# Description of the proposed research

## 6a Science: Background, research questions, approach, and innovation

The large-scale digitization of natural history collections is generating a wealth of image data that, if processed using image feature extraction and classification algorithms, will constitute an unparalleled knowledge base of forms seen in the natural world. Among the many ways in which extracted image features and image classification can be applied, is the potential to automate taxonomic identification using image data of one of the world's most important disease vectors: mosquitoes. Diseases spread by mosquitoes are usually specific to only some mosquito genera or species, and so advances in their correct identification will improve assessment of the risks posed by their presence. The research we here propose to this end thus contributes directly to the "top sector" Life Science and Health.



Figure Wings of four genera of mosquitoes that spread diseases. Photos taken using a €5 smartphone clip-on lens from HEMA.

In recent years, the consortium members have already assembled globally-sampled mosquito specimen collections (CML, Naturalis), digitized these and compiled image banks (Naturalis); developed prototype open source software frameworks for image feature extraction and classification (Naturalis); and developed advanced expertise in artificial intelligence applied to taxonomic classification of images (LIACS). Our preliminary tests of low-cost photography of mosquito wings using smartphone clip-on lenses (Figure 1), and our preliminary successes with the current state-of-the-art in supervised learning of convolutional neural networks applied to image sets of collection specimens, suggest that an automated identification system that operates on photos of mosquito wings taken 'in the field' using smartphones is feasible.

## 6b eScience: Technologies, methods, and expected impact of the research

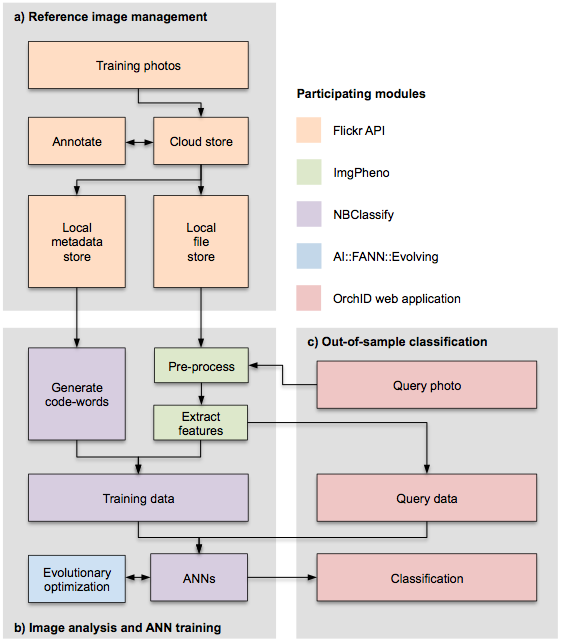


Figure Current architecture of the software framework

The proposed research will modify and extend an open source framework developed at Naturalis (<http://github.com/naturalis/img-classify-all>) consisting of the following components:

* **Reference image management and annotation** - In the current prototype, this is done using the image-hosting platform Flickr.com. Although this provides a convenient system for managing and annotating images and accessing them via an API, the facilities for semantic annotation of images are rudimentary and the API is not sufficiently scalable. One of the key contributions by the e-Science Center to the proposed research would be to replace this prototype with a fit-for-purpose solution for image management and annotation.
* **Image analysis and neural network training -** At present, image features are extracted by a library based on OpenCV, and classification is done using the library for fully connected artificial neural networks FANN. The key contributions by the requested PD will be to extend image feature extraction to edge detection (i.e. to detect veins) and to add the option to use convolutional neural networks based on VGGNet instead of FANN.
* **Out-of-sample classification interface** - Part of our framework is a web application suitable both for desktop and mobile usage that provides a graphical user interface with which end users can upload their photos for automated classification. A demo version that operates on photographs of slipper orchids, hosted on Naturalis's private cloud, can be found here: <http://orch-id.naturalis.nl>

With the proposed extensions and modifications, a ready-to-use system will come into existence with which end users can identify mosquitoes, and assess the health risks they pose, on their mobile devices.

## 6c Re-use, sustainability, dissemination and collaborations

The framework described in the previous section was specifically designed and is intended to be agnostic as to the specific problem set. For example, we have already applied this framework to the automated recognition of flowers of slipper orchids, an endangered group of orchids in which trade is restricted and which therefore need to be recognizable by customs officials in order to be able to combat wildlife crime, and we are also applying the framework to the recognition of butterfly specimens in a project that is currently and for the coming XXX years receiving funding from the Van Groenendael-Krijger Foundation. As such, the proposed technology is evidently applicable beyond the target use case. Given that the current use cases fit within Naturalis's capacity for hosting web applications and HPC virtualization, and given that the lion's share of the software development and workflow execution has been done by undergraduate students, maintenance and sustainability at present levels are secured. That said, we foresee a need for more sophisticated client side (i.e. smartphone end user) functionality. To this end, we have entered into an agreement with the not-for-profit Stichting Akvo.org, whose "Flow" platform for digital data capture under field conditions in developing nations we will pitch in upcoming CfPs that allow for field work components spearheaded by consortium partner CML (an example of such a CfP will be the 2018 call by the JRS

Foundation).

## 6d Data management

1. ***Is data generated or collected during the research that is appropriate for re-use?***Yes: reference image sets of mosquito specimens.
2. ***Where will this data be stored during the research?***Our current implementation uses Flickr.com. Section 6b discusses our needs for a better solution.
3. ***After the project has been completed, how will the data be stored for long-term and made available for the use by third parties? To whom will the data be accessible?***   
   Besides Flickr.com (which we prefer to replace), we are able to store image data long-term in Naturalis's "beeld bank". That said, we are keen to deploy a more fit-for-purpose solution in collaboration with the e-Science Center. All image data generated by Naturalis, including ours, are available to anyone under a CC0 license.
4. ***Which facilities (ICT, (secure) archive, refrigerators or legal expertise) do you expect will be needed for the storage of data during the research and after the research? Are these available?***We need a facility for storing, managing, annotating, and accessing (via API) our image data. As noted above, several options are available (Flickr.com, or at Naturalis), but the key contribution we request from the e-Science Center is in this area.

## 6e Software sustainability

1. ***Is software generated during the project that is appropriate for re-use? If so, please indicate which software will be appropriate for re-use.***The open source framework discussed under 6b and available at <http://github.com/naturalis/img-classify-all>
2. ***How will the software appropriate for re-use be licensed and made available to third parties?***The image feature extraction component is available under the terms of the MIT license. The classification system and the web application are available under the Apache license. All source code components are under version control hosted by Github.com, for which Naturalis has a commercial license and for which it makes local, permanent backups in its private infrastructure (a Ceph-based object store).
3. ***What measures are needed to make the software appropriate for long-term re-use for third parties?***None beyond the measures already taken.
4. ***In your expectation, how large is the expected community that will potentially use the software, and do you expect outside contributors to the software?***For the target use case of mosquito identification we have previously developed a conservative scenario based on the deployment of the "Flow" platform in malaria-prone areas of South Africa for which we projected XXX users (YYY data captures). However, we suspect that a functioning, easy-to-use, mobile system for mosquito identification by non-technical users may attract many, many more users.
5. ***What expertise do you expect to be needed to make the software appropriate for long-term re-use by third parties? Is this expertise available? Please state what your expectations are of the contribution from the e-Science Center in making the software appropriate for long-term re-use.***The key challenge is the management of annotated reference image data sets. We will need to be able to manage these collaboratively, have a flexible facility for semantic annotation, be able to access these through an API, and be able to grow these data sets with images uploaded by end users. For this we will need the help of the e-Science Center.

## 6f Use national e-infrastructure

As noted above, we will need an image management facility. Besides this, we currently perform neural network training, and are running the web application, on Naturalis's private cloud. At present, the resource use fits well within Naturalis's in-house capacity. However, under a scenario where the web app for mosquito identification works well and "goes viral", our e-Infrastructure needs may grow such that we may need to deploy the app image elsewhere.