



Classification vs Deep Learning in Cancer Degree on Limited Histopathology Datasets

Pedro Furtado^(✉) 

Universidade de Coimbra, 3000 Coimbra, Portugal
pnf@dei.uc.pt
<http://eden.dei.uc.pt/~pnf>

Abstract. Today deep learning techniques (DL) are the main focus in classification of disease conditions from histology slides, but this task used to be done by more traditional machine learning pipeline algorithms (MLp). The first can learn autonomously, without any feature engineering. But some questions arise: can we design a fully automated MLp? Can that MLp match DL, at least in some tasks? how should it be designed? Can both be useful and/or complement each other? In this chapter we try to answer those questions. In the process, we design an automated MLp, build DL architectures, apply both to cancer grading, compare accuracy experimentally and discuss the remaining issues. Surprisingly, a carefully designed MLp procedure (acc. 86.5%) compared favorably to deep learning (best acc. 82%) and to humans (acc. 84%) when detecting degree of atypia for breast cancer prognosis on limited-sized publicly available Mytos dataset, with the same DL architectures that achieved accuracies of 97% on a different cancer classification task. Most importantly, we discuss advantages and limitations of alternatives, in particular what features make DL superior and may justify that choice, but also how MLp can be almost fully automated and produce useful structures characterization. Finally, we raise challenges, identifying how MLp and DL should evolve to offer explainability and integrate humans in the loop.

Keywords: Digital pathology · Deep learning · Machine learning · Classification

1 Introduction

1.1 The Problem and Motivation

The definitions and procedures related to cancer prognosis based on histopathology analysis are well described in [6]. If a tumor is suspected to be malignant, a

Supported by U. Coimbra.

doctor removes all or part of it during a procedure called a biopsy. A pathologist then examines the biopsied tissue to determine whether the tumor is benign or malignant and the tumor's grade, identifying other characteristics of the tumor as well. The tumor grade is the description of a tumor based on how abnormal the tumor cells and the tumor tissue look under a microscope. It is an indicator of how quickly a tumor is likely to grow and spread. Lower grade tumors, with a good prognosis, can be treated less aggressively, and have a better survival rate. Higher grade tumors are treated more aggressively, and their intrinsically worse survival rate may warrant the adverse effects of more aggressive medications. An important aspect of expert analysis of the tissue is to detect variations in tissue and on its structures between different degrees of illness. Automatic grading of histopathology slides offers interesting challenges in terms of classification due to the convolved properties of the tissues and structures in histopathology slides. Deep learning (DL) using convolution neural networks (CNN) is state-of-the-art in this task, due to high accuracy and autonomous learning capabilities. Figure 1 shows an example from our own experiments classifying degree of atypia using DL on Mytos Atypia dataset [40], where the left image was correctly classified as grade 2 with 99.5% confidence, and the right image was correctly classified as grade 3 with 87.4% probability. DL has displaced techniques based on machine learning pipelines (MLp) that require custom-made code to segment, identify, extract and represent features of specific structures. The need to hand-code parts is usually identified as the main problem of MLp approaches, but in fact the crucial advantage of deep-learning approaches is end-to-end autonomous backpropagation learning, where a large number of iterations of gradient descent on error backpropagation allows the networks to adjust their weights until they have learnt how to estimate the required quantities as best as possible.

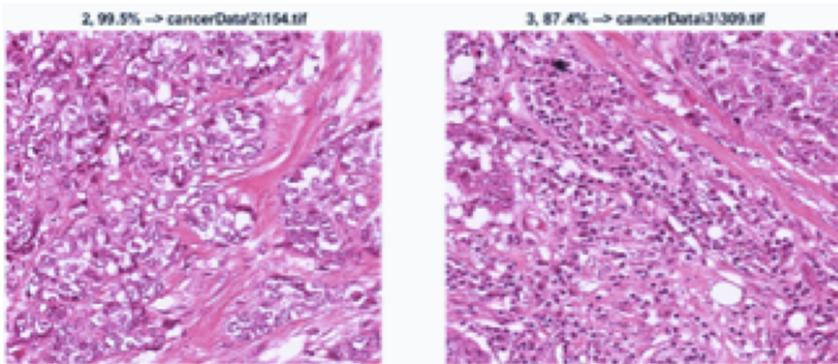


Fig. 1. Inception-V3 example classifications.

Given the lack of backpropagation learning and dependence on coding and operations choices in MLp, it is easy to design it in a very sub-optimal way, lacking the ability to extract and pick the best features for classification. This

raises two main questions: can we design a code-free fully automated MLP, or close to such, that is potentially competitive with DL, at least in some tasks? Can both be useful and/or complement each other? We also investigate future challenges in the context of classification of disease conditions from histology slides and in other medical imaging classification tasks. Those challenges include how the techniques should evolve to enhance their clinical/medical usefulness, how to provide explainability and how to integrate humans in the loop.

1.2 Contributions

In order to answer the questions posed we build state-of-the-art DL classifiers to be applied in the classification of histology slides, and we develop an automated MLP approach that is as optimized as we can possibly devise to maximize accuracy. The design of the MLP approach with its ability to find the appropriate features is an important contribution. Another contribution concerns applying the two experimentally and comparing the results in a specific problem of atypia grading using a publicly available dataset (Mytos Atypia), the same DL architectures also being applied to another well-known cancer detection problem (BreakHis) to confirm their capabilities. We conclude that the well-design MLP approach is competitive and even surpasses DL in the atypia grading problem with limited dataset, achieving (acc. 86.5%), versus deep learning (best acc. 82%) and humans (acc. 84%). Another contribution in this work is to discuss how the two paradigms can be used and/or complement each other. We expect this discussion to help in terms of clarifying the strengths of each paradigm and pointing directions for research. As part of future challenges, we highlight the need to make both DL and MLP more interpretable and explainable [1] and the need to work further in integration of humans in the loop [2]. We also discuss briefly how analysis of features in MLP can in the future be explored to help the objective of explainability. The study of MLP versus DL and the discussions on how to address future challenges based on these two techniques are very relevant in the definition of how future clinical/medical AI systems should be designed.

1.3 Structure

This chapter is structured as follows: Sect. 2 is the glossary, introducing used terms to ensure a common understanding. Section 3, state-of-the-art, describes both related work, the state-of-the-art DL architectures used and the design of the automated MLP approach. Then it discusses materials and methods for the experimental data on grading and classifying cancers, reports and compares results and concludes. Section 4 discusses challenges and open issues. In the light of the experience gained with the designs and experiments, we discuss the advantages and limitations of the two paradigms. Section 5, Future Outlook, proposes how automated MLP approaches and DL approaches may evolve to play together, complement or simply improve their capabilities.

2 Glossary

Artificial Intelligence, AI - a broad concept related to models and algorithms that make computer systems able to perform tasks normally requiring human intelligence.

Machine learning, ML - the term machine learning is defined as algorithms and statistical models designed to perform tasks without explicit instructions, relying on patterns and inference instead.

Machine learning pipeline, MLp - in this chapter we define MLp as a sequential pipeline consisting of the machine learning steps of segmentation, feature extraction, feature selection and classification.

Deep Learning, DL - Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher level features from the raw input. The word deep comes from having many layers and the word learning from the capacity to learn a model from data.

Convolution neural network (CNN) - The convolution neural network is the most frequent type of DL used to classify images into categories. It uses multiple layers with convolutions based on a large number of filters to capture properties automatically from more local fields of view, then it progressively extracts higher level features as it merges the feature maps into smaller, more generalized convolutions.

Segmentation - the process of partitioning a digital image into multiple segments or regions. The goal of segmentation is to delineate and locate structures or objects and boundaries in images.

Semantic segmentation - the task of classifying each and every pixel of an image as a class. The classes are the structures or objects that are to be discovered, such that all pixels belonging to those structures should be classified as such.

Feature extraction, features - given some image or data, feature extraction derives a set of values (features) intended to describe the main characteristics of the original data, to be used in subsequent learning, classification and generalization steps. In the case of image classification, the features are most frequently numeric quantities summarizing some properties of regions, e.g. colour, texture or shape properties.

Feature selection - feature selection means selecting a subset of all features that were extracted from the image. In general, in a classification problem, “good” features are features that contribute significantly to distinguish the class, and redundant features are features that are highly correlated. Feature selection should try to find the best describing features and drop redundancy as much as possible.

Dimensionality reduction - the process of reducing the number of features by obtaining a smaller set of principal variables. Dimensionality reduction can be obtained by either feature selection or feature projection. Feature selection was defined previously, feature projection involves a transformation of the variables into a space of fewer dimensions. An example of feature projection is principal

components analysis, where the variables are replaced by a smaller set of principal components that are computed from those variables and “summarize” the most relevant characteristics of those variables.

Classification - the problem of identifying to which of a set of categories an observation belongs, usually training from a set of data observations (or instances) whose category is known.

3 State-of-Art

3.1 Related Work

The field of cancer detection and classification using computerized techniques has gained increasing popularity during the last decade or so, given a large increase of computational power, the enormous advances in machine learning and the proposal and evolution of procedures that are able to analyze medical images automatically and classify or detect a degree of disease from those. Up to around 2013, most image analysis and classification techniques were machine learning pipelines (MLp), following a certain sequence of vaguely defined steps to segment, extract features and further analyze the images. Then Convolution Neural Networks (CNNs) started to gain popularity as highly accurate classifiers, new architectures were developed and beat previous approaches in terms of accuracy [16]. Looking at results from past works, accuracies on the order of 95% or 100% are very common in both MLp and DL paradigms. For instance, in [7] MLp using features describing characteristics of the cell nuclei present in the image result in accuracies of 96% to 97.5% using repeated 10-fold cross-validation. Likewise, state-of-the-art CNN approaches for classes cancer/no cancer on the BreakHis dataset [8] achieved 80 and 90% accuracy and, using patches and a myriad of modifications, others [9,10] report more than 95% accuracy on the same problem. But in [7], there seems to be a considerable amount of manual work to achieve the result. Can we automate all the steps, to make it less disadvantageous when compared with DL? And are there possible advantages in the use of MLp to complement DL? In MLp approaches segmentation and feature extraction that individualizes cells and measures a set of specific properties of those cells is needed. For instance, Loukas et al. [11] explored a technique that pre-selected 65 regions of interest to grade cancer into 3 degrees of malignancy (I-III), the neural network classifier achieving 90% accuracy; [12] achieved 80 to 95% accuracy grading prostate and breast cancers using three scales of low-level pixel values, high-level information based on relationships, and structural constraints; [13] used multi-wavelet grading of prostate pathological images, achieving a precision of 97% discriminating prostate tissue cancer/no cancer and 81% for different degrees of low and high Gleason. More related works exploring analysis of regions of interest or structures can be reviewed in [14] and [15]. In our own previous work we also explored regions characterization [4] and [5]. More recently Convolution neural networks (CNNs) achieved top accuracies [8–10] and replaced feature extraction by convolution layers applying

convolution operations [18]. Given the prior work on MLp and the highly practical and accurate CNN paradigm, a question arises on whether it is possible to design a fully automated MLp that might also be easy to apply and accurate. The closest to our intended MLp design would be CellProfiler in [21–23], since it at least offers some interface for deciding and collecting features, but using it in a complete MLp still requires a lot of manual human intervention to code, test and experiment with alternatives in steps of the pipeline, from segmentation to feature selection and classification. The automated MLp extracts objects, characterizes them and processes the extracted structures to build the classifier model automatically, and is applied automatically to classify new images as well.

Explainability/causability is another very relevant issue related to this kind of systems. In medicine there is growing demand for AI approaches that are trustworthy, transparent, interpretable and explainable for a human expert, and this is especially relevant in the clinical domain [1]. The authors in [1] argue that the best performing statistical approaches today are black-boxes and do not foster understanding, trust and error correction. This implies an urgent need for explainable models, explanation interfaces and causability. The systems should give answers to questions such as “Why did the algorithm do that?”, “Can I trust these results?”, “How can I correct an error?”, so that the medical expert would be able to understand why, learn and correct errors and re-enact on demand. The need to make both DL and MLp more interpretable and explainable should be answered in the future, and post-processing of the features extracted by MLp, DL or mixed approaches can be explored further to improve explainability. We call the attention to this future research challenge, at the same time that we briefly illustrate with a small example how MLp features can be useful for further analysis and explainability. We believe DL and MLp can “collaborate” in this issue using extracted features to explain better what is happening and why.

3.2 Deep Learning Architectures Used

Matlab™ 2018’s InceptionV3 [20] and Resnet-101 [19] networks are pre-trained implementations of the state-of-the-art InceptionV3 and Resnet architectures pre-trained on more than a million images from the ImageNet database to classify images into 1000 object categories. Not only each of these represents specific architectural details, as the number of layers increases as we move from InceptionV3 to Resnet. While InceptionV3 is a 48 layers deep network based on the Inception architectural features, Resnet-101 is a 101 layers deep network following the Resnet architecture.

3.3 Design of the Machine Learning Pipeline (MLp)

The building of the typical MLp has steps (1) segment, (2) extract features, (3) characterize the image, (4) reduce features space and (5) build classifier. The first part of MLp is an approach to characterize an image based in three main steps (segment, extract features, characterize the image). In each step it is necessary to take precautions to avoid losing information that is important for

accuracy of the approach. In step (1), the image is segmented, and structures are identified from the resulting regions. In the case of histopathology slides, examples of structures include cells of specific types, interstitial tissue, groups of cells, adipocytes and others. The outcome of this step is a set of image regions, $I = \{r_i\}$ and a mapping from regions to structures. In this step it is important not only to segment the structures well, but also to do it such that each image pixel will be assigned to some structure (semantic segmentation). This is important since tissue modifications related to disease conditions can occur in any type of structure that is present in the images. As an example, the fabric/texture of interstitial tissue is expected to change in a cancer condition, therefore the interstitial tissue should be one of the classes. Also, the classes should be aligned with the output of segmentation. For instance, since cells are frequently overlapped, class “cell cluster” is created to represent that structure. In step (2) a set of features $[F_j]$ are extracted per region, so that each region r_i is mapped into that set of features. While in DL end-to-end error backpropagation tunes feature extraction automatically, the only way to avoid losing important features in MLP is to define all potentially useful features and extract them all. Step (3) builds structure probability distribution functions (sPDF). Given the regions of each type of structure S_l , and for each feature F_j , the sPDF $_{l,j}$ is represented as a histogram $H_{l,j}$ where, for each interval of possible values, the probability of occurrence is recorded (F_{P_y}). This histogram represents the probability that some structure takes some value in an interval for some specific feature. The second part of MLP concerns reduction of the feature space and building of the classifier. Next we provide more details on the steps.

Segmentation of Histopathology Slides. Segmentation algorithms that could be used have been studied extensively in the past, including traditional unsupervised approaches (e.g. [24–28, 30, 32–35]) or semantic segmentation using in deep learning networks [36] (e.g. fully convolutional network [37], U-net [38] or deeplab [39]). For our purpose, the most relevant issue in the design of the MLP is not segmentation algorithm but a tool that separates the image into meaningful regions, such as what CellProfiler does [21–23]. As in CellProfiler, we created a tool for the user to obtain segmentations and tune segmentation parameters. Figure 2 shows an example output after we configured it to define 12 structures. Note that the different structures can overlap partially. The tool options are threshold intervals, morphological operations, geometric properties and grids (to divide regions that may extend over the whole image), and to individualize regions the tool uses labeling of connected components (bwlabel). Note that Fig. 2 structures include cells, clusters of cells, interstice, adypocits, but also halos or aureoles. A halo is also a structure but one which captures the vicinity of another type of structure.

After the user configures segmentation the MLP becomes autonomous segmenting any image of that type, and since all the remaining MLP steps are completely autonomous, the whole pipeline runs automatically for both training and use.

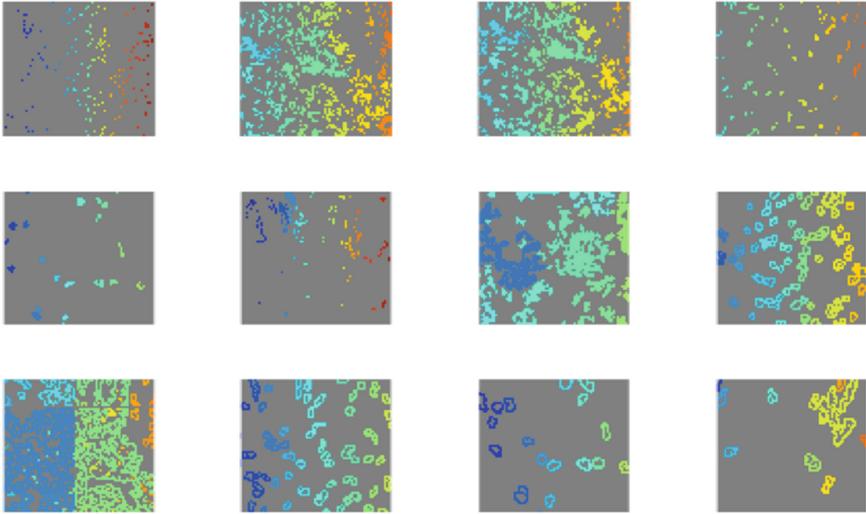


Fig. 2. A segmentation into 12 structures.

Describe Characteristics of Regions. The objective of this step is to characterize discovered regions using visual properties useful for distinguishing disease conditions. Since MLp does not learn which features to extract, it needs to extract all region features that could potentially be relevant for classification. The features that cover all properties that might be useful are counts and densities (D), shapes (S), geometries (G), texture (T) and color (C). Counts and densities (D) are aggregate measures counting the number of occurrences of each type of structure (number of regions of each structure) in images, and the number of occurrences per unit area in each of a number of grid divisions of the image (a $n \times n$ grid). A histogram describes the densities encountered. These details can capture for instance an abnormal concentration of small black cells, or any other abnormality in terms of densities of structures. Geometry (G) is a set of aggregate measures taken on each individual region that characterize the extent of the region (the pixels) as an aggregate. Shape (S) characterizes the form of the contours, not as aggregates (captured by geometry), but how the contour curves evolve. Texture (T) captures modifications in the general fabric of a specific structure such as interstitial tissue or cells. The ensemble of all extracted features is denoted as DSGTC features (Density, Shape, Geometry, Texture and Colour). All these features are extracted independently for each region, and since each region is of a specific structure type, structures are characterized by the distributions of those properties for all regions of the structure. The feature extraction process is completely automated, with no human intervention in any phase.

Feature Selection. Feature selection is needed to eliminate useless features, reduce the dimensionality of the dataset and to reveal the contribution of individual features to the outcome. Given n_s structures, n_f DSGTC features and n_{bin} bins per histogram, the final number of probability features, describing all structures for a specific disease condition, is given by $n = n_s \times n_f \times n_{bin}$. The value n is expected to be very large, since it represents all probability histogram bins for an already large number of features and for each structure. As an example, we had $n = 12 \text{ structures} \times 600 \text{ features} \times 10 \text{ bins} = 12 \times 6000$ probability features. Only a small fraction of those are relevant to determine the class (e.g. degree of atypia). Either feature projection or feature selection could be applied at this stage. We applied feature selection separately to each structure, reducing from 6000 to 50 most revealing features per structure. Three feature selection methods were used in conjunction: correlation analysis of each feature to the class (Y-correlation), correlation analysis between pairs of features (X1X2 correlation) and sequential feature selection (sequentialfs). Sequentialfs creates candidate feature subsets by sequentially adding each of the features not yet selected. For each candidate feature subset, sequentialfs performs 10-fold cross-validation by repeatedly evaluating classification with different training and test subsets, choosing the candidate feature with best accuracy. The number of features is reduced from 6000 to 100 by correlation analysis prior to calling sequentialfs.

Classification. Any classifier model can be tried in this step. We experimented a set of classifiers that included neural networks, random trees ensemble classifiers and nearest-neighbour. The classifiers were built using 5-fold cross-validation over the dataset, and the accuracy metrics collected over the test folds included precision, recall and F-Score, which were used to compare with DL.

3.4 Experimental Setup

The Mytos Atypia contest has provided a set of selected and annotated slides of breast cancer biopsy. The slides were stained with standard hematoxylin and eosin (H&E) dyes and they have been scanned by two slide scanners: Aperio Scanscope XT and Hamamatsu Nanozoomer 2.0-HT. In each slide, the pathologists selected several frames at X20 magnification located inside tumours. These X20 frames were used for scoring nuclear atypia. A X20 frame scanned using Aperio Scanscope XT is $755.649 \times 675.616 \mu\text{m}^2$, 1539×1376 pixels, one from Hamamatsu Nanozoomer 2.0-HT is $755.996474 \times 675.76707 \mu\text{m}^2$, 1663×1485 pixels. The number of frames is variable from slide to slide. In the training data set there are 284 frames at X20 magnification. Note that the dataset is limited in size, both patching and data augmentation were included as experiments to increase the size and variability in DL training. The frames are RGB bitmap images in TIFF format. The nuclear atypia score is provided as a number 1, 2 or 3. Score 1 denotes a low grade atypia, score 2 a moderate grade atypia, and score 3 a high grade atypia. This score has been given independently by two

different senior pathologists. There are some frames for which the pathologists disagree and gave a different score. We account those as incorrect classifications by human experts. In those cases a third senior pathologist would give the final score. Nuclear atypia score is a value, 1, 2 or 3, corresponding to a low, moderate or strong nuclear atypia respectively. Instead of focusing on segmenting and measuring the nuclei solely, both the DL and MLp approaches developed and tested in this work take the whole tissue images and detect atypical characteristics that dictate the degree of atypia. For experiments, the image datasets were collected and 5-fold cross-validation was applied. In 5-fold cross-validation 5 runs are ran with 80% training, 10% testing and 10% validation data. Patching refers to dividing the images of the dataset into multiple smaller images (e.g. 128×128 patches) to be fed to the DL. Those can better individualize structures and may provide a convenient degree of detail about regional structures, while also augmenting the dataset, by dividing an image into many patches. A stride (e.g. start a patch every 10 pixels) can be defined to obtain overlapping patches. Data augmentation is a different technique designed to increase the size and variability of the training dataset based on simple operations such as scaling, rotation, shearing or translation. This can contribute to increase the variability of training images, resulting in more and more diverse training instances.

Experimental Setup Details. We defined the following alternatives for experimentation: HExpert, MLp, CLASS, Iv3, R101, Iv3 augment, R101patch and Iv3patch. Hexpert is the accuracy of medical doctors, measured as the degree of agreement assigning the grades; MLp is the machine learning pipeline described in this work; CLASS is an ML classifier that does not differentiate structures, it simply applies segmentation and extracts a set of features (GLCM, LBP, gray level intensities) from all regions indistinctively, then applies feature reduction and a neural network classifier. CLASS represents a simpler ML pipeline; InceptionV3 (Iv3), Resnet-101 (R101), InceptionV3 with data augmentation (Iv3 augment), and versions of Iv3 and R101 with 128×128 , 10 pixels stride patching (R101patch, Iv3patch) are DL alternatives built using Matlab 2018 implementations of state-of-the-art CNNs, including InceptionV3 (Iv3) and Resnet-101 (Res). The imageAugmenter used for data augmentation applied Random X Reflection, and both X and Y translation. The DL training options included the following: (stochastic gradient descent with momentum ('sgdm'), miniBatchSize 10, maxEpochs 700, initial learn rate $1e-4$, validation frequency 3. In what concerns the MLp setup, after tuning with a few images the segmentation tool divided images into 5 intensity levels based on 3 thresholds (130, 180, 230), followed by a sequential set of operations to obtain the types of structures from those levels. The operations included removal of small noise (removal of very small regions inside larger regions resulting from thresholding), filling of holes and closing to fill and improve contours of small dark and mammarian cells, opening of interstitial tissue regions, individualizing regions by labeling connected regions (bwlabel), dividing white regions based on size, filling and closing those regions, applying circularity to distinguish rounded from non-rounded

large white regions, applying a grid to interstice. After these steps comes the step of creating halo structures, which are structures capturing the vicinity of the previously individualized regions for each structure type. Creating the halos involves dilating the regions (imdilate) and then retrieving only the dilation ring. The resulting regions were ‘darknCells’, ‘cells’, ‘cellsExtraFilled’, ‘fatSmall’, ‘fatLargeRound’, ‘fatLargeNotRound’, ‘interstice’, ‘darknCellsHalo’, ‘cellsHalo’, ‘fatSmallHalo’, ‘fatLargeRoundHalo’, ‘fatLargeNotRoundHalo’. Features of individual regions included area, solidity, major axis, minor axis, eccentricity, convex area, extent and others, contour slopes and variations of slopes (dslope), gray-level co-occurrence matrix (GLCM) [42], local binary patterns (LBP) [41], plus 2D texture histograms (spatial distance x colour intensity distance). Feature selection ran automatically and separately per structure. The first step involved pruning out features with a correlation less than 0.1 with the class. The second step involved removing features that were correlated above 95% with their class-correlation-ranked neighbours, followed by an additional class-correlation based pruning to keep only the top 100 features. Finally, sequential feature selection was used with the classifier F-score as the criteria to choose the subset of 50 best features for each structure. After this step all chosen 50×12 features representing all structures were again reduced into 100 final features using the same procedure. The last step of model building involved building of a classifier based on the 100 final features. We experimented neural network (see Matlab2018 patternet) with a configuration of two hidden layers of 10 neurons each, random trees ensemble classifier (see Matlab2018’s TreeBagger Bag of decision trees), with a default number of 10 trees, and k-nn with 3 neighbours. All the classifiers were built using 5-fold cross-validation over the dataset. From our experiments we report accuracy metrics that include precision, recall and F-Score.

Hardware Details. Experiments were ran in a PC running windows, with an intel core i5 at 3.4 GHz, 16 GB RAM and an SSD disk of 1TB. The PC had an NVIDIA GForce GTX 1070 GPU installed (Pascal architecture with 1920 cores, 8 GB GDDR5, 8 Gbps memory speed), and the experiments were all setup and ran in Matlab 2018a.

3.5 Experimental Results

Testing Accuracy of DL with BreakHis. In this experiment we tested Resnet-101 on the (yes/no) problem of detecting cancer on the BreakHis dataset [8], with 128×128 image patches. This served as a calibration test, resulting in a validation accuracy of 97%, as shown in the screenshot of Fig. 3 (and a test accuracy of 96.7% as well). This result coincides with results of other authors, giving us confidence that the DL approaches were well configured.

Results of DL with Mytos Atypia. Table 1 shows the results obtained by DL on the Mytos Atypia dataset. It reveals accuracies between 73% and 82.5%, the best approach being Iv3 with data augmentation.

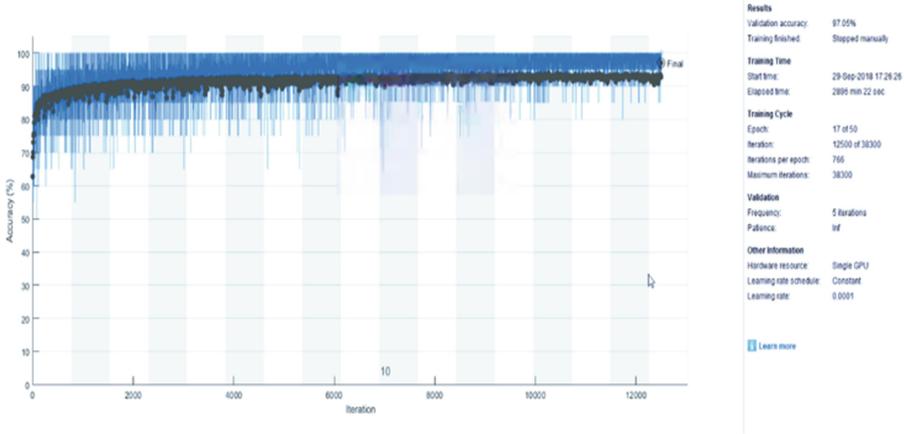


Fig. 3. Calibration run: Resnet on breakHis patches 128×128 .

Table 1. Comparison of DL results on Mytos Atypia.

Approach	Iv3+augm	Iv3	Iv3+patch	R101+patch	R101
Test accuracy	82.50%	81.20%	79.30%	77.60%	73.00%
C1 = low grade atypia	76.20%	67.10%	77.43%	79.80%	76.20%
C2 = moderate-grade atypia	71.40%	81.30%	73.20%	68.10%	57.20%
C3 = high-grade atypia	100%	95.20%	87.24%	84.90%	85.70%

Results of MLP on Mytos Atypia. Table 2 shows the results we obtained for MLP, including accuracy, precision and recall. The table shows the metrics for all classes and the metrics obtained for each class. Precision was 86.5%, and grade 2 has the lowest precision, probably because the boundary between grades is fuzzy.

Table 2. Accuracy, precision and recall of MLP.

Approach	Acc	Precision	Recall
All	86.5%	86%	87%
Grade 1	93%	94%	90%
Grade 2	79%	71%	91%
Grade 3	90%	99%	80%

Table 3 shows accuracy of MLP using different classifiers and also CLASS. The superiority of MLP is clear compared to CLASS, and all MLP classifiers had good accuracy, with random forests being the best.

Table 3. Accuracy, precision and recall of MLp.

Classifier	MLp	CLASS
knn-3	83.5%	74%
Random forest	86.5%	70%
Neural net	82.8%	73.5%

MLp Runtimes. Figure 4 shows image segmentation times (average 1.8 s) and Fig. 5 shows features extraction time per structure and per image (average 11.3 s). Since regions of 12 structures were extracted, each image took an average of $1.8 + 136$ s to be processed. This time is incurred during classifier construction. For classification of new images, it is possible to speedup execution by extracting only the 100 selected features. Feature selection takes a lot more time (average 52.5 min in five runs), due to the sequential feature selection step that calls the classifier on each step. Note however that this step is only necessary during model building.

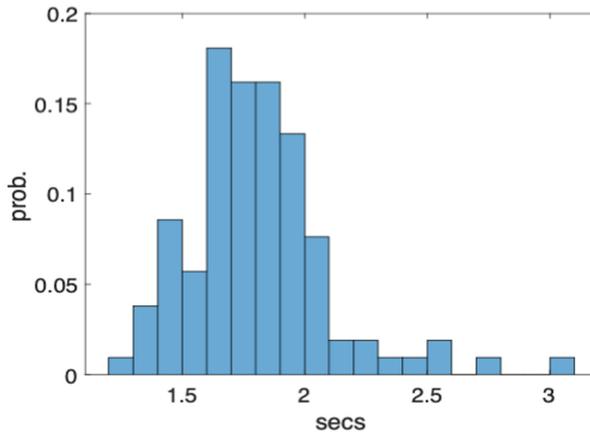


Fig. 4. image segmentation (mean = 1.8 s) and per-structure feature extraction (mean = 11.3 s).

Comparing MLp to DL. Table 4 compares the results of the two best DL approaches to those of MLp on the Mytos Atypia dataset. We also include the accuracy of human experts, measured as the fraction of agreement between the two pathologists labeling the images.

These results show that the state-of-the-art DL approaches did not achieve as high an accuracy as the well-designed MLp classifier. This can also be seen from the partial per-class accuracies, with MLp being superior in classifying low, moderate and high-grade atypia, with the exception of high-grade atypia by Iv3 with data augmentation, but at the expense of the remaining classes.

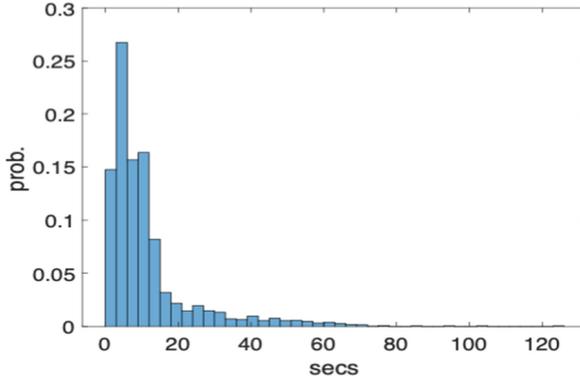


Fig. 5. Image segmentation (mean = 1.8 s) and per-structure feature extraction (mean = 11.3 s).

Table 4. Comparison DL to MLp and others on Mytos Atypia

Approach	MLp	HExpert	Iv3 augm	R101patch
Test accuracy	86.30%	84%	82.5%	77.6%
C1 = low grade atypia	93.10%	76.20%	79.80%	–
C2 = moderate-grade atypia	79.40%	71.40%	68.10%	–
C3 = high-grade atypia	90.20%	100%	84.90%	–

4 Open Problems and Challenges

There are open problems and challenges resulting from this study, both related to the design and use of MLp, the comparative study of MLp and DL, future improvements of DL and how to benefit from using both.

4.1 Degree of Automation of MLp

The MLp is able to run completely automatically over a dataset, but a prior human-based configuration of segmentation parameters to elicit structures was necessary, and the structures that were obtained are an important factor for accuracy. In the future this pre-configuration step can be replaced by deep-learning-based semantic segmentation using groundtruth segmentations. In that case MLp uses DL in one of its steps, and we need to provide the groundtruth segmentations, but the whole MLp procedure becomes totally automated with no need to pre-configure anything. Further work would also be required in improving semantic segmentation based in DL.

4.2 Accuracy Comparison DL to MLp

We created an automated MLp designed to recognize structures, characterize them and then classify disease conditions based on properties of those structures. We had to be careful in every step to avoid losing accuracy, by keeping as much information as possible up to feature selection, and “learning the best features” during feature selection. Our experiments have shown that MLp matched DL accuracy (and even improved it) in a specific experiment, with a limited sized dataset (Mytos Atypia). A large effort is welcome in the future to further compare well-designed MLp to DL in this and other contexts, with different and larger quantities of imaging data, and to evaluate under which circumstances one could be better than the other. Even in this experiment, although MLp achieved better accuracy than DL, it required prior manual configuration and tuning of segmentation parameters and some iterations of the whole pipeline to tune the segmentation to achieve top accuracy. Consequently, we have only shown that MLp can be competitive in some problems, but with some tuning still required.

4.3 Autonomous Learning Ability of DL Versus MLp

The advantage of DL compared with MLp is its capacity to learn end-to-end iteratively, where end-to-end means from the inputs (images) to the output, classification, based in error backpropagation [17]. Its main potential limitation is the difficulty to converge its inner weights to find the best solution. In contrast, MLp learns using a feature selection algorithm that is given a very large number of possible features, and a classifier building algorithm that is finding the most suitable parameters for the classifier. MLp does not learn segmentations iteratively currently, but that step could be replaced by deep learning-based semantic segmentation in the future. Still, MLp also does not backpropagate the classification error from output to input to improve accuracy along epochs of training. In spite of these limitations, MLp was still more accurate than DL in our experiment because it was able to find the most discriminating features among all relevant properties of structures. As a conclusion, DL seems to be the best choice in general, because it can learn more globally and adapt all of its weights automatically by back-propagating the error, to tune what it extracts and from where. Additionally, future work can bring improvements to DL that might make it more accurate still. But a well-designed MLp can still be more accurate, at least in some problems, and as long as it is fully or almost fully automated, it can always be used to provide extra information and characterization capabilities. The two paradigms can be applied simultaneously, and MLp can provide complementary information and characterizations.

4.4 Characterization of Disease Markers

MLp can be used to better characterize disease conditions based on which most relevant features of which structures are modified by disease conditions. We illustrate this by doing a short study using the MLp results. Figure 6 is taken

from our results and shows that texture features are important disease markers in interstitial tissue and in clusters of cells (groups of juxtaposed cells). This means detection of variations in the texture of the tissue itself and of agglomerates of cells helps distinguish disease conditions based on those structures. That is consistent with the hypothesis that modifications in tissue fabric, such as hypercellular and more irregular altered tissue architecture, are indicative of higher cancer grades; In contrast to that analysis, in what concerns mammarian cells, shape and geometry features gain a lot of importance. Additionally, the mix for cells clusters also indicates some relevance of shape features, although less prominent. These observations agree with the fact that increased irregularity of cells and contours, some bigger cells and more irregular shapes are indicative of higher grades of cancer. Finally, shape and geometry were also chosen as relevant discriminators in vacuoles and adipocytes, probably signaling the importance of detecting more squeezed structures and more irregularity in their shapes due to altered tissue architecture in higher cancer grades. This study could be enhanced in the future, with more detailed analysis of which features are most revealing and so on.

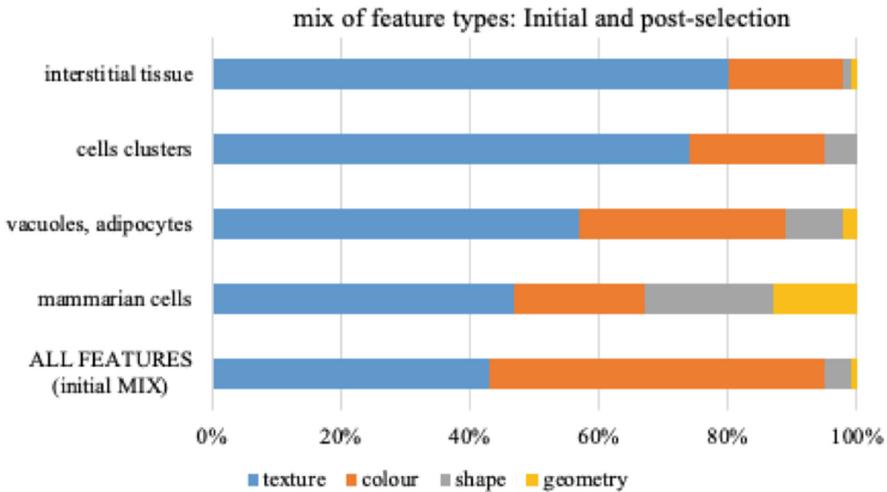


Fig. 6. Feature selection mixes for each structure.

4.5 Towards Explainability/Causability

As reviewed in the related work section, explainability/causability is one of the most relevant issues for adoption of AI models and techniques in medicine clinical practice [1]. The ability to characterize disease markers that we discussed in the previous subsection is one of multiple possible mechanisms that can be integrated to build explainable/causable AI systems. This means that MLp might

be used as part of automated procedures to explain classifications and to establish causability, and it can be mixed with DL in practical systems as well. Most importantly, one of the most relevant future challenges in either MLp, DL or a mix of both is how to achieve explainability/causability instead of just having a black box system that is unable to provide very relevant explanatory power to users.

5 Future Outlook

Deep learning has displaced prior approaches and is here to stay. Its superiority is not a guarantee that it achieves better accuracy, but the fact that it learns autonomously, iteratively and end-to-end is a great advantage. We should expect future research to improve DL approaches further and to apply them to most classification and segmentation problems related to medical imaging in general and digital pathology in particular. If we had to choose only one approach between DL and MLp, we would choose DL because of its end-to-end, backpropagation-based learning that tries to optimize accuracy completely autonomously. MLp also searches for the best features and classifier model, but it still lacks end-to-end error backpropagation. The future outlook for MLp seems grimmer than that of DL, but MLp can still have a role together with DL. Since MLp is fully automated after an initial configuration of segmentation step, we can apply both in a specific context and test their accuracy; Instead of deciding between the two, we can have the results of both to gain more evidence. Most importantly, MLp can complement DL by characterizing structures and providing explanatory information. It can tell us which types of features and which features are more discriminative to detect a disease condition, and how those features reveal the disease condition. More generically, a crucial future challenge is to design systems that have explainability/causability capabilities using MLp, DL or both. Another distinctive opportunity to apply MLp together with DL is to segment medical images into structures using DL and then apply the MLp pipeline we designed to identify the properties of structures that change with disease conditions and how they change. It would also be important to fully integrate humans in the loop in the future, as human experts should be able to interact with the AI system in much richer ways [2]. That includes integrating physicians high-level expert knowledge into the process, by acquiring his/her relevance judgments regarding a set of initial results [3]. Either MLp, DL or mixed systems should be designed to integrate the human-in-the-loop, adding interactivity and learning from both sides (humans and algorithms).

References

1. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. *Wiley Interdisc. Rev. Data Min. Knowl. Discov.* **9**, e1312 (2019)

2. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inf.* **3**(2), 119–131 (2016)
3. Akgul, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: current status and future directions. *J. Digit. Imaging* **24**(2), 208–222 (2011)
4. Furtado, P.: Objects characterization-based approach to enhance detection of degree of malignancy in breast cancer histopathology. In: *Medical Imaging 2019: Image Processing*, vol. 10949, p. 109491R. International Society for Optics and Photonics (2019)
5. Pereira, J. Barata, R., Furtado, P.: Experiments on automatic classification of tissue malignancy in the field of digital pathology. In: *Second International Workshop on Pattern Recognition*, vol. 10443, p. 1044312. International Society for Optics and Photonics (2017)
6. NCI Dictionary of Cancer Terms. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor-grade>. Accessed 4 Oct 2019
7. Wolberg, W.H., Street, W.N., Heisey, D.M., Mangasarian, O.L.: Computer-derived nuclear features distinguish malignant from benign breast cytology. *Hum. Pathol.* **26**, 792–796 (1995)
8. Spanhol, F.A., Oliveira, L.S., Petitje, C., et al.: A dataset for breast cancer histopathological image classification. *IEEE Trans. Biomed. Eng.* **63**(7), 1455–1462 (2016)
9. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv preprint: [arXiv:1606.05718](https://arxiv.org/abs/1606.05718) (2016)
10. Motlagh, N.H., et al.: Breast cancer histopathological image classification: a deep learning approach. In: *bioRxiv*, 242818 (2018)
11. Loukas, C., Kostopoulos, S., Cavouras, D.: Breast cancer characterization based on image classification of tissue sections visualized under low magnification. In: *Computational and Mathematical Methods in Medicine* (2013)
12. Naik, S., et al.: Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In: *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2008*. IEEE (2008)
13. Jafari-Khouzani, K., Soltanian-Zadeh, H.: Multiwavelet grading of pathological images of prostate. *IEEE Trans. Biomed. Eng.* **50**, 697–704 (2013)
14. Mitko, V., Pluim, J.P.W., van Diest, P.J., Viergever, M.A.: Breast cancer histopathology image analysis: a review. *IEEE Trans. Biomed. Eng.* **61**(5), 1400–1411 (2014)
15. Aswathy, M.A., Jagannath, M.: Detection of breast cancer on digital histopathology images: present status and future possibilities. *Inf. Med. Unlock.* **8**, 74–79 (2017)
16. Krizhevsky, A., Sutskever, I., Hinton, G. E. : ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
17. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature.* **323**(6088), 533–536 (1986)
18. Alom, Md.Z., et al.: The history began from AlexNet: a comprehensive survey on deep learning approaches. ArXiv Preprint [ArXiv:1803.01164](https://arxiv.org/abs/1803.01164) (2018)
19. He, K., Zhang, X., Ren, S., Sun., J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
21. Yu, K.-H., et al.: Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, Article number: 12474 (2016)
22. Carpenter, A.E., et al.: Cell profiler - image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**(10), R100 (2006)
23. Kametsky, L., et al.: Improved structure, function and compatibility for cell profiler: modular high-throughput image analysis software. *Bioinformatics* **27**, 1179–1180 (2011)
24. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. *Int. J. Comput. Vis.* **1**(4), 321 (1988)
25. Vincent, L., Soille, P.: Watersheds in digital spaces - an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(6), 583–598 (1991)
26. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Sys. Man. Cyber.* **9**(1), 62–66 (1979). <https://doi.org/10.1109/TSMC.1979.4310076>
27. Liao, P.-S., Chen, T.-S., Chung, P.-C.: A fast algorithm for multilevel thresholding. *J. Inf. Sci. Eng.* **17**(5), 713–727 (2001)
28. Zhu, N., Wang, G., Yang, G., Dai, W.: A fast 2D otsu thresholding algorithm based on improved histogram. In: Chinese Conference on Pattern Recognition, CCPR 2009, pp. 1–5 (2009)
29. Deing, Y., Manjunath, B.S.: Unsupervised segmentation of color-texture regions in images and videos. *Trans. Pattern Anal. Mach. Intell.* **23**(8), 800–810 (2001)
30. Hartigan, J.: *Clustering Algorithms*. Wiley, Hoboken (1975)
31. Achanta, R., et al.: Slic superpixels. In: EPFL Internal Report, No. EPFL-REPORT-149300 (2010)
32. Nock, R., Nielsen, F.: Statistical region merging. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(11), 1–7 (2004)
33. Ester, M., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231. AAAI Press (1996). ISBN 1-57735-004-9
34. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001)
35. Cheng, Y.: Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(8), 790–799 (1995)
36. Ouaknine, A.: Review of deep learning algorithms for image semantic segmentation. https://medium.com/arthur_ouaknine/review-of-deep-learning-algorithms-for-image-semantic-segmentation-509a600f7b57. Accessed October 2019
37. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

39. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2017)
40. MitoS-Atypia-14. <https://mitos-atypia-14.grand-challenge.org/Dataset/>. Accessed October 2019
41. Wang, L., He, D.: Texture classification using texture spectrum. *Pattern Recogn.* **23**(8), 905–910 (1990)
42. Haralick, R., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**(6), 610–621 (1973)