Detection, classification, and clustering of South African honey using YOLO, ViT and HDBSCAN

Lillian Mtumanje
University of Cape Town
South Africa
MTMLIL003@myuct.ac.za

ABSTRACT

Honey is a widely consumed natural product and a valuable agricultural commodity. Its floral origin can be determined through pollen analysis (melissopalynology), since pollen grains present in honey trace back to the plants visited by bees. Accurate identification is essential for consumer trust, trade, and protecting local apiculture, especially in biodiverse regions like South Africa. However, traditional pollen analysis methods are time-consuming and not scalable for modern demands in food traceability and quality assurance. In response, this paper presents an automated deep learning-based pipeline that integrates object detection (YOLOv11), classification (Vision Transformer), and unsupervised clustering (HDBSCAN) to identify known and novel pollen types from microscopy images. The study further investigates how different Vision Transformer configurations, specifically image resolution, patch size, and model size, affect the classification performance of South African pollen species. Experimental results highlight 3 main findings. First, the YOLOv11 detector achieved an mAP50 of 98.4%, highlighting its effectiveness in pollen grain detection. Secondly, the ViT classifier achieved an F1-score of 90.16%. However, when combined, the YOLO+ViT pipeline achieved a low overall F1-score of 22,87% in classifying South African pollen species. Interestingly, ViT-Small with 16×16 patches at 384×384 resolution achieved better performance than the pipeline model both in ViT configuration tests and on real honey sample evaluations, indicating that increasing input resolution can sometimes have a greater impact on performance than adjusting patch size or model size.Finally, HDBSCAN clustering of low-confidence predictions (<70%) produced 31 clusters with a 12.8% noise fraction, revealing meaningful groupings with a silhouette score of 0.58, indicating generally well-separated and internally consistent clusters.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Machine learning**.

KEYWORDS

Pollen analysis, YOLO, Vision Transformer, HDBSCAN, Deep learning

1 INTRODUCTION

Honey plays an important role not only as a dietary sweetener but also as a product with medicinal, cultural, and economic significance worldwide. The distinct chemical composition of each honey is shaped by the flowers from which bees collect nectar, with pollen grains serving as natural markers of its floral and geographical origin. Pollen analysis, therefore, provides a powerful tool for verifying honey authenticity, supporting quality control, detecting adulteration, and ensuring compliance with labelling regulations.

The standard traditional approach involves manual identification and counting of pollen grains under a microscope by trained experts. While effective, this method is labour-intensive, time-consuming, and susceptible to human error. For example, a single honey sample can contain hundreds to thousands of pollen grains, each of which must be manually identified and counted under a microscope. Preparing and analysing multiple samples can take several hours per sample, making the process time-consuming and mentally exhausting for trained experts.

Over time, new and efficient methods of pollen analysis have emerged, ranging from handcrafted feature-based methods [2,3,4] to more advanced techniques that utilize machine learning and computer vision. Earlier studies [5,6,14] have employed Convolutional Neural Networks (CNNs) for pollen classification. CNNs can automatically extract distinguishing features from images without the need for handcrafted features, handling variations in pollen size, shape, and surface patterns. They process large datasets efficiently, offering higher accuracy and scalability than traditional methods. CNNs detect local patterns such as edges and textures, which allows them to capture fine details like pollen surface structures. However, their focus on local features can be a limitation, as they may miss broader structural relationships, such as overall shape or feature arrangement, which are sometimes important for accurate classification.With this, transformers, which were initially applied to natural language processing (Vaswani et al. [11]), became adapted for image processing as Vision Transformers by Dosovitskiy et al. [12]. They became popular because they could capture long-range dependencies and relationships across the entire image, unlike CNNs which focus mainly on local features.

Notably, many studies on automated palynology have focused primarily on classifying individual pollen grains, often overlooking the crucial detection step required for comprehensive honey analysis. YOLO (You Only Look Once) addresses this gap by providing a real-time object detection system that can simultaneously identify and localize objects in images. Several studies have verified YOLO models as effective options for pollen grain detection [15, 16], highlighting their efficiency and accuracy for tasks where both speed and precision are essential. While many studies have been conducted on pollen analysis using deep learning, there is a notable lack of studies applying these techniques to South African honey, which is derived from the country's rich floral diversity. Although several studies have applied CNNs, YOLO, or Vision Transformers

to pollen detection and classification, these efforts have primarily focused on datasets from outside of South Africa. To date, there has been little to no work specifically addressing the diverse pollen species found in South African honey, despite the region's extraordinary floral biodiversity. Furthermore, most prior studies have treated detection and classification separately, without developing an integrated pipeline capable of handling both known and potentially novel pollen types in a single framework.

In response, this paper proposes a deep learning pipeline that combines object detection (YOLOv11), classification (Vision Transformer), and unsupervised clustering (HDBSCAN). The proposed pipeline begins with YOLOv11, which scans microscopy images to automatically locate individual pollen grains and draw bounding boxes around them, isolating each grain from the background for further analysis. The extracted grains are then passed through a Vision Transformer, which classifies each grain into a known pollen type by capturing both local details, such as surface textures, and global relationships between features. The classified pollen grains will be grouped by pollen species and counted to identify the type of honey by majority pollen species. Finally, to account for the presence of unclassified or novel pollen types, unsupervised clustering via HDBSCAN is performed on the feature embeddings extracted from the penultimate layer of the Vision Transformer. The study also examines how various Vision Transformer configurations such as image resolution, patch size, and model size impact the classification performance of South African pollen species.

The contributions are:

- The development of a deep learning-based pipeline that integrates YOLOv11 for pollen grain detection, a Vision Transformer for classification, and HDBSCAN for unsupervised clustering of South African pollen species.
- A study on the impact of Vision Transformer image resolution, patch size and model size on classification performance for South African pollen species.

2 RELATED WORK

Pollen detection and classification have evolved significantly over the years, moving from entirely manual microscopic analysis to advanced automated systems that utilizes deep learning. Early approaches relied on human expertise or handcrafted features, while recent developments leverage convolutional neural networks (CNNs), object detection models like YOLO, and Vision Transformers (ViTs) to improve speed, accuracy, and scalability.

2.1 Traditional Methods

Traditionally, pollen detection and classification were done manually using the ICBB method by Louveaux et al. [1]. The process required a trained expert to identify and count different pollen species under a microscope. This process was labour-intensive, time-consuming, and prone to human error. Over the years, advancements in pollen analysis have been made to automate this process, the most popular being the use of machine learning and computer vision.

2.2 Partial Automation

Before fully automated machine learning methods were applied, handcrafted feature-based methods were used. These were approaches where human experts designed specific features for a classifier to differentiate between pollen grains. These methods relied on predefined rules and statistical techniques rather than learning directly from raw images like deep learning models. The technique in Soares et al. [2] included image segmentation, feature extraction, and machine learning-based classification, achieving an F1-Score of 79%. Gonçalves et al. [3] used three feature extractors, Bag of Visual Words (BOW), Color, Shape and Texture (CST), and a combination of BOW+CST to classify 23 pollen types from the Brazilian Savannah, achieving a CCR of 64% with CST+BOW and C-SVC. Travieso et al. [4] focused on pollen shape, and with a dataset of 47 tropical honey plant species, achieved a mean success rate of 93.8%.

2.3 Machine Learning

Challenges with manual and semi-automated pollen analysis methods, combined with advancements in deep learning, image processing, and high-resolution microscopy, as well as the growing demand for large-scale analysis, led to fully automated approaches. These approaches included the use of CNNs, YOLO, and Vision Transformers. CNN approaches included several studies [5,6,7]. Tsiknakis et al. [5] conducted a comparative study on the Cretan Pollen Dataset v1, comparing four CNN models (Inceptionv3, Xception, ResNet, and Inception-ResNet) pretrained on ImageNet, applying a transfer and ensemble approach. The best-performing model was a soft voting ensemble of all base models, achieving an accuracy of 97.5%. Sevillano and Aznarte [6] compared three setups: Setup A used AlexNet for feature extraction with a linear discriminant classifier, Setup B applied transfer learning on the POLEN23E dataset, and Setup C combined both approaches. Setup C achieved the highest accuracy at 97.2%. Olsson et al. [7] applied transfer learning with ResNet-18, GoogleNet, and Xception on two pollen datasets, one with 83 species and another with 29 types, using both splitting (90/10) and leave-one-out cross-validation. The splitting experiment achieved up to 96% accuracy, while the leave-one-out experiment reached up to 86%, with larger training sets generally yielding higher recall rates.

Pollen analysis using YOLO included both detection and classification. Kubera et al. [8] used YOLOv5 to detect pollen grains from three highly allergenic taxa (Alnus, Betula, and Corylus) prevalent in Central and Eastern Europe, obtaining an mAP ranging from 86.8% to 92.4%.Zhang et al. [9] proposed a Swin-Transformer YOLOv5 (S-T-YOLOv5) model for detecting and quantifying a single pollen species, alfalfa (Medicago sativa L.), and compared its performance with four other YOLO models: YOLOv3, YOLOv4, YOLOR, and YOLOv5. S-T-YOLOv5 outperformed the others, achieving high precision (99.6%), recall (99.4%), F1-score (0.995), and mAP50 (99.4%). Jofre et al. [10] used bright-field microscopy with YOLOv8 to automatically count pollen and determine the floral origin of Guindo Santo honey, achieving an mAP of 97.6% for all classes and 94.4% for Guindo Santo honey.

Transformers were originally introduced for natural language processing in 2017 (Vaswani et al. [11]) to improve machine translation through self-attention and better handling of long-range dependencies compared to recurrent or convolutional models. They were later adapted for computer vision in 2020 (Dosovitskiy et al. [12]) as Vision Transformers (ViTs), which perform image classification by representing images as sequences of patches. However, pure self-attention does not capture local pixel relationships and relies on large-scale dataset pretraining to reach performance levels comparable to CNNs. Duan et al. [13] proposed a Vision Transformer that uses a FeatureMap-to-Token module and a CNN-like hierarchical design to combine local and global features, achieving the same accuracy as CNNs on electron-microscopy pollen images. On their custom 42-class electron microscopy (EM) pollen dataset, the small version of their model (Our-S) achieved 96.14% top-1 accuracy from scratch, improving to 97.16% with combined distillation.

In summary, recent advances in deep learning from CNN-based feature extraction to transformer-based classification and real-time object detection with YOLO have significantly improved the automation of pollen analysis. However, most prior research has focused on non-South African datasets and has primarily addressed either detection or classification in isolation. Very few studies have developed integrated pipelines that combine detection, classification, and clustering to handle both known and potentially novel pollen species. This gap is particularly relevant in the South African context, where the region's floral diversity presents unique challenges for honey authentication. Against this backdrop, the present study introduces a unified pipeline that leverages YOLOv11 for pollen grain detection, a Vision Transformer for classification, and HDBSCAN for unsupervised clustering of uncertain cases.

3 MATERIALS AND METHODS

3.1 Dataset

The dataset consisted of 77 pollen classes from South Africa. The dataset included two components: detection data (419 microscopy images) and classification data (7,594 microscopy images). The detection dataset contained microscopy images with multiple pollen grains on it. The classification dataset contained 77 folders, each representing a different pollen class, with multiple images of individual pollen grains stored inside. Both datasets were divided into training, validation, and test sets. To increase dataset variability and reduce overfitting, data augmentation techniques such as random horizontal flips, color jittering, and random rotations were applied.

3.2 Deep Learning Models

For detection, YOLOv11 was trained for 47 epochs with a batch size of 8 and an input image size of 512. Training employed stochastic gradient descent (SGD) with a learning rate of 0.00058 and momentum of 0.9504. For classification, a Vision Transformer with 77 output classes was fine-tuned for 50 epochs on 224×224 images, using a batch size of 16. The AdamW optimizer was applied with a learning rate of 0.00018, weight decay of 0.00168, and layer-wise learning rate decay, with 11 transformer blocks unfrozen during fine-tuning. All models were implemented in PyTorch.

3.3 Evaluation Metrics

YOLO and classification model performance was assessed using precision, recall, and F1-score[17,18]. Precision measures the proportion of correct positive predictions, recall measures the proportion of actual positives correctly identified, and F1-score provides a harmonic mean of precision and recall. These metrics were computed using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) . . Cluster performance was assessed by Silhouette Score and Noise Fraction[19,20]. The Silhouette Score measures how well-separated and compact clusters are, with values close to 1 indicating clear separation and values near 0 suggesting substantial overlap between clusters. The Noise Fraction represents the proportion of data points that HDBSCAN could not assign to any cluster, reflecting the presence of ambiguous or distorted samples.

3.4 Pipeline Setup

The proposed pipeline begins with YOLOv11, which detects and crops individual pollen grains from microscopy images. The cropped grains are passed to the ViT classifier. Predictions with a confidence score greater than 70% are grouped by species and counted to estimate the composition of the honey sample. A honey sample would be labelled as monofloral if a single pollen class contributed 45% or more to the total pollen count. Otherwise, it would be considered multifloral. Predictions below 70% confidence are clustered using HDBSCAN, which groups morphologically similar grains based on ViT feature embeddings. An overview of the pipeline is presented in Figure 1.

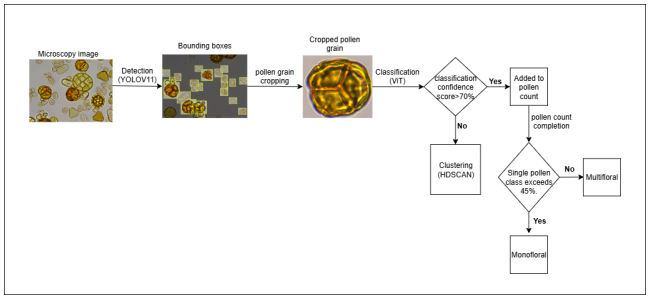

Figure 1: YOLOv11, Vision Transformer, and HDBSCAN pipeline for pollen analysis.

4 RESULTS AND DISCUSSION

4.1 YOLO

Figure 2 shows changes in precision, recall, mAP50 and mAP50-95 during the training process. There was an initial sharp increase in mAP50 (48.09%–94.28% over the first 4 epochs) and mAP50-95 (29.54%–80.86% over the first 10 epochs), after which performance plateaued with minimal further gains in later epochs. This indicated how the YOLOv11 model learned the patterns of the dataset quickly, reaching a high performance at an early epoch stage. The best validation epoch achieved an mAP50 of 98,40% and an mAP50-95 of 87,62%.This showed the YOLOv11 model's high accuracy in terms of bounding boxes. Precision and recall were 95,31% and 96,06%

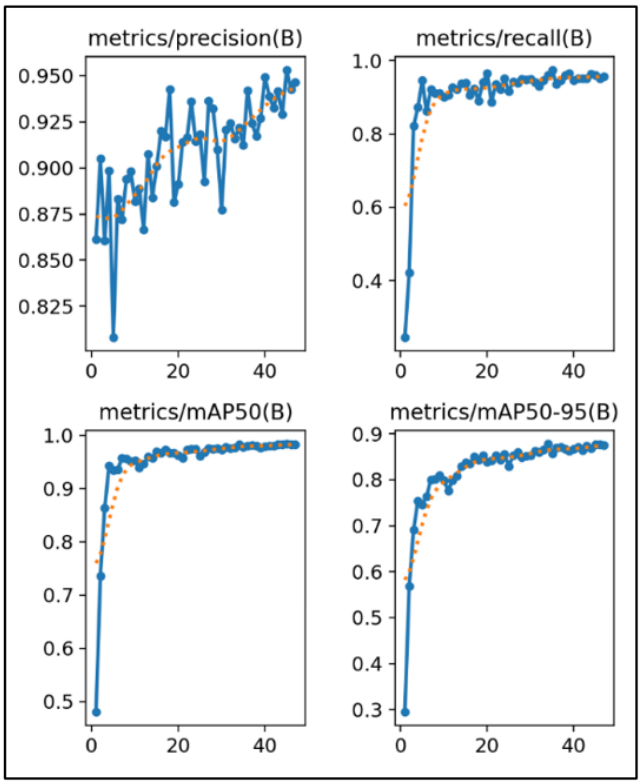

Figure 2: Performance of YOLOV11 model during the training process.

respectively. The high precision indicated that there were very few false positives and the high recall percentage also indicated that there were very few false negatives as well. This highlights how well the YOLOv11 model performed in detecting pollen grains in microscopy images. These results are consistent with previous YOLO-based pollen studies[8,9,10], confirming YOLO's robustness for object detection tasks. Strong performance can be attributed to the ability of YOLO to capture fine-grained local features, which are well suited to the textures and boundaries of pollen grains.

4.2 ViT Classification

The ViT classification model which was evaluated across 77 South African pollen classes achieved a macro average precision of 92,64%, a macro average recall of 90,05% and a macro average F1-score of 90,16%. To account for class imbalance weighted averages were also calculated. A weighted average precision, recall and F1-score of 94%, 93,7% and 93,4% was achieved. Out of 77 pollen species, 25 achieved perfect macro precision, recall and F1-score(See table 10 in appendix).At the other end, the six pollen classes with the lowest F1-scores (below 75%) were PAL0023, PAL0020, PAL0016, Combretum sp1, Eucalyptus sp1, and Daisy sp2. The PAL pollen classes (27 in total) were harder for the model to classify because their images weren't clearly labeled. They came from one of the other 50 South African pollen classes. Thus, it was reasonable to expect that the ViT model would have difficulty classifying them correctly. As for Combretum sp1, Eucalyptus sp1 and Daisy sp2, one reason for the low performance can be attributed to the number of training images. Having more training images usually helps a model perform better because it can learn more about each class. With fewer training images, the model may not learn enough and can make more mistakes on new data. This is evident for Daisy sp 2 which had 8 training images plus augmentation and only achieved an F1-score of 20% as illustrated in figure 3. Although Eucalyptus sp1 and Combretum sp1 had 36 and 37 training images respectively, they still achieved a moderate F1-score of 58,06% and 62,86%.While Lobostemon, which had the largest number of training images, had an F1-score of 95,88%. Hence figure 3 also shows that the number of training images is not the only reason for low performance as some pollen classes achieved 100% F1-score while having only a few training images. Thus the pretrained ViT model plus data augmentation may have benefited some pollen classes with a small number of training images, but not all.

Seven out of the 77 pollen classes consisted of more than one species that had to be considered for classification. For example, the Daisy pollen class, illustrated in Figure 4, had 7 species under it. This meant that the model had to distinguish the features of the different species of the same family. Out of 7 pollen classes that had more than one species to consider, only 1 of these pollen classes, Daisy sp2, had a low F1-score .This demonstrates that the model was able to distinguish between the different species of the same family with high accuracy as seen in figure 5(number of species indicated in brackets) .The highest average F1-score, 97,92%, was achieved by the Rhamnaceae pollen class with the lowest being an average of 77,68% for the Eucalyptus pollen class. Although the model performed well on the different species, intra-class similarity is another challenge that the model must tackle. Table 1 highlights the top ten pollen classes that were often misclassified by the model, with Eucalyptus sp 1 having the largest count. This misclassification exists because many pollen classes are present, 77 in total, and majority of these pollen classes share similar shapes and geometry like the example pollen classes in figure 6.This ultimately leads to misclassifications by the ViT classifier.

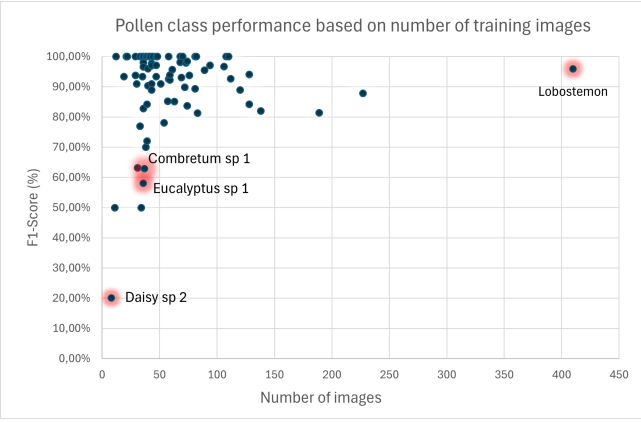

Figure 3: Relationship between F1-score and the number of training samples of South African pollen species.

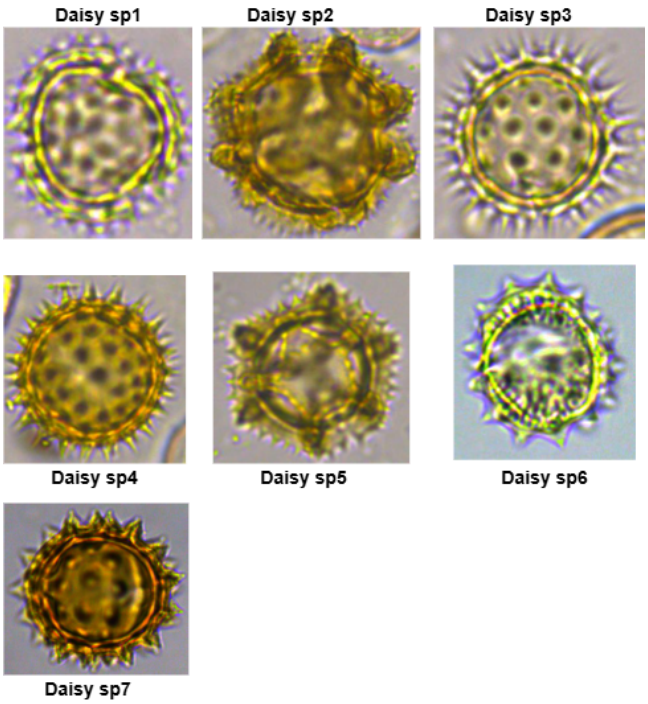

Figure 4: Seven different species of Daisy

Table 1: Misclassification counts of selected pollen species.

True Class	Predicted Class	Count
Eucalyptus_sp1	PAL0018	11
Vicia_sp_1	Apiaceae_sp1	7
PAL0023	Crassulaceae_sp1	6
PAL0014	PAL0019	5
Celtis	Citrus_sp1	4
Eucalyptus_sp3	PAL0024	4
PAL0020	Rhamnaceae_sp_1	4
PAL0026	PAL0018	4
PAL0004	Celtis	4

4.3 YOLO+ViT pipeline

Nine honey samples were tested on the integrated YOLO+ViT pipeline. Seven of these honey samples were sourced from different South African regions and thus consisted of different types of SA pollen classes. The other 2 honey samples(HS150 and HS189) contained some South African pollen classes along with other unknown pollen classes. A ground truth dataset listing the types of pollen classes along with the count within each honey sample microscopy image was used for comparison against what the model produced. Table 2 indicates the macro average precision, recall and F1-scores of the nine honey samples. The overall macro average performance across all honey samples was 29,65% precision, 5.37% recall, and 22,87% F1-score.These results are quite poor and reflect the failure

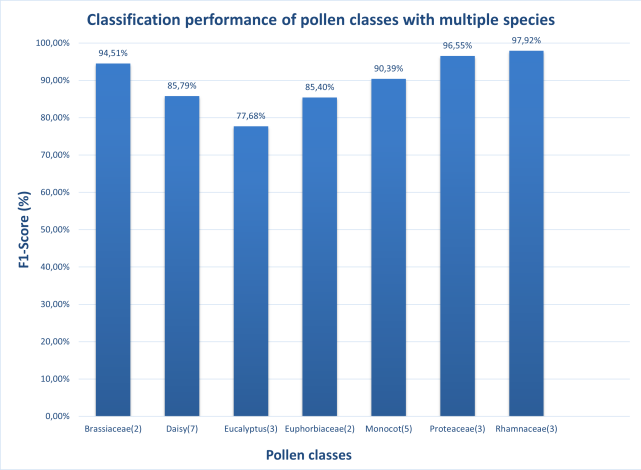

Figure 5: Performance of pollen classes with multiple species.

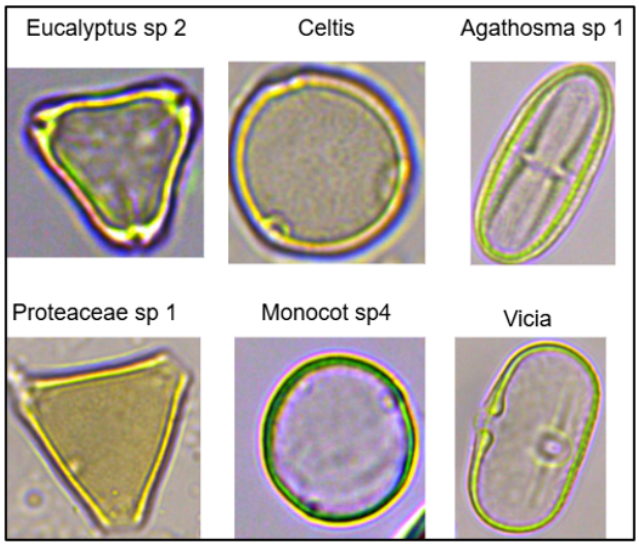

Figure 6: Similar pollen classes.

of the YOLO+ViT pipeline. The lowest F1-score was achieved by the Non Fynbos Angola honey sample(5,68%) with the highest being 56,61% for HS150 honey sample(the honey sample that contained some unknown pollen classes). Although the YOLOv11 detection model achieved impressive results during training with an mAP50 of 98,40% and the ViT classifier with an F1-score of 92,64%, the integration of the two did not lead to great results.

The observed underperformance of the YOLO+ViT pipeline may be linked to the discrepancy between the training data that the ViT classifier used (figure 7) and the extracted pollen grains that the ViT classifier received as input from the YOLO detector as shown in figure 8. YOLO generates bounding boxes around detected pollen grains, which are then cropped and passed to the ViT classifier. However, these crops can be imperfect: some grains may be overlapped with neighboring grains, or include background

noise, removing key morphological features such as shape as seen in figure 8.The model, having only been trained on perfect pollen images, would struggle with the imperfect ones. Because of this the model's predicted pollen count percentage would differ from the ground truth pollen count percentage. For example, according to Table 3, the Non-Fynbos Magoebaskloof honey sample was predicted as multifloral by the pipeline, but was actually monofloral for Celtis pollen with a percentage of 62.20%. Similarly, the Non Fynbos Angola honey sample was predicted as monofloral for PAL0020 with a pollen composition of 56.94%, whereas the ground truth indicated it was monofloral for PAL0019 with 49.40%.The only prediction that the pipeline got correct was for the Fynbos Du Toit's Kloof honey sample. It predicted the honey sample as monofloral for the Lobostemon pollen class with a percentage of 63,06%, compared to ground truth which was 76%. Although the pipeline seemed accurate in its labelling of multifloral honey samples, the problem lied with the predicted pollen composition percentages. In the Fynbos Stellenbosch honey sample (Table 4), the pipeline predicted 30 pollen classes, while the ground truth contained only 16. Of these, only 9 classes from the ground truth (Eucalyptus sp2, Lobostemon sp1,Monocot sp4, Aizoaceae sp1,Lycopodium,Eucalyptus sp1,Monocot sp 2, PAL009 and Citrus sp1) were present in the YOLO+ViT predictions. However, percentages greatly differed to the ground truth. For example Eucalyptus sp2 was predicated with a pollen composition percentage of 31,05%, as opposed to ground truth which was 11,88%.These results suggest that the combination of imperfect YOLO crops, domain differences between training and real honey samples, and low-confidence exclusions may have limited the ViT classifier's ability to generalize, contributing to inaccuracies in predicting pollen composition.

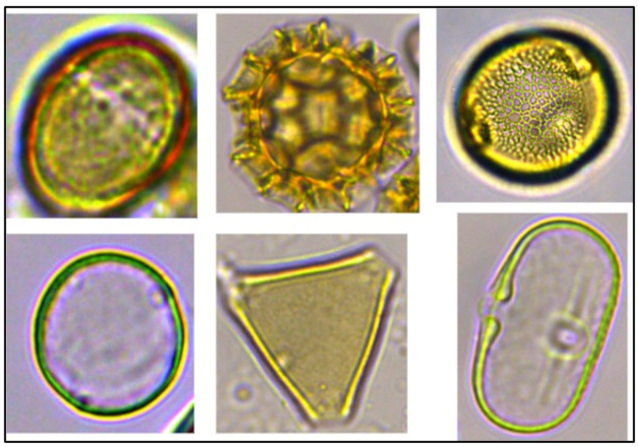

Figure 7: ViT model classification training data.

Table 2: YOLO+ViT model performance on honey samples.

Honey sample	Precision (%)	Recall (%)	F1-score (%)
Non-Fynbos Magoebaskloof	16.04	12.91	18.47
Fynbos Du Toit's Kloof	36.11	1.76	18.33
Non-Fynbos-Angola	16.58	2.71	5.68
Fynbos-West coast Langebaan	11.80	1.48	9.73
Fynbos-Eendekuil, near Citrusdal	15.34	5.37	13.61
Fynbos-Table Mountain	22.38	8.88	25.45
Fynbos-Stellenbosch	21.21	3.22	23.13
HS189	66.97	3.18	34.81
HS150	60.42	8.85	56.61
Macro averages	**29.65**	**5.37**	**22.87**

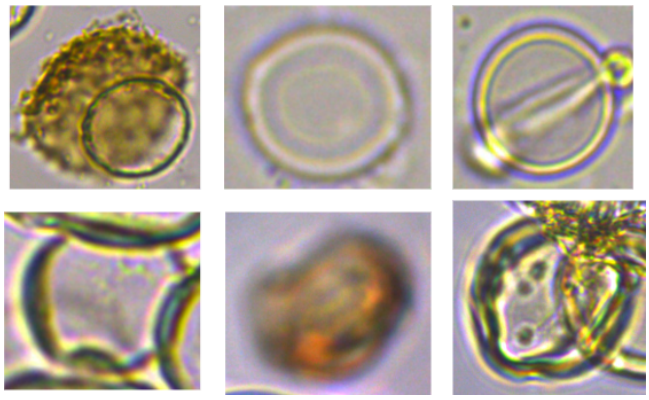

Figure 8: Extracted pollen grains from YOLOv11.

Table 3: Pipeline predictions vs. ground truth for honey samples.

Honey sample	Pipeline	Predicted %	Ground truth	Ground truth %
Non-Fynbos Magoe-baskloof	Multifloral	–	Monofloral: Celtis	62.20
Fynbos Du Toit's Kloof	Monofloral: Lobostemon	63.06	Monofloral: Lobostemon	76.00
Non-Fynbos-Angola	Monofloral: PAL0020	56.94	Monofloral: PAL0019	49.40
Fynbos-West coast Langebaan	Multifloral	–	Multifloral	–
Fynbos-Eendekuil, near Citrusdal	Multifloral	–	Multifloral	–
Fynbos-Table Mountain	Monofloral: Eucalyptus sp2	46.27	Multifloral	–
Fynbos-Stellenbosch	Multifloral	–	Multifloral	–
HS189	Multifloral	–	Multifloral	–
HS150	Multifloral	–	Multifloral	–

(– indicates a list of pollen classes and percentages predicted by the YOLO+ViT model, similar to Table 4).

Table 4: YOLO+ViT model predictions vs. ground truth for Fynbos-Stellenbosch honey sample.

YOLO+ViT prediction	Percentage (%)	Ground truth	Percentage (%)
Eucalyptus_sp2	31.05	Eucalyptus sp. 3	26.09
Lobostemon sp. 1	16.13	**Lobostemon sp. 1**	25.80
PAL0020	8.87	**Eucalyptus_sp1**	18.55
PAL0014	4.84	**Eucalyptus_sp2**	11.88
PAL0019	4.44	**Lycopodium**	6.09
Monocot sp. 4	4.03	**Monocot sp. 4**	5.80
Brachystegia	3.63	Uncertain	1.74
Combretum_sp1	3.63	**Citrus_sp1**	0.58
Passerina sp. 1	3.63	Daisy sp. 6	0.58
Aizoaceae sp. 1	2.82	PAL0026	0.58
Lycopodium	2.02	Vicia sp. 1	0.58
Eucalyptus_sp1	1.61	**Aizoaceae sp. 1**	0.29
Proteaceae sp. 2	1.21	Brassicaceae sp. 1	0.29
PAL0011	1.21	Daisy sp. 1	0.29
Scrophulariaceae	0.81	**Monocot sp. 2**	0.29
Monocot sp. 2	0.81	**PAL0009**	0.29
PAL0018	0.81	Poaceae	0.29
Rhamnaceae sp. 2	0.81		
PAL0015	0.81		
Brassicaeae_sp2	0.81		
PAL0009	0.81		
PAL0027	0.81		
Cichoriodae	0.81		
Plantago	0.81		
Apiaceae sp. 1	0.81		
PAL0022	0.40		
Citrus_sp1	0.40		
Euphorbiaceae_sp2	0.40		
Carpobrotus	0.40		
Erica_sp1	0.40		

4.4 Clustering

HDBSCAN identified a total of 31 clusters along with 238 noise points, corresponding to a noise fraction of about 12.8%. The clustering quality was supported by a silhouette score of 0.58, indicating that most clusters were reasonably well separated and internally consistent. As shown in Figure 9, some clusters were well isolated, such as clusters 0, 1, 4, and 5, while others appeared close to one another for example, clusters 16, 17, and 18, suggesting a closer relationship between those pollen classes according to the model. Inspecting the images within each cluster revealed that many were dominated by a specific pollen class as shown in table 5 (see table 11 for further details in the appendix). Notably, Cluster 0, located on the far left of figure 9, contained images from different Daisy species, pointing to a certain distinctness within the Daisy class itself. It was also interesting to note that some pollen classes dominated multiple clusters. This can happen because low-confidence images capture variation within a class, such as differences in shape, size, or orientation, and the ViT embeddings reflect this variation. For example, the Lobostemon class dominated clusters 13, 16, and 21, while Eucalyptus dominated clusters 6, 8, and 29, showing how the model split each class into subgroups based on subtle differences or uncertainty. Out of the 31 clusters, 20 were strongly associated with a single pollen class. This shows that some pollen classes were quite distinct and easily separated. However with 77 pollen classes evaluated overall this highlighted the strong similarity between pollen classes and the model's difficulty in distinguishing them.

While clustering revealed meaningful groupings, the limited quantitative evidence and reliance on qualitative inspection meant its reliability for practical honey authentication remains uncertain. Taken together, these results suggest that while individual models perform strongly, their integration into an end-to-end honey authentication pipeline remains a challenge.

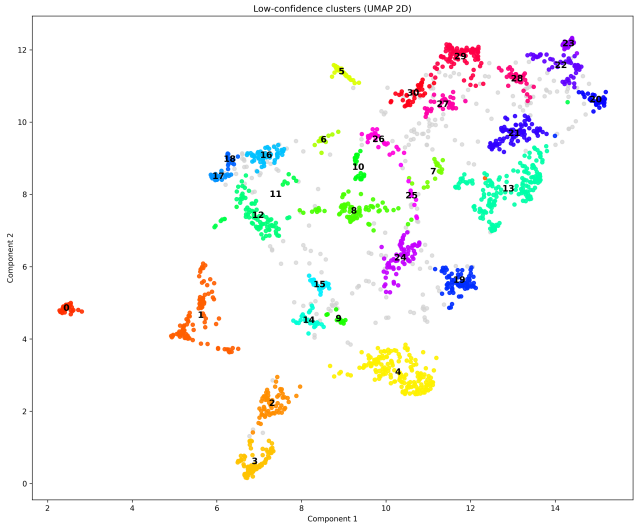

Figure 9: UMAP of clustered images with a confidence score of less than 70% using HDBSCAN.

5 VIT CONFIGURATION

The above pipeline utilized a ViT classifier configured with a small ViT model, a patch size of 16 and an image size of 224. A further experiment was carried out to investigate the effect of different

ViT configurations on classification performance. The same South African pollen dataset was used during the tests runs.Three key Vision Transformer (ViT) design parameters were varied: model size, patch size, and image size as indicated in table 6. ViT-Tiny, ViT-Small, and ViT-Base were selected, progressively increasing in depth, embedding dimension, and number of attention heads to capture the trade-off between computational efficiency and representational capacity. Patch sizes of 8×8 and 16×16 were explored to evaluate token granularity, with smaller patches generating more tokens and capturing finer morphological details, while larger patches reduce sequence length and computational cost. Input resolutions of 224×224 and 384×384 pixels were tested to assess the impact of resolution, as higher image sizes preserve subtle structural features of pollen grains that may be lost at lower resolutions, albeit at greater computational demand. This experimental design enabled an evaluation of the interplay between model capacity, token granularity, and input resolution, helping identify configurations that balance accuracy with efficiency for this specialized dataset.

Model size had a modest impact as shown in table 7. ViT-Base achieved the highest F1-score (91.83%), while ViT-Tiny slightly outperformed ViT-Small (91.31% vs. 90.16%), likely because smaller models were less prone to overfitting on the limited data. Patch size as seen in table 8 had minimal effect, with 8×8 patches slightly outperforming 16×16 (90.17% vs. 90.16%), suggesting that finer tokenization offers only marginal gains while increasing sequence length and computational cost.

Image resolution had the most significant effect as indicated in table 9. Increasing from 224×224 to 384×384 improved F1-score from 90.16% to 91.97%, as higher-resolution inputs allowed each token to encode richer morphological details without substantially increasing token count. Overall, the results indicate that, for fine-grained pollen classification on a limited dataset, higher input resolution provides more performance benefit than simply increasing model size or using smaller patches., while smaller ViT models can still capture most discriminative features efficiently.

The ViT-S/16 (384) model, which achieved the highest accuracy among the tested configurations(table 6) on the main dataset, was evaluated on the same nine honey samples previously used for the pipeline. Interestingly, ViT-S/16 (384) model slightly outperformed the ViT-S/16 (224) pipeline model, achieving a macro-average precision of 30,82%, recall of 5,83%, and an F1-score of 25,50%, compared to 29,65% precision, 5,37% recall, and 22,87% F1-score for the pipeline model. This suggests that increasing input resolution can capture finer morphological details thereby translating into better performance on real honey samples.

Table 5: Cluster assignments for pollen classes.

Cluster	Pollen class
0	Daisies
1	Lycopodium
2	Brassica
3	PAL0011
4	Celtis
5	Rhamnaceae sp. 1

Table 6: Performance comparison of different ViT models

Model	Precision (%)	Recall (%)	F1-score (%)
ViT-T/16 (224)	92.74	91.54	91.31
ViT-S/8 (224)	90.72	91.22	90.17
ViT-S/16 (224)(pipeline)	92.64	90.05	90.16
ViT-S/16 (384)	94.30	92.38	**91.97**
ViT-B/16 (224)	93.11	92.56	91.83

Table 7: Effect of Model Size on F1-Score (Patch = 16×16, Image = 224×224).

Model	F1-Score (%)
ViT-Tiny	91.31
ViT-Small	90.16
ViT-Base	91.83

Table 8: Effect of Patch Size on F1-Score (Model = ViT-Small, Image = 224×224).

Patch Size	F1-Score (%)
8×8	90.17
16×16	90.16

Table 9: Effect of Image Resolution on performance (Model = ViT-Small, Patch = 16×16).

Image Size	F1-Score (%)
224×224	90.16
384×384	91.97

6 CONCLUSIONS

This study aimed to develop and evaluate a deep learning pipeline for South African honey authentication using pollen analysis. The pipeline integrated YOLOv11 for pollen grain detection, a Vision Transformer (ViT) for classification, and HDBSCAN for clustering low-confidence predictions, with the goal of improving accuracy, scalability, and robustness in identifying floral origins of honey samples.

The YOLOv11 model performed strongly on pollen grain detection, achieving an mAP50 of 98.40%, a precision of 95.31% and a recall of 96.06%. The Vision Transformer classifier also performed well on individual pollen grain classification, with macro-averaged F1-scores just above 90% across 77 pollen classes, and demonstrated that model size, image resolution, and patch size impact classification performance. However, when integrated into the YOLO+ViT pipeline for honey sample analysis, performance dropped significantly, with macro F1-scores around 22,87%. The poor results were potentially due to imperfect pollen grains passed from YOLO to the classifier, leading to inaccurate predictions of pollen composition in honey. Clustering with HDBSCAN provided some value by grouping low-confidence predictions and highlighting morphologically related classes.

The main limitation of the project was the gap between strong model performance in isolated detection/classification tasks and poor performance in the full end-to-end pipeline on real honey samples. This was largely due to inconsistencies between the training data (well-prepared individual pollen images) and test data (YOLO-cropped grains from honey samples). Additionally, class imbalance in the dataset, with some species represented by very few samples, led to weaker classification performance for rare pollen types. Although clustering revealed useful groupings, its role in practical honey authentication was not assessed in detail because clustering was specifically applied to low-confidence predictions.

7 FUTURE WORK

Future research should focus on improving the integration between detection and classification by training Vision Transformers directly on YOLO-generated crops, ensuring that the classifier is exposed to the same imperfections present during inference. Expanding and balancing the dataset with more representative samples of South African pollen species is also essential, as the current imbalance limited performance for rare classes. In particular, collecting larger numbers of samples for underrepresented species and incorporating regional variations of the same pollen types would strengthen model generalization. Building such a comprehensive dataset would also enable experiments with larger transformer architectures and hybrid CNN–ViT models, potentially unlocking higher classification accuracy. Ultimately, these improvements would not only enhance the reliability of the pipeline for honey authentication but also contribute to the development of scalable automated palynology tools suited for South Africa's unique biodiversity.

8 ACKNOWLEDGEMENTS

The pollen dataset was provided by the University of Cape Town's Chemistry Department.Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za.

REFERENCES

[1] J. Louveaux, A. Maurizio, and G. Vorwohl. 1978. Methods of Melissopalynology. *Bee World* 59, 4 (Jan. 1978), 139–157. https://doi.org/10.1080/0005772X.1978.11097714

[2] D. Soares, D. Silva, L. Quinta, H. Pistori, and M. Borth. 2014. Application of wavelet transform in the classification of pollen grains. *African Journal of Agricultural Research* 9 (2014), 908–913. https://doi.org/10.5897/AJAR2013.7495

[3] A. B. Gonçalves, J. S. Souza, G. G. Da Silva, M. P. Cereda, A. Pott, M. H. Naka, and H. Pistori. 2016. Feature Extraction and Machine Learning for the Classification of Brazilian Savannah Pollen Grains. *PLoS ONE* 11, 6 (June 2016), e0157044. https://doi.org/10.1371/journal.pone.0157044

[4] C. M. Travieso, J. C. Briceno, J. R. Ticay-Rivas, and J. B. Alonso. 2011. Pollen classification based on contour features. In *Proc. IEEE Int. Conf. on Intelligent Engineering Systems (INES)*. IEEE, Poprad, Slovakia, 17–21. https://doi.org/10.1109/INES.2011.5954712

[5] N. Tsiknakis, E. Savvidaki, G. C. Manikis, P. Gotsiou, I. Remoundou, K. Marias, E. Alissandrakis, and N. Vidakis. 2022. Pollen Grain Classification Based on Ensemble Transfer Learning on the Cretan Pollen Dataset. *Plants* 11, 7 (Mar. 2022), 919. https://doi.org/10.3390/plants11070919

[6] V. Sevillano and J. L. Aznarte. 2018. Improving classification of pollen grain images of the POLEN23E dataset through deep learning CNNs. *PLoS ONE* 13, 9 (Sept. 2018), e0201807. https://doi.org/10.1371/journal.pone.0201807

[7] O. Olsson, M. Karlsson, A. S. Persson, H. G. Smith, V. Varadarajan, J. Yourstone, and M. Stjernman. 2021. Efficient, automated and robust pollen analysis using deep learning. *Methods Ecol Evol* 12, 5 (May 2021), 850–862. https://doi.org/10.1111/2041-210X.13575

[8] E. Kubera, A. Kubik-Komar, P. Kurasiński, K. Piotrowska-Weryszko, and M. Skrzypiec. 2022. Detection and Recognition of Pollen Grains in Multilabel Microscopic Images. *Sensors* 22, 7 (Mar. 2022), 2690. https://doi.org/10.3390/s22072690

[9] C. J. Zhang, T. Liu, J. Wang, et al. 2024. DeepPollenCount: a swin-transformer-YOLOv5-based method for pollen counting. *Aerobiologia* 40 (2024), 425–436. https://doi.org/10.1007/s10453-024-09828-8

[10] R. Jofre, J. Tapia, J. Troncoso, et al. 2025. YOLOv8-based on-the-fly classifier system for pollen analysis of Guindo Santo honey. *Journal of Agriculture and Food Research* 19 (2025), 101665. https://doi.org/10.1016/j.jafr.2025.101665

[11] A. Vaswani, N. Shazeer, N. Parmar, et al. 2017. Attention is all you need. arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.

[13] K. Duan, S. Bao, Z. Liu, et al. 2023. Exploring vision transformer: classifying electron-microscopy pollen images. *Neural Computing and Applications* 35 (2023), 735–748. https://doi.org/10.1007/s00521-022-07789-y

[14] T. Mahmood, J. Choi, and K. R. Park. 2023. AI-based classification of pollen grains using attention-guided feature aggregation. *J. King Saud Univ. – Computer and Information Sciences* 35, 2 (Feb. 2023), 740–756. https://doi.org/10.1016/j.jksuci.2023.01.013

[15] J. S. Khalane, N. D. Gawande, S. Raman, and S. Sankaranarayanan. 2025. Advanced Pollen Classification of Indian Medicinal Plants through SEM and Computer Vision. bioRxiv (2025). https://doi.org/10.1101/2025.01.08.631879

[16] C. He, A. Gkantiragas, and G. Glowacki. 2018. Honey Authentication with Machine Learning Augmented Bright-Field Microscopy. arXiv:1901.00516. https://doi.org/10.48550/arXiv.1901.00516

[17] D. Hoiem, Y. Chodpathumwan, and Q. Dai. 2012. Diagnosing Error in Object Detectors. In *Computer Vision – ECCV 2012*, Springer, 340–353. https://doi.org/10.1007/978-3-642-33712-3_25

[18] M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[19] P. J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[20] R. J. G. B. Campello, D. Moulavi, and J. Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Proc. PAKDD 2013*, LNCS Vol. 7819. Springer, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

[21] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit. 2021. Understanding Robustness of Transformers for Image Classification. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 10211–10221. IEEE. https://doi.org/10.1109/ICCV48922.2021.01007

[22] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng. 2021. DeepViT: Towards Deeper Vision Transformer. *arXiv preprint arXiv:2103.11886.* https://arxiv.org/abs/2103.11886

A APPENDIX

Table 10: Precision, Recall, and F1-Score for top 25 pollen classes.

Class	Precision	Recall	F1-Score
Acacia	1	1	1
Aulax	1	1	1
Brachystegia	1	1	1
Campanulaceae	1	1	1
Carpobrotus	1	1	1
Daisy_sp1	1	1	1
Daisy_sp6	1	1	1
Daisy_sp7	1	1	1
Erythrina	1	1	1
Lycopodium	1	1	1
Monocot_sp_2	1	1	1
Monocot_sp_3	1	1	1
PAL0002	1	1	1
PAL0003	1	1	1
PAL0011	1	1	1
PAL0012	1	1	1
PAL0015	1	1	1
PAL0022	1	1	1
PAL0025	1	1	1
Poaceae	1	1	1
Rhamnaceae_sp_1	1	1	1
Rhamnaceae_sp_2	1	1	1
Searsia	1	1	1
Solanaceae	1	1	1

Table 11: Most common taxa for each cluster of low-confidence predictions.

Cluster ID	Most common taxon
0	Daisies
1	Lycopodium
2	Brassica
3	PAL0011
4	Celtis
5	Rhamnaceae sp. 1
6	Eucalyptus sp. 3
7	Rhamnaceae sp. 1
8	Apiaceae sp. 1
9	PAL0003
10	Proteaceae
11	Unknown
12	Unknown
13	Lobostemon
14	Aizoaceae sp. 1
15	Crassulaceae sp. 1
16	Lobostemon
17	Unknown
18	Vahlia-type sp. 1
19	PAL0010
20	Scrophulariaceae sp. 1
21	Lobostemon
22	Vahlia-type sp. 1
23	Monocot sp. 5
24	PAL0019
25	Eucalyptus sp. 1
26	Eucalyptus sp. 1
27	Eucalyptus sp. 2
28	Eucalyptus sp. 3
29	Eucalyptus sp. 3
30	Eucalyptus sp. 2