*Pattern Recognition Letters*

This is the author post-print (ie. final draft post-refereeing) accepted version of the paper. Publisher (Elsevier) version will be available in Pattern Recognition Letters. `http://www.journals.elsevier.com/pattern-recognition-letters/`

# Towards Generic Image Classification using Tree-based Learning: an Extensive Empirical Study

Raphaël Marée[a,*], Pierre Geurts[a], Louis Wehenkel[a]

[a]*Systems and Modeling, Department of Electrical Engineering and Computer Science and GIGA-Research, University of Liège, Belgium*

**Abstract**

This paper considers the general problem of image classification without using any prior knowledge about image classes. We study variants of a method based on supervised learning whose common steps are the extraction of random subwindows described by raw pixel intensity values and the use of ensemble of extremely randomized trees to directly classify images or to learn image features. The influence of method parameters and variants is thoroughly evaluated so as to provide baselines and guidelines for future studies. Detailed results are provided on 80 publicly available datasets that depict very diverse types of images (more than 3800 image classes and over 1.5 million images).

*Keywords:* Image classification, machine learning, random subwindows, extremely randomized trees, feature learning.

## 1. Introduction

The aim of supervised image classification is to automatically build computerized models able to predict accurately the class (among predefined ones) of new images, once trained from a set of labelled images. In the real world, this generic problem encompasses well-known tasks such as the automatic recognition of images of handwritten characters, faces, cells, and road signs, to name but a few.

Since the early days of computer vision practice, when a researcher approaches a new image classification task, he or she often develops a dedicated algorithm to implement human prior knowledge as a sequence of specific operations, also known as a hand-crafted approach. Such an approach often involves the design and calculation of tailored filters and features capturing expected invariant image characteristics. In our preferred field of application, life science imaging, although several specific works have proved effective, the design choices are rarely straightforward hence such a strategy requires a lot of research and development efforts for each specific problem, and it might require major adjustments when parameters of the problem vary (e.g. sample preparation protocols, imaging modality, phenotypes to recognize, ...). In other words, this engineering approach does not scale well as there are hundreds of thousands of biological entities that can be screened using many different sample preparation techniques and imaging modalities. Hence, scientific studies are often limited in scale, or still partially performed by hand (e.g. 50 millions of galaxies were manually labeled into morphological classes by almost 150000 humans within one year through the GalaxyZoo web-based project (Lintott et al., 2008)), while others required very large computing infrastructures because they relied on dense feature computations (e.g. computers of the members of the Help Conquer Cancer project have contributed over 100 CPU-millenia for the automated classification of tens of millions of protein crystallization-trial images at a rate of 55 CPU-years per day (Kotseruba et al., 2012)).

*Corresponding author: Tel.: +32-4-366-2644; fax: +32-4-366-2988;

*Email address:* raphael.maree@ulg.ac.be (Raphaël Marée)

### 1.1. This work

Following and extending previous works (Marée et al., 2003, 2004, 2005, 2007), we consider the generic problem of supervised image classification without any preconception about image classes, ie. it encompasses the recognition of numerous types of images under various image acquisition conditions. Indeed, with the design of a general-purpose yet simple and easily applicable image classifier in mind, we proposed earlier an appearance-based, learning method, relying on dense random subwindow extraction in images, their description by raw pixel values, and the use of ensembles of extremely randomized trees to classify these subwindows hence images. Despite its conceptual simplicity and its rather low run-time complexity, it yielded interesting results on a few datasets. Subsequently, variants of the method were proposed in (Moosmann et al., 2008; Marée et al., 2009; Dumont et al., 2009; Stern et al., 2011) for object categorization, image segmentation, interest point detection, and content-based image retrieval.

In this paper, we extend and thoroughly evaluate our generic framework for image classification. Our contributions are as follows:

- While the main building blocks of the framework, subwindows extraction and extremely randomized trees, have been proposed in our earlier research, several algorithmic variants have not yet been considered and deserve to be tested. In particular, extending the work of (Moosmann et al., 2008), we explore in this paper several novel variants of the feature learning approach, corresponding to different ways to derive features from trees. We also consider yet unexplored parameter ranges (e.g., subwindow size intervals) and several simple pre-processing strategies (e.g., filters), which both turned out to be very beneficial on several datasets. These new algorithmic variants therefore greatly extend the range of image classification tasks that can be addressed by our framework and improve its generality.

- To assess our framework, we perform an extensive, systematic study of its performances on 80 publicly-available datasets (among which 25 bioimaging datasets). By conducting such a large-scale study, we are able to characterize the performances of the method and its recent variants, to study rigourously the influence of its parameters and classification schemes, to bring out the most influential design choices, and to draw general guidelines for future use so as to speed its application on new problems.

- To the best of our knowledge, no other image classification method has been evaluated so extensively. We deeply believe that generic methods can only be fully and fairly assessed by confronting them to several representative tasks and by extensively studying the influence of their parameters. By summarizing publicly available databases and by providing our positive and negative results, our hope is thus also to foster research in generic methods, by encouraging other researchers to evaluate and compare their methods on a wide range of imagery.

## 2. Experimental setup

We work with a large variety of datasets from many application domains. Our hypothesis is that by considering the image classification problem as a whole, it will possible to derive trends that are generally valuable, ie. applicable in several areas. For example, observations derived from experiments related to the recognition of traffic signs (captured with onboard cameras) or galaxies (captured during wide-field sky surveys) might be helpful for the recognition of cells (captured by microscopes) as these datasets are sharing some essential characteristics (they consist in different classes of shapes and they exhibit illumination and noise variations due to the acquisition process). Similarly, observations derived from material classification datasets might be of interest for biological tissue recognition (as their images have textured patterns).

### 2.1. Datasets and evaluation criteria

Our experimental setup comprises 80 image datasets that were previously published and are publicly and freely available. They sum up roughly to 1.5 million images depicting approximately 3850 distinct classes. The choice of datasets was made a priori and independently of the results obtained with our method. More details about these datasets are given in Supplementary Material. In particular, a summary of their characteristics is given in

Supplementary Table I, and an overview of image classes for all datasets is given in Supplementary Figures 1, 2, 3, and 4. Images were acquired worldwide, in controlled or uncontrolled conditions, using professional equipments in laboratory settings, individuals' digital camera in the real-world, various biomedical imaging equipements (fluorescence or brightfield microscopes, plain film radiography, etc.), robotic telescopes, synthetic aperture radars, etc. For a given dataset, image classes possibly exhibit subtle or prominent changes in their appearance due to various sources and levels of variations including possible changes in position, illumination, scale, and viewpoint, and/or presence of background clutter, occlusions, and noise. Moreover, either significant intra-class variations or high similarity between distinct classes could be present. Several of these datasets are synthetic and therefore variations are controlled (e.g. backgrounds are uniform) and well characterized, while many others contains real-world images so variations are mixed. Note that we only included in our experiments two widely used face datasets among tens of existing ones, given that face databases were recently summarized and evaluated thoroughly (Huang et al., 2007; Shamir, 2008; Pinto et al., 2008). Also, we did not include the Pascal VOC challenge datasets (Everingham et al., 2010) whose evaluation criteria (precision/recall curves for each object class) does not fit well into our evaluation framework (see below).

Our evaluation protocols are summarized in Supplementary Table I. Our evaluation metric is the misclassification error rate evaluated on independant test images. If a precise dataset protocol was defined in the literature and was adopted in several papers, we also used it. However, for many datasets (e.g. those where the protocol was not rigourously described, or different between papers, or where the number of test images was rather small), we performed 10 runs where each run uses a certain number of images randomly drawn for the learning phase (e.g. 80% of the total number of images) and the remaining images for the testing phase (e.g. 20%). The misclassification error rate is then averaged over all the test sets which allows to have a reliable insight into the effects of method parameters.

## 3. Methods

We present the two key components of our image classification variants. The method involves the extraction of random subwindows described by raw pixel values and the use of ensemble of extremely randomized trees by different means.

### *3.1. Random subwindows*

We introduced previously different random subwindow sampling schemes (Marée et al., 2003, 2005, 2007). Random subwindows are square patches of random sizes extracted at random positions within images. They are subsequently resized to a fixed patch size whose pixels are used as input of the machine learning algorithm (see next subsection). The resizing step improved robustness to scale changes and it allows one to use generic machine learning methods that work with fixed-size feature vectors. This procedure also introduces in the training set subwindows with slight pixel intensity variabilities through multiple over- or sub-sampling, a process that can help the algorithm learn to be more robust to such changes that could occur naturally in unseen test images. Variants also include the activation of right and straight angle rotations and mirroring to subwindows, so that the model can learn to be robust to rotations.

In this work, we first study systematically the influence of subwindow size intervals and the way random subwindows are encoded on all 80 datasets. Default tests are made using a total of $N_{ls} = 1$ million training subwindows (previous works (Marée et al., 2003, 2005, 2007) used only one hundred thousand subwindows) while a few others more intensive tests are performed with up to 50 millions subwindows. For a given dataset, the same number of subwindows are randomly drawn from each image, it equals $N_{ls}/N_{img}$ where $N_{img}$ is the number of training images. One can see subwindows as pixel context, support regions, or receptive fields of different sizes/scales whose intervals are systematically tested: we consider single pixels $1 \times 1$ as baseline, and 13 different configurations of square subwindows ranging from small image regions $[0\% - 10\%]$ to large ones $[90\% - 100\%]$, and including the default unconstrained size $[0\% - 100\%]$ used in (Marée et al., 2005). Constraining sizes to e.g. $[25\% - 50\%]$, means that the size of each subwindow is randomly choosen between 25% and 50% of $min(width, height)$ in

each image, then the position is randomly choosen in order to guarantee square subwindows are always fully contained within images. Note that in configurations with zero minimum ([0% − $x$%]), the minimum size is actually $1 \times 1$. For all configurations (except baseline $1 \times 1$ where no resizing is performed), each subwindow is subsequently resized by bilinear interpolation to a patch of fixed size ($8 \times 8$, $16 \times 16$ (default) or $32 \times 32$) and its pixel values encoded in HSV or graylevels are used as the subwindow descriptors. Whereas more elaborated or specific sampling schemes could be designed and might improve results on specific datasets (e.g.: localized sampling for datasets where positions of patterns of interest are known, rectangular subwindows for elongated objects, adaptive sampling (Moosmann et al., 2008), ...), we want here to investigate how far a basic, systematic, and generic random sampling could lead us in terms of accuracy on many datasets so as to provide baselines before developing more complex sampling schemes.

### *3.2. Extremely Randomized Trees for direct image classification or for feature learning*

Ensembles of randomized trees are increasingly used in machine learning and computer vision (see Criminisi and Shotton, 2013, for their recent developments in computer vision and medical imaging applications).

The Extra-Trees algorithm was proposed in (Geurts et al., 2006) where the reader can find a precise algorithm description. In this work, we evaluated the use of ensemble of extremely randomized trees by two different means: as direct classifiers, ET-DIC, or as feature learners, ET-FL. As we observed overall better performances when they are used as feature learners we only describe here this variant while the other variant is described in Supplementary Material. In the ET-FL classification scheme, instead of retaining probability estimates at terminal nodes and use trees to perform subwindow classification, and hence image classification, each terminal node (leaf) of a tree is considered as a "codebook" or "visual word". This latter approach is inspired by previous works using visual codebooks. In this setting, after propagating subwindows down the trees, each image is described by a single global feature vector which dimensionality equals the number of terminal nodes in the ensemble of trees, and where features are quantitative frequency values (they correspond to the number of image subwindows that reach a given terminal node divided by the total number of subwindows extracted in the image, i.e. a bin value is included in [0,1] and the sum over all terminal nodes equals to 1 in a given tree for a given image). Such a "bag-of-features" representation can then be fed into any classifier to build the final image classification model. In our case, we use a linear support vector machine classifier, as illustrated by Figure 1. To predict the class of a new image, its random subwindows are propagated into the ensemble of trees to build its global feature vector subsequently classified by the SVM classifier.

For both variants, we study systematically on all 80 datasets the influence of the minimum node sample size $n_{min}$ by picking a few of its possible values (from 1 to 1000 in ET-DIC and from 1 to 50000 in ET-FL), the number of random tests $k$ (from 1 to the maximum number of input variables), and the number of trees $T$ (from 1 to 20 in ET-DIC and from 1 to 40 in ET-FL although more extensive tests use up to $T = 1000$ trees). In ET-FL, we also study systematically the influence of the encoding of the global feature vector: We evaluate our quantitative frequency representation as well as binary encoding (where a feature equals to 1 if at least one of its subwindow was propagated to that terminal node, and 0 otherwise), either only at tree terminal nodes or in all the tree nodes (internal and terminal nodes). We use in ET-FL a linear SVM classifier to perform the final classification whose parameters were set to default values (see Supplementary Material for implementation details).

## 4. Results

### *4.1. Overall results*

Regarding overall performances, we achieve more than 80% recognition rate for 52 datasets among 80, and more than 90% recognition rate for 30 datasets (see Figure 2). However, our results are much lower for some other datasets or recently published ones that exhibit a lot of variabilities. In particular we achieve less than 50% recognition rate on 13 datasets, most of them containing images from the web that depict coarse-grained categories (natural scenes or various object/face classes with complex backgrounds and strong intensity and illumination changes). Overall, the mean of the best error rate computed over all 80 datasets is 22.22%. Interestingly, on the

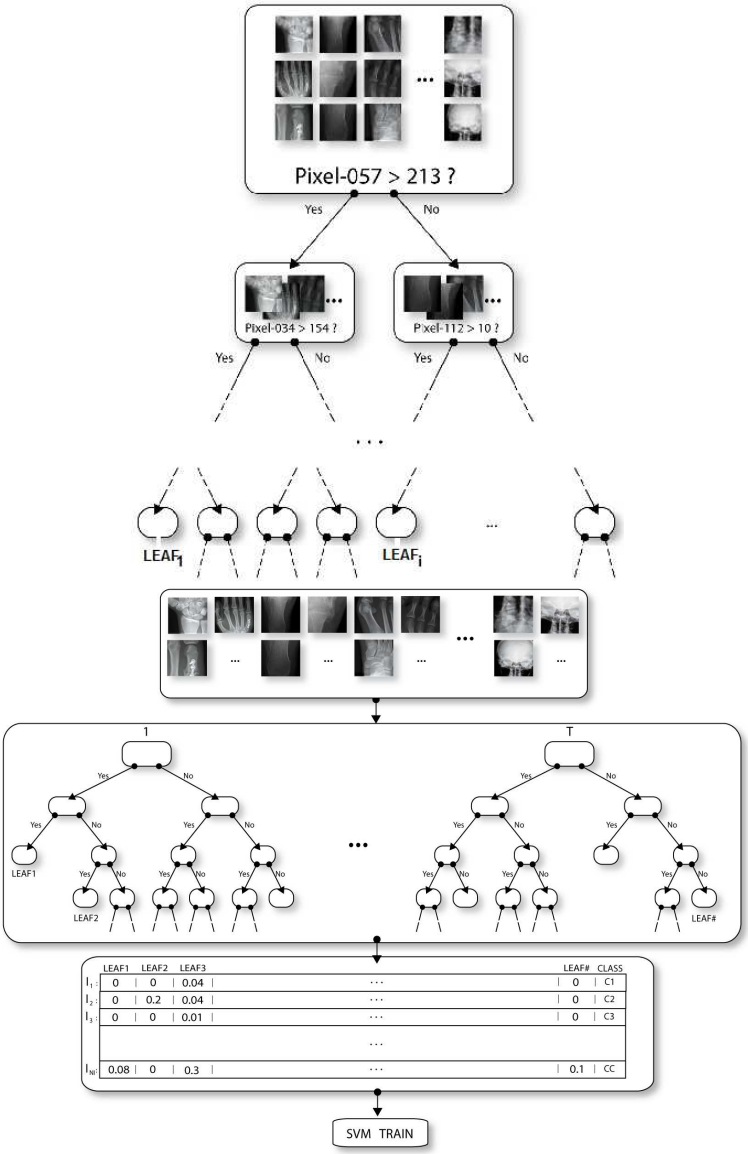


Figure 1: Left: A single tree induced from a training set of random subwindows, using node tests with single pixel thresholding, for the ET-FL scheme. Right: An ensemble of $T$ trees, the derived, quantitative frequency global representation for training images, and training of a final linear SVM classifier in ET-FL mode.

subset of 25 bioimaging datasets, the mean best error rate is 12.03% (an overview of these latter datasets is given in Figure 5.2). Life scientists working with images with visual appearances similar to one of these datasets should consider applying as a first try our classification algorithm with parameter values similar to those that yields the best recognition rate on this dataset (See Supplementary Material for detailed results on each dataset).

#### *4.1.1. Comparison of ET-DIC and ET-FL*

ET-DIC is slightly better for a quarter of the datasets, including particular object identification datasets in controlled conditions, but ET-FL yields better results on others (60 datasets among 80). These results show that on a majority of datasets, the construction of a global image representation based on tree terminal node frequencies subsequently classified by a linear classifier (ET-FL) yields better results compared to the direct classification of individual subwindows (ET-DIC). Although individual subwindows can be strongly predictive with respect to the class of the image they come from (when ET-DIC is performing well), ET-FL allows to describe images by a higher-level representation than raw pixels. It learns image features (from small or large patterns), as each tree leaf contains subwindows that fulfills a serie of tests on pixel intensities in (small to large) subwindows. The final classification model that combines such feature "responses" is more discriminative than the combination of individual predictions for every subwindows (ET-DIC).

### *4.2. Parameter influence study*

The influence of all parameter values was thoroughly evaluated for both variants. Our main results regarding the best method (ET-FL) parameter influences are summarized in Figure 3 and summarized below (Detailed results are available in Supplementary Tables II to XII).

Regarding the random subwindow extraction scheme, the most influential parameter is the size interval of subwindows that allows the method to be adapted to very different types of problems. The optimal sizes could be very small or very large proportionally to image sizes. We observed that small subwindows allow to capture fine details and generally perform best for images with highly repeatable patterns i.e. textured images (e.g. histological tissues, man-made materials, or assays with populations of cells, see Figure 4), while larger subwindows

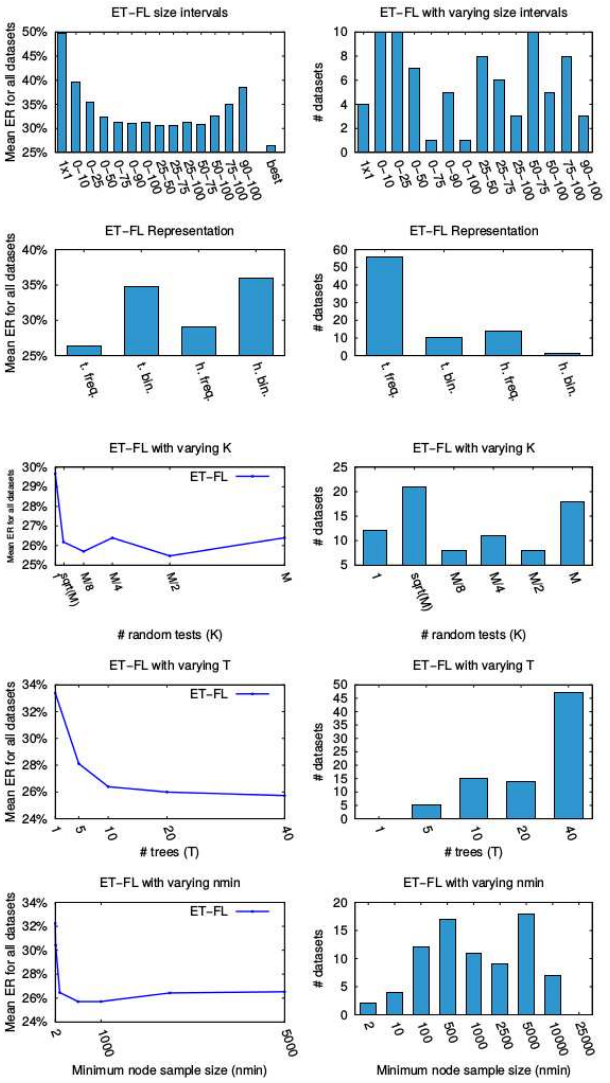


Figure 3: Results averaged over all 80 datasets for ET-FL variant: First column: Average of error rates for all datasets with subwindow size intervals (1st row), image representation (2nd), number of random tests (3rd), number of trees (4th), minimum node sample sizes (5th). Second column: Number of datasets for which the parameter values yield the best error rates. See Supplementary Tables II to XII for detailed results.

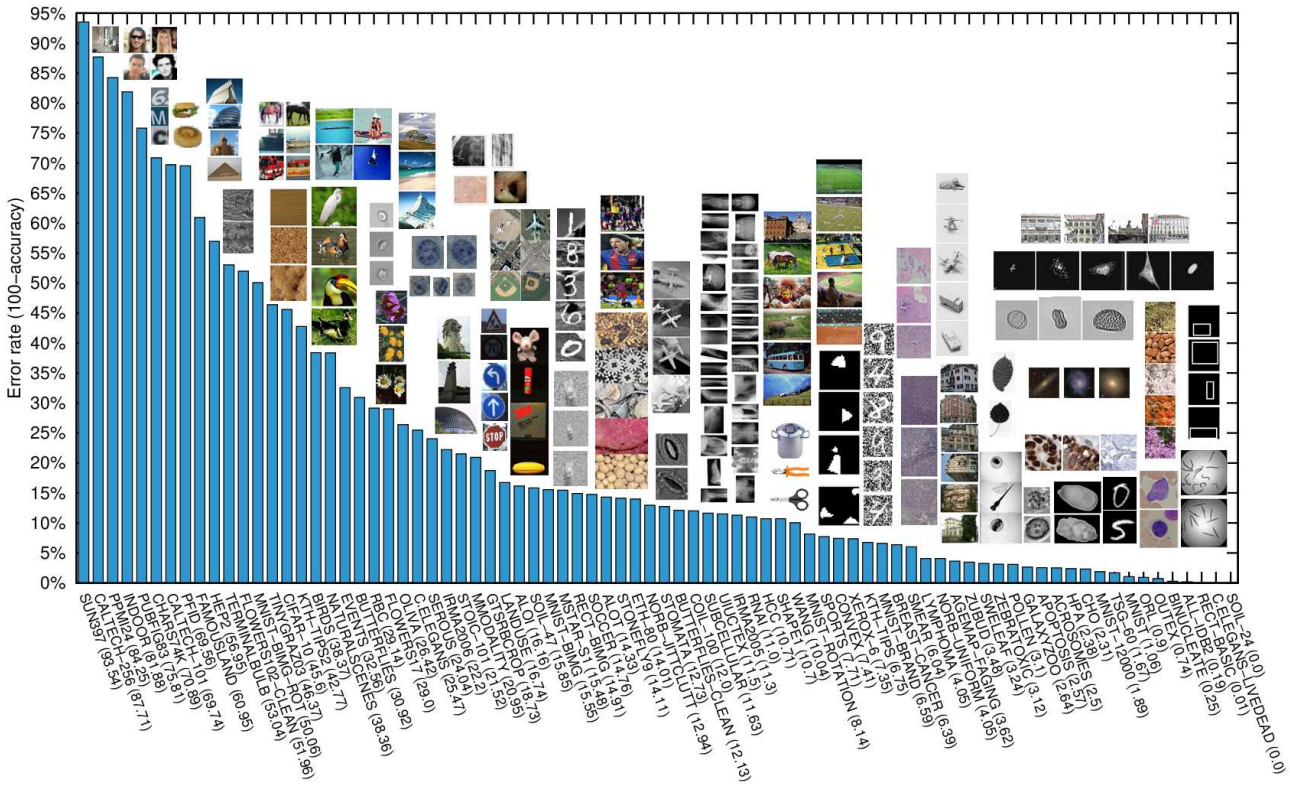


Figure 2: Summary of the best error rate obtained for each dataset without optimizations, gathered from Supplementary Tables II to XII. Illustrative image examples from ~ 50 datasets are positionned (approximately) according to dataset recognition performances. Note that some results (e.g. on CIFAR-10, GTSRB, and IRMA-2005) are significantly improved using further optimizations, see Section 4.3 and Supplementary Table XIII.

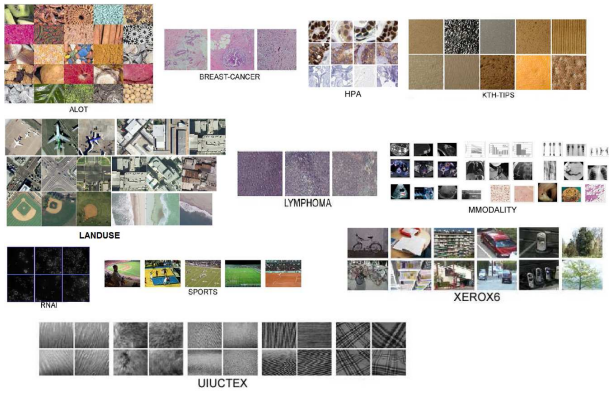


Figure 4: Several datasets for which smallller subwindow sizes yield lowest error rates.

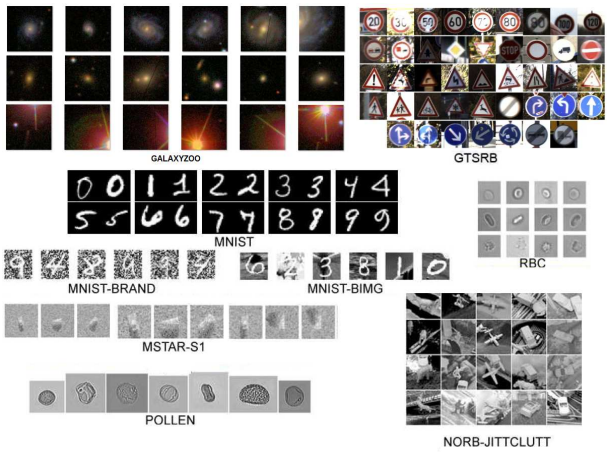


Figure 5: Several datasets for which larger subwindow sizes yield lowest error rates.

yield better results for shape-like datasets (e.g. red-blood cells, leaves, handwritten characters, and traffic signs, see Figure **??**). For these latter type of datasets, extracting large subwindows augments the training set with (small) scale and translation variations, and allows models to directly capture global patterns. Concerning the number of extracted subwindows, we observed a total of 1 million training subwindows performs well, but using a denser sampling can still improve results on several datasets.

In ET-FL variant, increasing the number of trees (hence the number of features for the final linear classifier) up to 40 brings improvement although the improvement is not always important compared to using only 10 trees. Trees should be pruned i.e. $n_{\min}$ value should be roughly one thousandth of the total number of subwindows of the training set (in order to build features that are not too much specific), except for a few problems including object identification tasks in controlled conditions (for which specific features work best). Terminal quantitative frequency yields better results than binary or hierachical encoding. On average, the default value of the filtering parameter (equals to the square root of the total number of pixels that describe a subwindow) achieves better results than unsupervised feature construction, but increasing that parameter to higher values does not seem so important, although for several problems (e.g. noisy, shape-like images) it is still beneficial to do so.

### *4.3. Further optimizations*

In practice, if the method does not achieve satisfactory results on a specific problem, it is possible to further optimize its parameters and implement slight algorithm variations to get better results. Although this alters somewhat the generality of the method, we believe these optimizations (some requiring only a few lines of codes) are simpler than designing a completely new, specific, approach. Although further work is needed to assess if some of these variants could be generalized and applied successfully on a larger number of datasets, Supplementary Table XIII present promising results obtained with several (combinations of) simple optimizations for a dozen datasets. These optimizations include extension of parameter ranges (e.g. increase the number of trees), use synthetic data (e.g. data augmentation by adding in the training set right and straight angle rotated and mirrored subwindows), normalization of random subwindow descriptors (e.g. by substracting the mean and then dividing by the standard deviation for each subwindow channel), evaluation of different node tests in Extra-Trees (e.g. node tests that threshold the difference of a pixel and one of its 8 direct neighbours), applying filters to original images (e.g. using linear filters and spatial pooling operations), adding statistical features to subwindow descriptors (e.g. using features of (Orlov et al., 2008)). These optimizations and their evaluation are discussed in Supplementary

Material.

## 5. Comparison with other methods

Without a centralized repository of results, gathering state-of-the-art results from the wide computer vision literature for all the datasets included in our study could hardly be up to date. To the best of our knowledge, no other image classification method was evaluated on so many datasets. We will therefore only draw general trends from what we observed. Detailed comparisons for a subset of datasets are provided in Supplementary Material.

First, we compared on several datasets our approach with other approaches using Extremely Randomized Trees. On a few datasets with fixed image sizes, we first compared our approach to the direct application of Extra-Trees without subwindow extraction, ie. where each image is represented by a single input vector encoding all its pixel values. Our results (see Supplementary Tables XIV) were significantly better using our approaches based on subwindow extraction, in particular on datasets where small subwindows yield better results (e.g. on immunostaining patterns) but also on datasets when large subwindows performed best. Compared to (Marée et al., 2005) using unconstrained subwindow size intervals and ET-DIC on a few datasets, we observe that adjusting parameters (such as the subwindow size intervals, the number of subwindows, the number of random tests, and the classification scheme) can yield very important accuracy improvements. Compared to (Moosmann et al., 2008) that uses ET-FL with binary encoding at terminal nodes and used a fixed number of features (by post-pruning) on a few object classes, we observed that quantitative encoding and problem-dependent numbers of learned features (from a few thousands up to millions of features) have a significant influence on results.

Second, we observed the method often performs better than previously published baselines used in original publications presenting several datasets. This is particularly true for global approaches e.g. using classifiers (nearest neighbor classifier with euclidian distance, logistic regression, or SVMs) applied on down-sampled images (see Supplementary Table XV). It also sometimes performs better than first specific methods developed once new datasets were published, e.g. for a building recognition dataset, a sport categorization dataset, a leaf recognition task, a dataset about land uses from overhead imagery, and several bioimaging datasets (See Supplementary Material). On several datasets, our approach is also on par with, or better than, methods using application-specific features (e.g. on galaxy recognition, leaves, zebrafish phenotypes, ...), and better than many other methods (e.g. proposed during international challenges), while not reaching state-of-the-art performances on each and every problem (e.g. on cells in immunofluorescence).

On several other problems (especially datasets with images from the web depicting e.g. wild animals, faces of celebrities, or natural scenes or actions), our results using raw pixel values from original images are not satisfying. On most of these datasets, our approach without optimizations yields worse results than GIST (Oliva and Torralba, 2001), and it is also significantly inferior than more elaborated approaches, e.g. methods combining numerous image descriptors (Gehler and Nowozin, 2009), or multi-stage (deep) architectures that combine various steps of normalization, filtering and spatial pooling (Pinto et al., 2009; Ciresan et al., 2012; Quattoni and A.Torralba, 2009; Xiao et al., 2010). On the web-scale object recognition dataset on which we evaluated optimizations using filtered images (see Section 4.3), our approach then becomes better than GIST (Ranzato et al., 2010) and also slightly better than other multi-stage approaches e.g. tiled convolutional neural networks (Le et al., 2010) and factorized third-order Boltzmann Machines (Ranzato et al., 2010), but still significantly inferior to the best known method on this dataset (Ciresan et al., 2012). In addition, we observed that on other problems (such as traffic sign recognition, and synthetic images of object categories), it seems not necessary to perform image filtering to be competitive with a variety of multi-stage approaches. These various results suggest that although deep learning is often presented as a unified framework (LeCun et al., 2015), there are in fact plenty of "deep learning" architectures and methods which yield very different recognition performances when evaluated on various datasets. We provide a few additional comparisons in Supplementary Material but we have considered that the comparison with deep learning variants is well beyond the scope of this paper.

## 6. Guidelines

Given its good overall performance, we believe our approach is a very good off-the-shelf image classification method. It will obviously not provide the best performance on each and every problem but, without too much tuning effort, it should give some good indication of the performance one could expect for any new problem.

Summarizing the extensive analysis carried out in this paper, we suggest to adopt the following procedure when applying our method on a new image classification task:

- Without any prior knowledge about the problem at hand, we suggest using the following default setting of the method parameters: 1 million training subwindows encoded by $16 \times 16$ patches in $HSV$ colorspace, 1000 subwindows per test image, ET-FL mode with $T = 10$, $k = 28$, $n_{min} = 1000$, and terminal frequency encoding. Regarding subwindow size intervals, we suggest first trying three settings: small ($0\% - 10\%$) subwindows, medium ($25\% - 50\%$), large ($75\% - 100\%$) then refining size intervals according to these first results. A better strategy could also be obtained by deriving these sizes from the most similar datasets to the one at hand in the pool of 80 datasets used in this paper (see supporting Tables $VII$ for the best subwindow parameter settings using ET-FL on each problem).
- If the results obtained with default settings are not satisfactory, we suggest then to try tuning some parameters. As discussed earlier, the number of trees and the number of subwindows should be chosen only taking into account the available computing ressources (since the higher they are, the better). To enrich the training set, we recommend to consider data augmentation (rotation, mirroring) if the classes are not orientation-dependent. As shown in Section 4, after subwindow size intervals which plays a major role, the more problem specific parameters are the filtering parameter $k$ and first tuning efforts should be focused on this parameter. Tuning $n_{min}$ and switching to ET-DIC might also be explored eventually but, given our experience, one should not expect a huge improvement.
- Finally, if results are still not good enough, we suggest to enrich subwindow feature descriptors by considering filtering images (with linear filters and spatial pooling operations), or by extracting explicitely new features. One could either rely on generic image feature extractors (e.g. those extracted by Orlov et al. (2008)) or on more problem specific feature extractors if such features can be derived from prior knowledge.

Whether or not to go through these three steps is of course application dependent.

## 7. Conclusions

This paper addressed the generic problem of supervised image classification without any preconception about image classes. An extensive empirical study has been conducted to evaluate overall performances of variants of a simple and brute-force method using random subwindows extraction, raw pixel intensity descriptors, and extremely randomized trees either to classify directly images or to learn features.

While our method does not reach state-of-the-art results on each and every problem, it is rather easy to evaluate and it achieves good performances for diverse image collections including images from real-word applications that exhibit various factors of variations. We therefore suggest it could be used as a first try on any new image classification problem and we provided guidelines to do so. We already successfully applied these guidelines and variants of our approach in practical biomedical applications including (Delga et al., 2014; Jeanray et al., 2015).

Finally, a Python implementation of our algorithms will be published in the near future under an open-source license and distributed with CYTOMINE, a rich internet application for the collaborative analysis of multi-gigapixel biomedical images (Marée et al., 2015).

## Acknowledgments

R.M. was supported by the CYTOMINE research grant of the Wallonia (DGO6, WIST3, 1017072), and by the GIGA interdisciplinary cluster of Genoproteomics of the University of Liège with financial support from the Wallonia and the European Regional Development fund.

## References


Ciresan, D.C., Meier, U., Schmidhuber, J., 2012. Multi-column deep neural networks for image classification, in: Computer Vision and Pattern Recognition, pp. 3642–3649.

Criminisi, A., Shotton, J. (Eds.), 2013. Decision Forests for Computer Vision and Medical Image Analysis. Advances in Computer Vision and Pattern Recognition, Springer.

Delga, A., Goffin, F., Marée, R., Lambert, C., Delvenne, P., 2014. Evaluation of cellsolutions bestprep(r) automated thin-layer liquid-based cytology papanicolaou slide preparation and bestcyte(r) cell sorter imaging system. Acta Cytologica 58, 469–77.

Dumont, M., Marée, R., Wehenkel, L., Geurts, P., 2009. Fast multi-class image annotation with random subwindows and multiple output randomized trees, in: Proc. VISAPP, pp. 196–203.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision 88, 303–338.

Gehler, P.V., Nowozin, S., 2009. On feature combination for multiclass object classification, in: Proc. IEEE International Conference on Computer Vision (ICCV), pp. 221–228.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning 36, 3–42.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49. University of Massachusetts, Amherst.

Jeanray, N., Marée, R., Pruvot, B., Stern, O., Geurts, P., Wehenkel, L., Muller, M., 2015. Phenotype classification of zebrafish embryos by supervised learning. PLos ONE 10, e0116989.

Kotseruba, T., Cumbaa, C., Jurisica, I., 2012. High-throughput protein crystallization on the world community grid and the gpu. Journal of Physics: Conference Series 341.

Le, Q.V., Ngiam, J., Chen, Z., Chia, D., Koh, P.W., Ng, A.Y., 2010. Tiled convolutional neural networks, in: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 23, pp. 1279–1287.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J., 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society 389, 1179–1189.

Marée, R., Geurts, P., Piater, J., Wehenkel, L., 2004. A generic approach for image classification based on decision tree ensembles and local sub-windows, in: Hong, K.S., Zhang, Z. (Eds.), Proceedings of the 6th Asian Conference on Computer Vision, pp. 860–865.

Marée, R., Geurts, P., Piater, J., Wehenkel, L., 2005. Random subwindows for robust image classification, in: Proc. IEEE CVPR, IEEE. pp. 34–40.

Marée, R., Geurts, P., Visimberga, G., Piater, J., Wehenkel, L., 2003. An empirical comparison of machine learning algorithms for generic image classification, in: Coenen, F., Preece, A., Macintosh, A. (Eds.), Proc. 23rd SGAI AI,, Springer. pp. 169–182.

Marée, R., Geurts, P., Wehenkel, L., 2007. Random subwindows and extremely randomized trees for image classification in cell biology. BMC Cell Biology supplement on Workshop of Multiscale Biological Imaging, Data Mining and Informatics 8.

Marée, R., Geurts, P., Wehenkel, L., 2009. Content-based image retrieval by indexing random subwindows with randomized trees. IPSJ Transactions on Computer Vision and Applications 1, 46–57.

Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J.M., Kainz, P., Geurts, P., Wehenkel, L., 2015. Collaborative analysis of multi-gigapixel imaging data using cytomine. Under revision URL: `http://www.cytomine.be/`.

Moosmann, F., Nowak, E., Jurie, F., 2008. Randomized clustering forests for image classification. IEEE Transactions on PAMI 30, 1632–1646.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M., Goldberg, I., 2008. Wnd-charm: Multi-purpose image classification using compound transforms. Pattern Recognition Letters 29, 1684–1693.

Pinto, N., Dicarlo, J., Cox, D., 2008. Establishing good benchmarks and baselines for face recognition, in: ECCV 2008 Faces in 'Real-Life' Images Workshop.

Pinto, N., Doukhan, D., DiCarlo, J., Cox, D., 2009. A high-throughput screening approach to discovering good forms of biologically-inspired visual representation. PLoS Computational Biology 5.

Quattoni, A., A.Torralba, 2009. Recognizing indoor scenes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 413–420.

Ranzato, M., Krizhevsky, A., Hinton, G.E., 2010. Factored 3-way restricted boltzmann machines for modeling natural images. Journal of Machine Learning Research - Proceedings Track 9, 621–628.

Shamir, L., 2008. Evaluation of face datasets as tools for assessing the performance of face recognition method. International Journal of Computer Vision 79, 225–230.

Stern, O., Marée, R., Aceto, J., Jeanray, N., Muller, M., Wehenkel, L., Geurts, P., 2011. Automatic localization of interest points in zebrafish images with tree-based methods, in: Proc. 6th IAPR International Conference on Pattern Recognition in Bioinformatics, Springer-Verlag. pp. 179–190.

Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010), pp. 3485–3492.

**Supplementary Material**

Supplementary material are available online.