# Deep Learning in palynology

### A use case for automated visual classification of pollen grains from honey samples

Philipp Viertel[1] and Matthias König[1]

**Abstract:** In this work, we will show a use case for visual pollen classification from honey samples. We discuss the current state of the art in pollen analysis, highlight the importance of data quantity and quality, and elaborate on how to transfer promising Deep Learning methods to the analysis of honey samples. A first experiment with a public data set is shown as well as samples from our work-in-progress data set. Our recommendations and methods show which steps are necessary in order to successfully deploy an automated pollen analysis solution for honey products.

**Keywords:** Deep Learning, Machine Learning, Palynology, Pollen analysis, Automation

## 1 Introduction

Deep Learning (DL) is a promising technology that offers a large field of applications. The field of palynology, the study and analysis of pollen grains, is an area that can greatly benefit from this technology. Palynology plays an important role for various disciplines, such as climate research, geology, allergy studies, and especially honey production. The study of the latter is also called melissopalynology.

A serious problem for beekeepers concerns the labelling of honey products. Beekeepers usually have to use generic names, such as summer honey, to label their produce. Due to the fact that honey yield can come from a large variety of plant sources it is necessary to identify the leading pollen, if one wants to label the honey by its majority pollen source. This procedure is offered by specialised laboratories and institutes which charge a fee; however, since most beekeepers are avocational, these services are usually not used, also due to the fact that it has to be done frequently for each honey yield. The high fees are justifiable, however, due to the nature of its process. A pollen perpetration sample has to be created by dilution and followed by centrifugation followed by a highly trained palynologist identifying the pollen grains visually and counting them. One gram of honey contains between 2,000 and 1 million pollen grains which can stem from more than 100 different plants. This entire process is very time-consuming, labour-intensive, and costly. To determine a honey product, e.g. canola honey, at least 80 % of the pollen have to come from canola. The required numbers vary from honey to honey.

[1] Fachhochschule Bielefeld, Campus Minden, Artilleriestr. 9, 32427 Minden, [Firstname].[Lastname]@fh-bielefeld.de

To address this issue, we worked with local beekeepers from the district of Minden-Lübbecke, to identify the issues that occur with local beekeeping and identifying and classifying the origin of their honey produce. We developed a use case that shows the individual steps that are necessary in order to make the application of DL models feasible in such scenarios.

## 2 Related Work

The need and the benefits of an automated pollen recognition system were described as early as in 1996 [SF96]. Stillman and Flenley mention the time-consuming process of extraction and identification by manual preparation and analysis. The time of such an analysis is estimated with 2 to 10 hours. The authors identify specific requirements for applicable solutions, such as increased speed, objectivity, and determination. With the advent of Machine Learning (ML), the requirements for automated pollen classification can be met. The majority of methods for a visual pollen classification task can be put into two categories: ML methods that require a manual definition of the features that describe pollen classes distinctively. Features can be derived from morphological features [Ha18] or based on colour or texture features. Although a large number of defined features exist, as shown in [Re15], it still requires work and the result is more static than the second option; the DL approach, which works similarly to a black box principle. A deep neural network finds the determinative features and patterns in the pollen on its own. Only the hyper parameters of the network itself and the training process as such are adjustable. Theoretically, such methods can operate with an infinite number of new pollen classes.

The largest obstacle is the absence of quality data sets. Most research is done on proprietary data sets that the researchers created or obtained themselves. This is problematic in multiple ways: it is difficult to validate the work of the authors and to compare different methods with each other since they are trained and evaluated on differing data sets. Therefore, it is not possible to make general statements about the performance and quality of the proposed methods, especially in its applicability. The data sets POLEN23E [Go16], by Duller et al. [Du99], and POLLEN13K [Ba20] are available and make up some of the exceptions. POLEN23E e.g. contains 805 images from 23 pollen classes, with a minimum resolution of 250x250 for each RGB image. The original authors of [Go16] use a Support Vector Machine (SVM) method and features based on colour, shape, and texture. An accuracy[2] of 64 % was achieved. [SA18] used the same data set, however with a Convolutional Neural Network (CNN) and a Linear Discriminant Classifier and achieved an accuracy of 97 %. Similar results, also using CNNs, but with proprietary data sets, were achieved in [SHA20] (97.86 %) and [Kh18] (95.9 %). But again, these methods use different data as well as different numbers of pollen.

[2] Accuracy is the fraction of all predictions that a model has correctly classified.

# 3 Method

## 3.1 Data requirements and acquisition

The largest problem, as already indicated in the introduction as well as in the related work, is the absence of freely available data. In order to train a DL model it is necessary to have a large number of images with specific qualitative standards. As of 2020, such data sets are sparse, as they are largely proprietary. POLEN23E e.g. fulfils the criteria, but it cannot be used in Germany, since the flora is specific to the Brazilian Savannah. Therefore, it is necessary to create a training data set that includes the common native species for the region, in which the application is intended to be deployed in. For this purpose, we worked

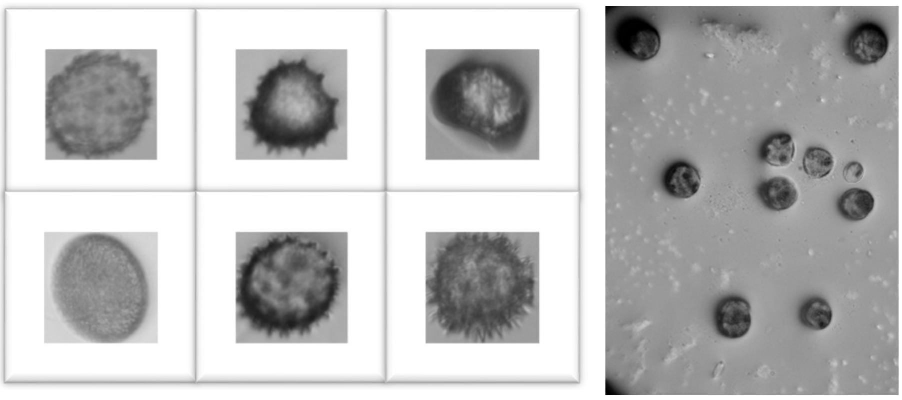

Fig. 1 Left: six different pollen classes from our work-in-progress data set. From left to right: *Solidago canadensis, Tanacetum vulgare, Alnus glutinosa, Sinapis arvensis, Symphyotrichum novae-angliae, Helianthus annuus.* Right: example of a pollen extraction from honey via sedimentation at 400 X magnification.

together with local beekeepers from the Minden-Lübbecke region. A set of specific requirements have to be met to create images that can be used in a DL model:

- The images have to be captured with a light microscope (LM) and at a magnification of at least 320 X to 1000 X (in accordance with the German DIN 10760).
- Pollen have a spheroid shape, i.e. ellipsoids with two semi-diameters. This means that pollen are 3D objects and can therefore appear differently under the LM, depending on their position. It is necessary to obtain a large variety of positions and foci of the pollen grains to assure a high generalisation effect.
- Pollen grains are best visible when isolated. However, pollen usually do not come in this ideal state. When pollen grains are cramped together in groups, adding liquids such as water on the object slide can disperse the pollen grains. It is important to consider the harmomegathic effect, i.e. the change of shape of the pollen when in a hydrated state.

- Each individual pollen grain has to be labelled, this is done manually, however, segmentation or detection algorithms can support this process if required.

Our own data set is work in progress, as shown in Figure 1. The data is collected from typical plants that grow in the area and are commonly found in honey yields.

### 3.2 Deep Learning model

If an adequate number of training images is acquired, the Machine Learning method has to be selected. Although not the only ones possible, but the most promising, DL methods achieve high accuracies in classifying pollen correctly. For visual tasks, CNNs are used. In our own experiments, we utilized the POLEN23 data set in order to validate the results. The data set contains 23 different pollen classes of which we removed 5 images per class for testing. The remaining images were randomly split into 80 % training data and 20 % validation data. The model achieved a validation accuracy of 96 %. When tested on isolated random samples (without cross-validation, in order to ensure real-life conditions) 78 % were achieved. For this test we used a pre-trained ResNet-101 architecture [He16] with transfer learning.

When tested on our own data (see Figure 1) with five classes (*alnus glutinosa, helianthus annuus, solidago canadensis, symphyotrichum novae-angliae* and *tanacetum vulgare*) and 25 images per class (80-20 split) we achieved a validation accuracy of 84 % and when tested on three random samples per class we achieved a classification accuracy of 80 %. All of the pollen samples were correctly classified, except *helianthus annuus.* The random samples contained grains that were not as illuminated as the training data, therefore, the characteristic spike-like features on the exine surface are not visible[3].

The final decision from the network output can also be supported by an expert system that includes knowledge about e.g. the current season or the frequency of occurrence of certain pollen in the local area, so that an informed decision can be made in the case of uncertain classification results.

Figure 2 illustrates all the necessary steps that our proposed solution requires. It is common practice, due to the complexity of DL models, to orientate oneself around best practices, in this regard, to models and networks that have proven to yield best results in the category of object detection and classification, e.g. the ResNet architecture, since the results from these areas are usually performed on uniform and standardized data sets.

[3] This shows that individual features, such as shape, are not enough to determine the class correctly. However, in this case, the classification can be corrected by adding more data to the training set.

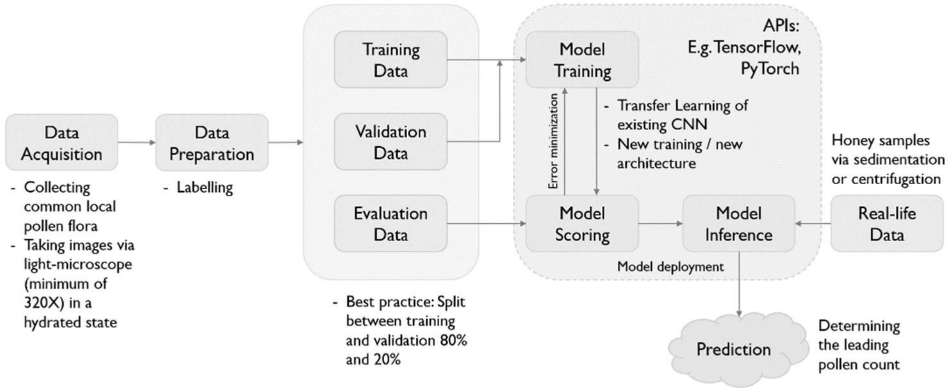


Fig. 2: Workflow for a DL solution to honey pollen analysis, from data acquisition to actual deployment, where images of honey pollen samples are analysed

## 4 Conclusion

This work showed that DL methods make it possible to classify pollen samples with high accuracy. This applicability can be used to classify pollen quantities in honey samples, given that an adequate number of images of local pollen types is available. Our own experiments showed that it is possible to recreate high accuracy pollen classification results and possibly transfer them to our case of application. For that purpose, the creation of a local data set for the area of Minden-Lübbecke is in progress, of which we showed some examples. The data acquisition process is of critical importance, since most DL methods work only as well as their training input.

## 5 Future Work

The pollen collection for the creation of a local data set is still in progress. Quality standards have been set and when enough data is acquired, labelling and training of various state of the art DL networks to evaluate the highest performance will be performed. Work on an autonomous, light-weight hardware solution for pollen analysis, that is intended to take an object slide with a prepared pollen sample and perform the methods described in this work, is also in progress. We believe that in the future our proposed method can help to increase the speed and reduce the costs of manual pollen analysis performed in laboratories and institutes or even allow beekeepers to perform their own analysis, when the DIN requirements are met.

## References

[SF96] Stillman, E. C.; Flenley, J. R.: The needs and prospects for automation in palynology. In: Quaternary Science Reviews, vol. 15, no. 1, pp. 1–5, 1996.

[Re15] Redondo, R. et.al.: Pollen segmentation and feature evaluation for automatic classification in bright-field microscopy. In: Computers and Electronics in Agriculture, vol. 110, pp. 56–69, 2015.

[Du99] Duller, A. et.al.: A pollen image database for evaluation of automated identification systems. In: Quaternary Newsletter, vol. 89, pp. 4–9, 1999.

[Go16] Gonçalves, A. B. et.al.: Feature Extraction and Machine Learning for the Classification of Brazilian Savannah Pollen Grains. In: PLOS ONE, vol. 11, no. 6, 2016.

[Ba20] Battiato, S. et.al.: Detection and Classification of Pollen Grain Microscope Images. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4220–4227, 2020.

[SA18] Sevillano, V.; Aznarte, J. L.: Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. In: PLOS ONE, vol. 13, no. 9, 2018.

[SHA20] Sevillano, V.; Holt, K.; Aznarte, J. L.: Precise automatic classification of 46 different pollen types with convolutional neural networks. In: PLOS ONE, vol. 15, no. 6, 2020.

[Kh18] Khanzhina, N. et.al.: Pollen grain recognition using convolutional neural network. In: ESANN European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 409-414, 2018.

[Ha18] Halbritter, H. et.al.: Illustrated Pollen Terminology. 2nd ed., Springer International, 2018.

[He16] He, K. et.al.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.