UNIVERSITY OF SOUTHERN DENMARK  
FACULTY OF SCIENCE

DS807: Applied Machine Learning

Winter 2023

1. Assigner: Christian Møller Dahl.
2. Hand-out: January 24, 2023, 12:00 (noon).
3. Hand-in: January 31, 2023, 12:00 (noon).
4. All pages, incl. the front page, should contain the following: Full name and SDU username (**not** CPR-number).
5. All pages must be numbered.

Form of examination for the certificate:  
Take-home assignment.

Supplementary information for the form of the exam:  
The exam may be solved in groups of up to 5 students or individually. Working in groups is encouraged. You should make the definition of your group in System DE-Digital Exam before the start of the exam. Follow this [guideline](https://mitsdu.dk/-/media/sdunet/filer/vaerktoejer/brugeradgang/digitaleksamen/digitaleksamen-stud/uk+vejledninger/uk_s02_guide_stud+gruppeaflevering.pdf).

Further:

1. In your report, be sure to state explicitly who is responsible for which parts to facilitate individual assessment.
2. Location: Home assignment.
3. Internet access: Necessary.
4. Hand-out: System DE-Digital Exam.
5. Hand-in: System DE-Digital Exam.
6. Extent: No longer than 30 pages, excluding references, appendices, and code.
7. Exam aids: All exam aids are allowed.
8. File format: The report must be submitted as a **.pdf** file. The code may be submitted as one of: **.ipynb**, **.html**, or **.py**.

Grading according to the Danish 7-point scale. Grading based on the performance of the individual student compared to the learning goals.

**Exam questions**

During the semester you have become very excited about the field of digital pathology which is an area that is developing rapidly due to advancements in microscopy imaging hardware. These advancements have allowed digitizing glass slides into whole-slide images. You have recently read the paper by [Veeling et al (2018)](https://paperswithcode.com/paper/rotation-equivariant-cnns-for-digital) and you are thrilled to see that the authors have derived a novel dataset, denoted PatchCamelyon (PCam), that will enable you to develop and benchmark your own machine learning models. As Veeling et al (2018) you are primarily interested in developing machine learning models that based on patches of whole-slide images of lymph node sections can assist pathologist in tumor detection.

The primary objective of this exam is to perform image classification using the PCam dataset. The dataset consists of 327,680 color images (96x96pxs) extracted from histopathologic scans of lymph node sections. Each image is annotated with a binary label indicating presence of metastatic tissue. Specifically, the database is split into 262,144 training images, 32,768 validation images and 32,768 test images. You must use the training/validation images to train/validate models that perform well at classifying the test images.

Importantly, you are not required to use the full dataset. Use the amount of data that is feasible for you and your hardware configuration. The PCam dataset is available from many online sources but some of them are very slow. I therefore recommend getting the dataset from my SDU repository at: [Link to PCAM](https://syddanskuni-my.sharepoint.com/:f:/g/personal/cmd_sam_sdu_dk/EiWD2LmuxCJBp-_tfGK7aL8Bair7l5z8FU5sp5pLjlhKwg?e=FLzWno) . Further instructions and hints on how to load the data efficiently and flexibly by using the module [tfds|Tensorflow Dataset](https://www.tensorflow.org/datasets/api_docs/python/tfds) are available on the course site’s [itslearning](https://sdu.itslearning.com) platform.

For all questions in the exam, be sure to state how and whyyou prepare the data, including considerations for how to further split the data, scale the data, reshape the data etc.[[1]](#footnote-1) In particular, it might be necessary to consider rescaling the images down due to the hardware limitations on your system. This is perfectly fine, but please be sure to discuss the potential implications and consequences. In general, if you use the same method for multiple questions, it is sufficient to describe the procedure once and refer to it in subsequent questions (however, it must still be motivated).

Question 1  
Use non-deep learning to perform image classification (tumor detection). Specifically, you must:

1. Discuss how the problem can be solved using support vector machines, random forests and boosting (discuss each method separately).
2. Use one of the methods in Part 1. above to solve the problem. A combination of two or all three of the methods may also be used, if you believe this is better (regardless of whether you use one or multiple methods, this must be clearly motivated). Calculate and report the method's performance on the training, validation, and test data. Does the performance differ between the different sets? If yes, does this surprise you (explain why or why not)?
3. How important is the choice of image resolution for the methods you are using in Part 2. above?

Question 2  
Use deep learning to perform image classification (tumor detection). Specifically, you must:

1. Discuss why convolutional neural networks could be an appropriate type of model architecture to use for this task. [[2]](#footnote-2)
2. Train CNNs to solve the problem and, as a minimum, present the final results from one trained CNN model and one trained spatial transformer model. Here you must explicitly:
   1. Consider and discuss alternative CNN-model architectures including spatial transformer networks.
   2. Discuss different optimization methods and motivate your final choice. Use visualizations to show the relative performance of the optimizers.
   3. Visualize how regularization (such as dropout, weight regularization, or early stopping) impacts the training of your model. Here, be sure to visualize plots of train and validation losses and accuracies both with and without the use of regularization. Discuss regularization and its relation to overfitting.
   4. Visualize how data augmentation impacts the training of your model. Here, be sure to visualize plots of train and validation losses and accuracies both with and without the use of data augmentation. Discuss data augmentation and its relation to overfitting.
   5. Discuss and apply transfer learning. Motivate what type of transfer learning you use and how you apply it, including considerations for how to prepare the data for this. Here, be sure to visualize plots of train and validation losses and accuracies.
   6. Following up on the points made by [Tan and Le (2020, EfficientNet)](https://arxiv.org/pdf/1905.11946) investigate and discuss how important image resolution is relative to network width and network depth.
3. Having run the experiments above, select your preferred models (motivate why it is your preferred models). Calculate and report their performance on the test data.

Question 3  
To solve this problem, restrict attention to your preferred CNN model and your preferred spatial transformer network (also if they do not perform very well) from Question 2. The objective of this question is to get a better understanding of your models and to investigate if the model(s) potentially actually can assist the pathologist in localizing the tumor(s) in the image. Specifically, you must:

1. Discuss visualizations of activations, in particular, heatmaps of the activations. Illustrate examples of the heatmaps based on you preferred CNN model. Does these heatmaps appear to be useful for localizing tumors (as you are not required to know how to visually identify a tumor this is just to verify if the heatmap gets more excited in images that are labelled as having a tumor)
2. Based on your preferred spatial transformer network, perform a visualization of the spatial transformations. Based on your possibly very limited domain knowledge of identifying tumors visually discuss - in your role as a data scientist - if the “visualizations” seem to be able to localize tumor(s) in the image.

**General hints for the exam**

This section provides a list of “best practices” for answering exams – not specific to this one, but in general.

1. Be sure to explicitly answer everything that is asked. This sounds obvious, but you may have missed something! Be very critical here – read carefully through the exam and be absolutely sure you have answered every question, discussed what needs discussing, and so forth. Also carefully study the front page and its list of requirements!
2. Make your answers as short and precise as possible.
3. Stay on topic! You are welcome to discuss topics further than what is explicitly asked for, *if it is relevant*. Do not start discussing unrelated topics! If a specific part of the curriculum is not asked for in an exam, it does not improve your exam if you start discussing it.
4. The objective of the exam is *not* to get the highest test performance, but rather to show you have understood the different concepts you are asked to discuss and use. I expect *reasonable* values (both with respect to parameters and performance), not optimal values!
   1. To expand, this means that while it is often a very good idea to search to some extent for the best parameters of your models, I am not interested in seeing a test of thousands of values.

**Loading data (important!)**

As already mentioned above I will provide hints/instructions on how to download and prepare the data. I have posted these on the course site’s itslearning platform under the plan named: “**Exam: Instructions/hints on how to load data**”. It is important that you take a look at the notebooks.

1. Even if you do not perform one or more of these steps, motivate why you choose not to. [↑](#footnote-ref-1)
2. If you disagree, you need to motivate why but still use CNNs to solve the subsequent questions. [↑](#footnote-ref-2)