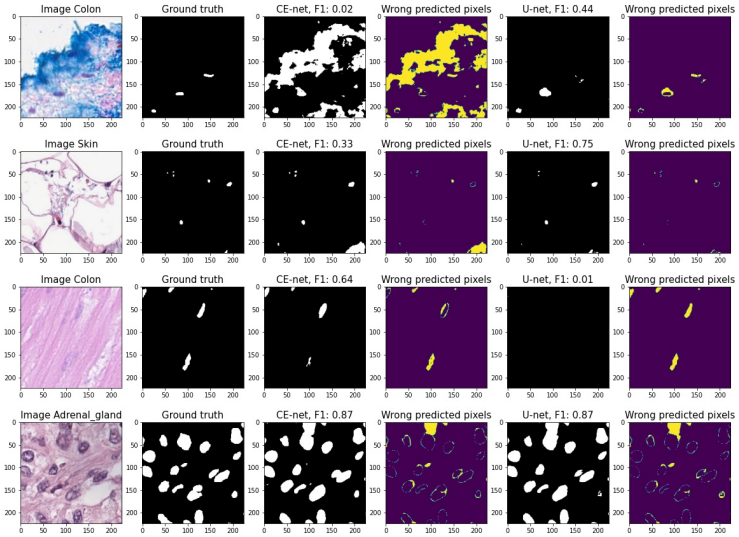
Fully-automated nuclei detection and segmentation in medical images using Deep-Learning techniques

submitted in partial fulfillment for the degree of master of science

Matty Vermet
11320524

master information studies
data science
faculty of science
university of amsterdam

Your Date of defence in the format 2021-07-05



	Internal Supervisor	2nd Internal Supervisor
Title, Name	Dr Maarten Marx	Lester van der Pluijm
Affiliation	UvA, FNWI, IvI	UvA, FNWI
Email	maartenmarx@uva.nl	l.c.vanderpluijm@uva.nl

Universiteit van Amsterdam

Amsterdam Data Science

Abstract

Nuclei segmentation plays a crucial role in the analysis of biomedical images. It enables determination of the distribution of certain cell types that are related to diseases such as cancer. Eventhough, it has become a hot topic for research, finding a suitable model that is able to generalize and accurately segment nuclei still remains a difficult problem. In this paper, CE-Net is presented as the model that could potentially be suitable for this problem as it is able to capture high-level information from the features due to pretrained ResNet blocks in the encoder. The model is trained and tested on the PanNuke dataset, which consist of >7000 images from 19 different tissue types. The outcomes on the evaluation metrics of the CE-Net is compared to the medical state-of-the-art model U-Net. The experimental results show that U-Net achieves an average gain over CE-Net of 0.03 on the Jaccard index score and the F1-score, 0.14 on the Cell-Counting Accuracy, and 0.03 on the Panoptic Qualtiy metric. The results also show that U-Net produces more robust predictions and is able to handle a small number of training samples of approximately 250.

Keywords: Deep-learning, nuclei, CE-Net, U-net, Segmentation, Detection

1 Introduction

Deep learning is a branch of machine learning and plays an important role in the shift to an automated world. It is primarily used for complex tasks such as image analysis, natural language processing and voice recognition. Also, in the healthcare sector deep learning is the main choice for biomedical image analysis due to its accuracy and ability to generalize [31]. Within the biomedical image analysis the image segmentation has become one of the most important task in healthcare researches [3]. In image segmentation a digital image is partitioned into multiple segments that capture particular objects in a set of pixels. The overall goal of image segmentation is to convert an image into an image that is easier to interpret and analyse. Through the use of image segmentation several kinds of pathologies, biological structures and tissues can be extracted in order to support e.g. surgical planning, the type of treatment, and the medical diagnosis [3].

The pathological sections are currently considered of great importance in diagnosing cancer as they are able to extract useful information concerning tumors [35]. The pathological microscopic images are widely used in order to make predictions such as prognosis, diagnosis or metastasis [35]. Cell nuclei segmentation is an indispensable step in biomedical image analysis, because by the segmentation of nuclei it is possible to determine the distribution of certain cell types that are related to a disease (such as cancer cells or immune cells) whereafter the tissue can be classified as malignant or healthy [35]. Apart from cell type classification, accurate nuclei detection and segmentation also provides valuable information for biological and medical analysis such as cell counting, or phenotype analysis [17].

A nucleus is the core of a cell which carries the desoxyribonucleideacid, or better known as DNA, which contains genetic information about the organism [22]. When detecting the nuclei, every separate cell can be identified and thereafter monitored to see how the individual nuclei respond to different cures. Due to the automation of detecting nuclei biologists and doctors are able to focus ever more on the cures, outcomes and solutions to improve insight and research. In the field of biomedical image analysis, nuclei detection is even stated as the most important task when it comes to diagnosing and to understand disease mechanics [22].

1.1 Automated nuclei segmentation

Image segmentation can be separated by different techniques into three categories: 1) Manual segmentation (MS), 2) Semi-automatic segmentation (SAS) and 3) fully automatic segmentation (AS) techniques [7].

(1) In order to be able to semi-automate or fully-automate nuclei segmentation, first the nuclei need to be segmentated manually. This process is conducted by experts on the subject which determine the so-called region of interest (ROI). Thereafter, the ROI is outlined as precise as possible in order to provide every pixel with the correct annotation. This way, every pixel is given a ground truth for further development of the SAS and AS of the nuclei [7].

(2) SAS is a combination of MS and AS. This method still required several low-level actions of the user (MS) in addition to the automated algorithm (AS). One of the possible interactions by the user may be the approximation of the location of the ROI which is thereafter used to segment the whole image. It can also require the user to manually check and re-write the ROI to lower the segmentation error [7].

(3) Fully AS do not involve user actions. It simply uses the annotated images from the MS as training data and trains the models based on supervised learning methods, e.g. deep neural networks (deep learning method). During training the model tries to capture patterns of the images by using a map of virtual neurons defined by layers between the input and output. The model assigns numerical values, called weights, to connections between the neurons. First, it assigns these weights randomly and check if it accurately captures the patterns, if not, the model will adjust the weights by a learning rate that is set beforehand. Challenges that may arise when working with Fully AS (or deep learning) are the high variation in e.g. size, shape, or color of the medical images

[7]. This complicates the application of the trained model to other datasets.

Even though a lot of models have been applied to a nuclei segmentation tasks the results are still not as high as would be expected. For example, in 2018 Kaggle organised the Data Science Bowl competition in which over 3500 teams participated including experienced data scientists. The challenge was to detect/segment nuclei across varied conditions of cell types or image modalities to find the best generalizing model. The best performing model had a F1-score of around 0.70. Meaning that for every 60 segmented nuclei, roughly 40 nuclei are either wrongly segmented (Type I error) or were not detected and segmented at all (Type II error). This indicates that finding a suitable model for the automated segmentation of nuclei is a complicated task, mostly due to the wide variety of pathological microscopic images and different nuclei types which complicates finding matching patterns throughout the dataset.

Multiple researches have tried increasing the performance on either the same dataset or a similar dataset with various conditions of the nuclei. Punn et al. (2020) implemented an updated version of U-Net but got similar results of approximately 0.70 on the F1-score [22]. Johnson (2019) used a different state-of-the-art model, Mask-RCNN, for the automated nuclei segmentation of the Kaggle dataset, and got a F1-score of 0.76 [10]. HoVer-Net implemented by Graham et al. (2019) [5] showed a F1-score of 0.83 on a different dataset. However, Gamper et al. [4] questions the validity of the results that are based on datasets used in challenges or contest, because of the lack of image diversity.

This research is based on the PanNuke dataset by Gamper et al. [4], which contains different forms of nuclei across 19 different tissue types but is not so extensively researched yet. They state that this dataset more accurately represent the variations in medical images than most nuclei datasets. Therefore, this would be a good dataset to use for finding a model that is able to generalize well while capturing the complicated patterns of the medical images. Gamper et al. [4] applied DIST, Mask-RCNN, MicroNet, HoVerNet to the PanNuke dataset, where HoVerNet gained the best results. In this research the PanNuke dataset will be examined by applying it to two state-of-the-art models, namely CE-Net and U-Net. Gu et al. [6] showed that their developed CE-Net was able to outperform U-Net in different medical segmentation tasks. Nuclei segmentation was not one of these tasks, but a similar task namely cell contour segmentation was.

CE-Net net has been applied to nuclei segmentation in only a few researches [23][9], where Huang et al. [9] showed that CE-Net is able to outperform U-Net. U-Net, however, is discussed and applied in multiple different researches and medical segmentation tasks. In multiple researches [35][29][22] U-Net is used as a baseline model in nuclei segmentation tasks or medical image segmentation in general to compare models performances with. To verify the performances of the models used in this research and enable comparison with the outcomes of the research by Gamper et al. [4], the Panoptic Quality (PQ) metric will be implemented.

The research question is as follows: How does CE-Net perform in detection and segmentation of nuclei compared to the medical baseline U-Net?

1. How sensitive are the models to outliers and noise?
2. Are the F1-scores or training times of the models influenced by a decrease in training samples?
3. How robust are the predictions of both models?
4. How is CE-Net performing in the segmentation of nuclei compared to U-Net looking at the F1-score and the Jaccard index scores?
5. How accurate is CE-Net in detecting the nuclei compared to U-Net looking at the Cell-counting accuracy metric?
6. Are the results on the PanNuke dataset of this research similar to the results of Gamper et al. [4], when comparing the Panoptic Quality metric?

The goal of this research is to assess the performance and the generalizability of CE-Net compared to the U-Net model by applying it to the multi-organ PanNuke dataset. First, this paper will describe the state-of-the-art models and put the models used in this research into context in the related work section, followed by an explanation of the used methods in the methodology section. Thereafter, the results will be discussed by examining each of the subquestions stated above. Finally, the results of this study will be reflected upon in the discussion section, and the key findings will be presented in the conclusion.

2 Related work

This section will give insights into the models that have been used for medical image segmentation and nuclei segmentation, and the additions or changes that have been made throughout the years to fill in the research gap or challenge of capturing the complicated patterns while remaining the ability to generalize between different medical images. To put the models used in this research, U-Net and CE-Net, into context of the state-of-the-art models and their history, this section will start with relevant work on medical image segmentation in general whereafter state-of-the-art models on nuclei segmentation will be discussed.

Important to know is that there are two types of segmentation; instance segmentation and semantic segmentation. The difference between the two is that instance segmentation identifies objects and stores it in separate classes, while semantic segmentation recognizes the class it belongs to and

Table 1. Overview of models and their scores on different datasets used for nuclei segmentation tasks.

Model	Kumar [14]			CoNSeP [5]			CPM2017 [28]			PanNuke [4]		
	AJI	F1	PQ	AJI	F1	PQ	AJI	F1	PQ	JI	F1	PQ
1) DIMAN (Xie et al. 2020 [30])	-	**0.89**	-	-	**0.87**	-	0.57	-	-	-	-	-
2) CNN3 (Kumar et al. 2017 [14])	0.51	0.76	-	-	-	-	-	-	-	-	-	-
3) DIST (Naylor et al. 2018 [20])	0.56	0.78	0.44	0.50	0.80	0.40	0.62	0.83	0.55	-	0.73	0.53
4) Triple U-Net (Zhao et al. 2020 [35])	**0.62**	0.84	**0.61**	**0.58**	0.84	**0.56**	**0.71**	**0.89**	0.69	-	-	-
5) HoVer-Net (Graham et al. 2019 [5])	0.62	0.83	0.60	0.57	0.85	0.55	0.71	0.87	0.70	-	0.8	0.66
6) Mask-RCNN (He et al. 2017 [8])	0.55	0.76	0.51	0.47	0.74	0.46	0.68	0.85	0.67	-	0.72	0.55
U-Net (this research)	-	-	-	-	-	-	-	-	-	0.70	**0.82**	**0.67**
CE-Net (this research)	-	-	-	-	-	-	-	-	-	0.67	0.79	0.64

assigns it to this class. Since there are two kinds of segmentation the techniques also differ, and are much dependent on the particular task and corresponding data.

2.1 Medical image segmentation

Medical images can be separated by the way they are generated into CT, MRI, ultrasound, whole slide images etc. The latter can then be divided into two categories: cell-level and tissue-level whole slide images [35].

Classical machine learning models such as SVM, Multilayer Perceptron, Random Forest, Markov Random Field, were the standard in medical image segmentation tasks before deep-learning methods had their breakthrough [26][15][29]. Recent literature showed that deep-learning methods outperform the conventional machine learning models in this particular area of medical image segmentation. However, deep-learning methods need a larger set of training data and also require more computational power and time than the conventional machine learning models [2]. The main bottleneck in applying deep learning to medical segmentation task is the lack of usable training data [35].

In the last years artificial neural networks (CNN) have become the main technique in medical image segmentation as CNN based models can match or even exceed human levels of performances in segmentation tasks [33]. One state-of-the-art deep-learning technique that is discussed often in relevant literature is U-Net.

U-Net is a framework based on Fully Convolutional Network (FCNs) and it uses an encoder-decoder structure [25]. The encoder-decoder structure, like U-Net, requires less training samples and overcomes the challenge that most deep-learning models face, having too few training data [35]. Cai et al. [1] review different techniques used for medical image classification and segmentation, and states that the main difference between CNNs and FCNs is that CNNs lose details and information about features in the pooling phase which results in difficulty of classifying a pixel. Where in FCNs (the encoder-decoder structure) the feature information is collected at the encoder path through downsampling and the image is restored at the decoder path by deconvolution, to remain the information in each pixel [23][1]. Also, the parameters that are trained in the training phase are far less which makes FCNs faster than CNNs [18].

A different extension of the encoder-decoder structure is CE-Net proposed by Gu et al. [6]. It adds a context extractor module between the encoder module and the decoder module, this way more high-level information of the image is captured and it is able to achieve higher performances than the state-of-the-art U-Net model [6][22].

Another method that enhances the information stored in each pixel and, therefore, preserves spatial information throughout the image is EncNet [34]. It introduces a layer that is able to capture the coding semantics and predict scaling factors [3].

2.2 Nuclei segmentation

Before the implementation of learning techniques, the common nuclei segmentation techniques were Otsu thresholding [21], which uses a color threshold to segment the background from the nuclei, or watershed algorithms, or watershed-markers [32]. Watershed-markers or watershed algorithms treat each pixel as topography; the grey scaled image shows the relief and gradient which can be seen as the elevation[26]. The brightness of the pixels relate to the height, this way lines can be identified [14]. By "flooding" the image the pixels can be classified into zeros "not flooded" and ones "flooded". Nevertheless, these algorithms often over-segment the image, especially for noisy images such as medical images [26][14].

Just like the trend in medical image segmentation, the learning models used for nuclei segmentation started with models such as Random Forest, Support Vector Machines, Multilayer Perceptron etc., where they were fed with features that are based on color and spatial filtering [13]

The application of CNNs to nuclei segmentation has been extensively researched and multiple different models are applied to various datasets. Most of them have used U-Net [17] or an improved version of U-Net [17][35], or have used U-Net as a benchmark model to compare their model [22]. Anyhow, U-Net is also in the nuclei segmentation task the one that is mostly discussed and analysed. Moreover, Mask R-CNN model [14] has been applied to a nuclei dataset. He et al. [8] had proposed a model that combines Mask R-CNN with U-Net and showed that it outperforms these models individually. DeepLab and its improvement, DeepLab V3+, have been suggested as a potential replacement of U-Net since it could slightly outperform U-Net. However, the results

of Rashid et al. [24] in a nuclei segmentation task has shown that Multiscale Dilated U-Net, another version of U-Net, outperforms DeepLab v3+.

A challenge in nuclei segmentation is to segment the nuclei that are touching or overlapping in an image. By adding watershed-markers to CNNs, Naylor et al. [19] showed a possible solution to overcome this challenge. Xie et al. [30] also implemented this technique of combining a deep neural network with watershed-markers and showed that it is able to outperforms state-of-the-art models such as U-Net or Mask-RCNN.

In 2019, Gu et al. [6] developed CE-Net which is able to capture more complex and high-level information from the features in medical images. They showed that CE-Net is able to outperform U-Net. One year later, Qin et al. [23], Huang et al. [9] applied CE-Net to nuclei segmentation tasks, where Huang et al. [9] showed that CE-Net is able to outperform U-Net.

2.3 Datasets

Some of the discussed models have been applied to the same datasets, and therefore, allows to do a quantitative comparison. Table 1 shows an overview of the Aggragated Jaccard Index (AJI) or Jaccard Index (JI), the F1-score (F1), and the Panoptic Quality (PQ) of six previously discussed models.

3 Methodology

In this section the used methods and implementation details will be explained. Firstly, the dataset that is used in this research will be examined by providing information of its origin, looking at the distribution of classes within the dataset, and look at the dataset splits that are used. Secondly, the architecture of the used models will be discusses, followed by an explanation of the chosen evaluation metrics and the loss functions. Lastly, the experimental set-up is provided with the actual implementation details of this research.

3.1 Data

PanNuke The dataset used for this research is an open source dataset called PanNuke, edited and extended by Gamper et al. [4]. The nuclei instance segmentation in this dataset is semi automatically generated for 19 different tissue types [4]. It consists of aggregated datasets containing nucleus classification. This dataset is the initial dataset for the ground truth generation. The aggregated datasets are:

1. Kumar [14]
2. CPM2017 [28]
3. TCGA [16]
4. dataset of bone marrow visual fields [11]

Every one of these datasets contains visual fields of multiple kinds of tissue. For the dataset TCGA alone, over 2000 visual fields are sampled from over 20000 Whole Slide Images (WSIs). Images containing frozen tissue were excluded. Also,

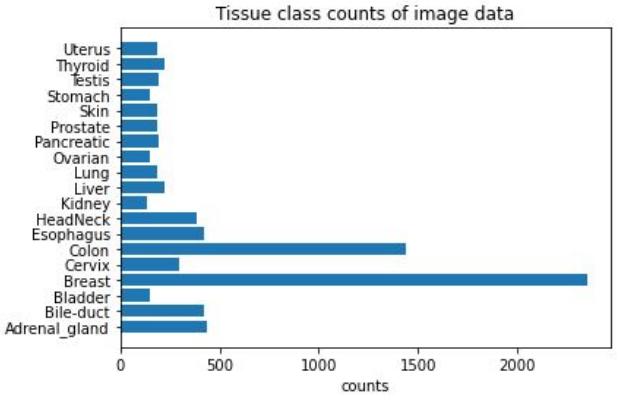


Figure 1. Class counts for tissue types in PanNuke dataset.

images with a resolution other than 40x were re-sized to 40x. In total, over 7000 images were labeled and thereafter double checked by pathologists. After the domain experts verified the labels, 189744 nuclei were annotated in total. These nuclei are found in tissue such as liver, prostate, kidney, breast, stomach, colorectal and bladder tissue. Figure 1 shows that there is a class imbalance in the PanNuke dataset, where 'colon' and 'breast' types are most common.

Furthermore, in the dataset the nuclei are categorized by the type of nuclei; Neoplastic, Non-Neo Epithelial, Inflammatory, Connective, and Dead nuclei. These are all represented by as their own layer in the masks, thus five different layers. The masks of the dataset contains six layers in total, where the last layer combines all the nuclei categories into a binary layer. The last layer labeled the pixels that correspond to nuclei in the images as 0 while non-nuclei pixels are labeled as 1. This research focuses on detection and segmentation of nuclei in general, thus the binary layer of the mask is used. In the prepossessing part of this research the 0 and 1 labels of the binary layer are swapped to ensure the boolean mask identifies the pixels containing nuclei on the label of 1 (Figure 3).

The binary layer of this dataset also shows some imbalance, where on average an image contains 21% true and 89% false labeled pixels. Indicating that the images consist of more background than foreground pixels. This should be taken into account when choosing accurate evaluation metrics, more on this in section 3.3.

In figure 2 the distribution of the total number of nuclei in an image and the distribution of the sizes these nuclei are presented. The number of nuclei per image ranges between 0 and ± 250, where the bulk of the images contain 0 to ± 40 nuclei. Figure 2 also shows that the sizes of most nuclei are between 1 and 1000 pixels, with outliers of sizes that reach up to 14000 pixels.

train/validation/test split In the PanNuke dataset, researched by Gamper et al. [4], the train, validation, test split

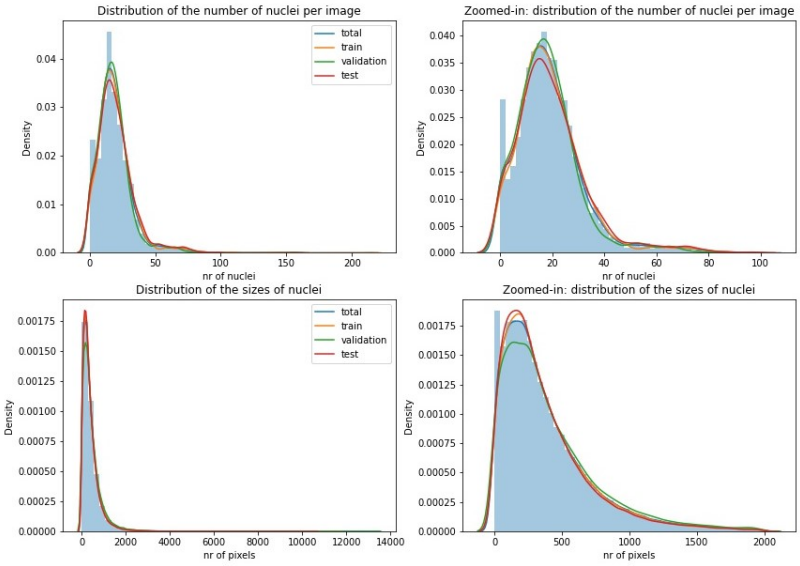


Figure 2. This figure shows: on the left, the distributions of the total number of nuclei in an image and the distribution of the sizes of these nuclei, and on the right, a zoomed-in version of the left.

is already incorporated, and each set contains respectively 2656, 2523, 2722 images. The splits are pre-extracted patches taken randomly to evaluate the models accurately. In each split it is ensured that every tissue type is represented equally based on the proportion of its occurrence in the total dataset (Figure 1). Also, the distribution of the number of nuclei in the images and their sizes are approximately equal across the splits (Figure 2). This research takes the same split to ensure accurate comparison between the scores on the performance metrics of Gamper et al. [4] and this research.

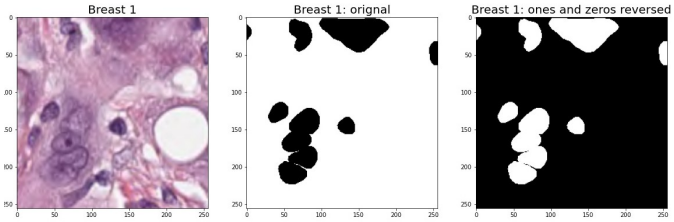


Figure 3. An example of the data where the left shows the image, and the right is the corresponding mask

3.2 Models

The proposed model for the detection and segmentation of nuclei is CE-Net [6], which will be examined by comparing it to U-Net. Figure 4 shows the architecture of the model designed by Gu et al. [6]. The CE-Net model is a modified U-Net model with pretrained ResNet block. It adds a dense atrous convolution block (DAC) in the middle, that can be seen in Figure 4, which could capture deeper or wider feature information, instead of the original 3x3 convolution and pooling operations in the U-Net model. Moreover, different to U-Net, CE-Net applies a residual multi-kernel pooling block (RMP) after the DAC block which further encodes the high-level feature information by performing different-sized pooling operations. A 1x1 convolutional layer ensures the reduction of the dimension of the feature maps flowing from the RMP block. These up-sampled features are concatenated with the original features, before they are fed to the decoding part which is the same as in the U-Net architecture.

Due to the increase of operations and complexity of the CE-Net model, the number of trainable parameters are significantly higher than U-Net: 40M of CE-Net compared to 2M trainable parameters of U-Net.

3.3 Metrics

Evaluation metrics To quantitatively assess the models' performances two metrics have been applied that were most used in related studies, including the F1-score (or DICE-coefficient) (Equation 4), and Jaccard Index, also known as Intersection-over-Union (Equation 2). These metrics measure at pixel-level and match the predicted pixels with the ground truth pixels from which the number of true positive pixels (TP), true negatives (TN), false positives (FP), and false negatives (FN) can be calculated. The F1-score is the harmonic mean of the precision (TP/(TP+FP)) and recall (TP/((TP+FN)) of the segmentated image and is a measure of accuracy (Equation 1). The Jaccard index measures the similarity between the predicted segmentation and the ground truth, in such a way that it divides the area of overlap by the area of union.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$
$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (1)$$

$$Jaccard\ Index = \frac{TP}{TP + FP + FN} \quad (2)$$

Literature on this subject also applied Pixel-Accuracy as one of the metrics to evaluate the segmentation of nuclei [9][22][29]. However, this metric is not suited for the task at hand since most images contain relatively more background than foreground pixels. To give an example, Figure 3 shows the image and corresponding mask of an example from the data. When the model would classify each pixel as none nuclei (0) it would already give an accuracy of approximately 86%, because 7167 pixels of the 50176 (224x224) are nuclei so (50176-7167)/50176 = 0.857.

Instead of measuring the Pixel-Accuracy, this research uses Nuclei-counting Accuracy (Equation 3). This metric

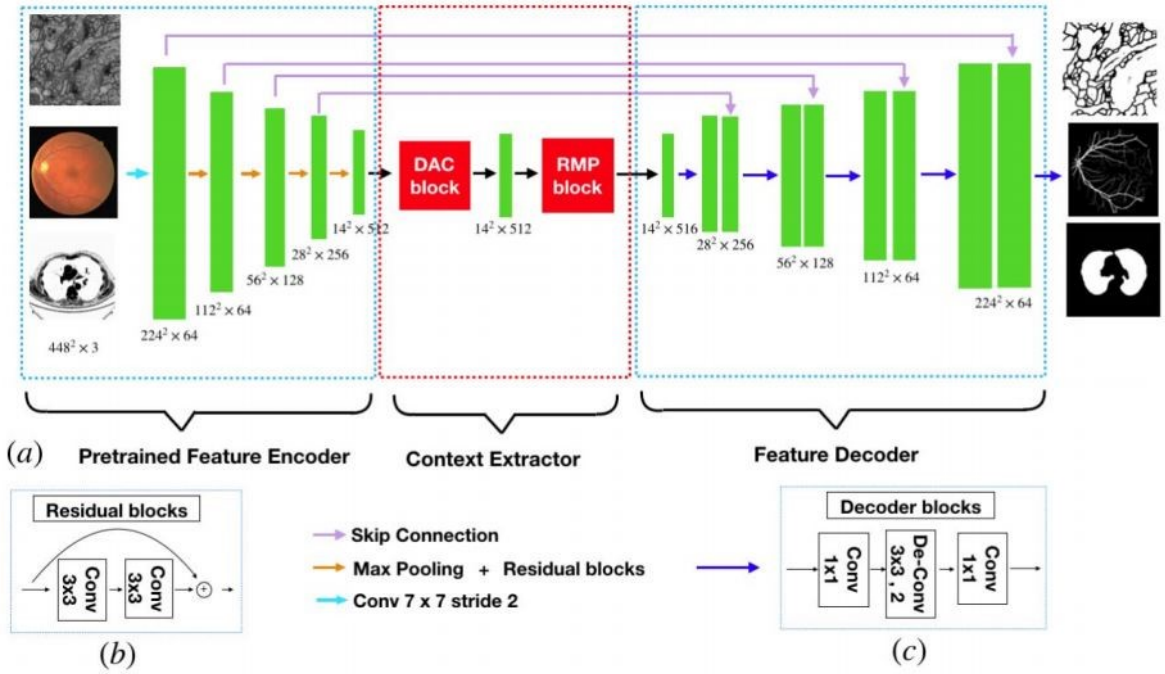


Figure 4. This figure shows an illustration of the CE-Net architecture as designed by Gu et al. [6]. In the first part of the model the images are fed to a pretrained feature encoder model, whereafter the context extractor with the dense atrous convolution block (DAC) and a residual multi-kernel pooling (RMP) block tries to capture high-level feature information. Lastly, these extracted features will enter the decoding part of the model. Here the feature size will be enlarged, and the output presents the mask as the segmentation prediction map [6].

measures at cell-level where the Mean Absolute Difference (MAE) is calculated for the absolute difference between predicted number of nuclei and ground truth number of nuclei for each image. The number of nuclei is calculated by counting the contour lines of each nuclei detected with the fuction findContours() of cv2 package. Subtracting the MAE for the number of nuclei from 1 results in an accuracy measure. This measure is sensitive to outliers, thus the standard deviation is given to get some insight in how the distributions of the MAEs look like for each model. This metric provides an insight in how good the models detect the nuclei in an image rather than how good it segments these nuclei.

$$\begin{aligned} Nuclei\ counting\ Accuracy &= 1 - MAE \\ MAE(nr\ of\ nuclei) &= \frac{\sum_{i=1}^{n} |G_i - P_i|}{n} \end{aligned} \tag{3}$$

The last evaluation metrics that is used in this research is the Panoptic Quality metric (PQ) as originally proposed by Kirillov et al. (2018) [12](Equation 4). This metric is believed to be the latest most optimal performance measure for nuclear segmentation by Graham et al. (2019) [5]. It is a combination of the detection quality (F1-score), which indicates the instance detection performance, and the segmentation quality, where p stands for the predicted and g for ground truth of a segment (Equation 4). The segmentation quality matches the correct predicted segment to the corresponding ground truth, only if the Jaccard index of that predicted segment is >0.5. If not, then this predicted segment is considered as false positive. Thus, in the task of nuclei segmentation a predicted individual nucleus is matched based on ID labels with the ground truth of that particular nucleus. The Jaccard index (IoU) will be calculated for this nucleus and only be considered true positive if the Jaccard index is >0.5. This will be done for all the nuclei in the image and summed up and divided by the true positives before multiplying it with the F1-score (Detection Quality in Equation (4)).

$$PQ = \underbrace{\frac{\sum_{\langle p,g \rangle \epsilon TP} \text{IoU}(p,g)}{TP}}_{\text{Segmentation Quality(SQ)}} \times \underbrace{\frac{TP}{TP + \frac{1}{2}FP + \frac{1}{2}FN}}_{\text{Detection Quality(DQ)}} \tag{4}$$

Loss functions

In most biomedical image segmentation is a pixel-wise classification problem where the goal is to classify each pixel as foreground or background. The loss-function that is most commonly used in these segmentation tasks is binary cross-entropy (Equation 5). Because the foreground pixels in nuclei images occupy only a small region of the image, a different loss-function is proposed [23]. The Dice coefficient loss is suggested to perform better in tasks like nuclei segmentation

since it uses a measure of overlap, and its equation is shown in Equation 6.

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log x_i + (1-y_i)\cdot \log\left(1-x_i\right)\right) \quad (5)$$

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} x_i y_i}{\sum_{i-1}^{N} x_i^2 + \sum_{i-1}^{N} y_i^2} \quad (6)$$

3.4 Experimental Set-up

These links show the notebooks and codes that are used in this study:

- https://github.com/TaarieqDinmohamed/Shared_thesis_design/blob/main/Matty/Final_thesis/unet-tensorflow.ipynb
- https://github.com/TaarieqDinmohamed/Shared_thesis_design/blob/main/Matty/Final_thesis/cenet-tensorflow.ipynb
- https://github.com/TaarieqDinmohamed/Shared_thesis_design/blob/main/Matty/Final_thesis/results-plots.ipynb

The data is stored in Kaggle, a web-based environment, which is also used as the main working directory in this research. On Kaggle there is free access to NVidia K80 GPUs which is used in this research for training and testing the models. While loading in the data into a Kaggle notebook, the images and masks are resized to 224x224 with the nearest neighbor method for the masks to remain the values of the original mask. In the preprocessing phase the '0' and '1' labels of the sixth layer of the masks are reversed to ensure the boolean mask contains 'True' for the nuclei pixels.

The models are implemented in Keras, with Tensorflow backend. In the training phase, image augmentation is applied to the train and validation set where the images are randomly adjusted by zoom, shear angle, rotation, width, and height. During training, an Adam optimizer is used with learning rate 0.001 and a batch size of 30 images. The steps per epoch and validation steps parameters are set to the outcome of dividing the length of the training or validation set by the batch size, to ensure that all images are analysed in the training phase. To create extra stochasticity in the training process shuffling of the images is enabled. The implementation of the evaluation metric Jaccard score and F1-score is done with builtin functions of sklearn package.

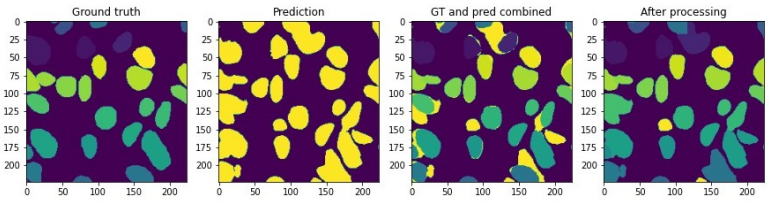


Figure 5. Example of post-processing the predicted mask before calculation of the PQ and Cell-Counting Accuracy metric.

Table 2. The table shows the sensitivity to outliers and noise of U-Net model and CE-Net model.

Removal of:	CE-Net		U-Net	
	F1-score	Jaccard	F1-score	Jaccard
	test set		test set	
-	0.63	0.49	0.66	0.52
Noise	0.73	0.60	0.77	0.64
Outliers area	0.62	0.48	0.66	0.52
Outliers PCA	0.70	0.58	0.74	0.62
Noise + all outliers	**0.74**	**0.61**	**0.79**	**0.67**

For the implementation of the Cell-Counting Accuracy and PQ metric some processing needs to be done. In the first five layers of the masks, which separates the various nuclei types, the nuclei are labeled with IDs that are indicated by different colors in the first plot in Figure 5. These IDs need to be matched to the binary predicted outcomes of the models. This is done by: 1) combining the five mask layers into one while preserving the IDs, 2) merging the predicted and this created ground truth mask on the pixels that are labeled as 1 in the predicted mask (third plot in Figure 5), and 3) change the remaining 1 values of the pixels in the merged mask with their nearest ID (fourth plot in Figure 5). The latter is only applied when the 1 values are not surrounded by zeros, because this could indicate that the model has found a different nucleus, therefore, a different ID label needs to be assigned to this nucleus. Thereafter, the evaluation metrics PQ and Cell-Counting Accuracy can be calculated as described in the Evaluation Metrics section.

4 Results

In this section the results are discussed, where the main question of -How does CE-Net perform in detection and segmentation of nuclei compared to the medical baseline U-Net?- lies central. Furthermore, the results are divided into subsections in which the results of the sub-questions are described.

4.1 Data exploration

An exploratory data analysis provided useful insights into the dataset that served as context to the results, and in particular to answer the first sub-question: How sensitive are the models to outliers and noise? First of all, it became clear that there is some noise in the data where some images are annotated as totally nuclei but in fact there is no nuclei visible on the image since it was completely blank. In total the data consists of 343 images that are not tissues images. Removal of the noise has shown an increase in F1-score and Jaccard index for both models (Table 2).

Furthermore, the outliers are selected based on the total area of nuclei in the images, and a Principal Component Analysis (PCA). The boxplots in Figure 6 show the thresholds for identification of the outliers. There are 36 outliers for

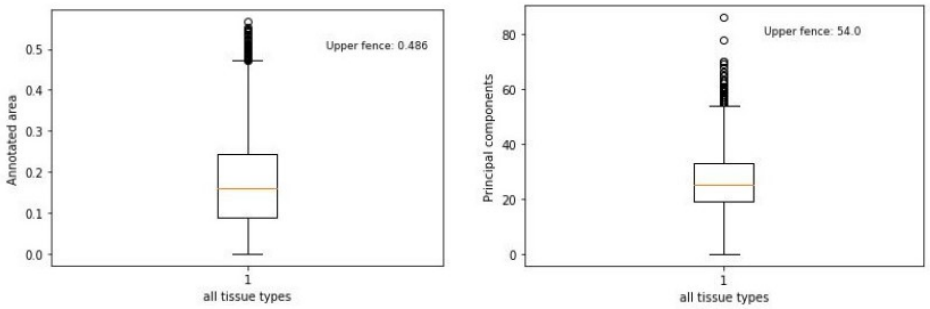


Figure 6. Outlier detection based on the total annotated area of the images and the principal component analysis.

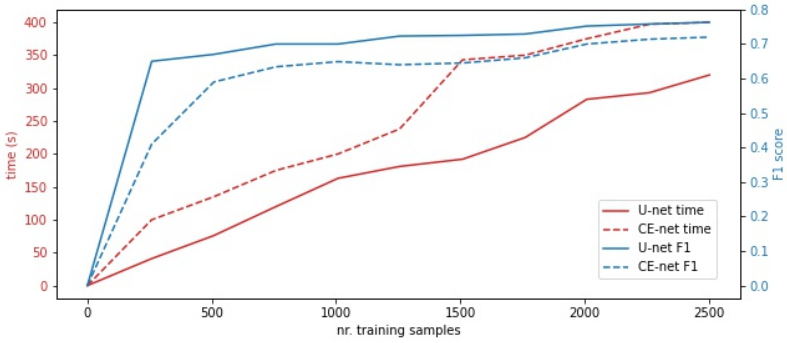


Figure 7. The figure shows the effect of increasing the training samples on the training time (red) and F1-score (blue) for U-Net (solid lines) and CE-Net (dashed lines).

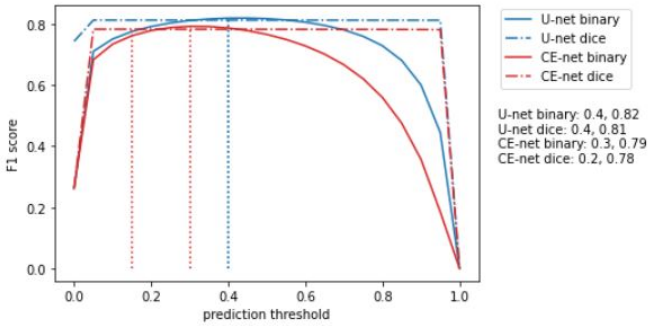


Figure 8. Prediction threshold for the test-set in relation to the F1-score for U-Net and CE-Net to compare robustness of the predictions.

the total area and 66 outliers in the principal component. Removal of the area outliers showed almost no difference while removal of the PCA outliers showed an increase in F1-score and Jaccard index for U-Net and CE-Net (Table 2).

4.2 Training samples

It is speculated that deep-learning models require a large amount of training samples. To see how sensitive U-Net and CE-Net are to a decrease in training samples, and answer the sub-question: Are the F1-scores or training times of the models influenced by a decrease in training samples?, Figure 7 shows the F1-score and training times in relation to the number of training samples. Because the dataset contains 19 tissue types, in which the 'colon' and 'breast' classes are most represented, the training samples are decreased by taking randomly 10% of each tissue type to overcome the problem of having only one or two tissue types in the train set. The models are run for 10 epochs in this experiment. In Figure 7 it is visible that U-Net is faster in training times and shows a higher F1-score compared to CE-Net. Furthermore, the F1-score for U-Net increases only slightly after 250 training samples, while this point for CE-Net lies at 500 training samples. Therefore, we can conclude that U-Net can handle a low amount of training samples a bit better, looking at the F1-scores. However, after 500 training samples the F1-score equally increases for both models.

4.3 Robustness

This subsection describes the results that help to answer the sub-question: How robust are the predictions of both models? The outcomes of the models return weights on which the mask for the test-set can be predicted. These predictions return a number between 0-1 for each pixel, where usually the pixels >0.5 are classified as 1. To check the robustness of the predictions of the models different thresholds are applied in relation to the corresponding F1-score of U-Net and CE-Net, as shown Figure 8. It is clear that the highest F1-score for both models is lower than a prediction level 0.5, namely 0.4 for U-Net and 0.2/0.3 for CE-Net, where U-Net is slightly more robust as it decreases less over an increasing prediction threshold. Also, the models with a dice loss function instead of binary cross-entropy loss show more robust predictions as the dashed line in Figure 7 stabilizes between a prediction threshold of 0.1 and 0.9. The same conclusions regarding the difference between the models and loss-functions can be drawn for the other metrics (Figure 12 and 13 in the Appendix). Note, that the Jaccard Index is not included since it shows the same plot as the F1-score in Figure 8 but all values approximately 0.1 lower.

4.4 Comparison of evaluation metrics

This subsection assesses the difference between the models on the chosen evaluation metrics, and it corresponds to two sub-questions: 1) How is CE-Net performing in the segmentation of nuclei compared to U-Net looking at the F1-score

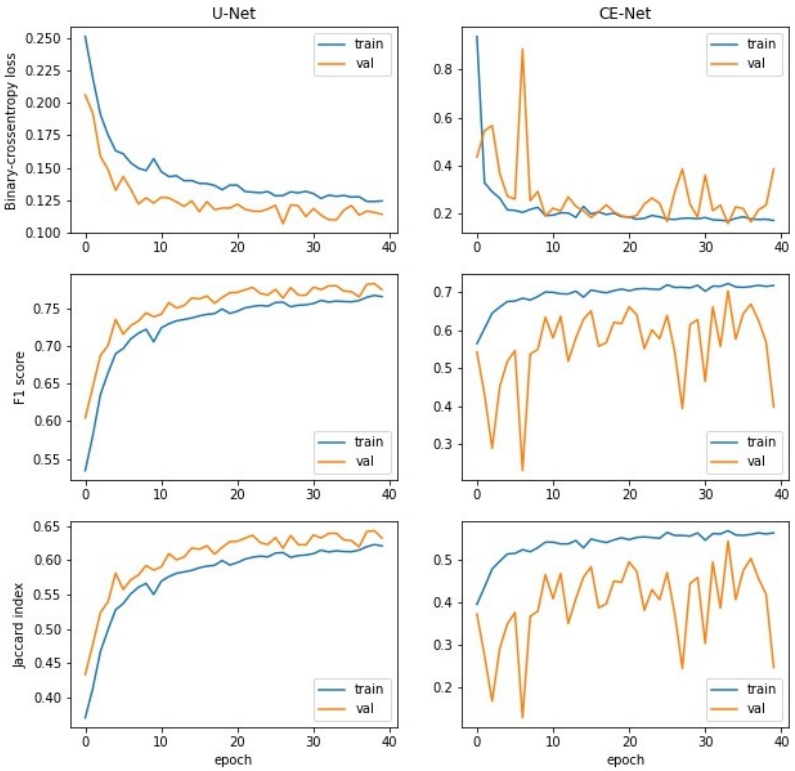


Figure 9. Training and validation assessment of the loss function, F1-score, and Jaccard index for U-Net and CE-Net.

and the Jaccard index scores?, 2) How accurate is CE-Net in detecting the nuclei compared to U-Net looking at the Cell-counting accuracy metric?. Firstly, Figure 9 shows the learning curves of U-Net and CE-Net over 40 epochs. The learning curves gives insight in how fast the models reach the optimal point. It also shows if the model is under- or over-fitting. The latter is the case for CE-Net, where the learning curve of the validation set is fluctuating over the epochs, while the learning curve of the train set is quite stable. In the plots for U-Net in Figure 9 the train and validation do not differ much (<0.05 for F1-score and Jaccard index) and remain relatively stable over the epochs. Moreover, U-Net keeps improving slightly over the epochs while CE-Net improves fast from the start and does not improve much between epoch 5 and epoch 40.

Secondly, Table 3 contains an overview of the performance metrics of the models and the non-trainable baseline where all pixels are labeled as nuclei. This table allows comparison between the two models based on the means and the standard deviation. Table 3 already shows that U-Net has higher outcomes on all evaluation metrics. To draw any statistical conclusion from these values, the confidence intervals are calculated using Equation 7, where n is the length of the test set (2581) after removal of noise and outliers and z corresponds to the confidence interval which in this case is 1.96 for a 95% confidence interval. The confidence intervals are displayed in Figure 10, which makes it clear that the scores on the evaluation metrics are significantly higher for U-Net.

$$Confidence\ Interval = \bar{x} \pm z\frac{s}{\sqrt{n}} \tag{7}$$

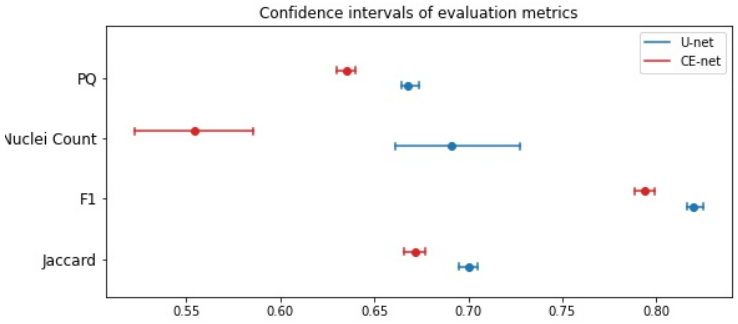


Figure 10. The 95% confidence intervals calculated with the mean and standard deviation of the evaluation metrics of U-Net and CE-Net in Table 3, using Equation 7

.

4.5 Comparison with relevant work

In this last subsection the results of the sixth sub-question will be discussed: Are the results on the PanNuke dataset of this research similar to the results of Gamper et al. [4], when comparing the Panoptic Quality metric? The data that is used for the experiments is the PanNuke dataset, which is also reseached by Gamper et al. [4]. They applied the Panoptic Quality (PQ) metric to evaluate the models' performances, where their lowest score was 0.53 for the DIST model and the highest score was 0.66 gained by a HoVer-Net model [4]. To put this into perspective, the results from this research showed a PQ of 0.67 for U-Net and 0.64 for CE-Net. Calculating the 95% confidence interval for the models from Gamper et al. [4] gives [0.534, 0.535] for DIST model and [0.659, 0.660] for HoVer-Net. Compared to the confidence intervals of the PQ in Figure 10 it is clear that: 1) the outcomes of CE-Net are lower than the HoVer-Net from Gamper et al. [4] but higher than their DIST model, 2) U-Net slightly outperforms HoVer-Net by 0.004, comparing the lower confidence boundary of U-Net (0.664) with the upper confidence boundary of HoVer-Net (0.660).

5 Discussion

A question that arises from the results is; why is CE-Net not outperforming U-Net as was expected? CE-Net is a more complex model than U-Net, therefore, could potentially capture more high-level information from the features. However, due to this complexity CE-Net is also sensitive to overfitting, in other word a reduction of the generalizability of the

Table 3. Performances comparison between U-Net, CE-Net, and a non-trainable baseline where all pixels are predicted as nuclei (mean / standard deviation).

Loss-function		Jaccard Index	F1 score	Nuclei-Counting Accuracy	Panoptic Quality
		test set	test set	test set	test set
U-Net	Binary crossentropy	**0.70 / 0.13**	**0.82 / 0.11**	0.69 / 0.77	**0.67 / 0.12**
	Dice loss	0.70 / 0.14	0.81 / 0.12	**0.69 / 0.75**	0.65 / 0.13
CE-Net	Binary crossentropy	0.67 / 0.14	0.79 / 0.14	0.55 / 0.82	0.64 / 0.13
	Dice loss	0.66 / 0.15	0.78 / 0.14	0.65 / 0.80	0.62 / 0.14
Non-trainable baseline: all nuclei		0.16 / 0.10	0.26 / 0.14	-	-

model. The first and second row of plots in Figure 14 in Appendix show that CE-Net captures more complex patterns but wrongly classifies the patterns as nuclei. This could be an indication that the model is still overfitting. Moreover, the significant difference between the Nuclei-Counting Accuracy metric could be explained by the plots in the first row of Figure 14 in Appendix, here the number of nuclei in the ground truth mask is 3 and CE-Net predicts 40 nuclei leading to an Nuclei-Counting accuracy of -11.0. On the other hand, the capture of high-level information as was expected from CE-Net is shown in the third row of plots in Figure 14 in Appendix where CE-Net is able to capture accurate information from the image while U-Net is not.

A limitation of this research is that the hyper parameters are tuned by trial and error, decreasing the learning rate from 0.001 to 0.0001 stabilizes the fluctuations of the validation set (Figure 9) but resulted in a lower outcome on the performance metrics after 40 epochs, but also after 70 epochs. An improvement to the learning rate hyper parameter is adding a scheduler which adjust the learning rate according to a predefined scheduler during the training phase.

Another hyper parameter that could stabilize the fluctuations of the validation set in the learning curve in Figure 9 and improve the outcomes on the performance metrics is the batch size. Smith et al. [27] argues that increasing the batch size during training achieves a similar learning curve to the learning curve with a scheduler learning rate. Thus, improvements could be made on automatically tune the hyper parameters to find the optimal settings that prevent overfitting and improving the learning curve while gaining high outcomes on the performance metrics.

During the experiments, some potential errors were found in the ground truth labels that could question the reliability of the results. Figure 11 visualizes the nuclei that were detected and segmented by both models, but were not indicated as nuclei in the ground truth label. The sensitivity of the models have been analysed in the results and showed that both models are sensitive to noise and outliers in the data, where removal of all outliers and noise resulted in an increase of 0.11 and 0.13 on the F1-score for CE-Net and U-Net respectively. This could be an explanation for the low outcomes on the Nuclei-Counting accuracy metric. However, the extent or the significance of these errors to the models' performance metrics have not been researched. Future research could give more insight in these potential errors. Also, applying the models to a multi-organ nuclei dataset that has been researched more extensively could add valuable insights onto the results of this research on the PanNuke dataset.

6 Conclusion

In this research, CE-Net is proposed as the model that could potentially outperform the state-of-the-art U-Net model in the detection and segmentation of nuclei in the PanNuke dataset, since it is able to capture more high-level feature information because of the different encoding blocks. Contrary to the expectations, U-Net had higher scores on all evaluation metrics that were used in this research. Furthermore, U-Net required less training time compared to CE-Net and is slightly better at handling <500 training samples. Furthermore, U-Net would be considered more robust as it is less responsive to different prediction thresholds.

Nevertheless, the reflection on the results has shown that CE-Net is in fact able to capture more complex patterns in the dataset. But due to overfitting it also captures patterns that are not relevant. Also, there were some potential errors found in the dataset which may have influenced the outcomes. In future research it would be interesting to determine the significance of these errors in the PanNuke dataset. Furthermore, future research could improve the tuning of the hyper parameters to determine if CE-Net is able to produce more accurate predictions. Also, it would be interesting to see how CE-Net is performing on a less diverse dataset to reflect on its generalizability.

References

[1] Lei Cai, Jingyang Gao, and Di Zhao. 2020. A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine* 8, 11 (2020).

[2] Juan C Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J Theis, et al. 2019. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A* 95, 9 (2019), 952–965.

[3] Junlong Cheng, Shengwei Tian, Long Yu, Hongchun Lu, and Xiaoyi Lv. 2020. Fully convolutional attention network for biomedical image segmentation. *Artificial Intelligence in Medicine* 107 (2020), 101899.

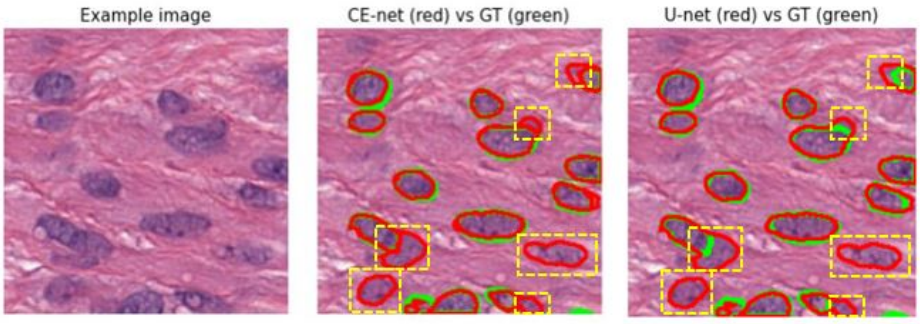


Figure 11. Nuclei contour predictions of CE-Net and U-Net on an example image from the test set in red, and the ground truth in green. The yellow boxes show predicted nuclei that are not present in the ground truth.

[4] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. 2020. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778* (2020).

[5] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis* 58 (2019), 101563.

[6] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. 2019. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging* 38, 10 (2019), 2281–2292.

[7] Intisar Rizwan I Haque and Jeremiah Neubert. 2020. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked* 18 (2020), 100297.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[9] Chuanbo Huang, Huali Ding, and Chuanling Liu. 2020. Segmentation of cell images based on improved deep learning approach. *IEEE Access* 8 (2020), 110189–110202.

[10] Jeremiah W Johnson. 2019. Automatic nucleus segmentation with Mask-RCNN. In *Science and Information Conference*. Springer, 399–407.

[11] Atilla P Kiraly, Clement Abi Nader, Ahmet Tuysuzoglu, Robert Grimm, Berthold Kiefer, Noha El-Zehiry, and Ali Kamen. 2017. Deep convolutional encoder-decoders for prostate cancer detection and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 489–497.

[12] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9404–9413.

[13] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, et al. 2019. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging* 39, 5 (2019), 1380–1391.

[14] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging* 36, 7 (2017), 1550–1560.

[15] SN Kumar, A Lenin Fred, S Muthukumar, H Ajay Kumar, and P Sebastian Varghese. 2018. A voyage on medical image segmentation algorithms. (2018).

[16] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. 2018. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 2 (2018), 400–416.

[17] Feixiao Long. 2020. Microscopy cell nuclei segmentation with enhanced U-Net. *BMC bioinformatics* 21, 1 (2020), 1–12.

[18] Conor McKeen, Fatemeh Zabihollahy, Jinu Kurian, Adrian DC Chan, Dina El Demellawy, and Eranga Ukwatta. 2019. Machine learning-based approach for fully automated segmentation of muscularis propria from histopathology images of intestinal specimens. In *Medical Imaging 2019: Digital Pathology*, Vol. 10956. International Society for Optics and Photonics, 109560P.

[19] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. 2017. Nuclei segmentation in histopathology images using deep neural networks. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE, 933–936.

[20] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging* 38, 2 (2018), 448–459.

[21] Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* 9, 1 (1979), 62–66.

[22] Narinder Singh Punn and Sonali Agarwal. 2020. Inception u-net architecture for semantic segmentation to identify nuclei in microscopy cell images. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 1 (2020), 1–15.

[23] Xiaofei Qin, Chengzi Wu, Hang Chang, Hao Lu, and Xuedian Zhang. 2020. Match Feature U-Net: Dynamic Receptive Field Networks for Biomedical Image Segmentation. *Symmetry* 12, 8 (2020), 1230.

[24] SN Rashid, MM Fraz, and S Javed. 2020. Multiscale Dilated UNet for Segmentation of Multi-Organ Nuclei in Digital Histology Images. In *2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*. IEEE, 68–72.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[26] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. 2020. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical physics* 47, 5 (2020), e148–e167.

[27] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. 2017. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489* (2017).

[28] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali

Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. 2019. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology* 7 (2019), 53.

[29] Haonan Wang, Yinhan Li, and Zhiyi Luo. 2020. An Improved Breast Cancer Nuclei Segmentation Method Based on UNet++. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*. 193–197.

[30] Lipeng Xie, Jin Qi, Lili Pan, and Samad Wali. 2020. Integrating deep convolutional neural networks with marker-controlled watershed for overlapping nuclei segmentation in histopathology images. *Neurocomputing* 376 (2020), 166–179.

[31] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*. Springer, 399–407.

[32] Xiaodong Yang, Houqiang Li, and Xiaobo Zhou. 2006. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy. *IEEE Transactions on Circuits and Systems I: Regular Papers* 53, 11 (2006), 2405–2414.

[33] George Zaki, Prabhakar R Gudla, Kyunghun Lee, Justin Kim, Laurent Ozbun, Sigal Shachar, Manasi Gadkari, Jing Sun, Iain DC Fraser, Luis M Franco, et al. 2020. A Deep Learning Pipeline for Nucleus Segmentation. *Cytometry Part A* 97, 12 (2020), 1248–1264.

[34] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. 2018. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7151–7160.

[35] Bingchao Zhao, Xin Chen, Zhi Li, Zhiwen Yu, Su Yao, Lixu Yan, Yuqian Wang, Zaiyi Liu, Changhong Liang, and Chu Han. 2020. Triple U-net: Hematoxylin-aware nuclei segmentation with progressive dense feature aggregation. *Medical Image Analysis* 65 (2020), 101786.

7 Appendix

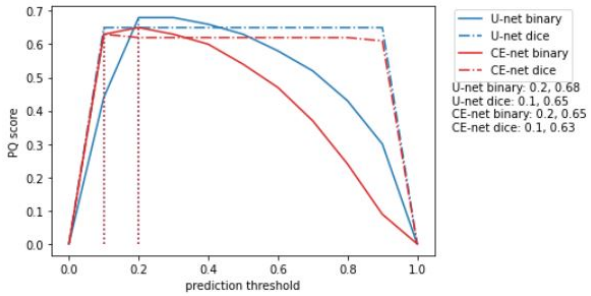


Figure 12. Prediction threshold PQ score to check robustness of predictions of both models.

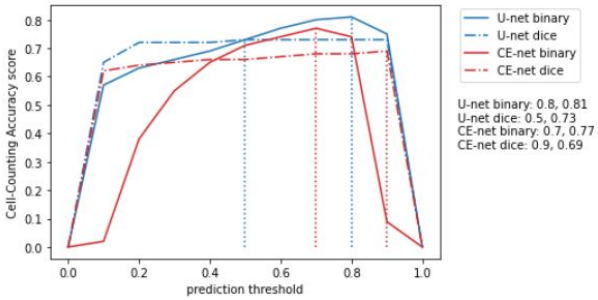


Figure 13. Prediction threshold PQ score to check robustness of predictions of both models.

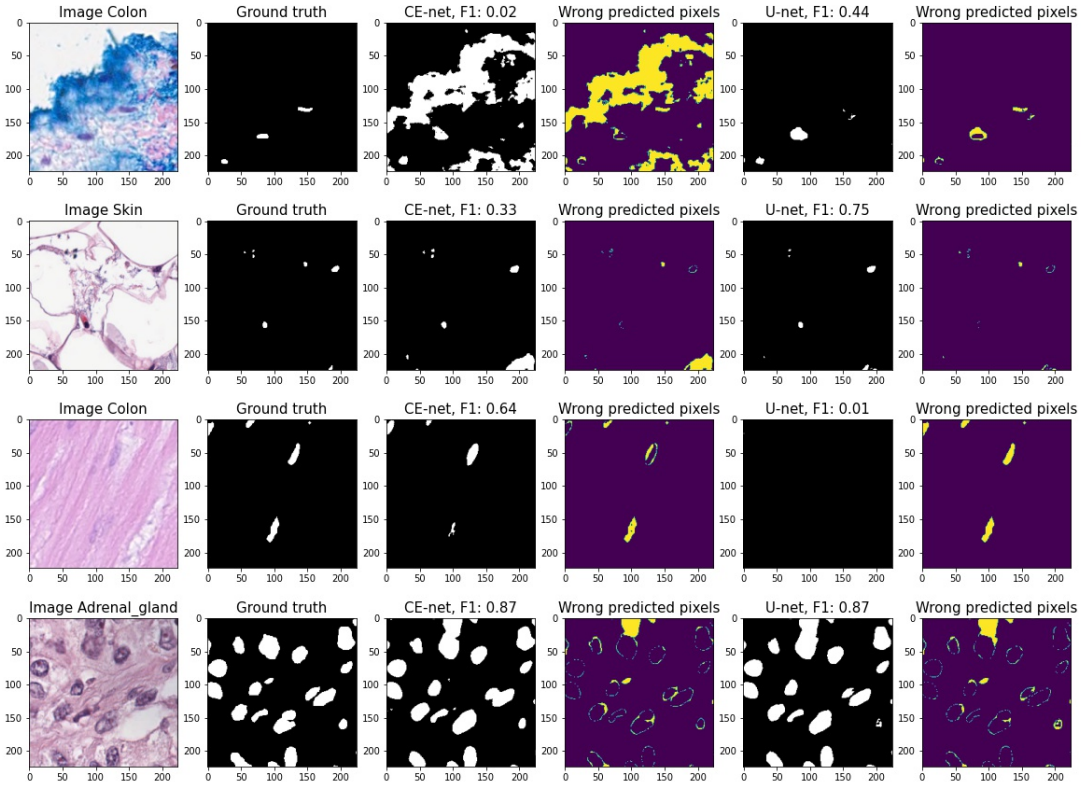


Figure 14. Some example images where in the first three rows of plots the difference between the predictions of CE-Net and U-Net is >0.40 based on the F1-score. And the last row of plots shows an example where the difference between CE-Net and U-Net is approximately 0.