UPPSALA
UNIVERSITET

UPTEC F 17039
Examensarbete 30 hp
Juli 2017

Deep Convolutional Neural Networks For Detecting Cellular Changes Due To Malignancy

Håkan Wieslander
Gustav Forslid

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Deep Convolutional Neural Networks For Detecting Cellular Changes Due To Malignancy

Håkan Wieslander, Gustav Forslid

Discovering cancer at an early stage is an effective way to increase the chance of survival. However, since most screening processes are done manually it is time inefficient and thus costly. One way of automizing the screening process could be to classify cells using Convolutional Neural Networks. Convolutional Neural Networks have been proven to produce high accuracy for image classification tasks. This thesis investigates if Convolutional Neural Networks can be used as a tool to detect cellular changes due to malignancy in the oral cavity and uterine cervix. Two datasets containing oral cells and two datasets containing cervical cells were used. The cells were divided into normal and abnormal cells for a binary classification. The performance was evaluated for two different network architectures, ResNet and VGG. For the oral datasets the accuracy varied between 78-82% correctly classified cells depending on the dataset and network. For the cervical datasets the accuracy varied between 84-86% correctly classified cells depending on the dataset and network. These results indicates a high potential for classifying abnormalities for oral and cervical cells. ResNet was shown to be the preferable network, with a higher accuracy and a smaller standard deviation.

Handledare: Sajith Kecheril Sadanandan, Ewert Bengtsson
Ämnesgranskare: Carolina Wählby
Examinator: Tomas Nyberg
ISSN: 1401-5757, UPTEC F 17039

Populärvetenskaplig sammanfattning

Cancer är en av de främsta anledningarna för dödlighet globalt. Att upptäcka cancern i tid är viktigt för att minska dödligheten och är anledningen till varför ett väl fungerande screeningprogram är viktigt. Information, screeningprogram och resurser för lämplig medicinsk vård är ofta bristfällig i utvecklingsländer. Denna brist kan vara anledningen till att 70% av all dödlighet på grund av cancer sker just i utvecklingsländer. Ett sätt att effektivisera screeningprocessen är att automatisera den genom att införa en datoriserad screening. Detta skulle bidra till att läkare enbart behöver undersöka en del celler från varje patient, vilket gör screeningen snabbare och därmed billigare. Ett sätt att utföra datoriseringen är att använda Convolutional Neural Networks (CNN). CNNs har visat ge hög noggrannhet inom bildklassificering och har de senaste åren vunnit de flesta bildklassificeringstävlingarna.

Två vanliga cancerformer är livmoderhalscancer och oralcancer. 2012 rapporterades 530'000 nya fall av livmoderhalscancer, vilket var 8% av alla rapporterade fall av cancer för kvinnor. Riskfaktorerna inkluderar bland annat rökning och infektion av HPV, som är ett sexuellt överförbart virus. Utvecklandet till livmoderhalscancer är en långsam process och för kvinnor med normalt immunförsvar kan det ta mellan 15 och 20 år. För kvinnor med nedsatt immunförsvar kan det ta mellan 5 och 10 år. Att upptäcka cancern i tid är ett av de viktigaste stegen för att minska dödligheten. Det har visats att tidig behandling av cancern kan rädda upp till 80% i jämförelse med dagens överlevnad vilken är 48%. I välutvecklade länder finns det screeningprogram och vaccination mot livmoderhalscancer. Screening görs med hjälp av skrapprov som tas från livmoderhalsen. Dessa prov undersöks manuellt för att se om det finns onormala celler som kan visa tecken på cancer.

Oralcancer definieras av cancer i munnen, vilket inkluderar munhålan, läppen och svalget. 2012 rapporterades det 529'000 nya fall runt hela världen. Riskfaktorerna är tobak och alkohol vilka tillsammans bidrar till 65% av alla nya fall. Som för livmoderhalscancer är infektion av HPV även en riskfaktor. I dagsläget finns det inget screeningprogram för oralcancer trots att det är en snabbt växande cancerform. För att upptäcka cancern måste det upprättas en misstanke av patienten eller av en läkare. Detta bidrar till att cancern ofta upptäcks i ett sent stadie. Om en misstanke om cancer finns kan läkaren välja att göra skrapprov i munnen för att undersöka om där finns onormala celler. För att säkerställa diagnosen måste en biopsi göras, vilket är ett vävnadsprov.

CNNs är en gren inom maskininlärning. De är konstruerade för att immitera hur människor och djur uppfattar bilder och objekt. Neuroner i människor och djurs syncentrum svarar mot ett specifikt område i synfältet. CNNs är uppbyggda av lager innehållande neuroner. Neuronerna sitter ihopkopplade i olika strukturer och består av vikter. I jämförelse med den mänskliga hjärnan måste nätverken tränas för att kunna skilja på bilder. I träningsfasen ändras vikterna i neuronerna vilket gör att nätverken blir bättre på att klassificera bilder. Träningsfasen består av två steg, först får nätverket se ett antal bilder och försöka identifiera vad bilden föreställer. I det andra steget får nätverket reda på vad bilden verkligen föreställer och räknar ut hur stort felet var. Detta felet används sen för att uppdatera vikterna i nätverket. Nätverket lär sig då hitta mönster i bilderna och blir bättre på att urskilja de olika klasserna från varandra.

Det här projektet syftar till att undersöka om CNNs kan upptäcka skillnader

mellan onormala och normala celler från livmodern och munnen. Detta skulle i så fall kunna användas som ett verktyg för att automatisera screeningprocessen både för oralcancer och livmoderhalscancer.

Projektet har resulterat i ett nytt dataset innehållande prover tagna från munnen som kan användas för forskning inom oralcancer. Projektet har också visat att det finns potential i att klassificera celler med hjälp av CNN. För cellprover tagna både från livmoderhalsen och munnen visades det på att över 80% av alla celler klassificeras korrekt.

Contents

1 Introduction

1.1 Background

Cervical cancer was for a long time one of the deadliest cancer types for women. In some developed countries women have the opportunity to become vaccinated as a prevention. For those who do not become vaccinated it is recommended that they become screened every few years. The screening program has proven to be very effective since early discovery of the disease leads to early treatment. The world health organization state that early treatment can prevent up to 80% of the mortality in the countries were women become screened [1]. The screening procedure is performed manually by a cytotechnologist, which makes it a time inefficient and a costly procedure. This results in fewer women having the opportunity of becoming screened.

Oral cancer is one of the larger cancer types worldwide [2]. However, it does not have a screening program unlike some other cancer types. Without a screening program it is hard to discover the cancer in time. Since it is crucial to discover oral cancer in time, most people who are affected by it suffer either by morbidity or mortality.

Convolutional Neural Networks (CNN) is a type of deep learning network constructed to mimic the animal visual cortex [3]. CNNs have been proven to produce high accuracy for image classification tasks and has won the ImageNet Large Scale Visual Recognition Challenge (ILSVCR) the last years [4]. The architecture is built up by layers which creates a volume with a width, height and depth. The layers consists of neurons with learnable parameters. During the training these parameters are updated with the purpose of increasing the performance of the networks predictions. When training a CNN the neurons starts to identify different features in the image which in the end results in the networks prediction [5]. The biggest part of CNNs are the convolutional layers. In these layers the input is convolved with the neurons in the layer to produce an output. The capability of classifying images in a high rate can be applied to many different fields.

1.2 Purpose

The purpose of this master thesis is to investigate if deep convolutional networks can be used to detect cellular changes due to malignancy. This would provide a tool to speed up the process in screening programs for oral and cervical cancer and thus make it more accessible.

1.2.1 Goals

- Create a dataset containing samples taken from the oral cavity.
- Investigate potential for CNNs to detect cellular changes due to malignancy in oral cavity and uterine cervix.
- Compare performance between ResNet and VGG for classifying cellular changes.

1.3 Delimitations

This study is investigating two different types of CNNs: ResNet and VGG. This decision is made because of the time limitations for a master thesis. The reader of this report should know that there are many more types of CNNs that can be used for this task.

This study is focused on binary classification, distinguish between malignant and non malignant cells. This decision was made due to shortage of time for this thesis. An important aspect when diagnosing a patient is also to consider what actual diagnosis each cell has and not just distinguish between malignant and non malignant cells.

1.4 Thesis information

This thesis was carried out under the Division of Visual Information and Interaction at the Department of Information Technology, Uppsala University. Due to the size of the project it was suited for two people. The work load was divided up by the two participants. Håkan focused on evaluating CNNs performance on oral cancer and Gustav focused on evaluating CNNs performance on cervical cancer.

2 Theory

2.1 Cancer

Cancer is a universal medical term for a large amount of different diseases that can affect and spread through out the whole body. The overall attribute for all forms of cancer is that abnormal cell clusters start growing in a fast and uncontrolled way. These cell clusters may even spread into other organs, which is often the reason a person with cancer dies. In a global view cancer is one of the leading causes of mortality. Many of the leading risk factors that can lead to cancer are controlled by humans e.g tobacco use and alcohol consumption etc. It is also widely known that early detection for all types of cancer is crucial to lower the mortality rate. This is the reason why a well working screening program is needed for all types of cancer. Information regarding risk factors, well working screening program and resources for adequate medical treatment are lacking in many low- and middle income countries. This might be the reason 70% of the mortality due to cancer occurs in developing countries [6].

2.1.1 Oral Cancer

Oral cancer is a type of cancer located in the oral area including lip, oral cavity and pharynx. It is most common in developing countries and is worldwide the seventh most common cancer form. In 2012 529'000 new cases and 292'000 deaths, due to oral cancer, were reported world wide [2]. The risk factors are tobacco usage and alcohol consumption which accounts for 65% of all cases. Both cigarettes and smokeless tobacco poses a risk for oral cancer and the risk increases with smoking rate and duration [2]. Another major risk factor is infection by the human papillomavirus (HPV) which is a sexually transmitted infection [2]. To discover oral cancer a doctor will first perform an exam of the oral cavity for possible signs of cancer. If suspicion is raised then a specialist is contacted, often a surgeon specialized in the oral area.

If the specialist thinks it is necessary a biopsy is taken. To find areas of where to perform the biopsy, scrapes of suspicions areas can be examined. The scrapes are smeared out on a piece of glass which is then dyed and examined under a microscope. If there is any presence of abnormal cells a biopsy can be performed. The biopsy is either performed in the doctors office or on a operating table depending on where the area of interest is located. The samples from the biopsy are sent for lab testing where a pathologist determines the diagnosis of the patient [7].

2.1.2 Cervical Cancer

Cervical cancer is a malignant disease in the uterine cervix. In 2012 530'000 new cases were reported which stood for approximately 8% of all reported cancer cases for women [8]. The same year 270'000 deaths, due to cervical cancer, were reported and 85% of those were reported from developing countries [1]. Cervical cancer usually starts in the transformation zone, which is the zone where the endocervix and the exocervix meet. This zone changes with age which is one of the reasons why older women have a greater risk of developing cervical cancer. Alongside age, family history of cervical cancer is a risk factor. Other risk factors, that are more controllable, include smoking, multiple sexual partners and HPV [9] [10]. Most HPV infections clear up within 2 years, either by the immune system or with medical help, but some HPV infections stay and progress into cancer. The development into cancer is a slow process. For women with normal immune systems it takes 15 to 20 years and for women with weaker immune systems it only takes 5 to 10 years. Early detection is a crucial step to lower the risks of morbidity and/or mortality rates. It has been shown that early treatment can save up to 80% in comparison with todays survival rate which is 48% globally [1]. Many developed countries provide vaccination and screenings for women, but since both are costly this service is not given in many of the developing countries. This leads to the above shown statistics where the developing countries are overrepresented in the mortality due to cervical cancer. The screening system that is mostly used in the developed countries is called pap-smears screening. A pap-smear is a smeared out specimen taken from the cervix. These are then dyed and sent to cytologists that preform a manual screening to determine if there are any abnormal cells present. Most cytologists uses the Bethesda system when determining the cervical diagnosis. Below follows a list of the most common diagnosis for cells from the cervix [11].

- NILM - Negative for Intraepithelial Lesion or Malignancy
- LSIL - Low-grade Squamous Intraepithelial Lesion
- HSIL - High-grade Squamous Intraepithelial Lesion
- SCC - Squamous Cell Carcinomas
- Adenocarcinoma - Adenosquamous Carcinomas
- ASC-H - Atypical Squamous Cells, which cannot exclude a High-grade lesion
- ASC-US - Atypical Squamous Cells of Undetermined Significance

2.1.3 Malignancy Associated Changes

Malignancy associated changes (MAC) refers to small changes in the morphology and chromatin structure of the nucleus in a cell. The changes appear in normal looking cells located in the presence of tumor-associated areas. The occurrence of MAC are usually explained by two general theories. The first is that MAC arise from normal tissue adjacent to tumors that have been exposed to the same carcinogens that caused the tumor. The other theory is that tumor tissue releases soluble factors that affects neighboring cells[12]. The changes are too discrete to reliably be caught by the human eye, but with the help of computers and image analysis one can start to discover these changes. The advantage of looking at MAC is that not every single cell needs to be individually examined and classified, instead a statistical sample from the cell population needs to be evaluated. The drawback is that not all normal looking cells in presence of tumor-tissue need to have MAC and creating a dataset for classification is hard since these changes can not be caught visually [13].

2.2 Convolutional Neural Network

Convolutional neural network (CNN) is a type of deep learning network that is constructed to mimic the architecture of an animal visual cortex. All the individual neurons in the visual cortex respond to a specific overlapping region in the visible field. CNNs are built up with a spatial architecture where a specific region in one layer is connected to a specific region in the next layer [3]. It is constructed by neurons and the spatial architecture of a layer creates a volume of these neurons with a width, height and depth. The width and height determines the size of the neuron and the depth determines the number of neurons. The neurons can be seen as kernels that are built up the learnable parameters. This means that during training these parameters are updated with the purpose of increasing the performance of the network prediction. During the training the updated kernels will start to identify different features in different parts of the input image [5]. In contrast to the depth of a layer one can talk about the depth of a network. The depth of the network can be seen as the number of stacked layers in the whole network. One can argue that using data containing advanced features requires a deeper network. While increasing the depth of the network the amount of parameters increases rapidly, this may result in overfitting. Overfitting occurs when the network becomes good at classifying the training images, but does not classify unseen images well. To reduce the risk of overfitting it is crucial that the amount of data in the training set is large enough.

The architecture of a CNN can vary a lot. Depending on the usage the architect can choose endless of combinations of layers and constructing each layer in endless ways. The most common layers are: convolutional layers, pooling layers and fully connected layers as can be seen in Figure 1. Other examples can be ReLU layers, batch normalization layers and dropout layers.

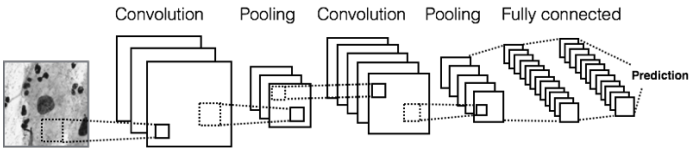


Figure 1: Example architecture of a simple Convolutional Neural Network

2.2.1 Convolutional Layer

The convolutional layer is the foundation of a CNN, where it executes the majority of the heavy computations. These layers have a spatial construction that is built up by filters with a width, height and depth. The width and height determines the size of the filter kernel and the depth is the number of kernels [5]. Each kernel is built up by learnable parameters. During the forward pass the kernels are convolved over the input image, where it preforms a dot product to produce the output as can be seen in Figure 2. During the backward pass the parameters are modified with the purpose of making the networks predictions better. These modification of the parameters leads to that the kernels will learn certain local features in the input image. The earlier the layer is located in the network the more primitive features it will learn to distinguish e.g vertical lines. The deeper the layer is located in the network the kernels will learn more advanced features. The convolutional layer have a size, a stride and a padding. The stride determines how many steps the kernel takes before performing a new dot product. The padding is used to control the size of the output from the layer and to control the boundary pixels [5].

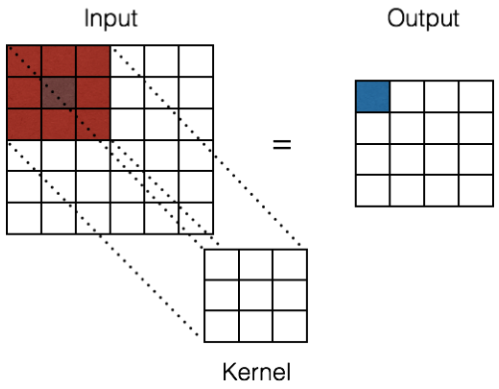


Figure 2: Convolving a kernel over the input image and producing the output image with a dot product.

2.2.2 Pooling Layer

Running a CNN demands a lot of computing power, by sequentially adding pooling layers in-between convolutional layers the need for computing power decreases. There are a few different types of pooling operations that can be used. The two most common ones are average pooling and max pooling. In practice a kernel is convolved over the input which produces an output depending on what type of pooling it is.

Average pooling calculates the average of the parameters as the output (Figure 3). Max pooling picks the maximum value of the parameters, inside the kernel, as the output (Figure 4). In practice pooling layers down-samples the input, which means that less computations are needed. For example using a kernel size of 2 by 2 and a stride of 2 the output size is reduced by 75%. By reducing the output size the CNN becomes less computationally demanding. Another advantage of using pooling layers is that the receptive field size increases after every pooling layer. This means that the following convolutional layers sees a larger region of the image. [5].

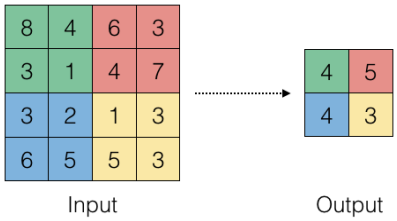


Figure 3: Average pooling

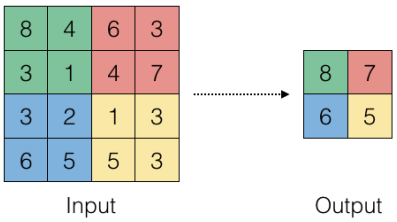


Figure 4: Max pooling

2.2.3 Rectified Linear Unit Layer

The rectified linear unit layer (ReLU) is a non-linear pixel by pixel operation. ReLU layers are systematically inserted after convolutional and fully connected layers in CNNs. The main task for a ReLU layer is to insert non-linearity in the network. Non-linearity has to be inserted since almost all real world data is non-linear. The operation that a ReLU layer preforms is

$$Output = max(0, Input) \tag{1}$$

Each pixel value is compared with 0 and the maximum of these is chosen. This means that a ReLU layer applies an element wise activation function on the input due to thresholding at zero [5]. ReLU is the most commonly used activation function for CNNs, mainly because it is not computationally heavy [14]. A reason for this is that both in the forward and backward pass it only uses an if statement. This advantage is important for deeper networks where the computational load is already large. Another advantage is that ReLU layers does not saturate the output like other activation functions.

2.2.4 Dropout Layer

Dropout layers are added to prevent overfitting in the network. This layers are often added in between the FC layers in the end of the network [15]. The dropout layers takes care of co-adaptions of features by breaking these during training. Co-adaptions are often created when a network is trained with too few images, multiple neurons can then learn the exact same features. This creates a fragile network with structures that only work for the training images. The break up creates dropped out units that can not impact the network during that specific training step, hence the name dropout layer. The dropout layers determine which units to break with a set probability. This break up of the units is only applied during training, which creates multiple thinner networks. This creates a stochastic regularization of the network and provides a good model averaging [16].

2.2.5 Batch Normalization Layer

During the training of a network the input values are modified by the weights to make the network preform better. Since the input of each layer is effected by the parameters of the previous layer, small changes might amplify through the network. This phenomena is called internal covariate shift [17]. This problem can be solved with batch normalization layers. These layers are often added after convolutional layers and before ReLU layers. Batch normalization layers performs a normalization, shifted to zero-mean and unit variance, in mini-batches during the training. This does not only fix the internal covariate shift the network can be trained with a higher learning rate and a less careful initialization of the parameters. It has been shown that adding batch normalization layers creates a network that need 14 times fewer training step but still provides the same accuracy [18].

2.2.6 Fully Connected Layer

In the end of a CNN the network needs to preform some sort of reasoning, this is often done in the fully connected layers (FC). All FC layers neurons are connected with the adjacent layers neurons, hence the name fully connected. These connections breaks of the spatial structure and creates a one dimensional array of numbers [19]. This deconstruction of spatial structure makes it impossible to preform further convolutional operations, which makes the FC layers the last layers in a CNN. FC layers can be stacked on top of each other, but the last FC layer must only have as many neuron as the data have classes.[5]

2.2.7 Softmax

To obtain class scores from the network a softmax function is often used. The softmax function is defined as,

$$P(y = j|z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{j=0}^{k} e^{z_j^{(i)}}} \tag{2}$$

where

$$z_j = \sum_{i=0}^{n} w_{ij} x_i \tag{3}$$

and w is the weight vector and x the input i.e one training sample. The z vector is the output from layer i to j i.e. the last layer in the network. The softmax function computes the probability of sample x belonging to class j given the weights w [20].

2.2.8 Backpropagation

For the network to become better at predicting the input, the error between the predicted output and the desired output needs to be minimized. This is done by adjusting the weights and biases in the network by backpropagation. The first thing the network does is calculating the loss in its output compared to the input label. The loss is calculated with some form of loss function J for example cross entropy

$$J(\bar{w}) = \frac{1}{n}\sum_{i=0}^{n} E(t_i, o_i) \tag{4}$$

where $\bar{w}$ is a weight vector used to weigh class imbalance, t the target, o the probability output from softmax and

$$E(t_i, o_i) = -\sum_{m} t_i log(o_i) \tag{5}$$

To minimize the loss, the gradient of the loss with respect to each of the networks weights needs to be calculated. First the gradient of the loss with respect to the output layer weights is calculated with the chain rule. Consider the error with respect to some hidden weights from j to i as $\frac{\delta E}{\delta w_{ij}}$ then

$$\frac{\delta E}{\delta w_{ij}} = \frac{\delta E}{\delta z_i} \cdot \frac{\delta z_i}{\delta w_{ij}} \tag{6}$$

where

$$\frac{\delta E}{\delta z_i} = \sum_{k}^{nClasses} \frac{\delta E}{\delta x_k} \cdot \frac{\delta x_k}{\delta z_i} \tag{7}$$

and

$$\frac{\delta z_i}{\delta w_{ij}} = x_i \tag{8}$$

From (6) you get an expression of how a change in the weight will effect the error. z and x comes from the softmax function (2) and (3). Considering the total input of any prior layer from i to j as

$$z_j = \sum_{i}^{n} o_i w_{ji} + b \tag{9}$$

where o_i is the non-linear output form the activation function. For any prior layers the effect of the error with respect to the weights can then be expressed as

$$\frac{\delta E}{\delta w_{ji}} = \frac{\delta E}{\delta o_i} \cdot \frac{\delta o_i}{\delta z_i} \cdot \frac{\delta z_i}{\delta w_{ji}} \tag{10}$$

This expression can then be used to calculate how to update the weights for earlier layers by computing $\frac{\delta E}{\delta w_{ij}}$ while backpropagating. Using some form of gradient descent

method the weights can then be updated in such a way that optimizes the network, for example stochastic gradient descent [20][21].

$$\bar{w}_i := \bar{w}_i - \alpha \frac{\delta E}{\delta w_{ji}} \tag{11}$$

2.2.9 Adam optimization

Adam (Adaptive momentum estimation) is a method for stochastic optimization using first order derivatives. The method used two adaptive momentum estimates (m_t, v_t) which is calculated as

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\nabla_W J(W) \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)(\nabla_W J(W))^2 \end{aligned} \tag{12}$$

where $\beta_1, \beta_2 \in [0, 1)$ is controlling the exponential decay rates of the momentum parameters. The update of the parameters is then performed as

$$W_{t+1} = W_t - \alpha \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{13}$$

where α is the learning rate. A good default choice for $\beta_1, \beta_2, \epsilon$ is proven to be $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$[22].

2.3 Network Architectures

2.3.1 VGG

The VGG network architecture became public after the Visual Geometry Group from Oxford University used it in ImageNet ILSVRC-2014. Compared with all the other networks at that time the VGG network was a lot deeper. A deeper network leads to an increase in the the number of parameters. This problem was solved by using small convolutional filters with a size of 3 x 3, stride of 1 and padding of 1 for all convolutional layers. This design made sure that the convolutional layers did not down-sample the input. VGG contains clustered convolutional layers in groups of two or three. After each group the input is down-sampled with max pooling. At the end of the network there are three FC layers and a softmax layer.

By both placing first and second in the ImageNet ILSVRC-2014 competition, the VGG creators showed that deeper networks worked well for image classification [23].

2.3.2 ResNet

ResNet is a deep learning network created by Microsoft Research Asia in late 2015 and won the ILSVRC 2015 classification task. The main difference between regular CNNs and ResNet is that ResNet is based on residual learning. The idea with residual learning is that the input for earlier layers is made available deeper in the network. A building block, called shortcut connection, for a residual network is described as,

$$y = F(x_i, \{W_i\}) + x \tag{14}$$

where y is the output, x the input and W the weights. For a ordinary CNN you would simply have,

$$y = F(x_i, \{W_i\}) \tag{15}$$

Basically the residual building block is calculating the term that should be added to the original input rather then only computing the transformation from x to F. The function F in a residual block can be built up by multiple convolutional layers. The addition is done element-wise and thus x and F needs to be of same dimensions. If the dimensions are not equal for instance when changing the number of output channels a linear projection (W_s) is used on the input to match dimensions.

$$y = F(x_i, \{W_i\}) + W_s x \tag{16}$$

A shortcut connection in ResNet is typically constructed with two convolutional layers having 3 x 3 kernels with a stride of 1 and a padding of 1. This results in the same output dimension as input dimension which is required for the element-wise addition. When the number of output channels increases the output size halves. These building blocks are built up by two convolutional layers. The first having 3 x 3 kernels, stride of 2 and padding of 1. The second one has 1 x 1 kernels with a stride of 1. To match the dimension the input is convolved with a 1 x 1 convolution having a stride of 2 (Figure 5) [24].

(a) Shortcut connection. (b) Halving shortcut connection.

Figure 5: Shortcut connections used in ResNet

2.4 Data Processing

2.4.1 Laplacian as focus measure

Finding the image with the best focus is important when working with cell classification. The differences that distinguish a cancerous cell from a normal cell can be small and easily lost due to bad focus in the image. One measurement of focus is to use the Laplacian. The second order derivative expressed in a Laplacian is known for passing high frequencies which can be an indication of sharp edges in an image. The second order derivative expresses how the rate of change changes. A faster change indicates a steeper edge. The Laplacian can be expressed using the mask

$$L = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{17}$$

One measure for the focus of an image can be to look at the variance of the Laplacian. A low variance indicates that there are few sharp edges which indicates blurriness. A

higher variance indicates a sharper image with better focus. The variance is calculated as

$$Var(I) = \sum_{m}^{M} \sum_{n}^{N} \left(|L(m,n)| - \bar{L} \right)^2 \tag{18}$$

where $\bar{L}$ is the mean of the absolute values [25].

2.4.2 Normalization

If the pixel intensities in an input image have a large variety in range the network might become discriminate. This due to the difficulties for a network to modify its weights in a way that suits all ranges of pixel intensities. To reduce the risk of a discriminating network normalization is done. By normalizing the dataset one scales down the range of pixel intensities [26]. First each pixel in the image is subtracted with the mean value for all pixels. In the second step the whole image is divided by the standard deviation of the original image (19).

$$\text{Normalized Image} = \frac{(\text{Image} - \text{mean(Image)})}{\text{std(Image)}} \tag{19}$$

2.4.3 Augmentation

Augmentation is used to reduce overfitting by synthetically enlarge the dataset [14]. There are many different transformation methods that can be used for augmentation e.g translation, mirroring and rotation. The basic concept of augmenting an image is to manipulate the image in such a way that do not change the original label. Each augmented version of the original image will be seen as a different image, having the same label, by the network.

2.5 Model Evaluation

2.5.1 *K*-Fold Cross Validation

A way to evaluate the accuracy of a CNN is K-fold cross validation [27]. First the dataset is divided into a training and testing set. The training set is then divided into K equally large folds. One of the K folds is kept as validation data and the rest of the $K-1$ folds are used for training. This process is then repeated K times which creates K different datasets with different mixture of training data and different folds for validation (Figure 6). After performing training and validation on all K folds the average of the results on the test set is calculated.

Figure 6: K-Fold Cross Validation with $K = 5$.

2.5.2 Performance Estimation

When testing a trained CNN, the network is fed with unseen images that the network classifies. One can look at the network predictions in different aspects, depending on what the aim is. One way is to divide all predictions into four categories, which can be seen in Figure 7. The four different categories are described as: Correctly classified samples (true positive, tp), correctly classified samples that do not belong to the class (true negative, tn), samples that were incorrectly assigned to the class (false positive, fp) and samples that belongs to the class but were not correctly classified (false negative, fn) [28].

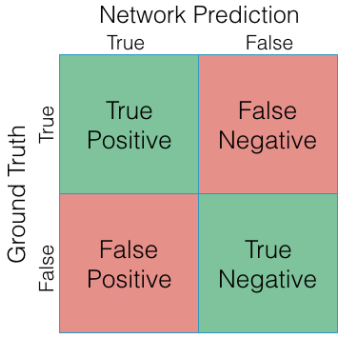


Figure 7: Ground truth vs. network prediction.

With these four classes one can now calculate four measurements on how well the networks predictions are. Average accuracy is an accuracy measurements of the prediction (20). Precision represents the rate of positive samples that are correctly classified as positive (21). Recall measures the rate positive samples are classified correctly (22). F score represents a harmonic mean of recall and precision (23). This measurement is high only when recall and precision are high.To get a high f score the network needs to have a low rate of false negatives and false positives [28].

$$\text{Average accuracy} = \frac{\sum_{k}^{NrClasses} \frac{tp_k + tn_k}{tp_k + tn_k + fp_k + fn_k}}{NrClasses} \tag{20}$$

$$\text{Precision} = \frac{\sum_{k}^{NrClasses} \frac{tp_k}{tp_k + fp_k}}{NrClasses} \tag{21}$$

$$\text{Recall} = \frac{\sum_{k}^{NrClasses} \frac{tp_k}{tp_k + fn_k}}{NrClasses} \tag{22}$$

$$\text{F score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{23}$$

3 Data

3.1 Oral Dataset

The cell samples were collected at Södersjukhuset in Stockholm. From each patient a sample is taken with a tiny brush that is scraped at different areas in the oral cavity. Each scrape is then smeared on a piece of glass which is stained to highlight important cellular structures. Images from each sample is acquired using an Olympus BX51 bright-field microscope with a 20x, 0.75 NA objective giving a pixel size of 0.32 μm. The microscope was equipped with an E-662 Piezo server controller and actuator which controls movement of the camera. From each smear, areas with a large number of cells are selected and a stack of images are taken with a step length of 0.4 μm. Each stack contains 15 images with different focus points. This so that each cell will be in focus in at least one of the images in the stack. 60-80 image fields are photographed from each smear. The coordinates are then extracted using a basic Matlab script where each cell is marked and the coordinates are saved. The resulting dataset can be seen in Table 1.

Table 1: Oral dataset

Patient	**Glass**	**Diagnosis**	**Nr. of cells**
1	3	Healthy	2123
1	4	Healthy	1993
2	5	Healthy	1454
2	6	Healthy	1024
3	7	Healthy	928
3	8	Healthy	777
4	12	Tumor	245
5	33	Tumor (healthy side)	1198
5	34	Tumor (healthy side)	1098
5	36	Tumor	519
6	37	Tumor	988
6	38	Tumor	828
6	40	Tumor (healthy side)	872
6	41	Tumor (healthy side)	912

3.2 Cervical Datasets

3.2.1 CerviSCAN Dataset

This dataset is a result from the CerviSCAN project at Uppsala University. From 82 graded pap-smears more than 900 images, with a focus stack of 41 images, were captured. The microscope used to preform the image acquisition was an Olympus BX51 bright field supplied with a 40x, 0.95 NA objective, resulting in a pixel size of 0.25 μm. To be able to capture focus stacks for each image the microscope was equipped with an E-662 Piezo server controller and actuator. This managed to capture the focus stack with a step length of $0.4\mu m$. After the images were captured they were manually examined by a cytologist with 30 years experience of screening pap-smears. The cytologist examined each image and marked individual cells and diagnosed these according to the Bethesda system. In the end each image focus stack was equipped with coordinates for all cells, diagnose for all cells and in which of the stack images each cell was in best focus [13]. The resulting dataset can be seen in the following list:
Normal - 9809 cells

- NILM - 9809 cells

Abnormal - 2421 cells

- LSIL - 766 cells
- HSIL - 718 cells
- SCC - 750 cells
- Adenocarcinoma - 53 cells
- ASC-H - 2 cells
- ASC-US - 132 cells

Other - 325 cells

- Occluded - 209 cells
- Distorted - 37 cells
- Degenarative - 20 cells
- Inflammatory - 53 cells
- Unknown - 6 cells

3.2.2 Herlev Dataset

The Herlev dataset is developed in cooperation between the department of Pathology at the Herlev University Hospital and the department of Automation at the Technical University of Denmark. The database containing single cell images was created from a large set of pap-smears. This procedure was done by professional cytologists using a microscope with a $0.201\mu m$/pixel resolution to capture the single cell images. The

database was then manually classified into seven different cell types. The classification was done by two different cytologists to be able to validate the classification. If the two cytologist did not agree on a cell classification, that specific cell was taken out of the database [29] [30]. The resulting database after the validation consisted of the following list:
Normal - 242 cells

- Superficial squamous epithelial - 74 cells
- Intermediate squamous epithelial - 70 cells
- Columnar epithelial - 98 cells

Abnormal - 675 cells

- Mild squamous non-keratinizing dysplasia - 182 cells
- Moderate squamous non-keratinizing dysplasia - 146 cells
- Severe squamous non-keratinizing dysplasia - 197 cells
- Squamous cell carcinoma in situ intermediate - 150 cells

4 Method

4.1 Data Processing

4.1.1 Oral and CerviSCAN Dataset

Both the Oral and CerviSCAN dataset have focus stacks for each cell image. The oral dataset contains 15 focus levels per image field and the CerviSCAN dataset contains 41 focus levels per image field. To find the image with best focus for each cell the variance of the Laplace transform was used. First all cells were cut out with a size of 100 x 100 pixels. The image size was chosen so that the nucleus would fit in the image for all cells. All images of the same cell in the stack was then convolved with the Laplacian (17) and the variance was calculated (18). The image with the highest variance after the Laplace transform was considered having the best focus and its stack number was saved. For the oral dataset all images in the stack were evaluated to find the best focus. In the CerviSCAN dataset each annotated cell is marked with which stack image it is in best focus in. Using this information we chose to only evaluate eight images below and eight images above the proposed focus image.

To obtain extra information about the cells and not lose information due to bad focus, smaller stacks of each cells were used for training and evaluation. These cell images were created using the focus information obtained with the Laplace transform. First the image with best focus was chosen as the middle image of the stack. The four most adjacent images were then stacked below and above to create an image with a depth of five. This means the images had a size of 5x100x100 (Figure 8)

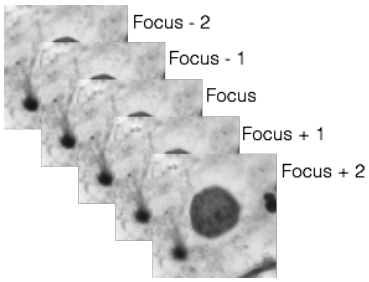


Figure 8: Five depth image used for training and evaluation.

The images were then normalized to scale down the range of pixel intensities (19). Each image in the image stacks were normalized separately. To expand the dataset the images were augmented. Since the diagnosis of a cell is depending on the relationship between neighboring pixels values we chose to augment without interpolation. This decision resulted in using two different transformation techniques. Initially the images were mirrored in 3 ways: Up-down, left-right and up-down-left-right. Then the original images and the three mirrored versions of it were rotated 90°. These two techniques combined increased the number of images eight times (Figure 9).

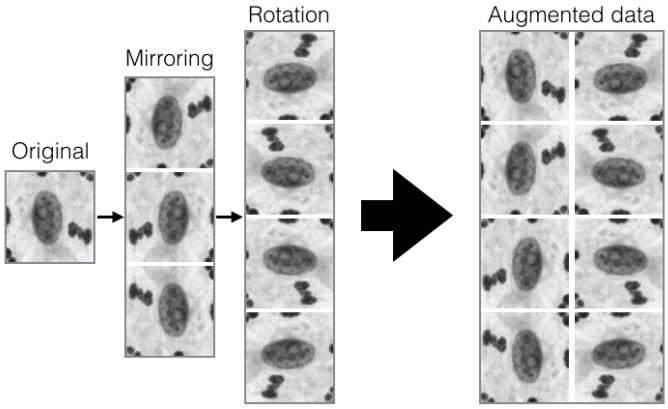


Figure 9: Original cell mirrored and rotated 90°.

4.1.1.1 Oral Dataset

The oral dataset is not annotated so the only information available is the patients diagnosis. To work with this dataset we had to make the assumption that all cells from a patient have the same diagnosis. The dataset was divided with respect to the patient diagnosis instead of the individual cell diagnosis. Two different sub-datasets were created, the first one (oral dataset 1) contained glasses from healthy patients and glasses with samples taken from the tumor side of patients with a tumor. The second dataset (oral dataset 2) contains healthy patients and samples taken from the healthy side of patients with a tumor. In oral dataset 1 there are three patients with tumors which gave us a 3-Fold cross evaluation (Table 2). Oral dataset 2 only contains

two patients with tumor which gave a 2-Fold cross evaluation. For both dataset we kept one healthy patient for evaluation and used the other two for training (Table 3). To balance the datasets for training, all cells from healthy patients were mixed and randomly chosen to have an equal amount of healthy cells and tumor cells.

Table 2: 3-Fold division of healthy patients and patients with tumor

Folder	Glasses for training		Glasses for evaluation	
	Healthy	Tumor	Healthy	Tumor
Fold 1	3,4,5,6	37,38,12	7,8	36
Fold 2	3,4,5,6	36,37,38	7,8	12
Fold 3	3,4,5,6	12,36	7,8	37,38

Table 3: 2-Fold division of healthy patients and patients with tumor (healthy side)

Folder	Glasses for training		Glasses for evaluation	
	Healthy	Tumor (healthy side)	Healthy	Tumor (healthy side)
Fold 1	3,4,5,6	33,34	7,8	40,41
Fold 2	3,4,5,6	40,41	7,8	33,34

4.1.1.2 CerviSCAN Dataset

The CerviSCAN dataset is annotated on a cell level and contains 12 different diagnosis of cells. Some of these are only visual explanation of the cell (e.g distorted) and some have few samples (e.g ASC-H). We chose to work with the normal cells and the three most common abnormal cells (Table 4). The dataset was divided into 2 parts, 80% for training and 20% for testing. The training set was divided into 5 equally large folds. To even out the training set we chose to remove normal cells until the two classes were equally large.

Table 4: Sub-dataset from CerviSCAN

Normal		Abnormal	
Diagnosis	Amount	Diagnosis	Amount
NILM	9809	LSIL	766
		HSIL	718
		SCC	750

4.1.2 Herlev Dataset

The Herlev dataset differs from the two other datasets in a couple of different ways. The biggest difference is that the Herlev dataset does not have focus stacks. Instead the dataset contains of 3 channel RGB images with the nucleus cropped out. Another discrepancy is that the size of each image varies. To solve this all the images were resized to 100 x 100 pixels, with bilinear interpolation, to have equally large sizes. After resizing the images they were normalized and divided into two classes (Table 5). Since the two classes differed in the amount of cell images the augmentation was

done different for the two classes. The normal cells were augmented by mirroring in three ways: Up-down, left-right and up-down-left-right. Then the original image and the three mirrored versions were rotated 90°. This increased the amount of normal cells eight times. The abnormal cells were only mirrored in one way, up-down-left-right, then rotated 90°. This increased the amount of abnormal cells four times. The dataset was divided into 2 parts, 80% for training and 20% for testing. The training set was then divided into 5 equally large folds.

Table 5: The Herlev Dataset

Normal

Diagnosis	Amount
Superficial squamous epithelial	74
Intermediate squamous epithelial	70
Columnar epithelial	98

Abnormal

Diagnosis	Amount
Mild squamous non-keratinizing dysplasia	182
Moderate squamous non-keratinizing dysplasia	146
Severe squamous non-keratinizing dysplasia	197
Squamous cell carcinoma in situ intermediate	150

4.2 Network Architecture

Two different networks were evaluated to compare performance. One was a version of a VGG network and the other a version of ResNet. To decide the size of the networks a parameter k was used. k controls the number of outputs in the first layers. After each sub-sampling (for instance after a pooling operation) the value of k was doubled. The value of k also controls the number of features in the fully connected layers. As loss function we used the negative log likelihood function which together with a softmax layer can be described as the cross entropy loss function (4). To optimize the network Adam optimization, with a learning rate of 0.01, was used. The best value of k for both networks were then evaluated using K-Fold cross validation.

4.2.1 VGG

The VGG architecture used was inspired from the original VGG16 network with 16 weight layers. The main difference is that one fully connected layer is removed and batch normalization and dropout are inserted to regularize the network. Batch normalization is inserted after every convolutional layer and FC layer. After batch normalization a ReLU layer is inserted as non-linearity. Between the two FC layers dropout layers are inserted with a probability of 0.5, one before batch noramlization and one after. At the end of the network softmax is used to calculate the probabilities (Figure 10).

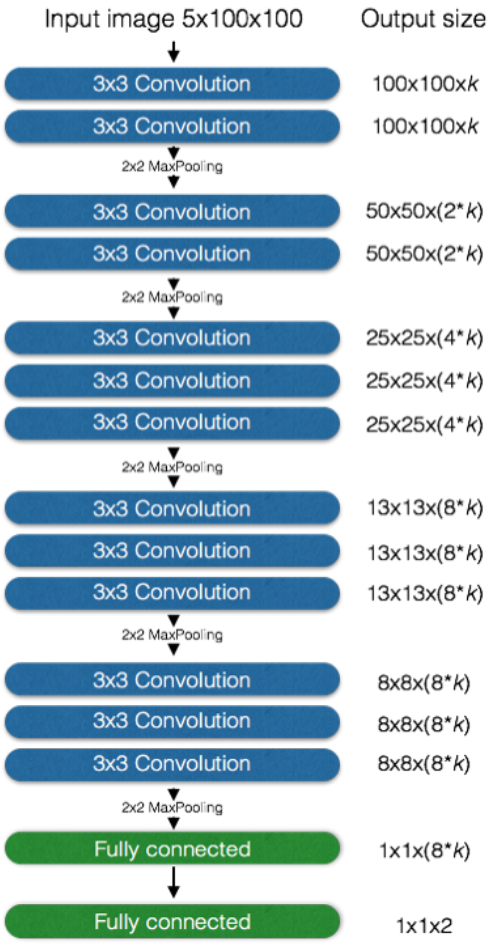
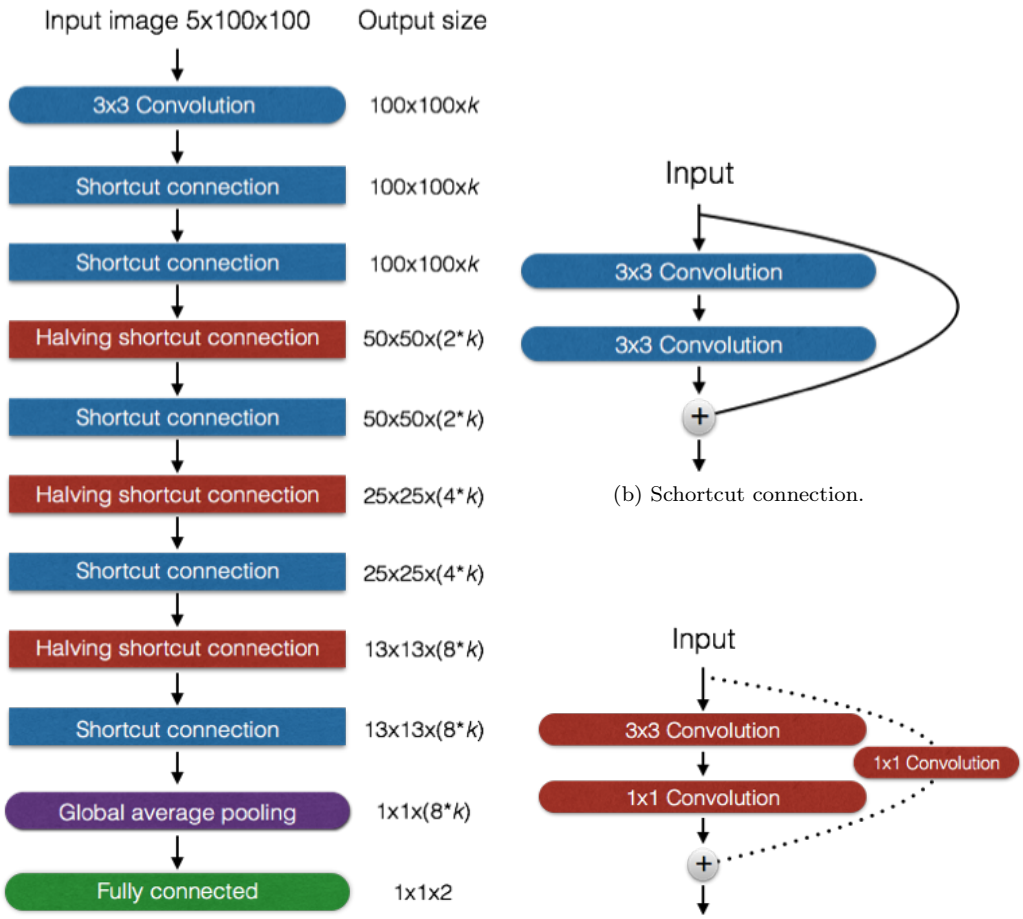


Figure 10: VGG inspired architecture.

4.2.2 ResNet

The ResNet architecture is inspired by the ResNet18 network created by Microsoft. In the shortcut connections batch normalization layers are inserted after both convolutional layers. ReLU layers are inserted after the first convolutional layer and after the addition. In the halving shortcut connections batch normalization is inserted after the two convolutional layers. ReLU is inserted after the first convolution and after the addition. Softmax is used at the end of the network to calculate probabilities (Figure 11).



(a) ResNet.

(b) Schortcut connection.

(c) Halving shortcut connection.

Figure 11: ResNet architecture with shortcut connections

5 Result

5.1 Oral Dataset

5.1.1 Oral Dataset 1

Investigating which network size to use led to the result in Figure 12. The networks were trained with the samples in Fold 1 (Table 2) which resulted in approximately 16,000 images per class after augmentation. The training was done for 40 epochs and validated once per epoch. At each validation the accuracy is compared to previous validation accuracies. If the accuracy is larger than all previous the network is saved for further evaluation.

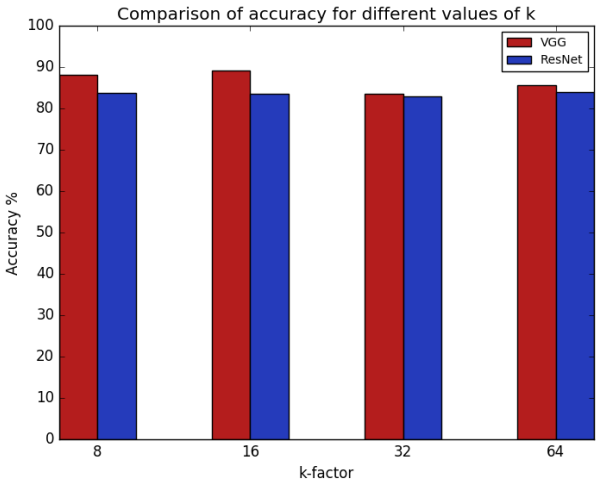


Figure 12: Comparison of accuracy for different values of k.

The k-value chosen for further evaluation was 16. The result after a 3-Fold evaluation for this k is presented in Table 6. Accuracy, precision, recall and f score are evaluated at cell level for all patients used for testing.

Table 6: 3-Fold evaluation for oral dataset 1

Network	Accuracy	Precision	Recall	F score
VGG	80.66 ± 3.00	75.04 ± 7.68	80.68 ± 3.05	77.68 ± 5.28
ResNet	78.34 ± 2.37	72.48 ± 4.46	79.00 ± 3.37	75.51 ± 3.17

The evaluation of the 3-Fold cross validation on patient level is presented in Figure 13 and 14. For glass 7 and 8 the mean accuracy is presented since there are three different results for these glasses.

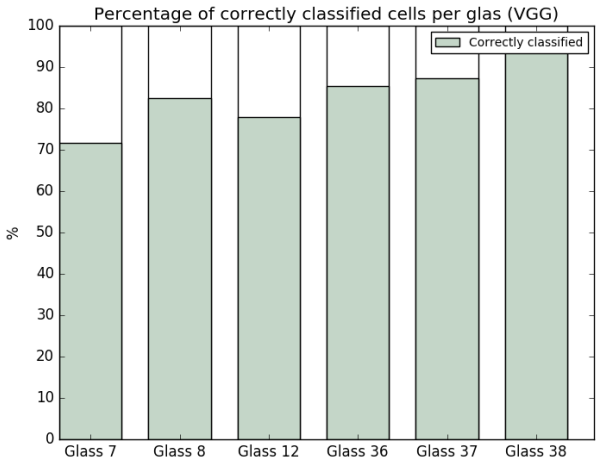


Figure 13: Percentage of correctly classified cells per glass for VGG.

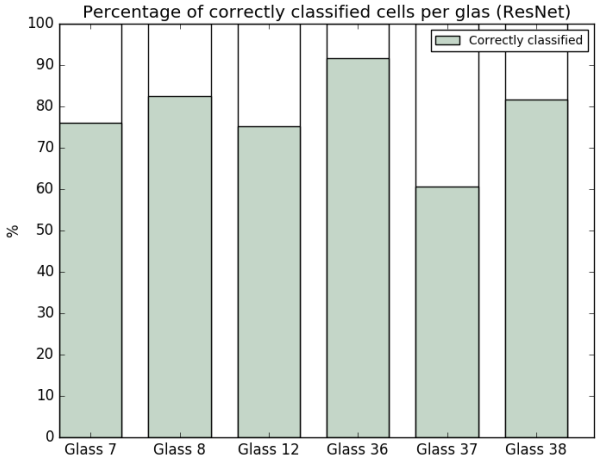


Figure 14: Percentage of correctly classified cells per glass for ResNet.

5.1.2 Oral Dataset 2

For the network size evaluation for the second dataset we used Fold 1 of oral dataset 2 (Table 3). The resulting training set contained approximately 18,000 images per class after augmentation. The networks were trained for 40 epochs and the best performing network was saved. Performance for different values of k is shown in Table 15.

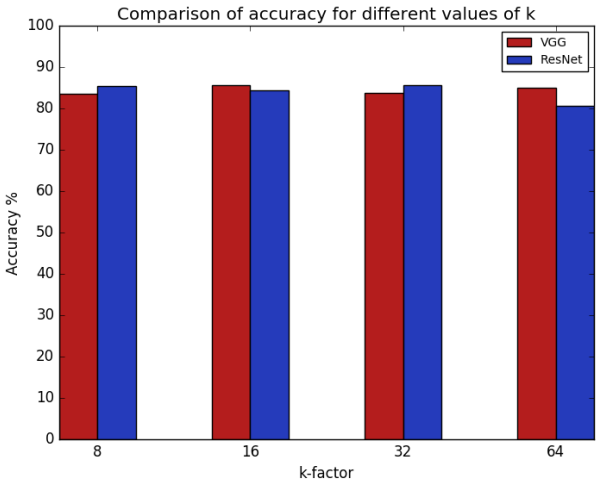


Figure 15: Comparison of accuracy for different values of k.

The chosen k-value for cross validation was 16. Table 7 shows the accuracy, precision, recall and f score at cell level after 2-Fold cross validation.

Table 7: 2-Fold evaluation for oral dataset 2

Network	Accuracy	Precision	Recall	F score
VGG	80.83 ± 2.55	82.41 ± 2.55	79.79 ± 3.75	81.07 ± 3.17
ResNet	82.39 ± 2.05	82.45 ± 2.38	82.58 ± 1.92	82.51 ± 2.15

The resulting accuracy at patient level is presented in Figure 16 and 17. For glass 7 and 8 the presented accuracy is the mean of the two folds.

Figure 16: Percentage of correctly classified cells per glass for VGG.

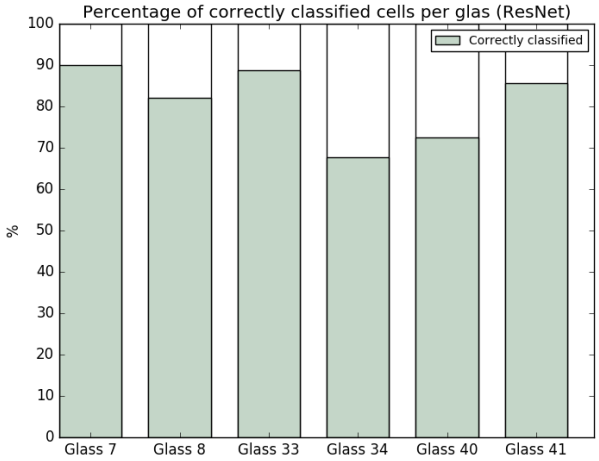


Figure 17: Percentage of correctly classified cells per glass for ResNet.

5.2 CerviSCAN Dataset

After augmentation the resulting dataset from CerviSCAN contained 11'000 images per class. The networks were trained for 40 epochs and the best preforming network was saved for testing. The result from training two different networks of different sizes is presented in Figure 18, the accuracy is for the test set.

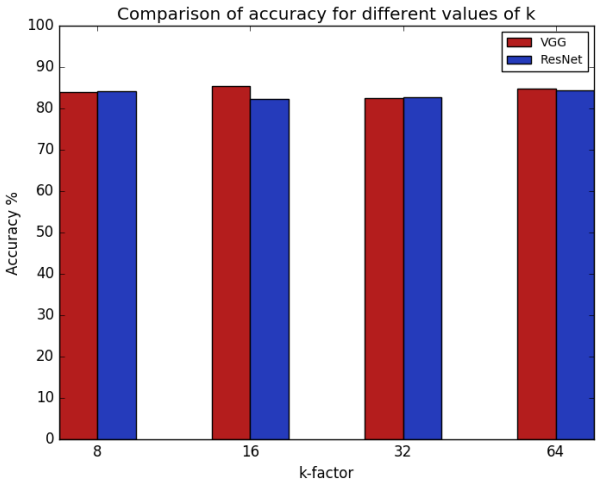


Figure 18: Comparison of accuracy for different values of k.

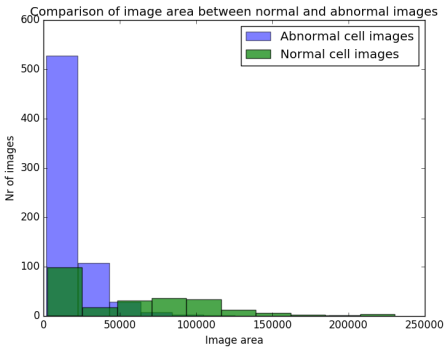
Further network evalutaion with 5-Fold cross validation was done for $k = 16$. The result is presented in Table 8 with the accuracy, precision, recall and f score. The evaluation is done on the test set presented with the mean of the 5-Fold cross validation with the standard deviation.

Table 8: 5-Fold evaluation for the CerviSCAN dataset

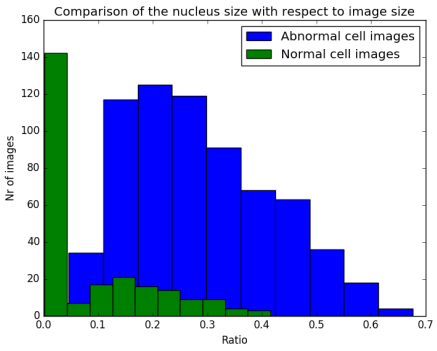
Network	Accuracy	Precision	Recall	F score
VGG	84.20 ± 0.86	84.35 ± 0.97	84.20 ± 0.86	84.28 ± 0.91
ResNet	84.45 ± 0.46	84.64 ± 0.38	84.45 ± 0.47	84.55 ± 0.41

5.3 Herlev Dataset

The first thing evaluated on the Herlev dataset is the distribution of image sizes. The distribution is presented in Figure 19 with the total amount of pixels and the ratio between the nucleus and image size.



(a) Image sizes.



(b) Ratio of nucleus size and image size.

Figure 19: Comparison of different image measurement for the Herlev dataset.

The Herlev dataset is evaluated with $k = 16$. The total amount of images for training was approximately 2600 with 1300 per class. The networks were trained for 40 epochs per fold and the resulting 5-Fold evaluation is presented in Table 9.

Table 9: 5-Fold evaluation for the Herlev dataset

Network	Accuracy	Precision	Recall	F score
VGG	86.56 ± 3.18	85.94 ± 6.98	79.04 ± 3.81	82.16 ± 3.85
ResNet	86.45 ± 3.81	82.35 ± 5.11	84.45 ± 2.16	83.36 ± 3.65

5.4 Comparison of Datasets

$k = 16$ was used for evaluating all four datasets for both type of networks. In Figure 20 the resulting f score with standard deviation is shown.

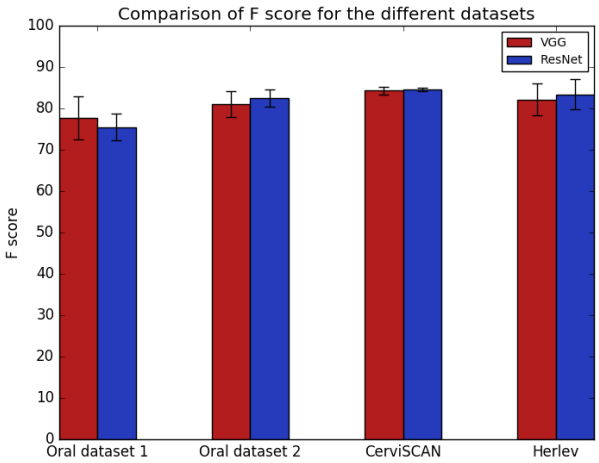


Figure 20: Comparison of f score for the different datasets.

6 Discussion

6.1 Oral Dataset

From each smear 60-80 image fields were captured. This means that cells from approximately 1/3 of the whole glass is extracted. Since no medical expert has gone through the glasses and selected interesting cells there is no way of knowing that important cells, indicating cancer, has been selected. The only information provided were the patients diagnosis, hence the assumption that all cells from a patient is considered having the same diagnosis. This problem can be solved by having the cells annotated by a medical expert. A way of expanding this dataset and get more cells would be to extract image fields from the rest of the glasses. Since the dataset is not annotated, the only possible classification is binary which leads to our results. Another problem in the dataset is the coordinate extraction. The coordinates for all cells in an image are not extracted by cytologist trained in distinguish between what is a cell and not. The cells used from this dataset is chosen by us with the criteria of what we think looks like a cell. Some errors could occur in this selection.

6.1.1 Oral Dataset 1

The first sub-dataset contained samples from healthy patients and samples taken from the tumor side of patients with a tumor. We chose to separate patients in training and evaluation since this would give a more accurate reflection of the reality. Some tests were done with having the same patient in training and evaluation and separating the individual glasses instead of patients. This gave a much higher accuracy which might indicate that the samples taken from the same patient looks too similar to draw any real conclusions. Since we only had three patients with tumors we chose to always have two patients for training and one for testing. This so that we could maximize the number of images used for training. The cost of this is that the networks were only

tested on one single patient per class and more extensive test on multiple patients would be to prefer to draw more general conclusions. We also chose not to rotate the healthy patient in the 3-Fold evaluation. We chose this under the assumption that there are no cells from healthy patient that indicate any form of cancer. On the other hand, in the samples taken from patients with tumor there could be a mix of cancerous cells and normal cells.

From Figure 12 we can see that there is only a small variation of accuracy for ResNet. A smaller value of k gives a slight boost in the accuracy. For VGG the smaller network size outperforms the larger by approximately 5% in accuracy. We chose the k value of 16 since it performed well for both VGG and ResNet.

Looking at the accuracy, precision, recall and f score for the two network with k=16 (Table 6). The main difference between the two networks is that VGG has larger overall values in the four categories, but ResNet has a lower standard deviation. This indicates that ResNet is a more stable network architecture and not as dependent on the training images. Both networks have a lower precision than recall which indicates that there are more false positive predictions which lowers the value of the f score.

For the evaluation of the individual glasses (Figure 13 and 14) the first thing to notice is that VGG have a more even classification results and over 70% correctly classified cells per glass. ResNet varies more and span from 60% to 90% correctly classified cells. The largest differences are for glasses 36, 37 and 38 where VGG outperforms ResNet on an average for those three glasses. One thing to notice is when glass 12 and 36 where used for training (Table 2) the number of images for training was lower. This leads to that the test result for glass 37 and 38 is taken from a network trained on fewer images than for the other glasses which can be an indication for the varying result.

Since the images in the dataset are normalized the networks will not be able to separate the two classes depending on the mean intensity values. This indicates that the networks have to learn cellular changes rather than other differences in the images, for instance differences that can occur in the coloring or capturing of the samples. Since we made the assumption that all cells in a glass have the same diagnosis this result strengthens the thesis that the network could find malignancy associated changes.

6.1.2 Oral Dataset 2

The second sub-dataset contained samples taken from healthy patients and samples taken from the non tumor side of patients with a tumor. This dataset was constructed with the assumption that malignancy associated changes could be present in patient with tumors. Cells taken from a patient with a tumor should have some form of indication of cancer even if the samples are not collected in vicinity of the tumor. Since only samples taken from the healthy side of patients with tumors were available from two patients, the evaluation of this dataset is limited. In the 2-Fold evaluation one patient was used for training and the other for testing, so the test results are limited to one patient.

From Figure 15 you can see that there is no large difference in accuracy for the two networks evaluated with different k values. Choosing to have as small network as possible, to minimize the training and testing times, we choose $k = 16$ as the best k-value.

For the accuracy, precision, recall and f score the different networks perform evenly (Table 7). ResNet has an overall smaller standard deviation which might indicate that it is a more stable architecture. ResNet also has a larger recall than VGG, which means that it has a lower rate of misclassification.

The evaluation of the network per glass (Figure 16 and 17) shows that both networks have over 60% per glass. VGG has a higher accuracy for glass 33 and 34 whilst ResNet has a higher accuracy for glass 40 and 41. ResNet also has a higher accuracy on glass 7 and 8. Since the glasses are not annotated it is hard to say if the networks becomes bias towards any cell types or what this difference depends on.

Since the images in the dataset are normalized the networks will not be able to separate the two classes depending on the mean intensity values. This is an indication that the networks looks for cellular changes rather then other less important differences. With the assumption that all cells from each glass is diagnosed the same way the network might find malignancy associated changes in the glasses.

6.2 CerviSCAN Dataset

During the collection of samples the pap-smears are stained and dyed. With 82 different pap-smears there is a chance that this process could have been different for some of them. For example some pap-smears might have been stained for a longer time than others. These possible differences might affect the network capabilities in a negative way. Some of the images also contains debris (for example blood cells and secretion). This might also affect the network capabilities in a negative way. There is always going to be a predominant chance that a pap-smear contains debris. Even though this could effect our results it displays how the real world works. The annotation process performed on the CerviSCAN dataset was done by one cytologist. Annotation by multiple cytologists would increase the reliability of this dataset.

In Figure 18 the comparison between the two networks, for different values of k, is displayed. This shows that VGG preforms slightly better overall than ResNet on the CerviSCAN dataset. VGG preforms better when k is lower meanwhile ResNet almost preforms equally well on all values of k. We chose to further evaluate the two networks with $k = 16$.

While looking at the 5-Fold evaluation in Table 8 ResNet surpasses VGG in all four measurements, though just barely. You can also see that the standard deviation is almost double for VGG than for ResNet. This indicates that ResNet is a more stable network than VGG, which means it is less depending on training data. One can also see that for both networks, accuracy, precision, recall and f score are similar. This means that the false positive rate is at the same level as the misclassification rate, which indicates that the networks does not discriminate either of the classes.

6.3 Herlev Dataset

The Herlev dataset was not created for CNNs, hence the variety of image sizes displayed in Figure 19a. This problem was solved by simply resizing all images to the same size. The resizing was done with bilinear interpolation which change the pixel values in the images. This can create a problem due to loss of information regarding the individual cell diagnosis. It was also noticed that the different diagnosis were cropped out differently. Most of the abnormal cells were cropped out in a way where

the nucleus covered the entire image. The normal cells were cropped out in a way where the whole cytoplasm and nucleus fitted the image. In Figure 19b the ratio between nucleus and image size is displayed. This showcases a problem where abnormal cell images and normal cell images are very different. This problem could create a discrimination during the training of a CNN. Where the CNN learns that abnormal cells have a large nucleus and normal cells have a small nucleus, rather than learning malignant changes in the cells.

The annotation of the Herlev dataset is done in a proper way. Two cytologists annotated the cells separately and if they did not agree on the diagnosis of a cell that cell was removed. By annotating in this way the human factor of error is small and one can assume that all annotations are true.

The Herlev dataset was used as a comparison to the results from the other datasets. Since $k = 16$ was used for all the other datasets fold evaluations, we chose to only use $k = 16$ for the Herlev 5-Fold evaluation. Table 9 shows that VGG have better accuracy and precision, while ResNet have better recall and f score. Since there were not an equal amount of cells in the two classes f score provides the most accurate result. With this in mind ResNet was the best choice of network for the Herlev dataset.

6.4 Comparison of Datasets

In Figure 20 we can see that the highest f score for both networks were achieved on the CerviSCAN dataset. The reason for this might be that the CerviSCAN dataset contains more images than the Herlev dataset, and that the cells are annotated on a cell level unlike for the oral datasets cells. The CerviSCAN dataset also has the lowest standard deviation which implies that a large annotated dataset is preferable for classifying cell images.

There is also a difference in the f score for the two oral datasets. This might be because oral dataset 1 contains three patients with tumor and the second one only two patients with tumor. With three different patients there might be more variations in the cells which might lower the result. One can also argue that the result for oral dataset 1 reflects the reality more since it could have a larger variety in the cells.

It is also noticeable that ResNet always have a lower standard deviation compared to VGG. This implies that ResNet is an overall more stable network regardless of which dataset it has been trained on.

7 Conclusion

7.1 Using Deep Convolutional Networks to detect changes due to malignancy

7.1.1 Oral Cancer

Since the oral dataset is not annotated on a cell level, the ground truth for each cell is missing. The conclusions that can be drawn is therefore limited. In order to classify the cells we made the assumption that all cells from a patient has the same diagnosis. For instance when looking at a patient with a tumor, all cells in the vicinity of the tumor is considered having some sort of change due to malignancy. Since malignancy

associated changes can not be caught by visual examination the obtained results need to be compared with other methods for finding these changes in cell images. For the two oral datasets we obtain an accuracy of over 80% and knowing that the mean intensity value per image does not differ between the two classes. This is an indication that the networks learns malignant changes instead of other, less relevant, differences between the images. There are potential in detecting cellular changes due to malignancy in oral cells with CNNs. An example is using a statistical threshold for the amount of classified malignant cells in a patient. If the network classifies fewer cells as malignant than the threshold the patient would be considered healthy. Patients with more cells classified as malignant would have to go through a manual examination. Another example is to let the network present a specific amount of the most malignant looking cells and let a doctor screen only those. Both these examples would reduce the manual examination time per patient and would be a more time and cost efficient alternative to manual screening.

Our research have also resulted in a dataset with samples taken from the oral cavity with the patient diagnosis available. This dataset could be used for further evaluation of classifying and diagnosing cells from the oral cavity.

7.1.2 Cervical Cancer

The results regarding cervical cancer are obtained on two different datasets, which are both annotated on a cell level. Both datasets have an accuracy and f score over 84%, which indicates that there is a high potential of detecting cellular changes due to malignancy. Using CNNs as a tool in the screening process would speed up the examination time per patient. Regarding how this tool would be implemented, further investigation must be done on which types of cells are most commonly misclassified etc. With this information it might be possible to showcase a specific amount of cells that, the network predicts, are most malignant which a doctor then could screen. This would decrease the time spent on each sample and thus make it cheaper and more efficient.

7.2 Choice of network architecture

The ResNet architecture has a lower standard deviation than VGG on all datasets. This implies that it is a more stable network and less dependent on the training images. ResNet also outperforms VGG on f score on three of the datasets and would therefore be the preferable choice for implementation. The choice of k-value for both networks does not affect the accuracy in a predominant way. Choosing a lower value speeds up the training and execution time which is why it is the preferable choice.

7.3 Future work

The major limitation for this projects were the datasets. The oral dataset contained few patients and no annotation on a cell level. By obtaining more samples and expanding the dataset, the results would be a more accurate reflection of the reality. Regarding the cell level annotation it would be preferable to have the dataset annotated by a well experienced cytologist. The CerviSCAN dataset could also be improved by letting more than one expert annotate the cells. This would reduce the risk of errors that could occur when a single person annotate the cells.

Another interesting part that could be examined is to classify each malignant cell in what type of cell it is. Implementing this after the binary classification would result in a tree structure. The network first classifies each cell as malignant or non malignant, then classifies the malignant cells in what actual diagnose it has (e.g NILM, HSIL etc.). This would increase the knowledge of the diagnosis of the patient.

One area that have potential for improving the results is the network architecture. The number of different types of networks that are available and the countless ways these can be constructed creates many choices for the user. It would be interesting to keep on trying new networks and modified versions of them.

8 Acknowledgments

First and foremost we would like to express our appreciation to our supervisors Sajith Kecheril Sadanandan and Ewert Bengtsson for all your help. To Sajith for the endless amount of help and inspiration with the neural networks and for guiding us forward in our project, thank you. To Ewert for inspiring us and passing on your experiences and knowledge in the subject, thank you. We would also like to thank our subject reviewer Carolina Wählby for making this project possible for us. You helped us stay on the right track and inspired us to explore new possibilities in the project. Also thanks to Petter Ranefall for all your help during this project.

A special thanks to Christina Runow-Stark for helping us collect samples and for the inspiring visit at Södersjukhuset in Stockholm. We would also like to thank Jan Hirsch for making this project possible, starting it up and pushing it forward.

References

[1] WHO. *Human papillomavirus (HPV) and cervical cancer*. 2017. URL: http://www.who.int/mediacentre/factsheets/fs380/en/ (visited on 04/26/2017).

[2] Stewart W.B. Wild P.C. *World Cancer Report 2014*. International Agency for Research on Cancer/World Health Organization, 2014.

[3] deeplearning.net. *Convolutional Neural Networks (LeNet)*. 2017. URL: http://deeplearning.net/tutorial/lenet.html (visited on 04/22/2017).

[4] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252. ISSN: 0920-5691. DOI: 10.1007/s11263-015-0816-y.

[5] A.Karpathy. *CS231n Convolutional Neural Networks for Visual Recognition*. 2016. URL: http://cs231n.github.io/convolutional-networks/.

[6] WHO. *Cancer*. Feb. 2017. URL: http://www.who.int/mediacentre/factsheets/fs297/en/ (visited on 05/05/2017).

[7] The American Cancer Society medical and editorial content team. *How Are Oral Cavity and Oropharyngeal Cancers Diagnosed?* Aug. 2016. URL: https://www.cancer.org/cancer/oral-cavity-and-oropharyngeal-cancer/detection-diagnosis-staging/how-diagnosed.html (visited on 04/26/2017).

[8] WHO. *Cervical cancer*. 2017. URL: http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/ (visited on 04/26/2017).

[9] American Cancer Society. *What Is Cervical Cancer?* Nov. 2016. URL: https://www.cancer.org/cancer/cervical-cancer/about/what-is-cervical-cancer.html (visited on 04/26/2017).

[10] American Cancer Society. *What Are the Risk Factors for Cervical Cancer?* Nov. 2016. URL: https://www.cancer.org/cancer/cervical-cancer/causes-risks-prevention/risk-factors.html (visited on 04/26/2017).

[11] Wisconsin State Laboratory of Hygiene. *Pap Test FAQ*. 2017. URL: http://www.slh.wisc.edu/clinical/cytology/resources-for-patients/pap-smear-faq/ (visited on 04/26/2017).

[12] M Schwab. *Encyclopedia of cancer*. Springer, 1-01-2012, pp. 2140–2142.

[13] Patrik Malm. "Image Analysis in Support of Computer-Assisted Cervical Cancer Screening". PhD thesis. 2013. ISBN: 978-91-554-8828-4. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-212518.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[15] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[16] Xiaodong Gu Haibing Wu. "Towards Dropout Training for Convolutional Neural Networks". In: *Neural Networks* 71 (Nov. 2015), pp. 1–10.

[17] Francis. *batch normalization*. Apr. 2015. URL: https://standardfrancis.wordpress.com/2015/04/16/batch-normalization/ (visited on 05/03/2017).

[18] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *CoRR* abs/1502.03167 (2015). URL: http://arxiv.org/abs/1502.03167.

[19] Andrew Gibiansky. *Convolutional Neural Networks*. Feb. 2014. URL: http://andrew.gibiansky.com/blog/machine-learning/convolutional-neural-networks/.

[20] Sebastian Raschka. *Python Machine Learning*. Birmingham, UK: Packt Publishing, 2015. ISBN: 1783555130.

[21] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (Oct. 1986), pp. 533–536. DOI: 10.1038/323533a0. URL: http://dx.doi.org/10.1038/323533a0.

[22] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014). URL: http://arxiv.org/abs/1412.6980.

[23] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556 (2014).

[24] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385.

[25] José Luis Pech-Pacheco et al. "Diatom Autofocusing in Brightfield Microscopy: a Comparative Study". In: *ICPR*. 2000.

[26] Andrej Karpathy. *Neural Networks Part 2: Setting up the Data and the Loss*. 2016. URL: http://cs231n.github.io/neural-networks-2/ (visited on 05/05/2017).

[27] Tadayoshi Fushiki. "Estimation of prediction error by using K-fold cross-validation". In: *Statistics and Computing* 21.2 (2011), pp. 137–146. ISSN: 1573-1375. DOI: 10.1007/s11222-009-9153-8. URL: http://dx.doi.org/10.1007/s11222-009-9153-8.

[28] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information Processing and Management* 45.4 (2009), pp. 427–437. ISSN: 0306-4573. DOI: http://dx.doi.org/10.1016/j.ipm.2009.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0306457309000259.

[29] Jonas Norup. "Classification of pap-smear data by transductive neuro-fuzzy methods". MA thesis. Technical University of Denmark, 2005.

[30] G. Dounias. *PAP-SMEAR (DTU/HERLEV) DATABASES & RELATED STUDIES*. July 2008. URL: http://mde-lab.aegean.gr/downloads (visited on 05/05/2017).