Weakly Supervised Semantic Cell Segmentation Using Knowledge Distillation

Ivana Háberová[*]
Ivan Vykopal[†]
Supervised by: Dr. Lukáš Hudec[‡]

Faculty of Informatics and Information Technologies
Slovak University of Technology
Bratislava / Slovak republic

Abstract

This study proposes a new approach for semantic cell segmentation that combines the use of neural networks and involving humans in the loop with the aim of improving the current state of digital pathology. The goal is to obtain cell segmentation and classification from heart biopsy images based on inaccurate data and simultaneously to reduce the demands on domain experts - doctors. In the first step, the approach utilizes a segmentation model and a combination of different datasets to detect the nuclei of cells in the patches of whole slide images, which are used to increase the amount of data. The proposed approach employs knowledge distillation, a technique that involves training a smaller "student" model to mimic the output of a larger "teacher" model and their chaining. This is done to overcome the limitations of having a small amount of accurate data and a high proportion of inaccurate annotations and to remove inaccuracies through chaining. The proposed approach is evaluated against traditional methods and shows that it achieves improved performance in terms of semantic cell segmentation. This demonstrates the potential for the approach to be applied in biomedical image analysis, where accurate and precise segmentation is essential for downstream analysis.

Keywords: Segmentation, Classification, Knowledge Distillation, Human-in-the-Loop, Weakly Annotated Data, Digital Pathology

1 Introduction

The analysis of whole slide images is one of the important components of pathologists' diagnosis of Cardiovascular diseases (CVDs). Research in this area is also progressing because $\sim$ 17.9M people die each year from CVDs, according to WHO[1], which is approximately one-third of all deaths worldwide. CVDs are heart or blood vessel diseases, such as coronary heart disease, cerebrovascular disease, and rheumatic heart disease. A heart biopsy is an effective way to detect changes in the heart muscle. On the other hand, this procedure is invasive, difficult for the patient - especially if heart problems occur - and requires sufficient time for sample collection, tissue processing and following evaluation by a doctor. Analyzing images after a heart biopsy can be a challenging task, as the tissue samples are often small and may be difficult to interpret.

Over the past 20 years, the field of pathology has made significant advancements in digital imaging through the development and improvement of whole-slide imaging. Digital pathology is a technology that can benefit from high-resolution digital images to aid in diagnosis and treatment planning. It is becoming increasingly popular in pathology departments, offering advantages over traditional, microscope-based methods of analyzing tissue samples. A combination of machine learning and digital pathology can automate image analysis and hence has the potential to revolutionise the field of pathology by improving diagnostic accuracy, increasing efficiency, and reducing costs. Currently, there are several automated tools providing a biomedical image or biomarker analysis like QuPath [4], MONAI [8], CellProfiler [14], or ImageJ [2].

This work presents a novel training strategy for weakly annotated data applied in semantic cell segmentation from histopathological heart biopsy images based on imprecise annotations. The motivation is mainly to reduce the demands on doctors, who can more easily detect problematic areas based on accurate classification or, in conjunction with rigorous quantitative analysis, detect small deviations earlier and thus bring new knowledge to the given area. The research is carried out in cooperation with experts from the Institute for Clinical and Experimental Medicine in Prague (IKEM).

To summarize, our main contributions are: (1) a robust model for nuclei segmentation in different organs and resolutions; (2) an approach that classifies cells in histopathological data using knowledge distillation; (3) a teacher model usable for training other students for different types of data and different types of annotations; (4) a relatively

[*] ivankahaberova@gmail.com
[†] ivan.vykopal@gmail.com
[‡] lukas.hudec@stuba.sk

small network that is well adapted to a specific task (cell classification in H&E images).

2 Related work

Identifying individual tissue types or small cells in a histopathological image is time-consuming and requires experienced doctors. The solution to this pathology image analysis challenge can be using deep learning algorithms that can process and evaluate the images quickly and segment effectively the needed areas - tumours, tissues or cells. The main task of (binary) segmentation is to distinguish the searched tissue from its surroundings, while there is a semantic segmentation, where individual segmented tissues have a different meaning. The encoder-decoder architectures with interconnections known as U-Net [17] proved effective, outperforming previously used methods - a combination of sliding window and convolutional networks. U-Net achieves quantitatively and qualitatively good results, even on a small amount of biomedical data with extensive augmentation. Olaf Ronneberger et al. [17] applied U-Net to a cell segmentation task in light microscopic images as part of the ISBI cell tracking challenge 2015, where they achieved on different partially annotated datasets average IoU (Intersection over Union) 77.5% and 92% and significantly outperformed other algorithms. The extension and improvement of the performance come with the deeply-supervised architecture of U-Net++ [20], which added more connections between the encoder and the decoder along with the intermediate outputs.

Classifying cells in histology images is challenging due to the high intra-class variability and inter-class similarity. Many papers deal with this problem using various modifications of convolutional neural networks (CNN). The first significant improvement in CNN results came with VGGNet [19], which demonstrated not only the positive influence of model depth on classification success but also the advantages of using relatively small reception fields (convolutional filters with size 3×3). VGG-16 and VGG-19 versions differ in the number of VGG blocks (16 vs 19), where one VGG block consists of several convolution layers followed by a max-pooling layer. In 2016 He et al. [12] proposed using residual blocks in the neural network. Applying skip connections or shortcuts made it possible to go deeper with the architecture and increased the network's learning ability. Similar to VGG, there are several versions of the Deep residual network architecture or ResNet, such as ResNet-18, ResNet-50 or ResNet-152. Xception [7] model outperforms on ImageNet classification dataset many state-of-the-art models such as VGG-16, ResNet-152 or Inception V3. This architecture is based entirely on depthwise separable convolution layers, which provide great computational efficiency. With the goal of application in diagnostics, where you cannot rely on high-performance computers, some authors try to design models "as simple as possible". This is the case with RCCNet [5], which was created with the aim of colon cancer nuclei classification and has 1.5M learnable parameters compared to VGG-16 with 138M parameters.

To deal with weakly annotated data, there is the human-in-the-loop method, based on domain experts' involvement in interacting with artificial intelligence to obtain more accurate annotations[16]. Most of the research involving experts in the process consists of three main phases: training the preliminary model, predicting unseen data with the preliminary model, and correcting predicted annotations using domain experts. Predictions and corrections are performed in the loop until certain conditions are met. Annotations can be corrected not only by experts but also through crowdsourcing, either by manual correction from a domain expert or by marking them as correct or incorrect [11, 3].

With weakly annotated data or small amounts of data, training a robust model that achieves the expected results is challenging. For this reason, different approaches are used to utilize the currently available resources optimally. A knowledge distillation approach works with weakly annotated data, transferring knowledge between two or more models. The principle of this approach, called Teacher-Student architecture, is to train a Teacher on a small amount of data or weakly annotated data and then train the student using the trained Teacher. In this case, the annotations obtained by the Teacher are used in training the student. Several methods are based on Teacher-Student architectures, including Teacher-Student chaining[18] or substituting Teacher and Student in the training process[6]. Both methods aim to use weak or insufficient annotations to train the best possible model well generalized to the desired task.

Pathologists in IKEM use QuPath for analyzing data - nuclei and higher morphological structures. QuPath can segment cells using parametric methods like color thresholding based on H&E staining for segmentation, whereas, for classification, there are three methods: K-nearest neighbors, Random Forest, and Artificial Neural Network (ANN). The disadvantage of this tool is the excessive dependence of the results on the initial setting by the doctor, which may differ each time based on different concentrations of staining color and therefore result insufficient.

3 Dataset

In our study, we work with two publicly available histological datasets Lizard[9], MoNuSeg[13] and custom dataset based on IKEM data. All three datasets contain histological images stained with hematoxylin and eosin staining. Images from each dataset are shown in Figure 1. By combining them, we created the largest dataset for nuclei segmentation in multiple organs with different magnifications to create a robust model for segmentation.

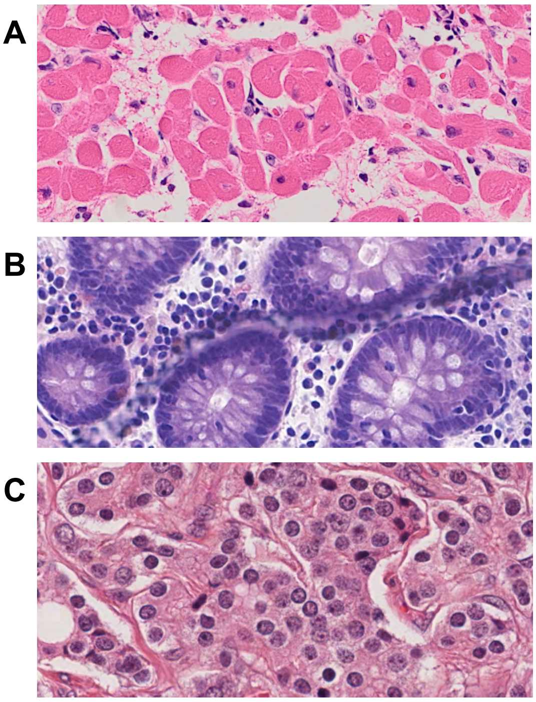

Figure 1: Comparison of datasets: A - IKEM, B - Lizard, C - MoNuSeg after normalization using Macenko method.

3.1 Lizard

Lizard dataset [9] consists of regions of interest (ROI) from Whole Slide Images (WSI) scans of the colon region. This dataset is designed to segment and classify nuclei and consists of six different datasets. Lizard is the largest histological dataset, with 238 ROI images and approximately 495K nuclei. The nuclei annotations were generated based on a multi-step approach consisting of segmentation and classification by the HoverNet [10] neural network to refine automatic and semi-automatic predictions, with the pathologist involved throughout the workflow to refine the segmentations and classifications. Table 1 shows the number of nuclei in the images.

3.2 MoNuSeg

MoNuSeg is a multi-organ dataset and contains 37 histological images together with their annotations. Similar to the Lizard dataset, we work with binary annotations to segment the nuclei. Table 1 shows the total number of nuclei in this dataset.

3.3 IKEM dataset

The IKEM dataset contains 25 WSI scans, each comprising three to five tissue sections (called fragments) from the heart region. The WSI format can store information about a given tissue in several resolutions with relatively small memory requirements, where the highest reaches dimensions up to 48724 × 17910.

Dataset	Nuclei count	Immune cells	Muscle cells	Other cells
Lizard	495 179	-	-	-
MoNuSeg	21 623	-	-	-
IKEM SSA	470 563	118 950	130 022	221 591
IKEM WSA	469 591	127 521	127 259	214 811
IKEM EA	6 834	1 947	2 794	2 093

Table 1: The number of nuclei in each dataset along with their classifications.

Cooperating pathologists provide us:

- QuPath project with trained object detectors and classifiers
- cell annotations as GeoJSON-s obtained by QuPath automatically
- several Artificial Neural Networks (ANN) classifiers
- 4 335 manual annotations on 6 WSI scans with 6 834 nuclei. We call these annotations expert annotations (EA).

The pathologists pre-trained the ANN classifier by iterative manual correction of cell classifications. We used this classifier to generate strong synthetic annotations (SSA). Then, we randomly selected one significantly weaker classifier from the previous iterative improvements and generated weak synthetic annotations (WSA). The resulting distributions of immune, muscle, and other cells are shown in Table 1.

4 Proposed method

Our study aimed to create a comprehensive approach for analyzing nuclei in histological images, from segmentation to classification and applicable to various tissue and organ types.

We focus on the problem of weak annotations that are generated by the QuPath tool using an Artificial Neural Network that has been trained by doctors. Our method's objective is to leverage weak annotations with minimal demands on doctor input effectively.

4.1 Data preprocessing

All data provided by IKEM, whether obtained automatically by QuPath or expert annotations, have the first processing step in common.

Preprocessing 1 (Fig. 2) consists of several steps to ensure efficiency and fast data processing by neural networks. The data are stored as multidimensional matrices, which greatly increases the memory requirements. For this reason, we chose to save in three resolutions - original, ½, and ¼. At the same time, with the aim of reducing memory requirements, images are divided into fragments, where

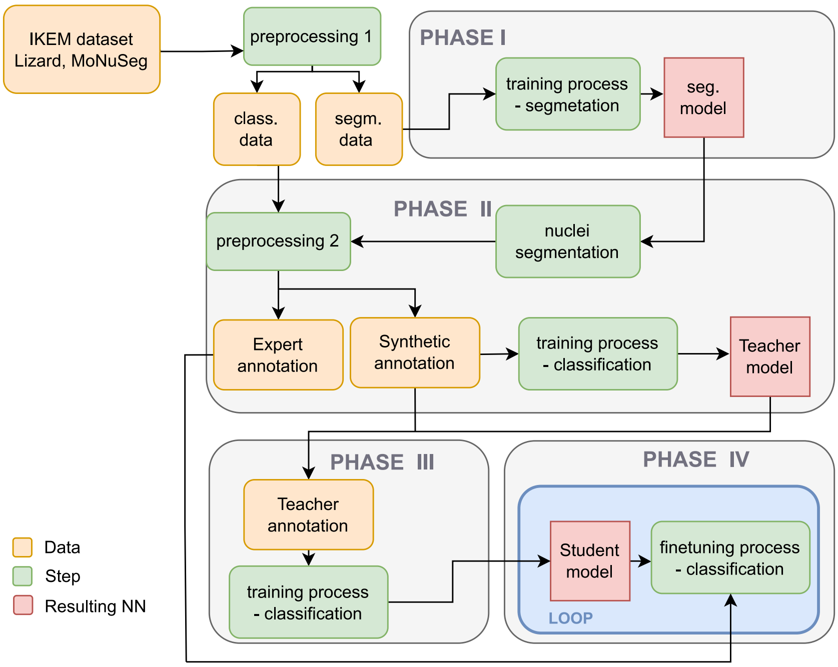

Figure 2: Overview of the proposed approach showing individual training phases, together with inputs, data flow and outputs. The result offers three trained models - one cell segmentation and two cell classification models.

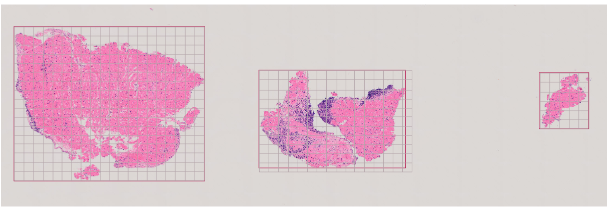

Figure 3: WSI scan with 3 fragments (red marked) and patch subdivision (grey marked). In preprocessing, only individual tissue fragments are stored based on their bounding boxes, so an unnecessary white area is omitted.

the position is chosen according to the smallest possible bounding boxes of individual parts of the tissue.

After saving the fragments as separate entities in different resolutions, optional image data normalization follows. We chose Macenko normalization [15] to reduce bias between datasets, which arose as a result of staining in another laboratory.

The proposed approach needs annotations in the format of multidimensional masks, so the original GeoJSONs are converted to multichannel images. Each channel contains information about one type of desired nuclei.

The images and corresponding annotations (of all datasets) are divided into patches of the selected size, in this case, 512×512, where we use the sliding-window approach without an overlay as shown in Fig. 3. Lizard and MoNuSeg images are stored as binary masks (1-nucleus, 0-background), while images of incompatible dimensions were zero-padded. IKEM data is converted to binary form when loading images during segmentation training with regard to their further use in classifications.

The second part of preprocessing (preprocessing 2 in Fig. 2) uses nuclei segmentation from Phase I. (Fig. 2). For each nucleus segmented, we identified a class using generated synthetic annotations and expert annotations. We assigned each nucleus to one of the classes: other (0), immune (1), muscle cells (2) and background (3). The background class was assigned to nuclei identified by our model but not segmented and classified by QuPath.

To identify the class of each nucleus, a patch is generated around it and SSA and EA are utilized to determine its classification. Patch size 16×16 is used for multiple and 32×32 for the original image size. For the original size, we used a larger patch size to contain only one nucleus, as opposed to multiscale data, where in some cases, a 16×16 patch contains more than one nucleus in the smallest magnification and only part of it in the highest.

When training models in all steps, the data is divided with a random distribution in the ratio of 70:15:15 into training, testing and validation parts.

4.2 Nuclei segmentation

In the initial stage of our comprehensive approach, we aim to segment nuclei in histological images.

In Phase I (Fig 2), we experimented with two traditional architectures, U-Net and U-Net++, commonly used in medical data segmentation. We modified both architectures by replacing the Upsampling layer with the inverse convolution layer, ConvTranspose. The benefit of using

ConvTranspose, a form of learning upsampling, is that it results in a larger number of trainable parameters, leading to a more robust model with a larger capacity.

We train models with all three datasets and a total of 1.93M nuclei in 512×512 patches. We trained our models using the Dice Loss function and Adam optimizer. The training process lasted for 30 epochs using 0.3 as dropout and 0.0001 as the learning rate. The best model was selected based on the validation loss function.

4.3 Classification

The following steps in this approach focus on obtaining the classification of the cells found in the images in the IKEM dataset using the Knowledge distillation approach. The task of the resulting model is to classify cells into three classes - muscle cells, immune cells and others. We chose teacher-student chaining with three models, where first, the weak ANN trains the complex ResNet-18 model in phase II. Subsequently, in phase III, the teacher transfers the information to the RCCNet model, fine-tuned in the phase IV by involving an expert.

The chosen loss function in all classification-trainings is CrossEntropy, optimized by Adam over 30 epochs.

4.3.1 Classification - Teacher

In Phase II, a network called a teacher is trained, which is necessary for the next process. After training on weakly annotated data (WSA and SSA), the goal is to obtain a robust classification model that can extract the essential information from weak annotations. The task consists in assigning each of the cells (based on our nucleus segmentation) to one of the classes required by the experts.

The specificity of this step also lies in the created dataset (described in 4.1 Data Preprocessing), where based on its different versions (various patch-size, normalization, various resolutions) more experiments were performed.

The selection of architecture focuses on state-of-the-art classification models with a large learning capacity, such as ResNet-16, ResNet-18, Xception or VGG. The ResNet architecture was changed for the needs of the chosen patch size, and the VGG architecture was modified (using only one VGG block with two linear layers) to process small images and preserve information efficiently. All these architectures are trained on multiscale data only, which leads to using the best architecture to classify data in the original magnifications to perform all the following steps in experiments.

In the training process, the demands on doctors are significantly reduced, especially by the processing of a weakly annotated dataset created by QuPath, but also by the fact that we do not require designing or setting parameters as QuPath does. The result is a classification model that can obtain enough important information and features from weakly annotated data and can thus be used for creating annotations in the next step.

4.3.2 Classification - Student

Our proposed method is training a network called Student to obtain a relatively small classification model specified for the given task. The goal is the same as when training the teacher (described above) - classifying cells based on small patches of images of a heart biopsy. However, the access to the provided data and the training process - aimed at transferring relevant knowledge from the teacher to the student, followed by specification by adding an expert to the loop - is different. We can divide this step into two phases: training (phase III) and fine-tuning (phase IV).

During the training process, the same patches of images are used for training as in the initial teacher training. The difference is that the teacher determines the label for each segmented nuclei. When training a student, we work with the highest resolution and use a teacher who has been trained on images with the highest resolution.

We chose RCCNet as the student architecture. The training duration depends on the network size, which is important to optimize as much as possible. The evaluation is done after each epoch against the SSA, where the model achieving the best results is saved.

In the last Phase IV, an already partially trained student model is trained again on expert annotations with the aim of precise specification for a given task - refining predictions by applying the human-in-the-loop approach.

Fine-tuning differs from training mainly in the data and its processing, where expert annotations are used in this step. When training, instead of using all the data, we look for a suitable lower limit of the count of cells (the same count from each cell class), which is necessary for a sufficient improvement of the results. To prevent overfitting, which could occur with a small amount of data, some experiments with different hyperparameters settings were performed with values for learning rate from interval $\leq 0.00005, 0.001 \geq$ and dropout rate from $\leq 0, 0.5 \geq$.

The student obtains more accurate information from reliable annotations, which should be reflected in better semantic segmentation and reducing or eliminating the error from the original weakly annotated data. The result is a relatively small network that is well adapted to a specific task, and at the same time, it can be quickly and efficiently modified by the next round of fine-tuning.

5 Evaluation

We ran all our experiments on a computer with a graphic card NVIDIA RTX 2080 Ti with 11GB of GPU RAM and 32GB of total memory RAM.

5.1 Nuclei segmentation

For segmentation, we trained U-Net and U-Net++, where the final nuclei segmentation model is chosen based on the metrics achieved on the test set. Based on the value of

Model	Accuracy [%]	Precision [%]	Recall [%]	Dice [%]
U-Net	97.75	72.19	90.56	79.66
U-Net++	97.44	67.92	92.71	78.04

Table 2: Evaluation of segmentation models U-Net and U-Net++ on test data.

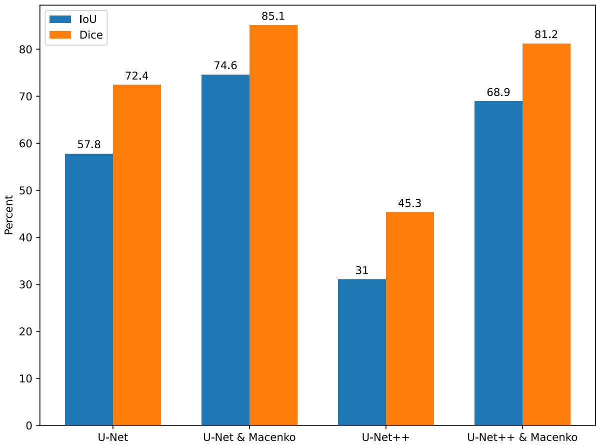

Figure 4: Segmentation results on WSI scans with and without normalization from the IKEM dataset against QuPath segmentation.

Dice over the test set, which was 79.66% for U-Net and 78.04% for U-Net++, we selected the trained U-Net model for further processing. Results are shown in Table 2.

The proposed method for nuclei segmentation was qualitatively evaluated and compared to the results obtained through automatic segmentation using QuPath. The results showed that the proposed method performed slightly better than QuPath. QuPath's approach for segmentation relies on defining threshold values for each staining and following statistical methods.

To further evaluate the segmentations, we quantitatively compared the segmentations produced by our models on WSI scans with the QuPath tool segmentations using the metrics Intersection over Union and Dice score. The results of U-Net and U-Net++ on data without and with normalization using Macenko's method are presented in Figure 4. The evaluation of WSI scans revealed that in both cases, the U-Net architecture performed better than U-Net++, both without image normalization and with normalization using Macenko's method.

5.2 Classification - Teacher

We evaluated strong and weak synthetic annotations against expert annotations, using metrics such as F1 Score, Accuracy, Precision and Recall. The evaluation was based on the cell segmented by the U-Net model. During this process, 574 segmented nuclei were identified as background. The results presented in Table 3 showed that weak synthetic annotations perform better than strong ones.

Data	F1 [%]	Accuracy [%]	Precision [%]	Recall [%]
SSA	82.01	81.82	82.44	81.82
WSA	82.34	82.23	82.56	82.23

Table 3: Comparison of classifications obtained using strong and weak synthetic annotations compared to expert annotations.

	Test set		Doctor set	
IKEM data	F1 [%]	Accuracy [%]	F1 [%]	Accuracy [%]
SSA	86.20	86.24	81.89	81.55
SSA & Macenko	82.90	82.66	81.54	81.24
WSA	85.64	85.34	82.04	81.75
WSA & Macenko	82.57	82.49	81.98	82.00

Table 4: Evaluation of Teacher architecture ResNet-18 on the test set and doctor annotations.

After evaluating the performance of ResNet-16, VGG-4 and Xception architectures on multiscale data using 16 × 16 patch, we identified that the ResNet-16 architecture achieved the highest F1 Scores. Therefore, we used the residual block-based architecture for our experiments on original-size data using 32 × 32 patch size. Due to the increase in patch size, we moved to architecture with a larger learning capacity - ResNet-18 as the Teacher model in all the following analysis steps.

We trained the selected ResNet-18 architecture on data without and with normalization using Macenko with strong and weak synthetic annotations. The results of training ResNet-18 as a Teacher model on these different data combinations can be found in Table 4.

When comparing the results of training the ResNet-18 model on data with and without Macenko normalization, the model performs better on data without applying normalization for both test data and expert annotations. In our analysis, we also compared the performance of the ResNet-18 model when training on SSA and WSA from QuPath. As shown in Table 4, the results indicate that training on SSA performs better on the test set. However, in the case of expert annotations, the results are better in the case of training using WSA. This may be caused by the fact that the doctors selected the ANN as the best based on a qualitative evaluation and visual comparison.

5.3 Classification - Student

Our evaluation of the trained models, called students, focused on the RCCNet architecture. We trained these models using the previously trained teachers from the previous step, utilizing different versions of the data, including both strong and weak synthetic annotations and data with and without normalization. After evaluating each model on the test set, we further evaluate the trained model using expert annotations. The results of the different data versions can be found in Table 5.

Based on the results presented in Table 5, the RCCNet

IKEM data	Test set		Doctor set	
	F1 [%]	Accuracy [%]	F1 [%]	Accuracy [%]
SSA	85.06	84.16	81.76	81.57
SSA & Macenko	81.84	81.03	82.02	82.02
WSA	85.13	84.96	81.71	81.42
WSA & Macenko	81.88	81.14	82.00	81.88

Table 5: Evaluation of Student architecture RCCNet on the test set and doctor annotations after training using Teacher architecture.

Labels count (per class)	F1 [%]		Accuracy [%]	
	Before fine-tuning	After fine-tuning	Before fine-tuning	After fine-tuning
100	81.88	80.47	81.90	80.80
250	81.92	80.73	81.95	81.30
400	82.17	79.85	82.25	80.56
550	82.64	81.21	82.79	82.29
750	83.17	80.15	83.43	81.19
850	83.73	78.40	84.13	79.13
1000	84.24	78.89	84.74	80.26
1250	83.55	82.47	84.24	82.62
1500	**83.30**	**84.93**	**84.23**	**85.58**
1750	82.57	83.99	83.88	84.50
1900	82.73	84.81	85.17	87.23

Table 6: Evaluating the performance of the student architecture on strong annotations with normalization before and after fine-tuning.

model performed better on data without normalization and using SSA within the test set. However, the normalized data performed better for expert annotations. Using WSA performed better for the test set and for expert annotations, the results were better using SSA. The results suggest that by training the student with the teacher, we achieved a higher level of generalization in the trained student model, leading to the improved classification of medical data using normalization.

Our last evaluation is focused on students fine-tuning based on gaining information from expert annotations. Fine-tuning aims to find the smallest number of annotations needed to improve RCCNet students trained by ResNet-18 teachers. For these experiments, we perform a grid search over the chosen nuclei counts for each class concerning the class that contained the smallest number of annotations. We experimented with nuclei counts of 100, 250, 400, 550, 750, 850, 1000, 1250, 1500, 1750 and 1900. Each of the above values represents the per-class count, which we divided into training and validation sets. The test set, over which we computed the metrics for fine-tuned models, is created from the remaining number of nuclei in the dataset.

As presented in Table 6, the results indicate that fine-tuning the pre-trained RCCNet model using SSA and normalized data led to improved performance compared to the initial training. Specifically, using 1500 nuclei per class during the fine-tuning process resulted in a higher F1 Score and Accuracy than before fine-tuning or using lower nuclei counts. However, utilizing a small sample of medical data to fine-tune the RCCNet model led to worse results.

6 Conclusion & Discussion

This study presents a novel robust approach for cell segmentation and classification, evaluated on WSI scans of heart biopsy. This approach can generally be applied to any histological images from different organs and different types of cells. Our method consists of traditional methods used in medicine combined with novel methods for working with weakly annotated data.

Using the Knowledge distillation and the Teacher-Student architecture for nuclei classification, we have identified that it is possible to improve results by using this approach under certain conditions. In this work, we specifically use Teacher-Student chaining. According to our results, the best use of this technique appears to be in the case of normalized data. We identified that there might be an improvement in the results after fine-tuning the student model.

An improvement in the results for normalized data was observed when the number of manually annotated cells per class reached a threshold of 1500, suggesting it may be a suitable cut-off point for expert annotations.

A potential limitation is a total number of cells manually annotated by pathologists. If more cells were annotated, further experiments could be performed to more accurately evaluate the impact of fine-tuning the student model on the medical annotations.

Future research may explore the potential of the Teacher-Student architecture without relying on ANN annotations. This could involve training an initial teacher model on a set of medical annotations and using it to train a student model on previously unseen data, allowing for further analysis and investigation of the approach.

References

[1] Cardiovascular diseases (cvds). `https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).` Accessed: 2023-03-23.

[2] Michael D Abràmoff, Paulo J Magalhães, and Sunanda J Ram. Image processing with imagej. *Biophotonics international*, 11(7):36–42, 2004.

[3] Saeed Alahmari, Dmitry Goldgof, Lawrence Hall, Palak Dave, Hady Ahmady Phoulady, and Peter Mouton. Iterative deep learning based unbiased stereology with human-in-the-loop. In *2018 17th*

IEEE International Conference on Machine Learning and Applications (ICMLA), pages 665–670, 2018.

[4] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

[5] SH Shabbeer Basha, Soumen Ghosh, Kancharagunta Kishan Babu, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Rccnet: An efficient convolutional neural network for histological routine colon cancer nuclei classification. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 1222–1227. IEEE, 2018.

[6] Sushovan Chaudhury, Nilesh Shelke, Kartik Sau, B Prasanalakshmi, and Mohammad Shabaz. A novel approach to classifying breast cancer histopathology biopsy images using bilateral knowledge distillation and label smoothing regularization. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[7] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Andres Diaz-Pinto, Sachidanand Alle, Alvin Ihsani, Muhammad Asad, Vishwesh Nath, Fernando Pérez-García, Pritesh Mehta, Wenqi Li, Holger R Roth, Tom Vercauteren, et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. *arXiv preprint arXiv:2203.12362*, 2022.

[9] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 684–693, 2021.

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

[11] Noah F Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565, 2022.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[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. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2019.

[14] Michael R Lamprecht, David M Sabatini, and Anne E Carpenter. Cellprofiler™: free, versatile software for automated biological image analysis. *Biotechniques*, 42(1):71–75, 2007.

[15] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.

[16] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, pages 1–50, 2022.

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

[18] Shayne Shaw, Maciej Pajak, Aneta Lisowska, Sotirios A Tsaftaris, and Alison Q O'Neil. Teacher-student chain for efficient semi-supervised histology image classification. *arXiv preprint arXiv:2003.08797*, 2020.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.