POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 961 **from** 16-11-2022

Ph.D. THESIS SUMMARY

Ing. Mihai BOLDEANU

AUTOMATIC POLLEN CLASSIFICATION USING DEEP LEARNING TECHNIQUES

CLASIFICARE AUTOMATĂ A POLENULUI FOLOSIND TEHNICI DE INVĂŢARE AUTOMATĂ

THESIS COMMITTEE

Prof. Dr. Ing. Gheorghe BREZEANU Univ. Politehnica din Bucureşti	President
Prof. Dr. Ing. Corneliu BURILEANU Univ. Politehnica din Bucureşti	PhD Supervisor
Prof. Dr. Ing. Corneliu RUSU Univ. Politehnica din Bucureşti	Referee
CS II Dr. Luminiţa MĂRMUREANU Institutul Naţional de Cercetare-Dezvoltare pentru Optoelectronică - INOE 2000	Referee
Conf. Dr. Ing. Horia CUCU Univ. Politehnica din Bucureşti	Referee

BUCHAREST 2022

Table of contents

Chapter 1

Introduction

In this thesis, we address the task of using machine learning to make automatic pollen classification. This is achieved using data from fully automated particle analyzers and deep learning algorithms. This sort of application can have multiple uses including in: agriculture, health and climate change monitoring.

Pollen is a part of plants reproductive system. It consist of individual fine grains that have different shapes and sizes and are produces by the male structures in seed-bearing plants. This fine powdery substance is transported by insects or by the wind and it interacts with the female plant structures, where fertilization occurs.

The differences in pollen shape and size are so large that a trained human can identify the species of plant by pollen alone. Nearly all angiosperms and gymnosperms can be identified by pollen with the field of study being called palynology or the study of pollen and spores.

This chapter is subsequently organized as such in section 1.1 we present the impacts of pollen on quality of life. In section 1.2 the difficulty in building such systems is presented. Section 1.3 is the motivation of pursuing this work. Section 1.4 deals with the scope and objectives of this thesis and finally, section 1.5 goes over the structure of the entire paper.

1.1 Pollen impact on air quality and human health

Pollen is one of the most common triggers of seasonal allergies in humans. Many people know pollen allergy as "hay fever", but experts usually refer to pollen allergy as "seasonal allergic rhinitis". Each spring, summer and autumn, and depending on the location event in winter, plants release millions of tiny pollen grains to fertilize other plants of the same species. While most pollen is harmless to humans and is transported by insects from one plant to another, there are certain species of trees, grasses and weeds that use the wind to transport their pollen. These plants make relatively small, light and dry pollen grains that

are easily picked up by wind and can find their way into eyes, noses and lungs, causing allergy symptoms for those with pollen allergies.

In recent years there has been a rise in the number of diagnosed cases of allergies to pollen and in Europe some studies suggest that the number of allergies will double from 33 milllion to 77 million by 2040-2060 [33]. While the exact cause is still not clear, some studies linking this rise to climate change [33] or pollution [57]. It is clear that more accurate monitoring of pollen concentrations and pollen seasons is crucial.

1.2 Problem Description

There are a number of problems that make pollen monitoring difficult with current devices and techniques.

The first problem with predicting or even identifying pollen in the air. This is difficult due to the very low concentration it is usually present in so large volumes of air have to sampled to identify pollen.

The second problem is that there isn't a simple or clear relationships between the period a plant will pollinate atmospheric parameters. This would allow for building predictive models. In order to identify all these relationships there is a need for large data-sets obtained by first monitoring pollen over many years.

The last problem with current pollen monitoring techniques, is that it is a very time-consuming endeavour that relies a lot on human oversight. The main types of pollen monitoring devices rely on humans for weekly or daily maintenance; this limits the size of a pollen monitoring network and only allows pollen monitoring in densely populated areas. Usually major urban population centers that have a university or a institution that can implement such networks.

While this work will not be able to address all of the problems in the pollen monitoring community, it will no doubt help push the topic to a more standardized and automated way of working.

1.3 Motivation

From the main problems surrounding the task of automated pollen monitoring and the impact pollen has on human health and well being, the motivation for developing artificial intelligence systems capable of detecting, identifying and classifying pollen becomes apparent. Such a work would have a great potential benefit for society at large and it would allow for a better understanding of a very complex phenomena.

Developing fully automated monitoring systems, and then monitoring networks, would allow for a much better data acquisition pipeline that could be useful in alerting people and building prediction models.

Taking this amount of data and providing another scope than the initial one is a huge opportunity for the palynology field. The second motivation of this thesis is strongly correlated to this; namely, using ML techniques to extract new and valuable knowledge.

After investigating further the topic of pollen classification I was surprised that there is still a heavy reliance on humans for counting and classification of pollen grains, using microscope imaging. This further motivated me in approaching the topic of pollen classification and all of the tasks this involves such as data acquisition and building new data-sets, finding architectures that could be applied to the unstructured pollen data sources, identifying new ways to train ML models using artificial sample generation and augmentation and new approaches for image segmentation.

1.4 Research Aims and Objectives

The aims of this work are:

Identifying some of the modern fully automated particle analyzing systems that could be used for pollen monitoring. This involves a comparison of multiple devices with the goal of finding several types of devices that provide reliable data, that can operate with minimum human interaction and that have minimal drift in data distribution to be able to use them to create historical data sets of pollen.

Create or find data-sets of pollen obtained from a selection of devices that pass the quality controls. The data-sets should ideally be created using data from multiple devices (of the same kind). The data should be standardized and cleaned to allow for a wide range of numerical models to be able to ingest the information as easily as possible.

Developing methods for handling all the different and heterogeneous types of data that is used to describe pollen particles. Finding ways to use unstructured data (eg. microscopy images, scattering images, fluorescence spectrum etc.) to identify and classify the plant that produced a certain pollen type. This would involve finding the best approaches for feature selection and feature engineering.

Building and training numerical models that can classify pollen to a degree similar to what human experts are capable of. While this aim is more ambiguous it is almost a standard first step in any machine learning task to first try to match the performance of humans and after that try to overcome it.

The final aim of this work is the creation of a number of open-source Python libraries that would allow the end-user to directly apply these methods and techniques to new data streams obtained from similar automated devices.

1.5 Structure of the Thesis

The thesis has the following structure: Chapter two, is a literature review of the study of pollen in general, than a more focused look at the available data sources and instru-

mentation and a historical overview at all the attempts so far made for automatic pollen classification and monitoring.

Chapter three, presents the theoretical machine learning frameworks used to construct and train models that can classify pollen. This chapter looks at the major sub-task related to pollen monitoring and at methodologies used in general to improve model performance.

Chapter four, is a case study on a automated particle analyzer that relies on microscope photography to capture pollen information. This chapter presents multiple ways of classifying pollen in images and all the difficulties this brings.

Chapter five, is a case study on a automated particle analyzer that takes advantage of fluorescence spectroscopy to obtain data about pollen. In this chapter we look into using multi-modal data to make pollen classifications.

Chapter six, looks at ways of analyzing and comparing large scale data sets of other atmospheric parameters that might have a good predictive capability related to the pollen seasons.

Finally, conclusions are drawn in chapter seven and a overview of all the published papers related to this work is shown.

Chapter 2

Literature Review

The Literature Section will touch on a number of topics that are important to the task of pollen classification and monitoring in general. It starts from a wide description of the field of palynology, it the moves on to the instruments used for pollen monitoring, it discusses the pros and cons of existing databases of pollen and finally it looks at the historical attempts at automated pollen classification using machine learning.

This chapter is structured as follows: First, a brief introduction to palynology is made in section 2.1. Then we present all the available automated devices capable of pollen monitoring work in section 2.2. We compare and contrast all of the existing data-sets of pollen data in section 2.3. And finally we look over the history of automatic pollen classification and the difficulties encountered so far.

2.1 Palynology

Palynology is literally the "study of dust" or of small particles, be they organic or inorganic. The classical approach involved the gathering of samples from the air, from water, or from deposits including sediments by a palynologist. These samples, were then analyzed to try to find clues to the life, environment, and energetic conditions that produced them. Palynology as an interdisciplinary science stands at the intersection of earth science (geology or geological science) and biological science (biology). With the rise in interest in automated or computer-aided approaches to palynology, the intersection also involves computer science and machine learning as fields that overlap.

2.2 Pollen monitoring instrumentation

Pollen monitoring instruments can trace their history to the initial experiments of Blakley in nineteenth century[63]. Blakley was a "hay fever" sufferer and started investigating his illness in 1859. His experiment were first published in 1873 [63].

These first attempts used glass slides coated in glycerine. These slides were exposed to the air and after a fixed period of time they were analyzed under a microscope to get a count for certain species of pollen. This methodology is still currently applied for pollen monitoring albeit the devices are a little more complex.

The creation of the modern manual pollen trap can be attributed to Hirst and his seminal paper An Automatic Volumetric Spore Trap in 1952 [26].

Hirst trap [26] was described as a suction trap in which spores enter a narrow orifice, directed into the wind, and impacted on a Vaseline-coated microscope slide moved across the orifice at 2 mm/hr. This setup allows for estimates of the spore content of air to be made with a higher efficiency than by previous traps and at different times of day and thus more closely correlated with weather variations.

Kramer-Collins spore sampler [31] was an improvement over the original Hirst trap because it allowed for continuous measurements over longer periods of time.

Burkard spore trap [47] is a more recent version of the Hirst type trap. In this instrument air is drawn into a 14 mm x 2 mm orifice at 10 liter per minute, and any airborne particles with sufficient inertia are impacted on either a greased tape or a greased microscope slide beneath the orifice.

The VPPS made by Lanzoni is the other main descendent of the Hirst trap. This device is a very versatile and reliable volumetric sampler, manufactured using high resistant materials to run under severe atmospheric conditions for long time.

The Durham trap [50] is a completely different approach to pollen capture. It is based on Erdtman's method [22] from two plexiglass discs 5 mm in width and 22.5 cm in diameter, with 10.5 cm separation between them. A microscope slide was held in place with duralumin holders. The slide surface was completely covered with Vaseline as an adhesive. I does not use any moving parts and relies on wind to bring in pollen particles.

Another type of pollen capture device is the Rotorod trap [66]. This type of device is a rotating-arm impactor that recovers airborne particles on two rapidly moving plastic collector rods. The driving priciples are described in [44].

The GRIPST-2009 is a rotational impactor type pollen trap. Rotational impaction samplers have become widely used devices when it comes to collecting air-borne particles. They have a proven to be effective at capturing particles as small as 2 microns which makes them excellent at capturing pollen or other spores.

Automatic-KH-3000 [60] is a particle counter especially aimed to measure the number of pollen by using a scattering of a semiconductor material laser beam. KH-3000 is designed to gather pollen particles effectively through a air-sheet column. The device measures and discriminates pollen by measuring a forward scattering and a side scattering with a semiconductor laser beam.

Rapid-E from Plair is a instrument that accurately and comprehensively analyze single aerosol particles in real time. Fully automated, it characterizes any airborne

particle in the range of 0.5-100 micrometers, matching and opening up numerous applications in environmental monitoring, and beyond.

The pollen monitor BAA500 is the first of a future line of fully-automated analysers for airborne particles and for particles in liquid media. It features automated sampling, optical detection and measurement, image analysis and archival storage of the samples. The device automatically extracts pollen grains from the environment with a virtual impactor, prepares microscopic specimens and analyzes and counts the extracted pollen grains under an automated light microscope with a dedicated image processing system.

SwisensPoleno is the latest generation of state-of-the-art measuring instruments for real-time pollen monitoring. With mature technology and network compatibility, SwisensPoleno enables fully autonomous and stable long-term measurement of local pollen concentrations [54].

Developed by Pollen Sense LLC, the APS-300 is a fully automated pollen imaging sensor that collects and images pollen and airborne particles down to less than 5 µm, in real-time (data reporting delay in <1 min). The APS-300 collects ambient air by an airflow system at a constant flow rate. The particles in the collected air adhered to the rotating tape medium, where a proprietary form of optical surface microscopy is performed. The collection service performs complex proprietary algorithms involving advancing, focusing, and lighting in order to obtain maximal information about each particle.

2.3 Pollen Data-sets

This section looks at all the pollen datasets that are publicly available.

The list of pollen data-sets previously used for classification tasks:

- POLLEN13K [8].
- POLEN23E [25].
- POLLEN73S [6].
- POLLEN20L-det .[29]
- Artemisia pollen dataset [39].
- Cretan Pollen Dataset [61].
- Classifynder 46 [58].
- ABCPollen [32].

Data-sets obtained from the automated Rapid-E devices:

Data-sets obtained from automated BAA-500 devices:

Dataset Name	Number of Classes	Number of Samples	Minimum Number Samples per Class
SAU-SRB [a]	14	85 k	985
SAU-LI [a]	11	399 k	16,114
SAU-CH [a]	10	50 k	1,075
MARS [b]	13	105 k	3,020

Table 2.1 Overview of available datasets from Rapid-E devices. [a] from [62]; [b] from [10].

1

- Data-set-15 [56] contains overall 51,277 samples and 15 classes and the originators obtained an unweighted average precision of 83.0 % and an unweighted average recall of 77.1 % across 15 classes of pollen taxa.
- Data-set-31 [55] is an expanded version of Data-set-15 but with some extra classes. In [55] the authors achieved an unweighted average F1 score of 93.8% across 15 classes and an unweighted average F1 score of 75.9% across 31 classes. While the result on these data-sets are very promising, the main issue still remains that these classifiers rely on classical image processing methods for segmentation. And these segmentation methods are not optimal for complex patterns formed by pollen particles on microscope slides.
- BAA-500 cropped dataset: contains over 45 thousand samples from 19 different types of particles, 16 pollen, 2 spore types and a class for debris [**?**].
- Alternaria segmentation dataset: contains over 3 thousand images with corresponding class mask for Alternaria spores [Citation needed].

2.4 Automatic pollen classification

This section presents an overview of past attempts at pollen classification. The data is compiled in Table 2.2.

Classifier	Features used	Pollen taxa	Reported Accuracy	Reference
C-SVC (SVM), CST+BOW	BOW, color, shape and texture	23	64%	[25]
Linear discriminant classifier, CNN	Automatic feature extraction	23	97%	[58]
Multivariate statistical classification	Texture analysis	6	94%	[34]
NN, leave-one-out classifier	Texture features	4	100%	[35]
Mahalanobis distance	Color, shape features	30	77%	[13]
Mahalanobis distance	Color, shape, geometric features	30	88.25%	[12]
Regression trees	Morphometric features	3	n.a.	[38]
SVM	Invariant gray-scale features (3D)	26	92%	[53]
Linear discriminant analysis	Texture/shape	13, 4	100%	[36]
MDC	Fourier descriptors	3	90%	[52]
MDC (majority voting)	Texture features (FSM)	5	85%, 87.4%	[14]
Nearest Neighbor, SVM	Group integration	26, 7	96.9%, 99.7% (SVM)	[49]
Linear normal classifier	Shape, stat. gray-level, pore/colpus	3	97.20%	[16]
SVM/MDC/MLP	Texture and shape features	3	89%	[51]
Bayesian classifier	Invariant features (local jets)	3	83%	[46]
Adaptive Bayesian Combination, LLC	Texture features, LLT	7	90.58%	[69]
MLP	Geometric, shape, texture features	3	90%	[2]
SVM	3D discrete spherical features	26, 33	96.3%, 91.8%	[68]
Mahalanobis distance	Color, texture, optical spatial frequency	3, 40	77%	[27]
KNN, Gaussian, SVDD	Morphological, shape, textural, color	5	92.30%	[17]
Linear discriminant analysis	Morphological, statistical, three space-frequency	15	99.40%	[48]
CNN	Automatic feature extraction	30	90%	[19]
SVM, Random Forest, Logistic regression	Color, texture	23	79%	[4]
CNN+RNN	Automatic feature extraction	10	100%	[20]
CNN	Automatic feature extraction	5, 11	99.8%, 95.9%	[30]
CNN	Automatic feature extraction	11	99.75%	[23]
SIFT	Local key points (3D)	27, 33, 28	88.25%	[64]
Random Forest	Geometric, textural	6	88.24%	[40]
NN + CNN	Flourescence Spectroscopy	11,13,14	74%,77%,80%	[62]
CNN + LDC	Automatic feature extraction	46	97.86%	[58]
NN + CNN	Flourescence Spectroscopy	11,13,14	77%,80%,84%	[10]
RBF SVM, CNN	LBP, HOG, Automatic feature extraction	4	87%, 90%	[8]

Table 2.2 Table of the automatic pollen classification attempts.

Chapter 3

Theoretical Framework

This chapter goes over all of the different types of algorithms and data processing methods used in this thesis. The role of this chapter is to be a glossary for all of the technical methods and procedures used through out this work. This chapter looks at generic algorithm and methods used in computer vision and object detection, with comments about the changes needed to make a algorithm work on pollen data.

In section 3.1, we discus computer vision; history and evolution.

In section 3.2, we look at classification algorithm in general and how they would apply to the task of pollen monitoring.

Section 3.3, deals with the procedures used to segment complex images to help in the detection or identification of pollen.

Section 3.4, presents ways of getting extra performance from ML models by fine-tuning there parameters and finding the best configuration for a specific task.

Section 3.5, shows the methods to do feature selection and engineering in order to make complex unstructured data more easy to use with a wide range of classification algorithms.

3.1 Computer vision

When talking about computer vision we are taking about the ability that allows computers or other computational systems to extract information from digital images, videos or other types of unstructured data (Lidar systems, radar systems, sodar systems). From the perspective of engineering, computer vision seeks to understand and automate tasks that the human visual system does.

Some of the typical task in the computer vision field are:

- Object recognition or classification – deals with analyzing an image or video for the presence of an object from a pre-specified catalogue of learned classes. This can be expanded to identifying the position in the image/video.

- Identification – is the individual identification of instances of objects. Identification of a specific persons face versus just identifying that a face is present in the image. Examples include identification of a specific person's face or fingerprint, identification of handwritten digits, or identification of a specific vehicle.
- Detection – is when an image is analyzed for the presence of certain learned conditions. Detection of abnormal cells or tissues in medical imaging, detection of vehicles at automatic road tolls systems. Detection systems can have multiple layers. A simple algorithm is used to find regions of interest and then a more advanced and compute heavy algorithms is used to further process only the areas of interest.

In the following section we will go over some of the algorithms used for the detection and classification of pollen.

3.2 Classification Algorithms

This section looks at classification problems in general and what are the main types of numerical models used in this work. The focus will be on classification algorithms that have been used for pollen classification.

Linear discriminant analysis (LDA), or normal discriminant analysis is a method used in statistics and other fields to find a linear combination of features that can be used to characterize or separate objects or events into classes. The resulting combination can be used as a linear classifier for new samples.

Quadratic discriminant analysis (QDA), is related to LDA, but it drops the assumption that measurements from each class are normally distributed, unlike LDA, in QDA there are no assumptions about having identical covariance between classes.

Support vector machines (SVM), are a class of supervised learning models that can be used to analyze data for classification and regression. SVMs are a robust prediction method that rely on finding a hyper-plane that separate elements from different classes. SVM works by mapping training examples to points in a high dimensional space so as to maximize the distance between the two categories.

Naive Bayes classifiers, are a family of probabilistic classifiers based on using Bayes theorem with strong or naive independence assumptions between the input features.

Decision Trees, are used as a predictive modeling approach in many fields. The method uses a decision tree to go from observations about an object or event to a class. In the tree structures, leaves represent class labels and the branches represent combinations of features that lead to those classes.

Random Forest, or random decision forest is an ensemble learning approach based on using multiple decision trees. These types of meta-models can be used for classification and regression. For the classification task the outout of the ensemble is class selected

by the most trees. Random forest correct the main problem of decision trees, that of overfitting the training data.

The perceptron, is an algorithm for supervised learning of binary classes. The model has the task of learning a linear threshold function between two classes by looking at a multinomial combination of the input features. As a linear classifier, the single-layer perceptron is the simplest feedforward neural network.

While there isn't a clear distinctions between what constitute a deep learning model compared to the classical ones. The distinctions relies more on the effect observed when training such models on large datastes. While the classical models are able to train on large datasets they suffer from diminishing returns with an increase in data set size. On the other hand deep learnign models have a much greater model capacity and can continue to improve significantly with the increase of the dataset.

As an example the multi layer perceptron is an extension of the original perceptron, with the main difference being that it introduces a non-linear function to act as an activation function between layers and a number of "hidden" layers between the input and output.

A multilayer perceptron (MLP) is a family of feed-forward artificial neural networks (ANN). This type of model is usually fully connected, meaning that all the weights of a layer are influenced by all the weight of the previous layer. The MLP is feed-forward, because when in use, the data flows from the input layers to the hidden layers and to the output layers without any way of higher layers to influence previous layers. While MLP is used generally to describe any feed-forward ANN, the strict definition refers only to networks build by stacking multiple perceptrons, with an activation function to introduce non-linearities. MLPs are sometimes considered basic neural networks and are the building block for more complex architectures.

An improvement over the MLP, when it comes to computer vision, is the convolutional neural network (CNN). CNN have been developed specifically to be used with unstructured data such as images, videos or other 2D/3D data. The main advantage over regular ANN is the significantly reduced number of parameters that have to be trained due to the shared-weight architecture of the convolution kernels or filters that slide along the input feature and provide some translation-invariant responses.

Another advantage to CNNs is that they can be used with images without much pre-processing compared to the methods presented previously. Because all the filters of a CNN are learned, the model learns also the best pre-processing or feature extraction for the specific task.

3.3 Segmentation Algorithms

This section looks at image or signal segmentation problems in general and what are the main types of approaches used.

When discussing about the segmentation of 2D/3D data there are several definitions depending on what the desired output is. Semantic segmentation is an approach for detecting all the pixels belonging to a specific class in an image. For example identify all the pixels that represent the sky or the ground in an image. Instance segmentation is an approach that identifies, for every pixel, a object instance it belongs to. It detects each distinct object of interest in an image but not what class it is. Panoptic segmentation is a combination of semantic and instance segmentation and it can be used to identify both the objects and their hierarchical classes.

Thresholding, is one of the simplest methods for image segmentation. This method is based on selecting a threshold value to turn a gray-scale image into a binary image. The key aspect of this methods is selecting the right threshold value for the specific task. Several methods for selecting the threshold value is using maximum entropy method, balanced histogram thresholding, maximum variance method or even k-means clustering. While these method work on really simple input images, it is much more difficult to effectively use for complex images such as microscope images of pollen.

Edge Detection, is well-developed branch of digital image processing. Region boundaries or objects and their edges are closely related, since there usually is a sharp adjustment in the intensity at the separation boundary. Edge detection techniques are usually the base for more complex image segmentation techniques. Edge detection usually find edges that are disconnected. To segment and object from an image, there is a need for closed region boundaries. These can be obtained by applying multiple morphological operations such as dilations and closings to the image with the edges detected.

Connected-component labeling or connected-components analysis, is the creation of a labeled image in which the positions associated with the same connected component of the binary input image have a unique label. This method relies on a pre-processing step to obtain a binary mask for the image and the it segments the binary mask into regions that are connected i.e., the flood fill algorithm in any image processing software.

There are also deep learning models capable of doing image segmentation, without the use of other digital image pre-processing. The advantage of such approaches is that it frees the user from having to select the best set of parameters for distinguishing between different objects in images. It moves this task to the model and the model learns the best way by looking at the input data.

The family of CNNs that are best known for their segmentation power are the fully convolutional networks (FCN). This type of network uses only convolutionary layers and its output is of the same size and shape of the input. This means that the network learns to label each individual pixel in the input image.

U-nets are one such type of network that were developed specifically for biomedical image segmentation. A U-net is built by adding skip-connections to a fully convolutional network. This allows for more information to flow from the input side to the output side

of the network and this makes the network learn faster using fewer training images while also improving the segmentation.

3.4 Hyper-parameter search

In this section, we talk about the difficulty in selecting the best configuration of parameters for specific model and task and how these settings can be found by doing different types of searches in the hyper-parameters space.

Hyper-parameters, in general are the configuration or setting that a user has to select when building a numerical model. These parameters range from the learning rate used to the numbers of layers in a MLP or the size of the convolution kernels in a CNN. Because these parameters heavily influence the performance of the final trained model their selection is a crucial step when developing a machine learning solution.

There are many approaches to finding the optimal setup for a model and task pair but this can be difficult to find because of the large number of knobs to tweak. If for example a model can be configured by selecting between 10 learning rates, 10 regularization parameters and 10 activation functions. The model will have a total of 1000 version that should be tested to find the best one. This gets incredibly difficult if the hyper-parameter can take continuous values and if we have a large number of parameters to select from.

The Grid search method, for finding the optimal hyper-parameter configuration relies on iterating over all the combinations of parameters from a fixed set and training a model with those parameters. This method realistically works only for simple models with few parameters or as a training exercises.

Random search, is similar to the grid search but it doesn't rely on going over all combination but selecting random values for hyper-parameters and training models with those settings. This approach does not guarantee that it will find the optimum for a specific model, task pair but in practice it usually find good enough configurations and can be considered a starting point for more advanced or guided types of searches.

Even better than a random search is a guided approach, such as Bayesian optimization [67]. This approach tries to optimize over the search space by first constructing a posterior distribution of functions that best describes the function to be optimized. Then, as the number of observations grows, the posterior distribution improves and the algorithm becomes more certain of which regions in the parameter space are worth exploring and which are not.

3.5 Feature selection, engineering and data augmentation

Data cleaning refers to the pre-processing steps related to making a data set more uniform and homogeneous before any analysis. This can involve editing and correcting images, structuring data and computing data set statistics.

After a data-set has been cleaned, the next step usually involves some amount of feature scaling.

Even after cleaning a data set and normalizing the features some steps are required when using classical machine learning. Dimensionality reduction, refers to a family of techniques that are used to distil the information present in a data set to the smallest number of features possible. This is done for multiple reasons such as, the more inputs features available the more difficult it is to find the feature that actually contain useful about the predicted value. Raw data are often sparse as a consequence of the curse of dimensionality [Citation Needed], and analyzing the data is usually computationally intractable.

Chapter 4

Microscopy images used in Pollen Classification

This chapter presents the pollen classification results obtained on microscope image data from the BAA-500 device, from Hund. The BAA-500, device, is an automated particle monitoring device that was designed to mimic the human approach to pollen monitoring. While this type of automation is very easy to understand and validate, it involves more complex steps then just trying to create a new method from the ground up with automation in mind.

This chapter will look at a number of task surrounding the main pollen classification task such as image segmentation, feature engineering of complex organic patterns, using pre-trained large deep learning models and developing new training approaches to improve performance.

The chapter is structured as follows: Section 4.1, presents setup and the standard operating procedures of the BAA500 instrument. The raw data and the built-in processing steps are also discussed here.

In section 4.2, a new microscopy image pollen data-set is presented along with all the steps done to ensure the quality of the data.

In section 4.3, the classification algorithms used are briefly presented along with some talk about their setup and training procedure. This section will also touch on the use of segmentation to enable the use of classifications models using microscope imaging data that is very heterogeneous.

Finally, section 4.4 presents the result of multiple classification algorithms using classical segmentation or deep learning enabled segmentation+classification of microscope images.

4.1 Instrument description

The BAA-500 device is capable of fully-automated analysis of airborne particles or particles in liquid media. In pollen monitoring, it is used as a feature complete system that fully automates all steps that were previously made by humans in Hirst type traps such as sampling, preparation of slides, optical detection and image analysis. In Figure **??** the device is shown in a outdoor environment, the device is housed inside a controlled atmosphere shelter.

4.2 Pollen Data-set creation

During this work, two data-sets were created. These datasets are related to the two task that must be tackled when working with the BAA-500 data. One obvious task is classification of pollen relying on the segmentation and the particles identified by the BAA-500 software and the second task is first segmenting the raw image data and then classifying.

The first data-set, contains cropped images that were initially obtained from the BAA-500 classification and that have been manually verified by pollen experts.

The second data-set, was created by for the segmentation task. More precisely for the identification of a type of spore that the BAA-500 had difficulties finding in images with its standard approach. In the following parts the data acquisition steps will be presented, after that the processing, data augmentation and feature engineering steps are discussed and finally, an overview of the data-sets is presented.

4.3 Architecture selection

In this section the model types used for pollen classification and microscope image segmentation are presented. The section is split along the lines of the two complementary task. We first discuss about the types of models used for pollen classification, with the assumption that there is a way to obtain the segmented or cropped images. The second discussion is related to the problem of segmenting the raw images and identifying where and what we have in those images.

This approach was taken because the BAA-500 device already does a segmentation of the images and we want a model that can use the cropped images. For the classification task we want better class accuracy. The segmentation task, is more of an extension to what the current BAA-500 is capable of identifying.

The final discussion, in this section will be about the metrics that have to be used when training models to solve the two tasks.

4.4 Results

For classical approaches to pollen classification we started from the work done by [8].

The data was first re-scaled from the 0–255 range to 0–1 range to make it easier for the models to train. Our cropped samples were 360×360 in size and a dimensionality reduction was necessary to accomplish this. We used histogram of gradients (HOG) and local binary pattern (LBP) as two different feature engineering steps.

Because they showed good results in previous works [8] some architectures were selected including Support Vector Classifiers (SVCs), Random Forest (RF), Decision Trees with Adaboost ensemble and Multi-Layer Perceptron (MLP). All of the selected models come from the sci-kit learn packages and were used with default settings.

Table 4.1 Classification F1 score for classical ML approaches.

Model type/Feat.Eng.	HOG	LBP
Linear SVC	0.46	0.46
RBF SVC	0.29	0.48
Random Forest	0.46	0.53
ADABoost	0.41	0.52
MLP	0.47	0.62

The experiments with CNN architectures are split into two groups: i) using a pre-trained model and fine tuning or ii) training from scratch. When using the pre-trained model the gray-scale images are treated as RGB images with identical information on all of the channels, to have the same input shape as the models trained on Imagenet [21]. The fully connected layers, on top of the convolutional part, of each model is replaced with a Global Pooling layer and a final soft-max layer used for classification. The convolutional layers are frozen and only the small fully connected part of the network is trained. The results are good and quite close for most of the models with an exception being the ResNet50 that suffers the most when pre-trained.

Table 4.2 Classification F1 score for deep architectures.

Architecture used	Pre-trained on ImageNet	Fully train on pollen
VGG-16	0.82	0.90
VGG-19	0.80	0.92
ResNet50	0.59	0.90
InceptionV3	0.85	0.93
Xception	0.86	0.93
DenseNet201	0.87	0.92

To have a comparison between the U-net model and the other CNN architectures used for classification, we propose an aggregation method to reduce an output mask to

just one label. The output mask of the U-net is added on the x and y spatial dimensions leaving a $1\times1\times20$ vector with the highest value in the class that is most present in the image. The first element of the vector is discarded because it is used for the background.

Using this approach we obtained a classification average unweighted F1 score of 0.95 for the U-net. This is a good way of validating that the model learned more from the pollen particles than other approaches.

Table 4.3 Classification class mean unweighted class IoU for U-net and variants at different widths.

U-NET VARIANT	W-4	W-8	W-16	W-32
FCN	0.61	0.72	0.71	0.79
U-NET	0.67	0.81	0.85	0.86
U-NET(ADD)	0.65	0.79	0.85	0.88
FCN + RES	0.64	0.78	0.82	0.82
U-NET + RES	0.66	0.78	0.83	0.87
U-NET(ADD) + RES	0.66	0.82	0.83	0.88

Chapter 5

Multi-Modal fluorescence and scattering data used in Pollen Classification

This chapter presents the pollen classification results, obtained on multi-modal data, from the Rapid-E device , from Plair. The chapter is structured as follows: Section 5.1, looks at the technical setup of the device and the experimental setup required to obtain good quality data.

Section 5.2, presents the experimental setup required when creating data-sets that can be used for developing ml models. This section will also look at the pre-processing steps, data cleaning and data augmentation available for this sort of configuration.

Section 5.3, looks at what types of models can be trained on the complex multi-modal data-set created from the Rapid-E particle analyzer. While also looking into maximizing performance by finding the best model type and model hyper-parameter configuration for the task of pollen classification.

Finally, in section 5.4 a comparison of all the results is presented. In this part the advantages and disadvantages of different model types are presented and discussions are made on what could be the best approach for developing a operational system for pollen classification using the Rapid-E device. Also some of the limitations of this system are discussed.

5.1 Instrument description

The Rapid-E device is a automated particle analyzer, developed by Plair. While the marketing around the device hints at plug-and-play capabilities for pollen classification, actual real-world usage shows that some setup and calibration is required. The device is setup in an enclosure that has a air-conditioning unit that protects the instrument from overheating in the summer and freezing in winter. The device can be placed in most

locations where it has a mains power source. The ideal setup also has a wired internet connection to allow easy access to the data and to the control software interface, but the device an operate without internet, as it has sufficient internal storage for multiple years of data.

5.2 Pollen Dataset creation

Under normal operations, the Rapid-E handles up to 10 thousand particles per minute, but realistically values never go over 2-3 thousand. This is because the flow rate of the device is quite low and the concentration of pollen is very small compared to other types of aerosols. To be able to only capture the information of only particles in the pollen size range the device uses to types of thresholding. While the Rapid-E, is very good at operating without human intervention or at the most minimal maintenance once every 2-3 months, to be able to create good quality data sets a good device is not sufficient and a strictly controlled experimental setup is also required. In this section we will look at multiple possible experimental setups for creating a pollen data-set, and compare what works and what doesn't.

The first set of experiments relied on plants samples that also had the flowering part of the plant. The initial setup uses a stream of inert N_2 gas to act in a similar way to how pollen gets lifted by the air in the real world.

In the second experimental setup, only pollen grains are used instead of having the entire plant sample. While this approach is more labor intensive on the pollen gathering side it guarantees cleaner data-sets because we remove almost all contaminants.

The results of this experiments in creating new pollen data-sets with a Rapid-E device were also documented and published as a conference paper in 2021 [11].

5.3 Architecture selection

When approaching any new machine learning task the first question that should be asked is what is the simplest model that is able to accomplish the task with good enough performance.

The main reasons to approach a problem in this fashion are: We always have a baseline to compare against, if we start from simple models and increase the complexity as needed. We can identify some problems with the data-set using simple architectures instead of trying to over-engineer models to compensate for poor data quality. We can more easily identify when large models start to over-fit, the training data, because we can compare against the training and testing performance of both small and large models. Finally, we might find models that provide sufficient performance early and we can stop the search for more complex models if we observe we have diminishing returns, with increase in model complexity.

For pollen classification, using Rapid-E data, models can be split into two types, those that need dimensionality reduction and those that can directly use the three features types from the Rapid-E, the scattering image, the fluorescence spectrum and the lifetime signal.

The models that need dimensionality reduction to work are: Naive Bayes (GNB), Quadratic Discriminant Analysis (QDA), Decision Trees(DT), and to a lesser degree multi-layer perceptron (MLP), which can use the raw data but loose some of the spatial information present in the features because we have to flatten the 2D arrays. The simple model implementation is from the Python library sklearn [45]. And the dimensionality reduction methods used are Principal Component Analysis (PCA), Independent Component Analysis(ICA), Gaussian Random Projection (GRP) and Sparse Random Projection(SRP).

Moving to more complex models the need for dimensionality reduction goes away as CNNs can directly use 2D arrays as inputs, i.e. images. In Figure **??** the overview of the CNN model is presented. This model has different convolutional branches for each of the three feature types. The roles of these branches is to extract the useful information from the scattering image, fluorescence spectrum and lifetime signal. The common dense layer part of the network has the role of using the information provided by each of the feature extractors to make a global classification using all features.

The feature extractors can be further broken down into convolutional blocks. Each block is built using one or more identical convolution layers with ReLU activations [1], followed by a batch normalization layer [28] and a max pooling layer [42]. After each max pooling layer the number of filters in the convolution layer is doubled to allow the model to cope with the reduction in the dimensionality introduced by the max pooling operation.

5.4 Results

In this section, we evaluate the results of the trained model and compare to the previous work performed on this type of classification for all four datasets.

The models used were all built using the insight gained after the hyper-parameter search. The case of ensembling the Decision Tree was treated separately because there exists a methods for building ensembles, Random Forest, that relies on the trees being non-identical.

In Table 5.2, we can see that all models gained accuracy using the best hyper-parameter configuration found. The gains were not equal across the board with most of the improvement having been obtained by the models trained on SAU-SRB and MARS. These were the most difficult datasets when looking at the number of classes. The increase in performance is more visible when looking at each individual feature extractor. Comparing rows 5–7 of Table 5.2 with rows 4–6 from Table **??**, we can see

Algo	Config/Dataset	RO	LI	SRB	CH
GNB	Best Conf.	31%	32%	34%	54%
	Ensemble	31%	32%	35%	54%
QDA	Best Conf.	44%	48%	53%	60%
	Ensemble	44%	48%	53%	61%
DT	Best Conf.	42%	51%	52%	63%
	Ensemble	47%	56%	57%	69%
	Rand. Forest	42%	55%	54%	63%
MLP	Best Conf.	63%	71%	67%	73%
	Ensemble	**66%**	**74%**	**70%**	**77%**

Table 5.1 Classification accuracy for all classical models.

that all the feature extractors were much better at making classifications. Examining the error rate for the scattering image feature across all four datasets, we obtained a drop from an average of 42% in the initial case to 38% using the best architecture found. For the fluorescence spectrum, the improvement was event greater, from an average error rate value of 46% to 37%. Finally, for the lifetime signal, the average error rate across the datasets decreased from 44% to 40%. In an ideal scenario, these improvements should translate to an similar improvement in the combined model. However, what we observed in practice is that the combined model error rate decreased but at diminishing returns. With the fine tuned architecture, compared to the initially proposed architecture, we obtained a relative reduction in the error rate of 13%, from 23% to 20%, for SAU-SRB; of 12%, from 16% to 14%, for SAU-LI; of 13%, from 15% to 13%, for SAU-CH; and of 20%, from 24% to 19%, for MARS.

Table 5.2 CNN Model performance after hyper-parameter tuning (accuracy).

Data-Set	**SAU-SRB**	**SAU-LI**	**SAU-CH**	**MARS**
Baseline [a]	74%	73%	80%	-
Initial architecture [b]	77%	84%	85%	76%
Best architecture	80%	86%	87%	81%
Scattering image only	58%	62%	61%	64%
Fluorescence spectrum only	56%	72%	66%	58%
Lifetime signal only	68%	59%	72%	41%

[a] Results from [62]; [b] results from [9].

Chapter 6

Pollen predictive models a future work

While the previous chapters focused on pollen detection, identification and classification, in this chapter we will look at the methods and methodologies used to predict the pollen season.

Accurate pollen concentration forecasts are an ideal solution to the problem of treating and managing allergies. If predictions could be made at a genus or species level people that suffer from allergies could start taking preventative treatment in advance of the actual pollen blooms. This would reduce the impact on the quality of life and prevent life-threatening events.

Pollen forecast try to predict a number of parameters including the entry dates of phenological phases and the start, peak and end of the pollen season. Also, given enough data, attempts have been made at predicting the day to day concentrations of pollen for multiple species.

There are two main types of models that are currently being developed to allow for pollen forecast: Observation-based models and Phenological models.

Observation-based models make no a-priory assumptions about the relationship between the pollen concentration and other atmospheric parameters. This class of models includes regression models, time-series modelling and applications of artificial intelligence methods to pollen data.

Phenological models try to model the entire life-cycle of the plant that will produce the pollen. These types of models use atmospheric parameters and observation on the spatial distributions of certain plant species to make predictions related to the pollen season and concentration. Phenological models are a type of process-based models, because they are built on assumptions rooted in experimental results on plant physiological responses to various environmental variables

In this chapter a number of atmospheric parameters that impact the pollen season will be discussed along with methods for analyzing long time-series of data. The parameters selected are important because they can be more easily monitored in an automated and continuous way compared to pollen monitoring.

A variety of different independent variables have been previously used used to predict daily average pollen counts, and include minimum, maximum and mean daily temperatures, rainfall, relative humidity, sunshine hours , wind speed and also direction and persistence, and the amount of pollen recorded in the previous days.

The chapter is structured as follows: Section 6.1 is a look at ways to monitor relative humidity or atmospheric water content. With relative humidity being an import an proxy variable for the start of the pollen season as detailed in [5], [7], [18], [70]. Section 6.2 is a long term analysis of temperature and UTI in particular. In this section we look at the rising trend in global temperatures. Section 6.3 discusses freezing rains and the impact they can have on vegetation. Section 6.4 is a case study regarding an event that involved snow forming around particles that have been transported over long distance.

6.1 Relative humidity impact on the pollen season

Water is an essential element in plant growth and development, but high levels of relative humidity and rainfall have a significant impact on the reproductive cycle of plants [5], [7]. Being able to accurately and consistently measure the atmospheric water content is an essential part in developing any pollen predictive models.

In this section we will look over some methods for continuous measurement of total precipitable water (TPW) using a Cimel sun photometer operating at a continental site in southeast Europe and compare against TPW obtained from a collocated microwave radiometer and nearby radiosondes during the 2007-2017 period.

6.2 Temperature and the effect on the pollen season

Temperature has a direct link to the start of most plants pollination season[24], [43], [59], [15]. On top of that the climate of a region heavily influences the spatial distribution of plant species.

This is why it is important to analyze the trend in the planet warming because this will impact the pollen season of many plant and will allow the spread of plants that have historically only been around hot areas to new territories further north.

In this section we will look over methods to analyze the bioclimatology of thermal stress on large areas and the impact it has on ecosystems. To be able to do this large datasets are required. These datasets have to be representative for large areas and for long periods of time. One such dataset is the Universal Thermal Climate Index (UTCI) derived from ERA5-HEAT reanalysis. While this type of index is used to describe the heat or clod stress felt by humans it is still a good proxy for the general effect of temperature on the ecosystem.

6.3 Freezing rain and impact on vegetation

Freezing rain and frost have a very powerful effect on most plants [37], [65]. Every autumn, winter and spring many plants run the risk of damage caused by cold weather. Depending on plant type the damage can be caused by anything from overnight frost or freezing rain events to prolonged periods with freezing temperatures.

In this sections we look at the atmospheric condition required for extreme cases of freezing rain in order to be able to predict future events and use this type of data to adjust pollen calendars. This study started by analyzing the event that took place on 24–26 January 2019, when a high-impact freezing rain event affected parts of southeastern Romania [3].

6.4 Orange snow and the long range transport of large particles

This section discusses an unusual phenomenon that was observed over Romania, on the morning of 23 March 2018. The event was observed centered in the southeastern part of the country, and it involved a fresh-layer of orange snow. The event was extensively reported in mass-media and social-media and raised questions about the origin and the possible impact of the orange snow (Figure **??**).

Even though Saharan dust intrusions are a common event in Romania and in Europe, their occurrence during negative temperature conditions is very rare. Saharan dust intrusion occurs over Europe mainly during spring and, in general, is not accompanied by snow at low altitudes.

In this study, [41], an analysis of both synoptic-scale conditions and the chemical and physical properties of the deposited dust particles was realized. The source of the dust was confirmed by both the elemental ratios of the main components (e.g., Al, Ca, Mg, Fe, K) and by using back-trajectories to see the origin of the intrusion. For example, the (Ca+Mg)/Fe ratio of 1.39 was characteristic for the north Sahara.

Chapter 7

Conclusion

This chapter has the role to bring together all the work presented in this thesis and to compare the results to the stated objectives, to put all the realizations in the broader context of the field and to highlight the personal contribution of the author, to list all of the published papers done during this PhD and finally, to look at future developments or future direction of research.

7.1 Objectives and Results

This section looks at the result obtained during the writing of this thesis. The particular results are split along the chapters where they are presented and discussed.

Chapter 4 is a comprehensive analysis on the ML models that can be used to classify or segment microscope image samples of pollen from BAA-500 devices.

(a) Building a new public pollen dataset using expert verified data from a BAA-500 device.

(b) Classification of pollen images using classical machine learning approaches.

(c) Classification of pollen images using pre-trained deep learning models.

(d) Segmentation of pollen particles in microscope images images deep learning models.

(e) New approach to construct artificial microscope images that can be used to train segmentation models.

(f) Building new public data set for segmentation of Alternaria spores.

(g) Validating the artificial data training approach on historical data and during a measurement campaign.

Chapter 5 is a comprehensive analysis on the models that can be used to classify multi-modal, pollen, data obtained from Rapid-E devices.

(a) Building a new public pollen dataset using data from a Rapid-E device.

(b) Building classical machine learning models capable of classification of pollen on Rapid-E data for low power use-cases.

(c) Improving on the current state of the art in classification accuracy. on existing Rapid-E datasets.

(d) Developing data augmentation methods that allow for improvements in performance.

(e) Identifying efficient setups for hyper-parameter search on large models.

Chapter 6 has the objective of finding independent atmospheric variables that could be used as either inputs for pollen concentration predictive models or as proxy data-sets for the pollen concentration directly. In this chapter multiple studies are presented that relate to these atmospheric variables.

(a) Performed the analysis for the sunshine duration and cloud coverage in order to determine the biases introduce to Cimel sun photometers. Useful in determining the Total Precipitable Water, an important parameter in predicting pollen concentrations.

(b) Performed the analysis on Universal Thermal Climate Index (UTCI) derived from ERA5-HEAT reanalysis to derive the trend for thermal stress over entire Europe continent.

(c) Performed the analysis on long term radiosonde data to identify all of the freezing rain events that appeared in Romania for the period 1900 -2000 to find a methodology to more accuratly predict such events and to be able to identify exceptional cases.

(d) Development a methodology for fusing LEM images for different chemical components to create visualisation and help in the identification of particles transported over long ranges.

7.2 Original contributions

1. Improvements to pollen classification using Rapid-E data in [1].

2. Creating new pollen data-set from a Rapid-E device located near Bucharest, Romania in [2].

3. Further improvements to pollen classification using Rapid-E data by using advanced hyper-parameter searches in [3].

4. Search for classical machine learning models that can classify Rapid-E data and an extensive hyper-parameter search to improve results in [4].

5. Pollen classification and image segmentation on pollen data from BAA-500 device in [5].

6. Creating new pollen data-set using data from multiple BAA-500 devices in [5].

7. Developing methodologies for training segmentation models using artificial data on BAA-500 samples in [5].

8. Building and training models capable of identifying Alternaria spores in Baa-500 data in [6].

9. Performed the analysis for the sunshine duration and cloud coverage in [7].

10. Visualization and analysis of long term data-set development if in [8].

11. Developed the methodology, the processing software and the visualization in [9].

12. Developed the methodology for image segmentation and visualization of electron microscope data for aerosol particles in [10].

7.3 List of published works

1. **Boldeanu, M.**, Cucu, H., Burileanu, C., and Mărmureanu, L. Automatic pollen classification using convolutional neural networks. In 2021 44th International Conference on Telecommunications and Signal Processing (TSP), pages 130–133.

2. **Boldeanu, M.**, Marin, C., Ene, D., Mărmureanu, L., Cucu, H., and Burileanu, C. Mars: the first romanian pollen dataset using a rapid-e particle analyzer. In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pages 145–150.

3. **Boldeanu, M.**; Cucu, H.; Burileanu, C.; Mărmureanu, L. Multi-Input Convolutional Neural Networks for Automatic Pollen Classification. Appl. Sci. 2021, 11, 11707. https://doi.org/10.3390/app112411707

4. **Boldeanu, M.**; Cucu, H.; Burileanu, C.; Mărmureanu, L. Pollen Classification using Classical Machine Learning Algorithms on Fluorescence and Scattering Imaging, Buletinul Polithenicii (Not yet Published)

5. **Boldeanu, M.+**; González-Alonso, M.+; Cucu, H.;Burileanu, C.; Maya-Manzano, J. M. and Buters, J. T. M., "Automatic Pollen Classification and Segmentation Using U-Nets and Synthetic Data," in IEEE Access, vol. 10, pp. 73675-73684, 2022, doi: 10.1109/ACCESS.2022.3189012. [1]

[1] Authors with + contributed equally.

6. González-Alonso, M.+; **Boldeanu, M.+**; Koritnik, T.; Gonçalves, J. ; Belzner, L. ; Stemmler, T. ; Gebauer, R. ; Grewling L.; Tummon, F.; Maya-Manzano, J. M. ;Ariño, A.H.; Schmidt-Weber, C.; and Buters, J. T. M., Alternaria spore exposure in Bavaria, Germany, measured by using artificial intelligence algorithms in a network of BAA500 automatic pollen monitors., Scientific Reports (Not yet Published)[2]

7. Fragkos, K., Antonescu, B., Giles, D. M., Ene, D., **Boldeanu, M.**, Efstathiou, G. A., Belegante, L., and Nicolae, D.: Assessment of the total precipitable water from a sun photometer, microwave radiometer and radiosondes at a continental site in southeastern Europe, Atmos. Meas. Tech., 12, 1979–1997, https://doi.org/10.5194/amt-12-1979-2019, 2019.

8. Antonescu, B., Mărmureanu, L., Vasilescu, J., Marin, C., Andrei, S., **Boldeanu, M.**, Ene, D., and Ţilea, A. (2021). A 41-year bioclimatology of thermal stress in europe. International Journal of Climatology, 41(7):3934–3952.

9. Andrei, S., Antonescu, B., **Boldeanu, M.**, Mărmureanu, L., Marin, C. A., Vasilescu, J., and Ene, D. (2019). An exceptional case of freezing rain in Bucharest (Romania). Atmosphere, 10(11).

10. Mărmureanu, L., Marin, C. A., Andrei, S., Antonescu, B., Ene, D., **Boldeanu, M.**, Vasilescu, J., Viţelaru, C., Cadar, O., and Levei, E. (2019). Orange snow—a saharan dust intrusion over romania during winter conditions. Remote Sensing, 11(21).

7.4 Future research

This section is a wish-list of research topics that I would enjoy working on after finishing the work on this thesis. This list could be structured as a post-doctoral project to add more to this body of work.

For future work on data from Rapid-E devices the main aspects that remain to be addressed are:

The identification of architectures that allow better generalization of models, from the training data to data from other devices of the same kind.

The development of transfer learning techniques between the existing data-sets.

The development of data-set from multiple devices/multiple regions that would allow for better overall performance.

A validation of the methodologies presented using co-located Hirst type traps as a baseline.

[2]Authors with + contributed equally.

For the BAA-500 device the next steps would involve:

A move from class segmentation for BAA-500 images to panoptic segmentation. This would allow for better pollen counts as it would be able to segment not just based on class but also based on instances.

An improvement of the generators used to create artificial images for training, provide more reliable examples(closer in distribution to the real data).

Develop some segmentation data-sets with multiple classes as opposed to the Alternaria one-class data-set.

A search for more and better image augmentation transformation that can be applied to pollen microscopy image data.

A multi year analysis of the performance of the segmentation/classification pipeline developed in this work on multiple devices from multiple locations.

References

[1] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375.

[2] Allen, G. P., Hodgson, R. M., Marsland, S. R., and Flenley, J. R. (2008). Machine vision for automated optical recognition and classification of pollen grains or other singulated microscopic objects. In *2008 15th International Conference on Mechatronics and Machine Vision in Practice*, pages 221–226.

[3] Andrei, S., Antonescu, B., Boldeanu, M., Marmureanu, L., Marin, C. A., Vasilescu, J., and Ene, D. (2019). An exceptional case of freezing rain in bucharest (romania). *Atmosphere*, 10(11).

[4] Arias, D. G., Mussel Cirne, M. V., Chire, J. E., and Pedrini, H. (2017). Classification of pollen grain images based on an ensemble of classifiers. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 234–240.

[5] Aronne, G. (1999). Effects of relative humidity and temperature stress on pollen viability of cistus incanus and myrtus communis. *Grana*, 38(6):364–367.

[6] Astolfi, G., Gonçalves, A. B., Menezes, G. V., Borges, F. S. B., Astolfi, A. C. M. N., Matsubara, E. T., Alvarez, M., and Pistori, H. (2020). Pollen73s: An image dataset for pollen grains classification. *Ecological Informatics*, 60:101165.

[7] Barnes, C., Pacheco, F., Landuyt, J., Hu, F., and Portnoy, J. (2001). The effect of temperature, relative humidity and rainfall on airborne ragweed pollen concentrations. *Aerobiologia*, 17:61–68.

[8] Battiato, S., Ortis, A., Trenta, F., Ascari, L., Politi, M., and Siniscalco, C. (2020). Pollen13k: A large scale microscope pollen grain image dataset. *CoRR*, abs/2007.04690.

[9] Boldeanu, M., Cucu, H., Burileanu, C., and Mărmureanu, L. (2021a). Automatic pollen classification using convolutional neural networks. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pages 130–133.

[10] Boldeanu, M., Cucu, H., Burileanu, C., and Mărmureanu, L. (2021b). Multi-input convolutional neural networks for automatic pollen classification. *Applied Sciences*, 11(24).

[11] Boldeanu, M., Marin, C., Ene, D., Marmureanu, L., Cucu, H., and Burileanu, C. (2021c). Mars: the first romanian pollen dataset using a rapid-e particle analyzer. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 145–150.

[12] Bonton, P., Boucher, A., Thonnat, M., Tomczak, R., Hidalgo, P. J., Belmonte, J., and Galan, C. (2011). Colour image in 2d and 3d microscopy for the automation of pollen rate measurement. *Image Analysis & Stereology*, 21(1):25–30.

[13] Boucher, A., Hidalgo, P. J., Thonnat, M., Belmonte, J., Galan, C., Bonton, P., and Tomczak, R. (2002). Development of a semi-automatic system for pollen recognition. *Aerobiologia*, 18(3):195–201.

[14] Carrión, P., Cernadas, E., Gálvez, J. F., Damián, M., and de Sá-Otero, P. (2004). Classification of honeybee pollen using a multiscale texture filtering scheme. *Machine Vision and Applications*, 15(4):186–193.

[15] Chaturvedi, P., Wiese, A. J., Ghatak, A., Drábková, L. Z., Weckwerth, W., and Honys, D. (2021). Heat stress response mechanisms in pollen development. *New Phytologist*, 231(2):571–585.

[16] Chen, C., Hendriks, E. A., Duin, R. P. W., Reiber, J. H. C., Hiemstra, P. S., de Weger, L. A., and Stoel, B. C. (2006). Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort. *Aerobiologia*, 22(4):275–284.

[17] Chica, M. (2012). Authentication of bee pollen grains in bright-field microscopy by combining one-class classification techniques and image processing. *Microscopy Research and Technique*, 75(11):1475–1485.

[18] Dąbrowska-Zapart, K., Chłopek, K., and Niedźwiedź, T. (2018). The impact of meteorological conditions on the concentration of alder pollen in sosnowiec (poland) in the years 1997-2017. *Aerobiologia*, 34(4):469–485. 30532345[pmid].

[19] Daood, A., Ribeiro, E., and Bush, M. (2016). Pollen grain recognition using deep learning. In Bebis, G., Boyle, R., Parvin, B., Koracin, D., Porikli, F., Skaff, S., Entezari, A., Min, J., Iwai, D., Sadagic, A., Scheidegger, C., and Isenberg, T., editors, *Advances in Visual Computing*, pages 321–330, Cham. Springer International Publishing.

[20] Daood, A. I., Ribeiro, E., and Bush, M. B. (2018). Sequential recognition of pollen grain z-stacks by combining cnn and rnn. In *FLAIRS Conference*.

[21] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[22] Erdtman, G. (1952). Pollen morphology and plant taxonomy. *Geologiska Foreningen i Stockholm Forhandlingar*, 74(4):526–527.

[23] Gallardo-Caballero, R., García-Orellana, C. J., García-Manso, A., González-Velasco, H. M., Tormo-Molina, R., and Macías-Macías, M. (2019). Precise pollen grain detection in bright field microscopy using deep learning techniques. *Sensors*, 19(16).

[24] Gehrig, R. (2006). The influence of the hot and dry summer 2003 on the pollen season in switzerland. *Aerobiologia*, 22(1):27–34.

[25] Gonçalves, A. B., Souza, J. S., Silva, G. G. d., Cereda, M. P., Pott, A., Naka, M. H., and Pistori, H. (2016). Feature extraction and machine learning for the classification of brazilian savannah pollen grains. *PLOS ONE*, 11(6):1–20.

[26] HIRST, J. M. (1952). AN AUTOMATIC VOLUMETRIC SPORE TRAP. *Annals of Applied Biology*, 39(2):257–265.

[27] Holt, K., Allen, G., Hodgson, R., Marsland, S., and Flenley, J. (2011). Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology*, 167(3):175–183.

[28] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

[29] Khanzhina, N., Filchenkov, A., Minaeva, N., Novoselova, L., Petukhov, M., Kharisova, I., Pinaeva, J., Zamorin, G., Putin, E., Zamyatina, E., et al. (2022). Combating data incompetence in pollen images detection and classification for pollinosis prevention. *Computers in biology and medicine*, 140:105064.

[30] Khanzhina, N., Putin, E., Filchenkov, A., and Zamyatina, E. (2018). Pollen grain recognition using convolutional neural network. In *ESANN*.

[31] Kramer, C. L. and Pady, S. M. (1966). A new 24-hour spore sampler. *Phytopathology*, 56(5):517–520.

[32] Kubera, E., Kubik-Komar, A., Piotrowska-Weryszko, K., and Skrzypiec, M. (2021). Deep learning methods for improving pollen monitoring. *Sensors (Basel)*, 21(10).

[33] Lake, I. R., Jones, N. R., Agnew, M., Goodess, C. M., Giorgi, F., Hamaoui-Laguel, L., Semenov, M. A., Solmon, F., Storkey, J., Vautard, R., and Epstein, M. M. (2017). Climate change and future pollen allergy in europe. *Environmental health perspectives*, 125(3):385–391. EHP173[PII].

[34] Langford, M., Taylor, G., and Flenley, J. (1990). Computerized identification of pollen grains by texture analysis. *Review of Palaeobotany and Palynology*, 64(1):197–203. The Proceedings of the 7th International Palynological Congress (Part I).

[35] Li, P. and Flenley, J. R. (1999). Pollen texture identification using neural networks. *Grana*, 38(1):59–64.

[36] Li, P., Treloar, W., Flenley, J., and Empson, L. (2004). Towards automation of palynology 2: the use of texture measures and neural network analysis for automated identification of optical images of pollen grains. *Journal of Quaternary Science: Published for the Quaternary Research Association*, 19(8):755–762.

[37] Lind, L., Nilsson, C., and Weber, C. (2014). Effects of ice and floods on vegetation in streams in cold regions: implications for climate change. *Ecology and evolution*, 4(21):4173–4184. 25505542[pmid].

[38] Lindbladh, M., O'Connor, R., and Jacobson, G. L. (2002). Morphometric analysis of pollen grains for paleoecological studies: classification of picea from eastern north america. *American Journal of Botany*, 89(9):1459–1467.

[39] Lu, L.-L., Jiao, B.-H., Qin, F., Xie, G., Lu, K.-Q., Li, J.-F., Sun, B., Li, M., Ferguson, D. K., Gao, T.-G., Yao, Y.-F., and Wang, Y.-F. (2022). *Artemisia* pollen dataset for exploring the potential ecological indicators in deep time. *Earth System Science Data Discussions*, 2022:1–48.

[40] Manikis, G. C., Marias, K., Alissandrakis, E., Perrotto, L., Savvidaki, E., and Vidakis, N. (2019). Pollen grain classification using geometrical and textural features. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6.

[41] Marmureanu, L., Marin, C. A., Andrei, S., Antonescu, B., Ene, D., Boldeanu, M., Vasilescu, J., Viţelaru, C., Cadar, O., and Levei, E. (2019). Orange snow—a saharan dust intrusion over romania during winter conditions. *Remote Sensing*, 11(21).

[42] Nagi, J., Ducatelle, F., Di Caro, G. A., Cireşan, D., Meier, U., Giusti, A., Nagi, F., Schmidhuber, J., and Gambardella, L. M. (2011). Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347.

[43] Natali, F., Cecchi, L., Torrigiani Malaspina, T., Barbano, F., and Orlandini, S. (2013). Impact of 2003 heat waves on aerobiological indices of allergenic herbaceous family pollen season in tuscany (italy). *Aerobiologia*, 29(3):399–406.

[44] Noll, K. E. (1970). A rotary inertial impactor for sampling giant particles in the atmosphere. *Atmospheric Environment (1967)*, 4(1):9–19.

[45] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[46] Ranzato, M., Taylor, P., House, J., Flagan, R., LeCun, Y., and Perona, P. (2007). Automatic recognition of biological particles in microscopic images. *Pattern Recognition Letters*, 28(1):31–39.

[47] Razmovski, V., O'meara, T., Hjelmroos, M., Marks, G., and Tovey, E. (1998). Adhesive tapes as capturing surfaces in burkard sampling. *Grana*, 37(5):305–310.

[48] Redondo, R., Bueno, G., Chung, F., Nava, R., Víctor Marcos, J., Cristóbal, G., Rodríguez, T., Gonzalez-Porto, A., Pardo, C., Déniz, O., and Escalante-Ramírez, B. (2015). Pollen segmentation and feature evaluation for automatic classification in bright-field microscopy. *Computers and Electronics in Agriculture*, 110:56–69.

[49] Reisert, M. and Burkhardt, H. (2006). Invariant features for 3d-data based on group integration using directional information and spherical harmonic expansion. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 206–209.

[50] Rodinkova, V., Kremenska, L., Palamarchuk, O., Motruk, I., Alexandrova, E., Dudarenko, O., Vakolyuk, L., and Yermishev, O. (2018). Seasonal changes in plant pollen concentrations over recent years in vinnytsya, central ukraine. *Acta Agrobotanica*, 71.

[51] Rodriguez-Damian, M., Cernadas, E., Formella, A., Fernandez-Delgado, M., and Sa-Otero, P. D. (2006). Automatic detection and classification of grains of pollen based on shape and texture. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36(4):531–542.

[52] Rodriguez-Damian, M., Cernadas, E., Formella, A., and Sa-Otero, R. (2004). Pollen classification using brightness-based and shape-based descriptors. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 212–215 Vol.2.

[53] Ronneberger, O., Schultz, E., and Burkhardt, H. (2002). Automated pollen recognition using 3d volume images from fluorescence microscopy. *Aerobiologia*, 18(2):107–115.

[54] Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., and Vasilatou, K. (2020). Real-time pollen monitoring using digital holography. *Atmospheric Measurement Techniques*, 13:1539–1550.

[55] Schaefer, J., Milling, M., Schuller, B. W., Bauer, B., Brunner, J. O., Traidl-Hoffmann, C., and Damialis, A. (2021). Towards automatic airborne pollen monitoring: From commercial devices to operational by mitigating class-imbalance in a deep learning approach. *Science of The Total Environment*, 796:148932.

[56] Schiele, J., Rabe, F., Schmitt, M., Glaser, M., Häring, F., Brunner, J. O., Bauer, B., Schuller, B., Traidl-Hoffmann, C., and Damialis, A. (2019). Automated classification of airborne pollen using neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4474–4478.

[57] Sedghy, F., Varasteh, A.-R., Sankian, M., and Moghadam, M. (2018). Interaction between air pollutants and pollen grains: The role on the rising trend in allergy. *Reports of biochemistry & molecular biology*, 6(2):219–224. 29766006[pmid].

[58] Sevillano, V., Holt, K., and Aznarte, J. L. (2020). Precise automatic classification of 46 different pollen types with convolutional neural networks. *PLOS ONE*, 15(6):1–15.

[59] Smith, L. M. (2019). The Heat Is On: Maize Pollen Development after a Heat Wave. *Plant Physiology*, 181(2):387–388.

[60] Takahashi, Y., Kawashima, S., Fujita, T., Ito, C., Togashi, R., and Takeda, H. (2001). [comparison between real-time pollen monitor KH-3000 and burkard sampler]. *Arerugi*, 50(12):1136–1142.

[61] Tsiknakis, N., Savvidaki, E., Manikis, G. C., Gotsiou, P., Remoundou, I., Marias, K., Alissandrakis, E., and Vidakis, N. (2022). Pollen grain classification based on ensemble transfer learning on the cretan pollen dataset. *Plants*, 11(7).

[62] Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., and Sofiev, M. (2019). Automatic pollen recognition with the rapid-e particle counter: the first-level procedure, experience and next steps. *Atmospheric Measurement Techniques*, 12(6):3435–3452.

[63] Waite, K. J. (1995). Blackley and the development of hay fever as a disease of civilization in the nineteenth century. *Medical history*, 39(2):186–196. 7739297[pmid].

[64] Wang, Z., Bao, W., Lin, D., and Wang, Z. (2019). A local feature descriptor based on sift for 3d pollen image recognition. *IEEE Access*, 7:152658–152666.

[65] Wassan, S., Xi, C., Jhanjhi, N., and Binte-Imran, L. (2021). Effect of frost on plants, leaves, and forecast of frost events using convolutional neural networks. *International Journal of Distributed Sensor Networks*, 17(10):15501477211053777.

[66] Wood, G. S. (1961). Sampling apparatus and method.

[67] Wu, J., Poloczek, M., Wilson, A. G., and Frazier, P. I. (2017). Bayesian optimization with gradients. *Neural Information Processing Systems 30 (NIPS)*.

[68] Xie, Y. and OhEigeartaigh, M. (2010). 3d discrete spherical fourier descriptors based on surface curvature voxels for pollen classification. In *2010 WASE International Conference on Information Engineering*, volume 1, pages 207–211.

[69] Yaslan, Y. and Cataltepe, Z. (2008). Co-training with adaptive bayesian classifier combination. In *2008 23rd International Symposium on Computer and Information Sciences*, pages 1–4.

[70] Zemmer, F., Dahl, Å., and Galán, C. (2022). The duration and severity of the allergenic pollen season in istanbul, and the role of meteorological factors. *Aerobiologia*.