MLP Coursework 3
Baseline Approaches for Breast Cancer Image Classification

Group G102 (s1722201 & s1640380 & s1778742)

Abstract

This coursework presents a first approximation to the breast tumor tissue image classification task. We implemented and compared the performance of one classical machine learning method, Support Vector Machines, and two deep neural network architectures: LeNet, and GoogLeNet. The experiments were conducted over the BreakHis dataset, we discovered that the best performing baseline was GoogLeNet, which achieved 83% validation accuracy on average. For evaluating the performance of these experiments we used accuracy, log loss, and cross-entropy metrics. We conclude that obtaining suitably high classification accuracy of microscopic images of breast tumor tissue is a challenging and an ongoing open research problem. However, this coursework has set the foundation for the direction of investigation to follow in the next coursework.

1. Introduction

Deep learning models are a subset of machine learning algorithms based on learning accurate representations from data, these type of algorithms have made impressive progress in solving a wide range of tasks. From speech recognition to computer vision and natural language processing, deep learning methods have become crucial to achieving state-of-the-art results for many open problems in many different areas, therefore applying these principles to new fields can help researchers to approach to a wide range of open and challenging problems.

According to figures from the World Health Organization (WHO), among the most common diseases in human beings, cancer remains the second highest cause of mortality in the world. For instance, this condition was responsible for 8.8 millions of deaths worldwide in 2015. Interestingly, cancer is one of the diseases that has shown accelerated exponential growth. For example, according to (Ferlay et al., 2015), in 2012 this condition presented 14 million new cases. Indeed, cancer remains a serious issue for the public health in our modern society. However, the impact of cancer is not only from a public health perspective, cancer has serious economic consequences, for example, according to (Stewart, 2014), the monetary damage derived from cancer can be estimated at almost US$ 1.16 trillion.

Although the medical community has achieved significant progress in the development of medical treatments and early detection methods, cancer remains a significant problem from an economic, health, and social standpoint. It is well known that currently, the best method for treating cancer is early detection.

Given that there is a significant interest in the medical community in developing methods that allow the early detection of cancer, we decided to experiment with the use of deep neural networks to classify a biopsy microscope image as malignant or benign. The development of this type of technology can help doctors to reduce the time in which a diagnosis of cancer can be made, allowing early treatment, which would presumably result in fewer medical complications or deaths. To understand how medical image analysis works, and how deep neural architectures can help to enhance current cancer detection systems, we decided to use a breast cancer image dataset. It is also an area where simple human error can have grave consequences, being able to abstract away from human errors made due to tiredness or distraction or the lack of experts in some geographical regions using a model would be desirable.

2. Research Questions

Researchers and medics often classify these tissue sample images by looking at the overall structure of an image with a microscope and zooming in and out over critical areas of the tissue. It is, therefore, possible that the analysis of these images could be enhanced by the use of deep neural network systems, particularly convolutional neural nets. However, this class of systems can benefit from suitable pre-processing of the data in such a way that the critical structure required to classify a tissue image becomes increasingly evident to the system.

The central research question addressed for this coursework is: " *How do Support Vector Machines and Convolutional Neural Network based architectures perform on a classification task on the Breast Cancer Histopathological Image Classification (BreakHis) dataset?*". Based on the previous coursework results, and inspired by the research work and dataset of (Spanhol et al., 2016a), we tried to answer the previous question by contrasting the performance of three baseline image classification architectures, a baseline Support Vector Machine, LeNet and finally GoogLeNet. We contrasted the performance of these deep neural networks and found that the problem is challenging, and it requires more time for experimenting with different configurations,

features, and infrastructure.

This coursework presents a first approximation for the breast cancer image classification task, which consisted of developing and testing deep learning baseline systems. The second part of this report formalizes the problem and narrows the scope of this coursework. Additionally, it describes the dataset used for this assignment. The third section describes the dataset. The 5th and 6th sections present the methodology followed in this coursework as well as a summary of the baseline systems design and implementation and a comparison of the results of the experiments. Finally, section 6 outlines plans for the next phase of experimentation and section 7 reports final conclusions.

Going forward, our research question will shift to "*Which pre-processing steps can improve the performance of these models and what does this tell us about the nature of the problem at hand?*".

3. Dataset and Task Description

The main goal and motivation for this coursework is to take a step further and study the behavior and implementation of deep neural network architectures with a new, inherently challenging dataset. Coursework 2 provided the basis to speculate that the use of convolutional neural networks could be a successful approach to classify tissue images into malignant or benign. Thus, we studied and tested this hypothesis over this interesting real dataset.

3.1. The BreakHis Dataset

The Breast Cancer Histopathological Image Classification (BreakHis) dataset compiled by the research work of (Spanhol et al., 2016a), contains 9109 images of breast tumor tissue from 81 patients. The images are of different magnifications of 40X, 100X, 200X and 400X. There are approximately 31% benign samples and 69% malignant samples.

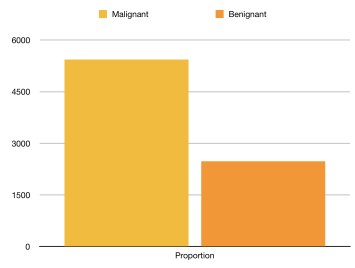
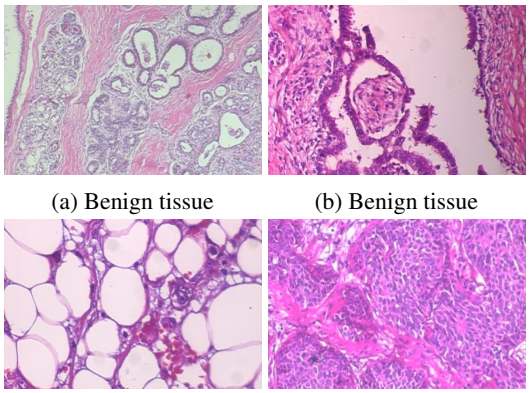


Figure 1. The proportion of images of the BreakHis dataset.

The samples of the BreakHis dataset were curated by (Spanhol et al., 2016a), and are composed by microscopic images of breast tumor tissue. Interestingly, according to the authors, the image samples of the dataset used in this coursework were collected by the SOB method or named partial mastectomy or excisional biopsy. This method removes the larger size of a tissue sample and is done in a hospital with a general anesthetic. Figure 1 shows some example images for this dataset (i.e., benignant and malignant tissues).



(a) Benign tissue (b) Benign tissue

(c) Malignant tissue (d) Malignant tissue

Figure 2. Example images of both Benign and Malignant tissue, image taken from (Spanhol et al., 2016a)

Each image has several associated labels containing information about the sample such as the method of biopsy procedure, patient id, tumor class, tumor type, and magnification factor. However for this task the label of interest is the tumor class of malignant or benign. There are images for the following sub-classes of tumor type:

1. Benign: Adenosis, Fibroadenoma, Tubular Adenoma or Phyllodes Tumor

2. Malignant: Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma or Papillary Carcinoma.

3.2. The Breast Cancer Image Classification Task

Among the most relevant research work that evaluates machine learning methods for image breast cancer classification is (Spanhol et al., 2016b), the authors performed a study and comparison of traditional machine learning. The authors found that traditional feature extractors like local binary patterns (LBP), completed local binary patterns (CLBP) (Guo et al., 2010), and (ORB) (Hamilton et al., 2007) had good performance with random forests and support vector machines. However, in matters of accuracy, one drawback of these traditional techniques is the fact that they are prone to confuse benign tissues with malignant tissues. For example, the authors of the above research mentioned that some benign tumors such as fibroadenoma were classified as malignant.

The research of (Spanhol et al., 2016b) mentions that histopathologic image classification is a challenging task, mainly because images present high variance, rich and complex geometrical structures. Contrary to traditional machine learning techniques which rely on the engineering of handmade features, one of the many advantages offered by deep learning methods is the ability to abstract representations with the most critical information from any dataset without any feature extraction step.

Among the most successful state of the art techniques, the usage convolutional neural networks (CNN) is gaining more popularity in the medical image analysis community. There are several examples of these applications, for instance in (Ciresan et al., 2012) researchers provided a deep learning method which performs automatic segmentation of neural structures depicted in stacks of electron microscopy images. Another example is (Cruz-Roa et al., 2013), where the authors presented a full cell carcinoma cancer image analysis pipeline that performs image representation and learning, classification, and results interpretability with 91.4% of accuracy.

For this coursework, we narrowed this task as predicting if a new tissue slide is malignant or benign. The images have not been normalized nor color standardized at this point. The split of Training to Validation to Test set is 70:20:10 with test set images chosen randomly from each and therefore there having a representative split with regards to both tumor type and Malignant/Benign, with various resolutions and parts of the tissue examined.

3.3. Data Set Preprocessing

We approached this problem as a supervised binary classification task. For the initial part of the experiment set up, there was no pre-processing involved apart from splitting the images into Training, Validation, and Test Sets. We created a script that loads the images as a scaled numpy matrix, and given the directory or location (i.e., Benignant or Malignant) of the sample, it automatically labels each image sample or row vector with 0 or 1. Additionally, the dataset was split based on patient ID, by doing so, we ensure that the samples are mutually exclusive. In other words, no samples from one patients are repeated across the training, testing, and validation set. The Convolutional Neural Networks were run over the above matrix of pixel values. This setting allowed us to explore the behavior the models with raw data, although the pixel values were divided by 225 in order to move values to the range 0 to 1 and avoid issues with numerical overflow which can otherwise arise, for example in the non-linear activation layer of a neural network.

4. Methodology and Experiment Setup

For implementing our deep baseline models and experiments we used MATLAB Neural Networks toolbox for fast prototyping, and Python with Keras (Chollet et al., 2015) a higher level interface of TensorFlow (Abadi et al., 2016), which allowed us to experiment quickly and design our image classification baselines. As we were working with high-resolution images and with a relatively large dataset, from coursework 2 we learned that often, deep learning models require time for training. Therefore, for reducing the computational complexity or cost, we configure Keras with Tensorflow-GPU 1.4 back-end over a GTX 970m GPU.

As mentioned previously, our hypothesis was that based on Coursework 2, CNNs provide a significantly high accuracy compared to feed forward neural networks and other simpler neural models in some image tasks. Additionally, we considered the research work of (Krizhevsky et al., 2012a), where the authors trained a CNN at scale with the LSVRC-2010 ImageNet dataset. More specifically, they stacked 5 convolutional layers achieving error rates of 39.7% and 18.9%. Thus, we speculated that deeper layers can perform well over a challenging real life data as the BreakHis dataset.

5. Baseline Experiments and Results

In this section we present the experiments and the results for our three proposed baseline models. We discover that images with full resolution were very expensive to run, therefore we re-sized all input data to 224x224 for all our experiments.

5.1. SVM + Keypoint Detection and Description

According to (Wu et al., 2008), support vector machines (SVMs) offer one of the most robust and accurate methods among all well-known algorithms. This algorithm has a good theoretical foundation and it requires relatively small amounts of examples for training, and is insensitive very sparse datasets. This algorithm operates by finding the maximum margin hyper plane that best splits the input data.

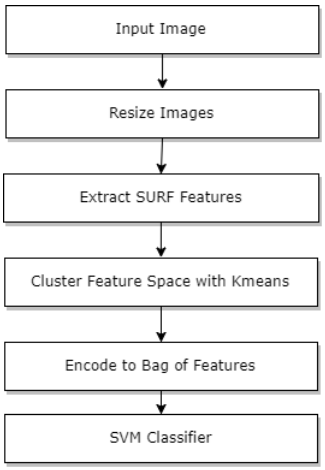


Figure 3. A description of our SVM image classification pipeline.

Figure 3, shows the configuration of our first baseline experiment. Then we extracted SURF features and performed cluster feature space with k-means. According to (Bay et al., 2006), SURF is a scale and rotation invariant interest feature extractor, that operates by fixing a reproducible orientation by taking into account the circular region of a point of interest, this reduces the effect of photometric changes. Finally, we encode the feature matrix using K-means clustering algorithm (Jain, 2010) to bag of features and then we fit and evaluate the SVM model.

The average performance of our model had 72% validation accuracy overall. In particular our model predicted benign

and malignant images of breast tumor tissue with 83% and 64% accuracy respectively.

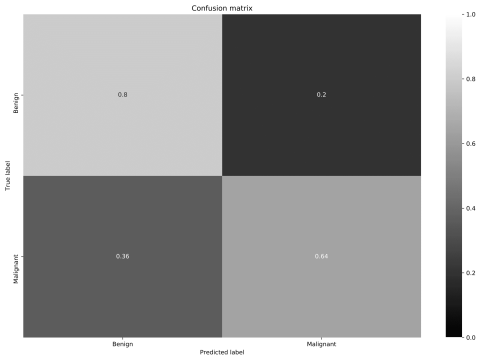


Figure 4. SVM + Local Descriptor confusion matrix

Figure 4 presents the confusion matrix for the SVM + Local Descriptor baseline. As we can see in the above image, this baseline system specifically has difficulty in correctly identifying images malignant tissue as the precision and recall are significantly worse for Malignant images than for Benign images.

Run	Training Accuracy	Training Loss
1	0.7219	0.7728
2	0.7219	0.7728
AVG (STD)	0.7219	0.7728

Table 1. SVM Results for 2 runs to show same results will be obtained

Additionally, we ran one hundred iterations of k-means, and selected the cluster with the best performing validation accuracy, this allowed us to ensure that we are using the best encoding features.

5.2. LeNet Baseline

Based on coursework2, convolutional neural network performed significantly well on image classification tasks. As a result, we propose LeNet as a baseline classifier. According to (LeCun et al., 1998), the LeNet architecture can be summarized as a CNN with 7 layers without considering the input. It contains a set of trainable weights.

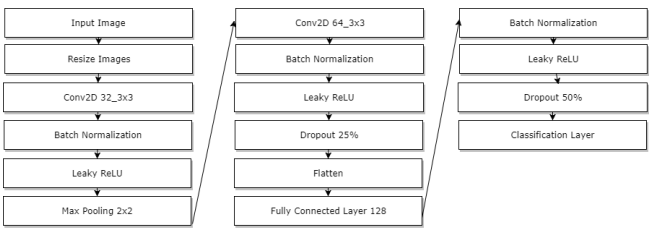


Figure 5. A description of LeNet architecture.

Figure 5, describes the LeNet architecture that we use for this baseline experiment. As we can see, we resized the input image and then we presented it to a convolutional layer of two dimensions, Leaky Relu was used as activation function, and the drop out rate was 25% and 50% respectively, finally it is passed to the classification layer.

As an indicator of the difficulty of this task for CNNs, the LeNet architecture as proposed by (LeCun et al., 1998) is not training in any meaningful way. As illustrated by figure 6, the training accuracy does not appear to improve over time. We suspect that diminishing gradient could be the problem. we captured the weights of the first convolutional layer form epoch 9 and compare it with weights of the same layer in epoch 10. The differences can be seen in figure 7.

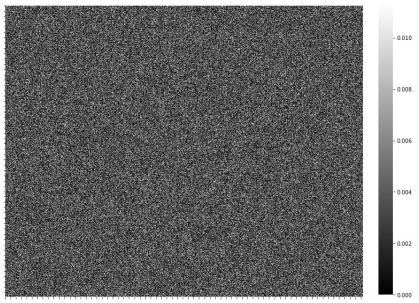


Figure 6. Heatmap comparing weights of first convolutional layer between epoch 9 and epoch 10 of LeNet.

The above figure shows that the weights of the first convolutional layer is changing, indicating that the model is learning. Therefore, the relatively low training accuracy of 71% led to the conclusion that this network lacks sufficient abstraction power to capture the complex relationships involved. The representation power of a network can be added to with more layers as seen in coursework 2. For this reason, the behaviour and performance of GoogleNet was examined in the next section.

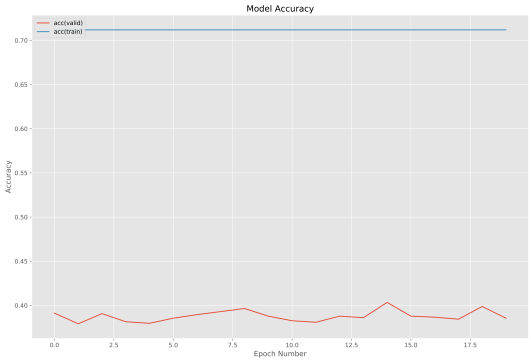


Figure 7. LeNet validation curve for the breast cancer image classification task, scoring accuracy.

We iterate with 20 epochs, and the configuration for LeNet baseline was: a learning rate of 0.0001 with ADAM optimizer. As a loss function we used cross entropy and the parameters β_1, β_2, and ϵ where fixed as 0.99, 0.9, 1e-08

respectively, with a decay of 0.0.

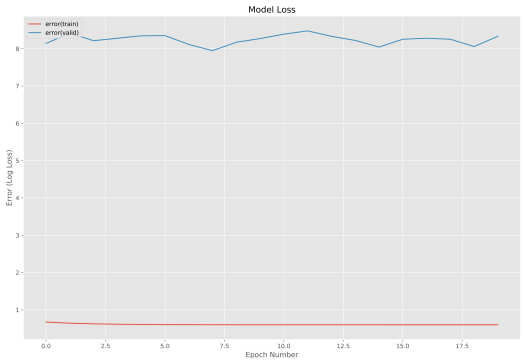


Figure 8. LeNet validation curve for the breast cancer image classification task, scoring loss.

Run	Training Accuracy	Training Loss
1	0.7119	0.4317
2	0.7119	0.4334
3	0.7119	0.4276
4	0.7119	0.4034
5	0.7119	0.4023
AVG (STD)	0.7119	0.4197
Mean	0	0.0155

Table 2. LeNet Results for each of 5 runs

Clearly, this baseline is performing very poorly on the Validation Set, especially as this class is binary classification. As the attained Training Accuracy is relatively low this suggests the model is unable to fully capture the complexity of the training data, a more powerful model is potentially needed. The plots show an inability to train.

5.3. GoogleNet

In order to examine the performance of a deeper and more powerful network, GoogLeNet was implemented. GoogLeNet achieved state-of-the-art results in the ImageNet Large-Scale Visual Recognition Challenge in 2014 by leveraging the idea of "Inception Modules" which focuses on efficiency. It has 22 layers thus assumably offering increased representative power over LeNet, while focusing on efficiency in terms of both memory and the number of parameters to be trained.

In proposing GoogLeNet (Szegedy et al., 2015) built on the previous work of (Serre et al., 2007) in stacking layers of Gabor Filters in which all filters are learned during training and combining this idea with the 1x1 convolutional layers used by (Lin et al., 2013). In this model the 1x1 convolutional layers serve as essentially a dimensionality reduction technique. This results in significant performance improvement as claimed by previosuly mentioned research work. This advantage led us to choose googLeNet over other architechture such as VGGnet (Simonyan & Zisserman, 2014) or AlexNet (Krizhevsky et al., 2012b).

The use of Gabor Filters was also a factor in choosing to implement this architecture as Gabor Filters are linear filters frequently used in texture analysis. It is possible that this vision problem can be framed as a texture problem.

The Inception Module is structured as follows using these 1x1 convolutional layers:

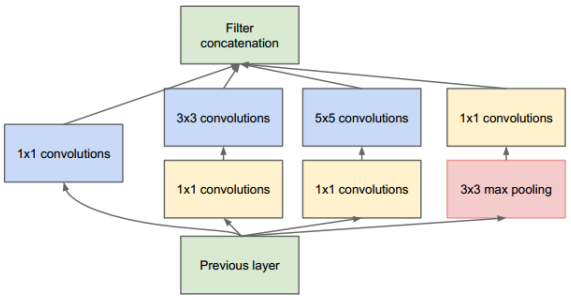


Figure 9. The inception module architecture presented in (Szegedy et al., 2015).

The results obtained were superior to the performance of LeNet as expected:

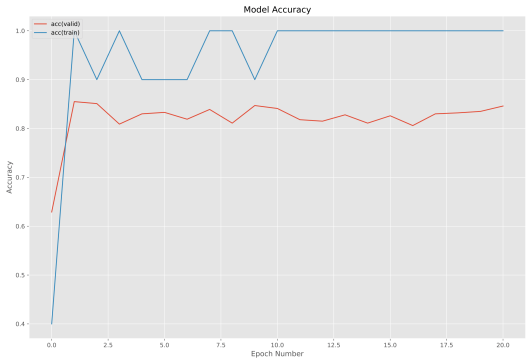


Figure 10. GoogLeNet plot for the breast cancer image classification task, scoring accuracy.

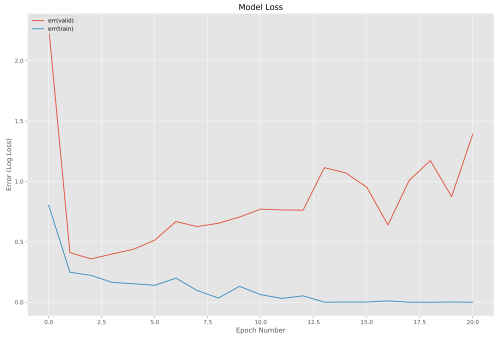


Figure 11. GoogLeNet plot for the breast cancer image classification task, scoring loss.

Interestingly, a comparison of the Confusion Matrices for the SVM and GoogLeNet show that the models are excelling at different classes with the SVM performing best on Benign images and the GoogLeNet performing best on the Malignant images.

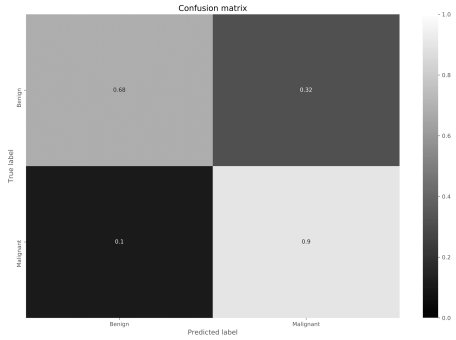


Figure 12. GoogLeNet confusion matrix

The configuration for GoogleNet baseline was a learning rate of 0.001, with stochastic gradient descent as an optimizer, we used as loss function cross entropy, we iterated 20 epochs.

Run	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1	1.0	0	0.8450	1.5758
2	1.0	0.0246	0.8369	0.0246
3	1.0	0.0006	0.8444	1.5939
4	1.0	0.0002	0.8236	1.6188
5	1.0	0.0050	0.8173	1.8161
AVG (STD)	1.0	0.0050 (0.004)	0.8334 (0.005)	1.3258 (0.328)

Table 3. GoogleNet Results for each of 5 Runs

As expected, GoogleNet outperformed LeNet considerably with an average training accuracy of 100% with a standard deviation of 0% and validation accuracy of 83% with a standard deviation of 0.5% over 5 runs. The increased training accuracy in particular indicates that this architecture was better able to capture the behaviour of the training data. Clearly, further increasing the validation accuracy will be the goal going forward as a 17% error rate would not be suitable for any application with the high-stakes involved in Cancer detection.

6. Future Work Planning

Following the results of these experiments, the next steps to be examined are heavily focused on pre-processing the data. Attempting to read some literature on the identification of malignant cancer from such images, it would seem that the focus is often on cell shape and the local structure such as irregular clumps of cells. Additionally, the different levels of color are due to the specific dye used in the process. Therefore the following pre-processing steps seem promising:

1. Data standardization
2. Color normalization
3. Localization (Detection and Segmentation)
4. Data Augmentation (flipping, rotation, dilation)
5. Experimenting with the usefulness of each magnification level

However, localization may be out of scope to implement in such a short time due to lack of labeled ground-truth data. We will attempt to explore the use of traditional methods from Computer Vision such as color thresholding, time permitting.

7. Conclusions

This coursework presented a first approximation to the breast image cancer classification dataset (BreakHis), we found that classifying tissue images for detecting benignant or malignant precence of cells is challenging. To summarize the main outcomes of this coursework were:

- We experimented with a completely different, real and new dataset.
- We created several baseline models to evaluate the performance of deeper neural network architectures.
- We analyzed and investigated how typical medical approaches tumors on slide microscopic images. This let us study and model the breast cancer image classification task more accurately.
- We learned to use Deep Learning libraries in two different programming languages, like Python and MATLAB. In the Python side we discovered Keras, which allowed us to implement production level deep learning pipelines. On the MATLAB side we learned to use the Neural Networks Tool Box, for quick deployment.
- We successfully implemented three image classification pipelines, SVM + Key Point Detection Description, LeNet, and GoogleNet.

In the next coursework we will try to enhance the performance of our classification pipeline with data preprocessing and augmentation techniques,in light of the research presented in (Wang & Perez, 2017), we are planning to improve the results presented in this coursework. Regarding the preprocessing, we will try to segmentate images and rearrange training data based on different magnification.

References

Abadi, Martín, Barham, Paul, Chen, Jianmin, Chen, Zhifeng, Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Irving, Geoffrey, Isard, Michael, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Bay, Herbert, Tuytelaars, Tinne, and Van Gool, Luc. Surf: Speeded up robust features. In *European conference on computer vision*, pp. 404–417. Springer, 2006.

Chollet, François et al. Keras, 2015.

Ciresan, Dan, Giusti, Alessandro, Gambardella, Luca M, and Schmidhuber, Jürgen. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pp. 2843–2851, 2012.

Cruz-Roa, Angel Alfonso, Ovalle, John Edison Arevalo, Madabhushi, Anant, and Osorio, Fabio Augusto González. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 403–410. Springer, 2013.

Ferlay, Jacques, Soerjomataram, Isabelle, Dikshit, Rajesh, Eser, Sultan, Mathers, Colin, Rebelo, Marise, Parkin, Donald Maxwell, Forman, David, and Bray, Freddie. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5), 2015.

Guo, Zhenhua, Zhang, Lei, and Zhang, David. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

Hamilton, Nicholas A, Pantelic, Radosav S, Hanson, Kelly, and Teasdale, Rohan D. Fast automated cell phenotype image classification. *BMC bioinformatics*, 8(1):110, 2007.

Jain, Anil K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012a.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012b.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lin, Min, Chen, Qiang, and Yan, Shuicheng. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Serre, Thomas, Wolf, Lior, Bileschi, Stanley, Riesenhuber, Maximilian, and Poggio, Tomaso. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29 (3):411–426, 2007.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. corr abs/1409.1556 (2014). *arxiv. org/abs/1409.1556*, 2014.

Spanhol, Fabio A, Oliveira, Luiz S, Petitjean, Caroline, and Heutte, Laurent. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016a.

Spanhol, Fabio Alexandre, Oliveira, Luiz S, Petitjean, Caroline, and Heutte, Laurent. Breast cancer histopathological image classification using convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 2560–2567. IEEE, 2016b.

Stewart, Bernhard W. Wild cp, editors. *World Cancer Report*, 2014.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, Rabinovich, Andrew, et al. Going deeper with convolutions. Cvpr, 2015.

Wang, Jason and Perez, Luis. The effectiveness of data augmentation in image classification using deep learning. Technical report, Technical report, 2017.

Wu, Xindong, Kumar, Vipin, Quinlan, J Ross, Ghosh, Joydeep, Yang, Qiang, Motoda, Hiroshi, McLachlan, Geoffrey J, Ng, Angus, Liu, Bing, Philip, S Yu, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.