

SLVVA: Scalable Land Viability via Vision-Language Architecture

Vishvam Porwal¹, Stacey D. Scott¹, Neil D. B. Bruce¹, Asim Biswas²

¹*School of Computer Science*

²*School of Environmental Science*

University of Guelph

Guelph, Canada

Emails: {porwalv, stacey.scott, brucen, biswas}@uoguelph.ca

Abstract—Digital soil mapping is a process of creating maps of soil properties and their spatial distribution. It plays a vital role in monitoring soil health and promoting sustainable and efficient land use. In the past, environmental data was used to guide the creation of soil property maps. However, the failure to consider the accessibility of locations has led to a bias in the mapping process. In our research, we utilize satellite imagery to evaluate location accessibility, leading to more balanced soil property mapping. We formulate land viability detection and introduce a scalable two-step framework for its detection. Initially, we classify land viability, followed by its segmentation. We leverage Convolutional Neural Networks (CNNs) for classification and a resilient and generalizable vision-language architecture for segmentation. Our most notable results stem from fine-tuning a pre-existing VGGNet for classification and employing a CLIP-based Segmentation method (CLIPSeg) for segmentation. We demonstrate the effectiveness of our approach through extensive experimentation on EuroSAT and OpenEarthMap datasets. Our work is the first to address the challenge of biased sampling in digital soil mapping by incorporating satellite images to assess the accessibility of locations, ensuring a more representative soil property mapping.

Keywords—Vision Language models, Multi-modal architectures, Convolutional Neural Networks, Digital Soil Mapping, Land Use Classification, Land Use Segmentation, Remote Sensing

I. INTRODUCTION

The advent of digital soil mapping has significantly streamlined the process of mapping soil characteristics, making it much more efficient [1]. However, due to the high variance in soil properties even in a small area, the process of creating maps is still quite challenging. Substantial prior research has been done to find the optimal number of samples needed and the coordinates of the samples to be collected to create a map that captures the most variance in the soil properties. Notably, Saurette et al. [2] proposed a method to find the optimal number of samples needed to create a map, and Zhang et al. [3] proposed a method to find the optimal coordinates of the samples to be collected. These approaches use environmental data and statistical techniques to accomplish the respective tasks. However, the environmental data fails to determine the accessibility of the location. This leads to a bias in the results, as samples are only collected from easily accessible locations. This bias

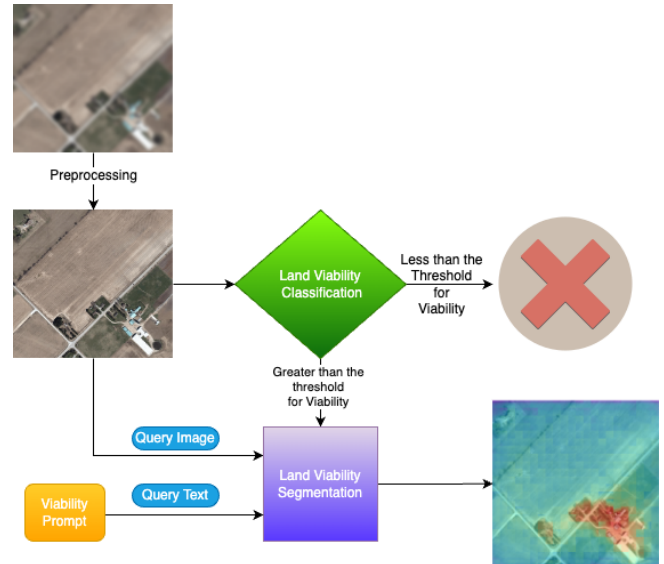


Figure 1. Generalized architecture for land viability detection. First, the satellite image is passed through a preprocessing step to enhance the image. Then, the enhanced image is passed through a land viability classification model to classify the land viability along with the confidence score. If the confidence score is above a threshold, then the image is passed through a land viability segmentation model to get the segmentation mask.

can be mitigated by using satellite images to determine the accessibility of different locations.

In this paper, we introduce a method to assess location accessibility using satellite imagery. The proposed method uses a combination of vision-based classification and segmentation techniques in a scalable manner. We utilize Deep Convolutional Neural Networks for classification. Additionally, we incorporate multiple modalities to improve the segmentation results. To accomplish this, we introduce a variation of a CLIP-based Segmentation approach (CLIPSeg) [4], [5] to generalize the approach to unseen entities in remote sensing. CLIP [4] was used as it has been shown to generalize well to elements not previously seen during training [38]. CLIPSeg originally used a U-Net-inspired [6] decoder to learn the activations of CLIP and produce a segmentation mask. Furthermore, during the training of their decoder, the CLIP model is frozen and either image or text is passed

through the CLIP model to produce the activations. In this paper, we find that only using text activations is sufficient to reach competitive results. Additionally, we find that fine-tuning the CLIP model on the EuroSAT dataset [7] improves the segmentation performance.

Our main contributions are:

- a novel, generalized multi-modal architecture (Fig. 1) designed specifically for detecting land viability,
- a highly scalable architecture capable of performing global land viability detection, and
- a proposed CLIPSeg variation in the context of remote sensing, that underscores robustness achieved through text modality activations.

II. RELATED WORK

Our work intersects two distinct fields of study: land use classification and land use segmentation, with the application of vision language models. In this section, we provide a brief overview of the related work in these fields.

A. Land Use Classification

Land use classification is a process of classifying the land use of an area into different categories. It is a well-studied problem in the field of remote sensing. In this field of study, entire images are classified into different categories. This problem has been approached using different techniques, which are as follows:

1) *Handcrafted Features*: One of the most popular techniques is to use handcrafted features and train a classifier using them. Notably, Chen et al. [8] used Gabor-filtering-based completed local binary patterns (GCLBP) to extract features from the image and then trained a Kernel Extreme Learning Machine (KELM) on top of them. Similarly, Yang and Newsam [9] compared Scale-Invariant Feature Transform (SIFT) [10] and Gabor texture features (Newsam et al.) [11] and trained a Support Vector Machine (SVM) on top of them. However, it is difficult to adapt these methods and they tend to underperform when applied to data distributions from geographical areas not covered in their training phase.

2) *Convolutional Neural Networks (CNN)*: Recent research has focused largely on the use of Deep Convolutional Neural Networks to classify the land use of an image. A study conducted by Li et al. [12] used AlexNet [13], VGGNet [14], GoogLeNet [15], and ResNet [16] to classify the land use of an image. They found that these models perform well in comparison to handcrafted features. Stoimchev et al. [17] used a unique method of fine-tuning each layer of pre-trained CNN models, trained on ImageNet [18], to extract relevant features. They then utilized these features to train a decision tree classifier.

3) *Vision Transformer (ViT)*: In addition to CNNs, Vision Transformers have also been used to classify the land use of an image. Transformers, known for their ability to capture long-range dependencies, have been widely used in Natural

Language Processing (NLP) since their introduction in 2017 [19]. Recently, they have been adapted to the field of Computer Vision [20] and have shown promising results. In one study, Bazi et al. [21] used ViT to classify the land use of an image. They found that ViT performs better than CNNs in this task. However, ViT is not as efficient as CNNs and requires substantial resources to train.

Overall, CNNs have been the most popular choice for land use classification.

B. Land Use Segmentation

Land use segmentation entails the same problem as land use classification, but instead of classifying the entire image, it assigns a class to each pixel of the image. Common applications of this problem are to separate roads from buildings, water from land, and identify disaster-affected areas. This problem has been approached using different techniques, which are as follows.

1) *Convolutional Neural Networks (CNN)*: Similar to land use classification, CNNs have been one of the most popular choices for land use segmentation. Different architectures have been proposed to solve this problem. Notably, Långkvist et al. [22] employed multiple parallel CNNs with varying contextual input sizes. Their outputs were then combined with a fully connected layer to produce the final segmentation. Another interesting approach, discussed by Brown et al. [23], involves using a Fully Convolutional Network (FCN) [24] to produce a segmentation mask. They perform a series of pre-processing steps to enhance the image before passing it through the FCN. Finally, the most popular approach, discussed by Singh and Nongmeikapam [25], involves using a U-Net [6] architecture to produce a segmentation mask. However, these approaches do not generalize well to unseen classes.

2) *Vision Transformer (ViT)*: Recently, ViT [20] has been adapted to the field of land use segmentation. Panboonyuen et al. [26] used a combination of ViT and mixed-scale convolutional feed-forward networks to produce a segmentation mask. Lin et al. [27] utilized the SegFormer [28] architecture to produce a segmentation mask. However, these approaches also do not generalize well to unseen classes.

C. Vision-Language Models

Vision-language models are multi-modal architectures that combine image and text modalities to perform a task. These models have been employed in diverse areas such as image captioning, visual question answering, and visual reasoning, and have recently found applications in remote sensing. For instance, Alsaleh et al. [29] utilized image and text encoders along with a decoder for visual question answering in remote sensing. Jiang et al. [30] leveraged a language-guided framework for semantic segmentation in remote sensing, using a linguistic mode with prior knowledge and visual-linguistic alignment in a shared space to generate a segmen-

tation mask. Beyond remote sensing, CLIPSeg [5], a vision-language model, has been used for semantic segmentation, using a U-Net-inspired decoder to learn CLIP activations and create a segmentation mask. Our study incorporates a variant of CLIPSeg.

III. LAND VIABILITY DETECTION METHOD

We implemented an end-to-end pipeline to detect the viability of an area using satellite images. The pipeline consists of two main components: land use classification and land use segmentation. The main reason to first classify the land use of an area is to narrow down the search space. Then, the land use segmentation can be done to find the viability of the area. The structure of this pipeline is depicted in Fig. 1, which we will discuss in more detail in this section.

A. Experimental Setup

To implement this pipeline and conduct our experiments, we utilized an M2 MacBook Pro equipped with 32GB RAM, 19 core CPUs, and an M2 max chip for both training and inference stages. The software environment for this research was set up with PyTorch version 2.1.1 and Python version 3.11.6.

B. Data and Preprocessing

A common problem in the field of remote sensing is the variation in the resolution of the images [31]. This inconsistency can be traced to several factors, including the fact that images are often sourced from a variety of satellites and aerial vehicles. Hence, we need to preprocess and augment the images to make them consistent for both training and testing phases.

1) *Data for Land Use Classification:* We used the EuroSAT dataset [7] for training and testing our classification models. It consists of 27,000 labeled images of 10 different land use classes (Table I), with an image size of 64x64 pixels. We split the dataset into 60% training, 20% validation, and 20% testing. We use the same split for all our experiments. We have aggregated the classes into two classes: viable and non-viable for our soil sampling goals. The viable class consists of AnnualCrop, Forest, HerbaceousVegetation, PermanentCrop, and Pasture classes. The non-viable class consists of the Highway, Industrial, Residential, River, and SeaLake classes. These aggregated classes are used for further steps in the pipeline.

2) *Data for Land Use Segmentation:* We used the OpenEarthMap dataset [32] for training and testing our segmentation models. It consists of satellite and aerial images across 6 continents. The images are labeled with 8 different land use classes (Table II), with an image size of 1024x1024 pixels. This dataset contains a total of 5,000 images. We split the dataset into 80% training, 10% validation, and 10% testing. This split is motivated by the small size of the dataset. Again, we aggregated the classes into two classes: viable

Table I
EUROSAT DATASET (27000 IMAGES)

Class	Aggregated Class	Total Images (%)
AnnualCrop	Viable	11.1
Forest	Viable	11.1
Pasture	Viable	11.1
PermanentCrop	Viable	9.2
HerbaceousVegetation	Viable	11.1
Highway	Non-Viable	9.2
Industrial	Non-Viable	9.2
Residential	Non-Viable	11.1
River	Non-Viable	9.2
SeaLake	Non-Viable	11.1

and non-viable. The viable class consists of Rangeland, Tree, and Agricultural Land classes. The non-viable class consists of Bareland, Developed Space, Road, Water, and Building classes. These aggregated classes are used for further steps in the pipeline.

Table II
OPENEARTHMAP DATASET (5000 IMAGES)

Class	Aggregated Class	Total Images (%)
Rangeland	Viable	22.9%
Tree	Viable	20.2%
Agricultural Land	Viable	13.7%
Water	Non-Viable	3.3%
Building	Non-Viable	15.6%
Bareland	Non-Viable	1.5%
Developed Space	Non-Viable	16.1%
Road	Non-Viable	6.7%

3) *Preprocessing during Training:* We used similar preprocessing steps before training for both datasets. For the EuroSAT dataset, we resized the images to 56x56 and for the OpenEarthMap dataset, we randomly cropped the images to 352x352. We then normalized the images using the mean and standard deviation of the pixels of images in the dataset.

Moreover, due to the small size of the OpenEarthMap dataset, we used data augmentation to increase the size of the dataset. We used random horizontal and vertical flips, and random rotations of 180 degrees to augment the dataset. No additional augmentation was applied to the EuroSAT dataset given its sufficient size.

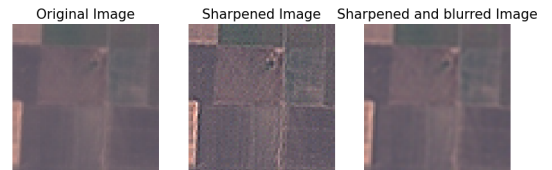


Figure 2. Preprocessing steps for enhancing the image during inference. First, the sharpening filter is applied to enhance the edges of the image. Then, the Gaussian blur is applied to remove the noise from the image.

4) *Preprocessing during Inference:* We used the same preprocessing steps during inference for both datasets. In our experiments, we found that applying these preprocessing steps improved the performance. We resized the images to 56x56 for classification and 352x352 for segmentation, normalized them using the mean and standard deviation of the pixels of images in the dataset, and then applied a sharpening filter and Gaussian blur. While a CNN could theoretically learn to perform similar tasks, in practice, we found that doing this preprocessing leads to better results. The steps we took for preprocessing are illustrated in Fig. 2.

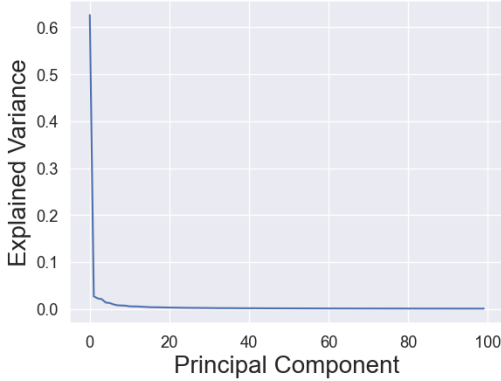


Figure 3. Explained variance ratio of the principal components of the EuroSAT dataset. The first ten principal components explain more than 90% of the variance in the dataset.

C. Land Viability Classification

We employ land viability classification to narrow down the search space for land viability segmentation. We did this by classifying the land viability of an area into two classes: viable and non-viable along with the confidence score. If the confidence score for land viability exceeds 80%, we forward the image to the land viability segmentation stage to generate a segmentation mask. Images falling below this threshold are discarded. We have strategically set the confidence score threshold at 80% to maintain a balance. This ensures we are focusing on images with a high probability of land viability, while also retaining enough images for the segmentation stage. This method allows us to direct our resources towards the most promising images, optimizing the overall process. We experimented with different classification models, as follows:

1) *Preliminary Experiments: SVM and a Bespoke CNN:* In our initial experiments, we explored the use of a Support Vector Machine (SVM) and a bespoke Convolutional Neural Network (CNN).

For the SVM, we used the first ten principal components as features, which captured the most variance in the dataset (Fig. 3). We trained the SVM with a Radial Basis Function (RBF) kernel and used 5-fold cross-validation to obtain the

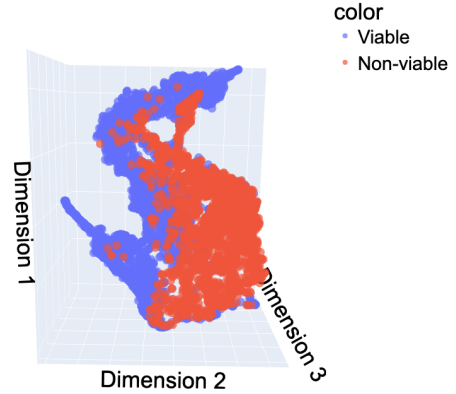


Figure 4. T-SNE [35] visualization of the activations from the last hidden layer of the CNN model. The model is not able to learn to get the best feature representation of the data as the classes are not separated.

confidence score. However, the SVM did not effectively learn the optimal decision boundary for the data.

We also built a CNN, considering the spatial characteristics observed in satellite images. The CNN had a bespoke architecture with 3 convolutional layers, 3 max-pooling layers, and 2 fully connected layers, all using the ReLU activation function. It was trained using the Adam optimizer [34] and an early stopping mechanism was implemented. However, it did not learn the optimal feature representation of the data (Fig. 4). Given these results, we decided to explore the use of a pre-trained VGGNet for our task.

2) *Fine-tuning a Pre-trained VGGNet:* We used VGGNet [14], a widely recognized CNN architecture frequently employed in land use classification [12] to assess the land viability of a given area. We take advantage of the pre-trained VGG16 model, which we subsequently fine-tune. The model is trained using the SGD optimizer [36], with a learning rate of 0.001 and a momentum of 0.9. We utilized the Cross-Entropy Loss function during the training process. The model is trained over 15 epochs with a batch size of 64. To optimize the process, we use the same early stopping mechanism that halts the training if there is no improvement in the validation loss over 5 consecutive epochs. This approach delivers the best results, leading us to choose this model for the following steps in our pipeline. Its superior performance can be attributed to its pre-training on ImageNet, which allowed it to learn a more sophisticated feature representation of the data.

D. Land Viability Segmentation

In our research, we devised a robust and generalized approach for land viability segmentation. We employed a multi-modal architecture that builds upon the foundational

capabilities of CLIP [4]. We adopted the CLIPSeg model [5] to delineate the land viability within a specific area. The CLIPSeg model utilized a U-Net-inspired decoder to learn the activations of CLIP and generate a segmentation mask. During training, the CLIP model was frozen and either an image or text was passed through it to obtain the activations. Only the activations of three specific layers (layers 3, 6, and 9) of the CLIP model were used for training the decoder, resulting in a three-layer decoder.

Following a similar approach, we used the activations of the CLIP model to train our decoder. However, we only utilized the activations from the text modality of CLIP. This choice is motivated by the fact that the text modality is more robust than the image modality. Small changes in the image can lead to different activations in the CLIP model, whereas the text modality is more resilient to such variations. As a result, we supply the segmentation target to the decoder based on the text linked with the image. In alignment with the CLIPSeg model, we employed CLIP ViT-B/16. The training of the decoder is carried out over 10 epochs with a batch size of 64, using the AdamW [34] optimizer and a learning rate of $3e-4$. We further employed Cross-Entropy Loss with logits for the decoder's training.

To improve our results, we fine-tuned the CLIP model on the EuroSAT dataset. This refinement process improved CLIP's capability to produce more accurate activations, specifically enhancing its understanding and differentiation between "viability" and "non-viability" prompts in the context of land assessment. We fine-tuned the CLIP model for 10 epochs with a batch size of 64. We used the AdamW [37] optimizer with a learning rate of $2e-6$. The choice of a smaller learning rate was strategic, and aimed at maintaining the wide-ranging knowledge that CLIP had already acquired. After fine-tuning the CLIP model, we followed the same approach as described in the previous section to train our decoder. We used the activations from the fine-tuned CLIP model, specifically from the text modality, to train the decoder.

IV. RESULTS AND DISCUSSION

Our approach was assessed using designated test image batches from the EuroSAT and OpenEarthMap datasets. Given that our methodology involves a mix of classification and segmentation models, we conduct separate evaluations for each.

A. Land Viability Classification

The performance of the models was assessed using several common metrics in classification tasks:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Area Under the Receiver Operating Characteristic Curve (AUC)
- Sensitivity: $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FP}$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. These metrics were reported for all our experiments and the results are presented in Table III. We find that the fine-tuned VGG16 model outperformed the other models in our experiments. The model's performance is illustrated through the confusion matrix, ROC curve, and precision-recall curve in Fig. 5.

Table III
LAND VIABILITY CLASSIFICATION RESULTS

Approach	Accuracy	AUC	Specificity	Sensitivity
SVM	0.83	0.82	0.80	0.84
CNN	0.93	0.92	0.94	0.90
VGG16	0.99	0.99	0.99	0.98

B. Land Viability Segmentation

In segmentation, Intersection over Union (IoU) is the most common metric used to evaluate the performance of the model. The IoU is calculated as: $\frac{\text{Area of Overlap}}{\text{Area of Union}}$

where the Area of Overlap is the area of overlap between the predicted mask and the ground truth mask, and the Area of Union is the area of union between the predicted mask and the ground truth mask. However, IoU needs a threshold to be set to determine whether a pixel belongs to a class or not. We experimented with different pixel thresholds and report the results for the best threshold. Moreover, since we are only interested in foreground pixels, we report the results for the foreground pixels only (IoU-F), and we also report the results for the average of foreground and background pixels (IoU-Bin). The results of these evaluations are presented in Table IV.

Table IV
LAND VIABILITY SEGMENTATION RESULTS

Layers of CLIP	CLIP Fine-tuned	IoU-F	IoU-Bin
3 and 9	No	0.38	0.37
3, 7, and 9	No	0.39	0.40
3, 6, and 9	No	0.60	0.58
3 and 9	Yes	0.46	0.45
3, 7, and 9	Yes	0.51	0.50
3, 6, and 9	Yes	0.79	0.78

We experimented with different layers of CLIP to find the best combination of layers. We found that the CLIPSeg model with layers 3, 6, and 9 of CLIP performs the best. It is interesting to note that CLIPSeg originally used layers 3, 7, and 9 of CLIP to get activations and we found that extraction of layers 3, 6, and 9 of CLIP performs better in this context. Moreover, we found that CLIPSeg results in an IoU-F of 0.53 on the OpenEarthMap dataset, whereas our approach resulted in an IoU-F of 0.79. The underlying reason for this is that we fine-tuned the CLIP model on the EuroSAT dataset, which allowed us to get better activations for the decoder.

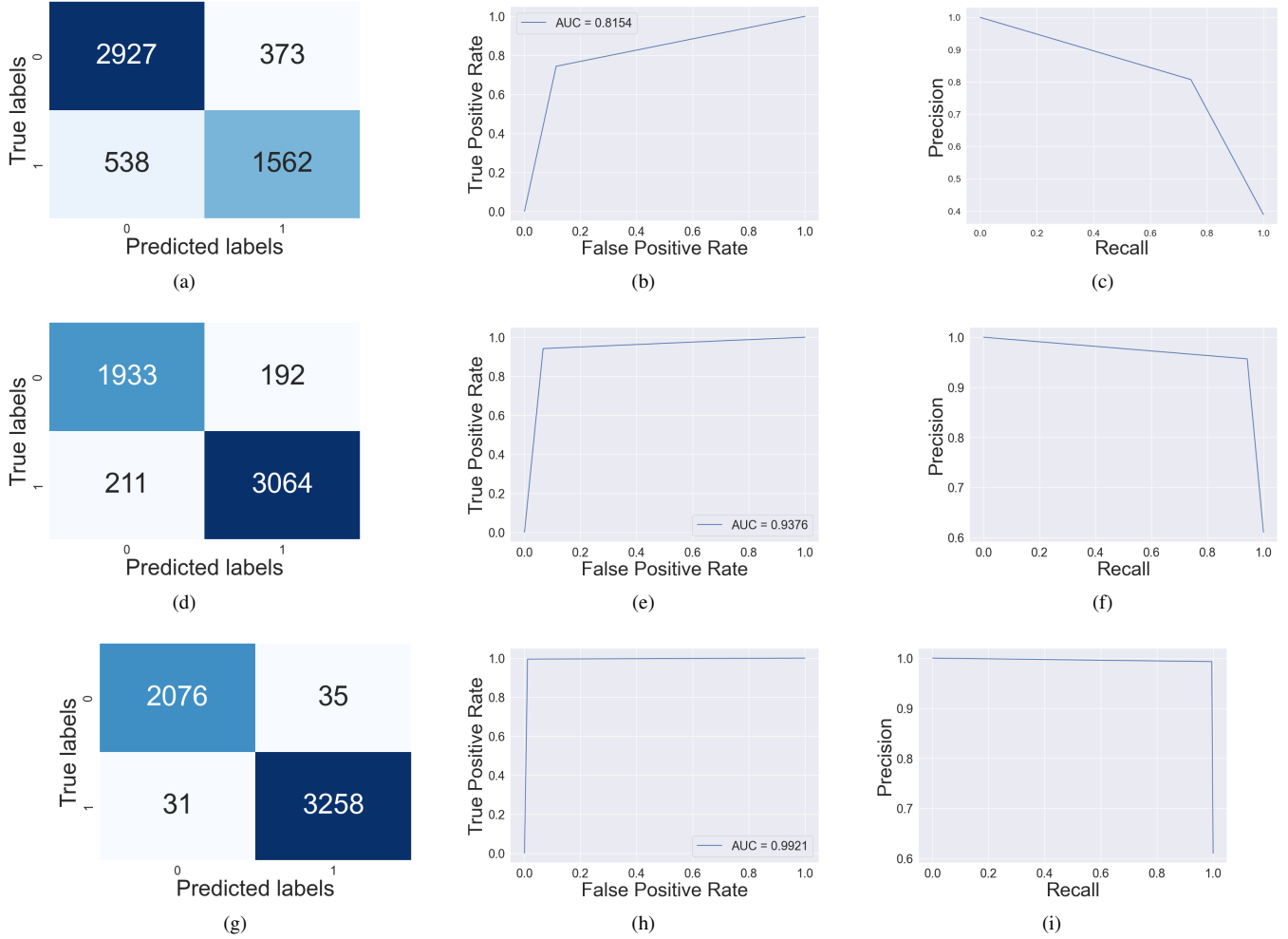


Figure 5. The EuroSAT dataset was used to evaluate SVM, CNN, and VGGNet models using Confusion Matrix, ROC Curve, and Precision-Recall Curve. The respective metrics for the models are shown in (a-c), (d-f), and (g-i).

C. Generalization to Unseen Classes

Given that CLIP is trained on a vast collection of internet-sourced images and text, it possesses the ability to generalize to classes it has not encountered before. This makes it a suitable choice as an activation generator for segmentation, particularly when dealing with classes outside the usual distribution. It is worth noting that CLIP's ability to generalize effectively is not hampered even when there is a minimal similarity between the training and testing classes [38].

D. Implications of Land Viability Detection

Land viability detection plays a crucial role in enhancing the accuracy of digital soil mapping. By assessing the viability of an area and selectively sampling it, we ensure that samples are collected only from viable areas. It is important to remember that there is a human element in digital soil mapping. This means that the results are reviewed by humans, ensuring a high level of accuracy before the

sampling process begins. In this context, an Intersection over Union - F score (IoU-F) of 0.79 is likely adequate but an in-field validation is still needed. This score represents a good balance of precision and recall, which is assumed to be sufficient when human verification is part of the process. This strategy is not only efficient but also scalable, making it well-suited for large-scale operations.

E. Limitations

A limitation of our approach is the dependency of the segmentation model on the image resolution. The quality of the segmentation mask inherently improves with higher-resolution images. Despite implementing preprocessing measures, this issue persists. Additionally, the presence of shadows in the images can potentially hinder the accuracy of our results. These challenges provide a valuable foundation for future research in this domain.

V. CONCLUSION

In this research, we have developed a new framework for detecting land viability using satellite images, which improves the accuracy of digital soil mapping. This achievement is marked by our introduction of a unique, generalized architecture. Our approach involves a two-step process: land viability classification and land viability segmentation. The first step uses Convolutional Neural Networks for classification, effectively narrowing down the search space and enhancing the scalability of our approach. This is followed by the land viability segmentation step, where we employ a Multi-modal Segmentation method.

Adding a novel twist to our approach, we have proposed a CLIPSeg variant that uses only text to generate the activations for training the decoder. This approach proves to be robust in the field of remote sensing, as even minor image alterations can result in different activations. The effectiveness of this approach was confirmed through tests conducted on the EuroSAT and OpenEarthMap datasets. The fine-tuned VGGNet excelled in classification, and the CLIPSeg decoder, trained on layers 3, 6, and 9 of the fine-tuned CLIP, was the top performer in segmentation. Our approach addresses digital soil mapping's biased sampling issue by using satellite images to assess location accessibility, ensuring balanced soil property mapping.

Future work could focus on improving the performance of the segmentation model by using state-of-the-art text and image encoders to generate better activations. Additionally, the model's effectiveness in handling unseen classes and diverse geographical regions could be further explored and enhanced. Future research could also involve comparing our segmentation approach with other segmentation methods to evaluate its relative performance. A post-processing step could also be added to the pipeline to remove the noise from the segmentation mask. In conclusion, our research offers a robust and scalable technique for land viability detection, contributing to the ongoing advancements in digital soil mapping. This opens up new opportunities for further research and improvements in this field.

ACKNOWLEDGMENT

The research was partially funded by the Mitacs Globalink Graduate Fellowship and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

- [1] B. Heung, D. Saurette, and C. E. Bulmer, "Digital Soil Mapping," *Usask.ca*, Aug. 12, 2021. Accessed: Feb. 02, 2024. [Online]. Available: <https://openpress.usask.ca/soilscience/chapter/digital-soil-mapping/>
- [2] D. D. Saurette et al., "Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon," *Geoderma*, vol. 425, p. 116054, Nov. 2022, doi: 10.1016/j.geoderma.2022.116054.
- [3] Y. Zhang, D. D. Saurette, T. H. Easher, W. Ji, V. I. Adamchuk, and A. Biswas, "Comparison of sampling designs for calibrating digital soil maps at multiple depths," *Pedosphere*, vol. 32, no. 4, pp. 588-601, Aug. 2022, doi: 10.1016/S1002-0160(21)60055-3.
- [4] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision." *arXiv*, Feb. 26, 2021. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/2103.00020>
- [5] T. Luddecke and A. Ecker, "Image Segmentation Using Text and Image Prompts," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 7076-7086. doi: 10.1109/CVPR52688.2022.00695.
- [6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation." *arXiv*, May 18, 2015. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [7] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 12, no. 7, pp. 2217-2226, Jul. 2019, doi: 10.1109/JSTARS.2019.2918242.
- [8] C. Chen, L. Zhou, J. Guo, W. Li, H. Su, and F. Guo, "Gabor-Filtering-Based Completed Local Binary Patterns for Land-Use Scene Classification," in *2015 IEEE International Conference on Multimedia Big Data*, Beijing, China: IEEE, Apr. 2015, pp. 324-329. doi: 10.1109/BigMM.2015.23.
- [9] Y. Yang and S. Newsam, "Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery," in *2008 15th IEEE International Conference on Image Processing*, San Diego, CA, USA: IEEE, 2008, pp. 1852-1855. doi: 10.1109/ICIP.2008.4712139.
- [10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
- [11] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Journal of Applied Optics: Information Processing*, vol. 43, no. 2, pp. 210-217, 2004.
- [12] H. Li et al., "RSI-CB: A Large-Scale Remote Sensing Image Classification Benchmark Using Crowdsourced Data," *Sensors*, vol. 20, no. 6, p. 1594, Mar. 2020, doi: 10.3390/s20061594.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84-90, May 2017, doi: 10.1145/3065386.
- [14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv*, Apr. 10, 2015. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1409.1556>

- [15] C. Szegedy et al., "Going Deeper with Convolutions." arXiv, Sep. 16, 2014. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1409.4842>
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 10, 2015. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [17] M. Stoimchev, D. Koccev, and S. Džeroski, "Deep Network Architectures as Feature Extractors for Multi-Label Classification of Remote Sensing Images," *Remote Sensing*, vol. 15, no. 2, p. 538, Jan. 2023, doi: 10.3390/rs15020538.
- [18] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL: IEEE, Jun. 2009, pp. 248-255. doi: 10.1109/CVPR.2009.5206848.
- [19] A. Vaswani et al., "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Apr. 29, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- [20] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/2010.11929>
- [21] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision Transformers for Remote Sensing Image Classification," *Remote Sensing*, vol. 13, no. 3, p. 516, Feb. 2021, doi: 10.3390/rs13030516.
- [22] M. Långkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks," *Remote Sensing*, vol. 8, no. 4, p. 329, Apr. 2016, doi: 10.3390/rs8040329.
- [23] C. F. Brown et al., "Dynamic World, Near real-time global 10 m land use land cover mapping," *Sci Data*, vol. 9, no. 1, p. 251, Jun. 2022, doi: 10.1038/s41597-022-01307-4.
- [24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 3431-3440. doi: 10.1109/CVPR.2015.7298965.
- [25] N. J. Singh and K. Nongmeikapam, "Semantic Segmentation of Satellite Images Using Deep-Unet," *Arab J Sci Eng*, vol. 48, no. 2, pp. 1193-1205, Feb. 2023, doi: 10.1007/s13369-022-06734-4.
- [26] T. Panboonyuen, C. Charoenphon, and C. Satirapod, "MeViT: Medium-Resolution Vision Transformer for Semantic Segmentation on Landsat Satellite Imagery for Agriculture in Thailand," In Review, preprint, May 2023. doi: 10.21203/rs.3.rs-2945208/v1.
- [27] X. Lin et al., "Semantic Segmentation of China's Coastal Wetlands Based on Sentinel-2 and Segformer," *Remote Sensing*, vol. 15, no. 15, p. 3714, Jul. 2023, doi: 10.3390/rs15153714.
- [28] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." arXiv, Oct. 28, 2021. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/2105.15203>
- [29] S. O. Alsaleh, Y. Bazi, M. M. Al Rahhal, and M. Al Zuair, "Open-Ended Visual Question Answering Model For Remote Sensing Images," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 2848-2851. doi: 10.1109/IGARSS46834.2022.9884295.
- [30] X. Jiang, N. Zhou, and X. Li, "Few-Shot Segmentation of Remote Sensing Images Using Deep Metric Learning," *IEEE Geosci. Remote Sensing Lett.*, vol. 19, pp. 1-5, 2022, doi: 10.1109/LGRS.2022.3154402.
- [31] M. Liu et al., "The Impact of Spatial Resolution on the Classification of Vegetation Types in Highly Fragmented Planting Areas Based on Unmanned Aerial Vehicle Hyperspectral Images," *Remote Sensing*, vol. 12, no. 1, p. 146, Jan. 2020, doi: 10.3390/rs12010146.
- [32] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping." arXiv, Oct. 19, 2022. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/2210.10732>
- [33] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273-297, Sep. 1995, doi: 10.1007/BF00994018.
- [34] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 29, 2017. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [35] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE", *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.
- [36] S. Ruder, "An overview of gradient descent optimization algorithms." arXiv, Jun. 15, 2017. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1609.04747>
- [37] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." arXiv, Jan. 04, 2019. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/1711.05101>
- [38] P. Mayilvahanan, T. Wiedemer, E. Rusak, M. Bethge, and W. Brendel, "Does CLIP's Generalization Performance Mainly Stem from High Train-Test Similarity?" arXiv, Oct. 14, 2023. Accessed: Feb. 02, 2024. [Online]. Available: <http://arxiv.org/abs/2310.09562>