designated livestock to rich pastures or new settlement areas. With no formal abodes, NP are among the world's poorest and most marginalized communities. Along with the myriad of health challenges that these groups face, they are systematically underrepresented in healthcare survey data due to difficulties in monitoring settlement locations. In this paper, I present a comprehensive dataset of NP settlements in Nyangatom Ethiopia for the development of deep learning classification models that focus on localizing NP settlements from satellite imagery. Furthermore, I present the results of a comprehensive deep learning study, where I trained and tuned a variety of state-of-the-art models on the classification of settlements from the presented dataset. Quantitative and qualitative results obtained from ablation tests of the best performing model establish promising baselines for future deep learning explorations and provide insights into useful training paradigms for NP detection models. Achieved results indicate potential in localizing NP settlements beyond Ethiopia, thus extending a helping hand to a population in urgent need.

1 Introduction

"Nonexistent," is the word that the World Health Organization (WHO) uses to characterize health data on nomadic pastoralists (NP) [1]. By definition, NP are individuals who travel seasonally with designated livestock to resource-filled pastures. With no formal abodes, NP are among the world's poorest and most marginalized populations. Pastoralist communities commonly face a myriad of health challenges ranging from ecological anomalies, civil conflict, and protracted humanitarian crises; these challenges are compounded with their systematic lack of representation in health survey data [2, 3, 4]. As a direct consequence, considerations of their health are often excluded from both the planning of health services and health campaigns.

The roots of NPs' lack in representation can be traced to data collection challenges associated with mobility. Since most large-scale data sources use sampling frameworks derived from census enumerations, mobile populations can frequently escape surveys undetected for several years. Carr-Hill et al. describes the bias introduced to health data as a result of this exclusion as the "denominator problem" [5]. Prime examples of this dynamic can be seen in Chad, where an estimated 70% of pastoralist households were no longer present in the same location during a one-year time frame [1]. To provide context, the 2016 Ethiopian Demographic and Health Surveys (DHS), a paragon of health surveys in East Africa, used a sampling framework based on the 2007 census, thus likely capturing only a small fraction of NP settlements from the decade prior [6].

CS 229, 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Recently, remote sensing and satellite imagery have emerged as a valuable resource for monitoring urban development [7], tracking deforestation [8], and surveying objects in local neighborhoods [9]. In 2019, Wild et al. performed a comprehensive qualititative assessment of satellite imagery to help localize active nomadic pastoralist settlements for health survey design and intervention. Reported features that distinguish these settlements in satellite imagery include the presence of livestock circles, village huts, and burned land. Along with these distinctive features, promising results in using machine learning to detect local objects in satellite imagery reveal that a notable opportunity exists to locate and monitor nomadic pastoralists from satellite imagery using deep learning methods. Two key studies serve as a valuable reference to the viability of deep learning methods in localizing objects in satellite imagery with staggering precision.

In 2020, Sheng et al. developed a comprehensive database of U.S. oil refineries and demonstrated the success of deep learning approaches in detecting associated facilities from satellite imagery. Furthermore, from a nation-wide, random ablation study, their best model was able to detect gas infrastructure facilities previously undocumented in state-of-the-art government datasets. This study documents valuable components in the construction process of satellite imagery datasets, namely attending to cultivating both model precision and sensitivity [7]. A similar study by Sethi et al. explored deep learning classification methods of objects in local neighborhoods, providing insight into common model iteration frameworks and training paradigms. For example, the authors in this study describe the use of valuable metrics to perform error analysis along with advantages offered by specific computer vision model architectures tailored to satellite imagery applications [9]. These studies informed large components of both the dataset construction and model iteration processes in my study.

3 Dataset Generation Pipeline

To commence, I met with experts from the Stanford School of Medicine and Geographical Information Systems departments to learn about common challenges encountered in manually identifying nomadic pastoralist settlements from satellite imagery. Following this, I aided in the design, screening, and revision of settlement annotations. Annotations were gathered through surveying composites produced in both ArcGIS Pro and QGIS map visualization services. Metadata of annotated tiles were stored in the service and extracted for further dataset processing

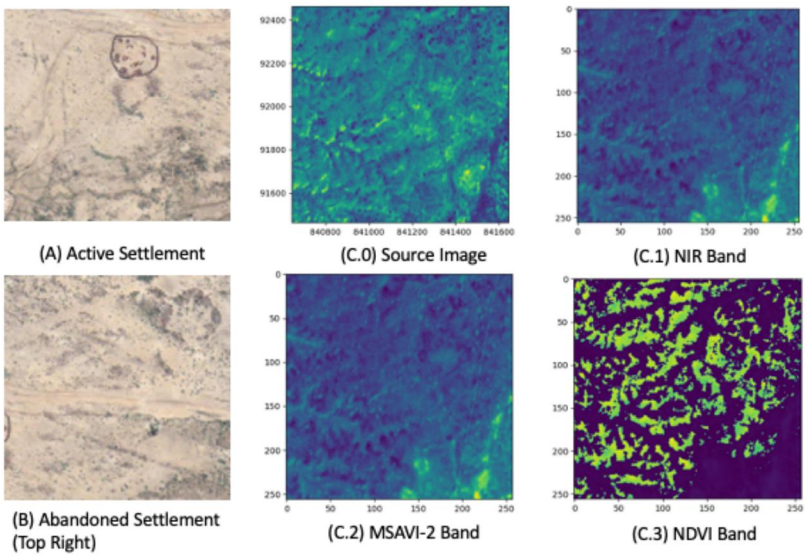

Figure 1: Comprehensive summary of engineered image features (C.0-C.3) and sample images of abandoned (A) and active (B) settlements.

Following the organization of settlement annotations, I engineered an end-to-end dataset generation pipeline to query images from select providers and process them for deep learning applications. Specifically, Planet Labs and QGIS were selected after a brief evaluation of satellite imagery services, in which raw images are supplied from the PlanetScope Dove Constellation (PDC) and the Sentinel-2 Satellite Missions (S2). Using the Planet Orders, Planet Data, and QGIS

Python APIs, I queried 700 images from periods of January, 2017 - February, 2021 directly from labels drawn around the Omo Valley of Ethiopia area of interest (AOI). I then composited, projected, and tiled these images into 1,200 256x256 4-band tiles used for model development. An additional 4,800 tiles were randomly sampled from neighboring regions to form a negative sample pool with similar terrain characteristics. Alongside this process, I compiled image metadata and engineered Normalized Difference Vegetation index (NDVI) and Modified Soil-Adjusted Vegetation Index (MSAVI2) features in final tile configurations. Tiles queried from Google Earth served as a human-in-the-loop component for dataset development, specifically due to the state-of-the-art resolution capabilities offered by Maxar imagery products. Finally, processed tiles were matched to their respective annotations using geographical projections and bounding area estimations.

Five classes, including 1) Inhabited Village, 2) Livestock Enclosure, 3) Building, 4) Uninhabited Village, and 5) Negative are present in our full annotations. To set up our binary classification task, I aggregated classes 1-4, thus pooling village features into one positive class. All code used to develop and parse the dataset is original work, and a visual summary of the dataset and its creation process can be seen in Fig. 1, 2, respectively.

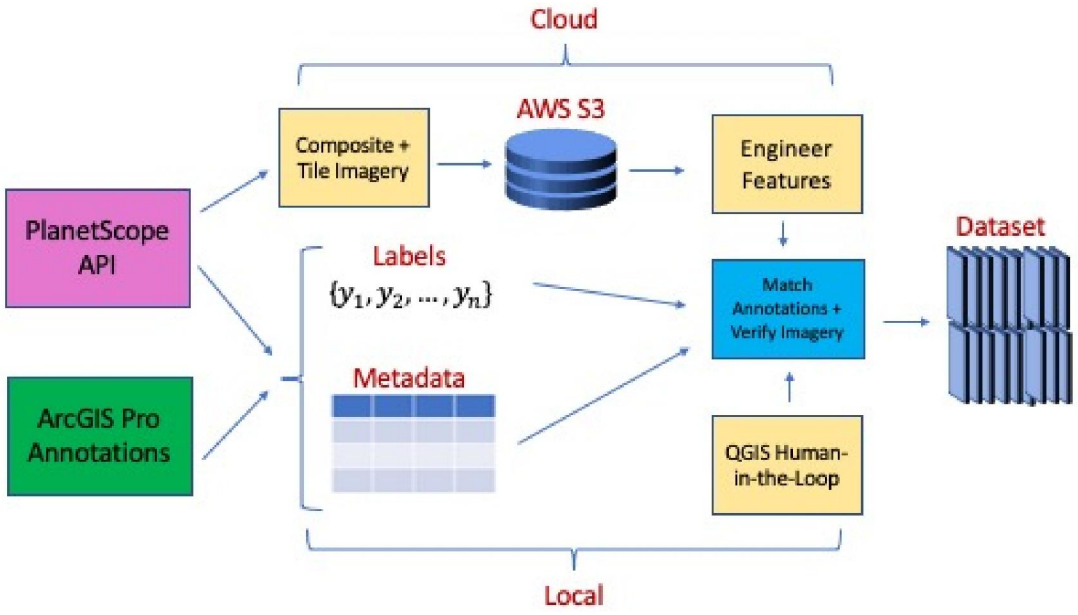

Figure 2: Diagram of dataset creation pipeline.

4 Methods

We performed an exploratory set of experiments on three state-of-the-art models, specifically DenseNet-121, ResNet-50, and EfficientNet-V2 using the RGB subset of bands from our images [10, 11]. To adapt these models to our binary classification task, we removed the top layer from each of these models and appended a four-layer fully connected network with dropout ($p = 0.5$) and batch normalization in between linear layers. In this set of initial experiments, we performed hyperparameter tuning before training on our full dataset. Following the evaluation of our trained models, we selected our best model for a follow-up ablation study.

During our ablation study, we created three variations of our initial dataset. Negative samples that we originally sampled from neighboring Ethiopian regions were characterized as "hard negatives," as they feature similar terrain and morphological features to those of NP settlements. We sampled an additional 4800 random tiles from around the United States midwest to form our "easy negatives" pool. This design choice was made to composite tiles filled with both vegetation and urban facilities in an effort to assess model sensitivity. With our original positive pool, we created three dataset variants: 1) H1: Positive samples and easy negatives; 2) H2: Positive samples and hard negatives; and 3) H3: Positive samples with both easy negatives and hard negatives. In our ablation study, we trained our best model on each dataset and evaluated subsequent performances

on dataset H3, our benchmark. We finally evaluated our highest performing model on H1, H2, and H3 to perform a fine-grained analysis of the model's sensitivity and specificity in real-world scenarios.

For training, we employed the Adam optimization algorithm [12] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We performed a randomized grid search to tune our learning rate, batch size, and pretrained weight inclusion hyperparameters with the Keras tuner class, over a three-fold, 5 epoch per configuration tuning setup. Optimization on batch size and learning rate were performed over a log scale interval of $(-2, -5)$ and a linear interval of $(8, 64)$ with a step size of 8, respectively. The model was trained on a binary cross entropy loss, $-\sum_i^C y_i log(f(x_i))$ for label y_i and example x_i, for 150 epochs with a patience of 25 epochs. Early stopping was evaluated on the basis of validation loss. Finally, horizontal and vertical random flipping along with random rotations were introduced at model run-time to reduce the risk of overfitting.

5 Results and Discussion

The results of my initial experiments are displayed in table 1. Following hyperparameter tuning, Resnet-50 achieved both the highest test accuracy and F1 score. Although the prevalance of positive examples in the test and train dataset are ≈ 0.1, notable precision and recall scores indicate that the model performed with reasonable specificity and sensitivity. In contrast, EfficientNet-B2 performed with notably lower accuracy and precision. Upon conducting error analysis, it was observed that the outputted predicted probabilities were extremely similar for many examples, thus signaling that the model struggled with differentiating examples with similar morphological and terrain-like characteristics. This might be explained by EfficientNet-B2's weaker relative, representational power, although it is noted that anomalies in training this model will be further explored in the future.

Model	AUROC	AUPRC	F1-Score	Precision	Recall	Accuracy
Resnet-50	0.937	0.869	0.829	0.792	0.860	0.927
DenseNet-121	0.888	0.796	0.756	0.783	0.721	0.904
EfficientNet-B1	0.620	0.244	0.416	0.289	0.730	0.588

Table 1: Table displaying results attained from initial experiments exploring the performance of state-of-the-art models on the original dataset. Results are reported for the best models following hyperparameter tuning of batch size, learning rate, and usage of pretrained weights.

Following initial experiments, Resnet-50 and its optimal hyperparameter configuration were selected for a comprehensive ablation study. In the first part of this study, our best model was trained on each of the dataset variants and evaluated on the H3 benchmark test set (table 2). As expected, the H1-trained model displayed the lowest accuracy with the highest recall, indicating its enhanced sensitivity to positive examples yet its inability to separate 'hard' negative examples. However, the out-performance of the H2-trained model on the H3-trained model indicates that the inclusion of 'easy' negative examples in training hinders the model's ability to discretize 'hard' negative examples overall. This finding suggests that deep learning models aiming to localize NP settlements must be exposed to negative examples with similar morphological characteristics to perform optimally. Furthermore the inclusion of negative examples with vastly different morphological features must be considered carefully in the trade-off of model precision.

Train Set	Train Prevalence	AUROC	AUPRC	F1-Score	Precision	Recall	Accuracy
H1	0.2	0.868	0.337	0.486	0.323	0.959	0.763
H2	0.2	0.957	0.846	0.829	0.838	0.811	0.960
H3	0.1	0.893	0.569	0.556	0.682	0.459	0.912

Table 2: Table displaying results from the first part of the ablation study. The best performing model in initial experiments, Resnet-50, was trained on each of the ablation dataset variants and tested on the combined H3 benchmark test set. Metrics are reported for each tested model.

In the second and final part of my ablation study, Resnet-50 trained on the H2 training set variant was tested on each test set variant to further evaluate its sensitivity and specificity. The optimal model performed well across all test set variants, notably achieving an accuracy of 0.96 and F1-score of 0.829 on the H3 test set (table 3). High recall and precision scores on the H1 test set indicate that the exposure of 'difficult' negative examples in training are effective in allowing the model to process 'easy' negative examples. High AUROC and AUPRC scores across all models provide further evidence of the model's precision and sensitivity robustness in imbalanced testing scenarios. Overall, the results obtained from initial experiments and the follow-up ablation study in this paper serve as promising baselines for the future development of NP localizing models.

Test Set	Test Prevalence	AUROC	AUPRC	F1-Score	Precision	Recall	Accuracy
H1	0.2	0.964	0.917	0.862	0.935	0.790	0.951
H2	0.2	0.937	0.869	0.829	0.792	0.860	0.927
H3	0.1	0.957	0.846	0.829	0.838	0.811	0.960

Table 3: Table displaying results from the second part of the ablation study evaluating the best performing model, Resnet-50 trained on the H2 dataset variant. As seen, the model achieves notable accuracies across all three dataset variants with high overall precision and recall metrics. Note that results obtained from the H1 test set row match those obtained from the initial experiment as H1 is composed of the original dataset. Similarly, results from the H3 row match those obtained from the benchmark test in the first part of the ablation study.

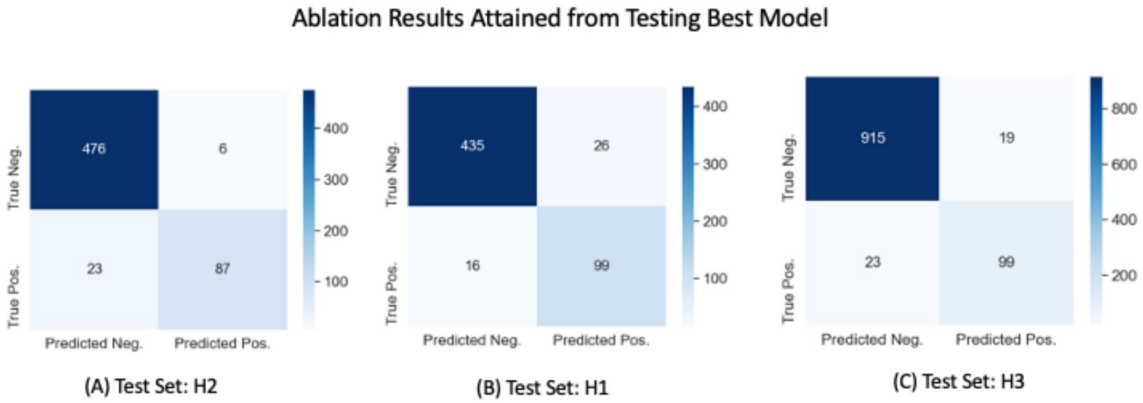

Figure 3: Confusion matrices attained from testing the best model achieved over the ablation study, tuned Resnet-50 model trained on H2 dataset variant, on all dataset variants.

6 Conclusion and Future Work

This study features three novel contributions to the field of deep learning for localizing nomadic pastoralist (NP) populations. Firstly, a novel dataset was created, aggregating ≈ 1200 NP settlements in Nyangatom, Ethiopia along with ≈ 4800 negative examples featuring similar terrain and morphological features. Comprehensive tests with state-of-the-art computer vision models and a follow-up ablation study provide strong evidence for the viability of this dataset as a source for future deep learning experiments and a basis for training models that aim to localize NP settlements globally. Secondly, a comprehensive ablation study was performed, providing insights into training paradigms that are specific to tuning effective models for NP localization. Specifically, it was discovered that the exposure to 'hard' negative examples in training is paramount to a model's precision across diverse satellite images, generalizing well to examples with more diverse morphological features (i.e. vegetation, urban features). In addition, it was observed that exposure to vastly different, 'easy' negative examples in training notably lowers a model's precision and recall, indicating that NP settlement features are fine-grained and sensitive to small perturbations in model interpretation. Our third and final contribution is the attainment of valuable model performance baselines for future studies across three dataset variants. In particular, our best performing model achieves an accuracy, precision, and recall of 0.96, 0.838, and 0.811, respectively, among a test set comprising a diverse makeup

of negative examples. These results are a promising step toward excelling deep learning efforts to localize nomadic pastoralists, a marginalized and largely forgotten population, now receiving a helping hand.

7 Contributions and Additional Comments

I wrote original code for the dataset creation pipeline, interfacing with both QGIS and planet-labs APIs. In addition, I wrote extensive code to parse geographical metadata and unify image metadata with label metadata to create the unified dataset. On the front of model testing, I devised all experiments, including the initial model exploration, hyperparameter tuning, and the ablation study. Finally, I conducted analysis of my results and formulated the idea to evaluate model sensitivity and specificity with dataset variants. I received guidance from mentor Stace Maples on all questions related to NP identification and deployment considerations. In addition, I received extensive help from Justin Kong of the Stanford Geogrphical Information Systems center in labeling NP settlements. This work is part of a research collaboration that I have with the Stanford School of Medicine, for which I am leading all computational efforts; all work in this write-up is reflective of work that I have done this quarter alone. As a note of clarification, I excluded results obtained from evaluating VGG-16, as initially indicated in my proposal, due to a result metric discrepancy that I was not able to debug.

References

[1] E Schelling, C Diguimbaye, S Daoud, J Nicolet, P Boerlin, M Tanner, and J Zinsstag. Brucellosis and q-fever seroprevalences of nomadic pastoralists and their livestock in chad. *Preventive veterinary medicine*, 61(4):279–293, 2003.

[2] Maria E Fernandez-Gimenez. The role of mongolian nomadic pastoralists'ecological knowledge in rangeland management. *Ecological applications*, 10(5):1318–1326, 2000.

[3] Rada Dyson-Hudson and Neville Dyson-Hudson. Nomadic pastoralism. *Annual review of anthropology*, 9(1):15–61, 1980.

[4] Hannah Wild, Luke Glowacki, Stace Maples, Iván Mejía-Guevara, Amy Krystosik, Matthew H Bonds, Abiy Hiruy, A Desiree LaBeaud, and Michele Barry. Making pastoralists count: geospatial methods for the health surveillance of nomadic populations. *The American journal of tropical medicine and hygiene*, 101(3):661–669, 2019.

[5] Roy Carr-Hill. Measuring development progress in africa: The denominator problem. *Canadian Journal of Development Studies/Revue canadienne d'études du développement*, 35(1):136–154, 2014.

[6] Sara Randall. Where have all the nomads gone? fifty years of statistical and demographic invisibilities of african mobile pastoralists. *Pastoralism*, 5(1):1–22, 2015.

[7] Hao Sheng, Jeremy Irvin, Sasankh Munukutla, Shawn Zhang, Christopher Cross, Kyle Story, Rose Rustowicz, Cooper Elsworth, Zutao Yang, Mark Omara, et al. Ognet: Towards a global oil and gas infrastructure database using deep learning on remotely sensed imagery. *arXiv preprint arXiv:2011.07227*, 2020.

[8] Jeremy Irvin, Hao Sheng, Neel Ramachandran, Sonja Johnson-Yu, Sharon Zhou, Kyle Story, Rose Rustowicz, Cooper Elsworth, Kemen Austin, and Andrew Y Ng. Forestnet: Classifying drivers of deforestation in indonesia using deep learning on satellite imagery. *arXiv preprint arXiv:2011.05479*, 2020.

[9] Adam Van Etten. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv preprint arXiv:1805.09512*, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.