UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MAIKEL MACIEL RÖNNAU

CNN-based Methods and Datasets for Segmentation and Counting of Nuclei and AgNORs in AgNOR-stained Images and for Segmentation and Classification of Cells in Papanicolaou-stained Images

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Advisor: Prof. Dr. Manuel Menezes de Oliveira Neto

Porto Alegre
July 2024

CIP — CATALOGING-IN-PUBLICATION

Rönnau, Maikel Maciel

CNN-based Methods and Datasets for Segmentation and Counting of Nuclei and AgNORs in AgNOR-stained Images and for Segmentation and Classification of Cells in Papanicolaou-stained Images / Maikel Maciel Rönnau. – Porto Alegre: PPGC da UFRGS, 2024.

76 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2024. Advisor: Manuel Menezes de Oliveira Neto.

1. Deep Learning. 2. Convolutional Neural Networks. 3. Image Segmentation. 4. AgNOR. 5. Papanicolaou. I. Menezes de Oliveira Neto, Manuel. II. Título.

“If I have seen further than others,
it is because I stood on the shoulders of giants.”
— Sir Isaac Newton

AGRADECIMENTOS

Agradeço a todos que contribuíram para a realização deste trabalho, em especial ao meu orientador, Prof. Dr. Manuel Menezes de Oliveira Neto, pela orientação, apoio e confiança. Agradeço também ao Prof. Dr. Pantelis Varvaki Rados e ao seu grupo de pesquisa pela disponibilização dos conjuntos de dados de imagens. Agradeço à minha família e amigos pelo apoio e compreensão.

ABSTRACT

Oral cancer is the sixth most common kind of human cancer. Early detection is crucial for lowering patient mortality. Two staining techniques, Argyrophilic staining of Nucleolar Organizer Regions (AgNORs) and Papanicolaou staining, can assist in early detection. However, manual counting and interpretation of these techniques are time-consuming, labor-intensive, and error-prone. This thesis proposes two convolutional neural network (CNN) based methods to address these shortcomings. The first method automatically segments individual nuclei and AgNORs in microscope slide images and counts the number of AgNORs within each nucleus. The second method automatically segments and classifies morphological features in Papanicolaou-stained microscope slide images. Both methods were trained and evaluated on new image datasets of epithelial cells from oral mucosa, with ground truth annotated by specialists. The effectiveness of our models was evaluated against a group of human experts. Our CNN-based joint segmentation and quantification of nuclei and NORs in AgNOR-stained images achieved an Intraclass Correlation Coefficient (ICC) of 0.91 for nuclei and 0.81 for AgNORs, indicating strong agreement with experts. Our CNN model for automatic segmentation and classification of cells in Papanicolaou-stained images also demonstrated expert-level performance, with ICC values above 0.84 for all cell types, showing excellent or good agreement for most cell types. Both methods were significantly faster than manual analysis, reducing the processing time from hours to minutes. These results highlight their potential to accelerate diagnostic workflows. Our trained models, code, and datasets are available on GitHub and can stimulate new research in early oral cancer detection.

Keywords: Deep Learning. Convolutional Neural Networks. Image Segmentation. AgNOR. Papanicolaou.

Métodos Baseados em CNN e Conjuntos de Dados para Segmentação e Contagem de Núcleos e AgNORs em Imagens Coradas com AgNOR e para Segmentação e Classificação de Células em Imagens Coradas com Papanicolaou

RESUMO

O câncer oral é o sexto tipo mais comum de câncer humano. A detecção precoce é crucial para reduzir a mortalidade dos pacientes. Duas técnicas de coloração, coloração argirofílica das Regiões Organizadoras Nucleolares (AgNORs) e Papanicolaou, podem auxiliar na detecção precoce dos sinais deste tipo de câncer. No entanto, a contagem e a interpretação manual dessas técnicas são demoradas, trabalhosas e propensas a erros. Esta dissertação propõe dois métodos baseados em redes neurais convolucionais (CNN) para resolver essas limitações. O primeiro método segmenta automaticamente núcleos individuais e AgNORs em imagens de lâminas de microscópio e conta o número de AgNORs dentro de cada núcleo. O segundo método segmenta e classifica automaticamente características morfológicas em imagens de lâminas de microscópio coradas por Papanicolaou. Ambos os métodos teiveram seus modelos (CNNs) treinados e avaliados em novos conjuntos de dados de imagens de células epiteliais da mucosa oral, com *Ground Truth* anotado por especialistas. A eficácia de nossos modelos foi avaliada em comparação com um grupo de especialistas humanos. Nossos modelos baseados em CNN para segmentação e quantificação conjunta de núcleos e NORs em imagens coradas por AgNOR, bem como nosso modelo para segmentação e classificação automática de células em imagens coradas por Papanicolaou, ambos alcançaram níveis de desempenho semelhantes aos de especialistas, com significância estatística verificada, sendo ordens de magnitude mais rápidos do que a segmentação e contagem/classificação manual realizada pelos especialistas. Tais resultados destacam seu potencial para acelerar os fluxos de trabalho de diagnóstico. Nossos modelos treinados, código e conjuntos de dados estão disponíveis no GitHub e podem estimular novas pesquisas na detecção precoce de câncer oral.

Palavras-chave: Aprendizado Profundo. Redes Neurais Convolucionais. Segmentação de Imagens. AgNOR. Papanicolaou.

LIST OF ABBREVIATIONS AND ACRONYMS

AgNOR	Argyrophilic Nucleolar Organizer Region
CNN	Convolutional Neural Network
CE	Cross Entropy
CAA	Contour Analysis Algorithm
CRF	Conditional Random Field
FCN	Fully Convolutional Network
GMM	Gaussian Mixture Model
HDR	High Dynamic Range
IOU	Intersection Over Union
ICC	Intraclass Correlation Coefficient
LBC	Liquid-Based Cytology
LBCP	Liquid-Based-Cytology Pap Smear
MLP	Multi-layer Perceptron
ROI	Region of Interest
SVM	Support Vector Machine
TSS	Temperature Scaling Softmax

LIST OF FIGURES

LIST OF TABLES

CONTENTS

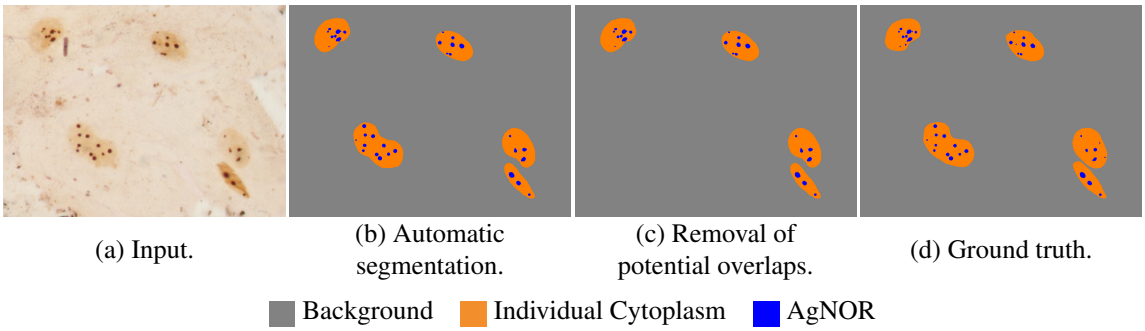
1 INTRODUCTION

Oral cancer stands as the sixth most prevalent human cancer globally and the most prevalent in the head and neck region (Vigneswaran; Williams, 2014), with an annual estimate of 657,000 new cases and 300,000 deaths (OCF, 2024b). The mortality rates associated with oral cancer have been steadily rising over the past decade, underscoring the urgency for improved diagnostic methodologies. Unfortunately, the diagnosis often occurs at advanced stages, as physical examinations and biopsies are typically sought only after symptoms manifest, leading to significant challenges in treatment efficacy and patient survival rates (OCF, 2024b). Moreover, survivors often endure considerable functional and cosmetic impairments due to the aggressive nature of the tumor and the treatment (OCF, 2024b). However, early detection of oral cancer signs through cytopathology offers a promising avenue for timely intervention. Two prominent techniques, AgNOR staining and Papanicolaou staining, have emerged as valuable tools for identifying cellular abnormalities indicative of malignancy.

AgNOR staining, though gradually supplanted by immunohistochemistry, remains cost-effective and accessible, particularly in resource-limited settings (Jajodia et al., 2017). This technique relies on quantifying the number of stained Argyrophilic Nucleolar Organizer Regions (AgNORs) in cell nuclei, serving as a marker for proliferative activity and malignant potential. The AgNOR count has been correlated with the degree of cellular proliferation, offering valuable insights into the progression of oral cancer (Tyagi et al., 2020). The technique has demonstrated utility in distinguishing between benign and malignant lesions, aiding in the early detection of oral cancer signs (Tyagi et al., 2020).

Papanicolaou staining has demonstrated success in early detection of cervical cancer, contributing to a progressive reduction in mortality rates (Bedell et al., 2019). Leveraging its ability to highlight cellular abnormalities, including changes in nuclear volume, shape, and staining properties, we believe that Papanicolaou staining holds promise for identifying early signs of malignant alterations in oral cells. The American Dental Association expert consensus group has advocated for the use of oral cytology tests in general dental practice, particularly when tissue biopsy is not feasible, underscoring the importance of reliable screening methodologies (Lingen et al., 2017).

Despite their utility, both AgNOR and Papanicolaou staining techniques necessitate skilled pathologists for evaluation, hindering scalability and timely diagnosis. To address this challenge, this work proposes leveraging deep learning, specifically convolu-



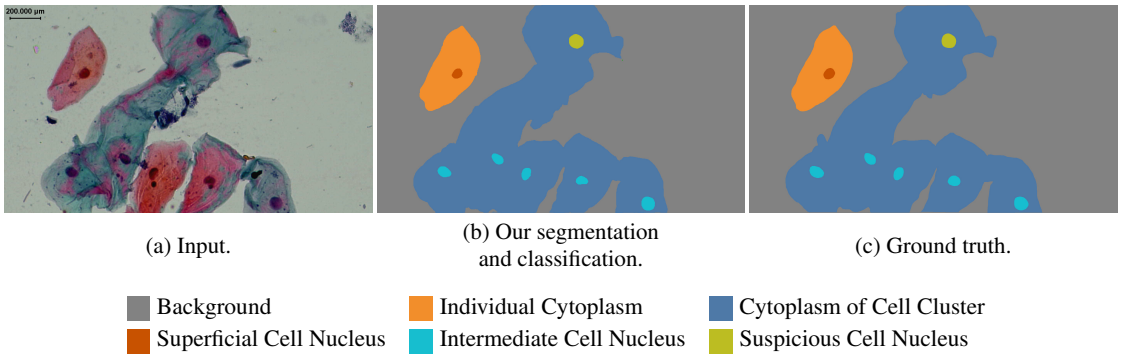
Source: The Authors.

Figure 1.1 – Example of automatic nuclei and AgNOR segmentation in oral cells using our CNN. (a) AgNOR-stained cytological slide image provided as input. (b) Automatic segmentation produced by our model, with nuclei, AgNORs, and background shown in orange, blue, and gray, respectively. (c) Result obtained after discarding potentially-overlapping nuclei, which tend to hide AgNORs. (d) Ground truth segmentation.

tional neural networks (CNNs), for automatic segmentation and analysis of slide images from both staining techniques. This thesis asserts the feasibility of employing CNN models to accelerate evaluation processes, achieving expert-level performance while taking less than a minute to process hundreds of images.

Contributing to this endeavor, our work introduces CNN-based methodologies tailored to each of these staining techniques. For AgNOR staining, we present a comprehensive approach encompassing automatic segmentation and quantification of nuclei and AgNORs, supported by a diverse image dataset annotated by specialists (UFRGS AgE-COM (Rönnau et al., 2023c)). This method utilizes the number of stained AgNORs within cell nuclei as an indicator of proliferative activity and malignant potential, facilitating early detection of oral cancer signs. Fig. 1.1 illustrates the use of our model to segment nuclei and AgNORs, where the cytological slide image and its corresponding ground truth are shown alongside the automatic segmentation results, demonstrating the efficacy of our approach. Similarly, for Papanicolaou staining, our methodology enables per-pixel segmentation and classification of cellular structures, complemented by a rich dataset comprising images from various oral mucosa conditions (UFRGS Pap-OMD (Rönnau et al., 2024)). By leveraging the enhanced cell contrast provided by Papanicolaou staining, our CNN model can accurately identify suspicious cells and clusters, facilitating early detection of malignant transformations. Fig. 1.2 further illustrates the application of our CNN-based model in segmenting nuclei, cytoplasms, and cell clusters, highlighting its ability to identify various types of nuclei and cellular structures with high precision.

Our CNN models offer efficient, scalable, and accurate solutions for assisting early



Source: The Authors.

Figure 1.2 – Segmentation and classification of a Papanicolaou-stained image from the test set of our dataset of epithelial cells from the oral mucosa. (a) Input image. (b) Our model's automatic segmentation and classification, with individual cytoplasms, cell clusters, and background colored in orange, blue, and gray, respectively. The different types of nuclei are shown in yellow (suspicious), red (superficial), and cyan (intermediate). (c) Ground truth. Note the proper classification of cell structures.

detection of oral cancer and improve patient outcomes. The availability of annotated datasets (Rönnau et al., 2023c; Rönnau et al., 2024) and model implementations (Rönnau et al., 2023b) underscores our commitment to fostering reproducibility and facilitating further research in this critical domain.

1.1 Thesis statement

The central idea behind this research can be stated as:

> *It is possible to develop deep learning models for efficient automatic segmentation and analysis of oral cytology images stained using AgNOR and Papanicolaou techniques, achieving expert-like performance levels. The resulting models should provide scalable solutions for assisting healthcare professionals in early detection of oral cancer signs.*

We demonstrate this statement by designing and training CNN models, building datasets, and comparing the results produced by our models with the ones provided by a group of expert cytopathologists. Our CNN models achieved expert-like performance level with verified statistical significance, while being orders of magnitude faster than the manual segmentation and counting/classification performed by the specialists.
The **contributions** of this thesis include:

- A methodology for developing and evaluating deep learning models for cytopathol-

ogy applications;

- A CNN-based approach for automatic joint segmentation and quantification (counting) of nuclei and AgNORs in AgNOR-stained images (Section 4.2);
- An AgNOR-stained image dataset of epithelial cells from oral mucosa containing 1,171 slide images from 48 patients, annotated by specialists (Section 4.1);
- A semi-automatic image annotation strategy to reduce the workload from specialists to produce ground truth image annotations (Section 4.1.1);
- An algorithm for identifying overlapping nuclei and excluding them from AgNOR counting (Section 4.2.3.1).
- A CNN model for automatic segmentation and classification of cells in Papanicolaou-stained images as suspicious, superficial, intermediate, or anucleate squamous. It also classifies cell clusters as suspicious or non-suspicious (Section 5.2). Ours is the first automatic solution to simultaneously perform both segmentation and classification of cells and cell clusters in Papanicolaou-stained images;
- A Papanicolaou-stained image dataset of the oral mucosa cells with 1,563 images from 52 patients, annotated by specialists (Section 5.1).

The results of the work on the CNN-based approach for segmenting and counting AgNORs has been published in the journal *Computer Methods and Programs in Biomedicine* (Rönnau et al., 2023a). The work on the Papanicolaou method was submitted for publication and is currently under review.

1.2 Structure of the thesis

The remaining of this thesis is structured as follows: Chapter 2 provides background information on oral cancer, AgNOR and Papanicolaou staining techniques, and on deep learning methodologies for image segmentation. Chapter 4 presents the model for segmenting and counting AgNORs in AgNOR-stained images. Chapter 5 describes the the model for segmenting and classifying cells and cell clusters in Papanicolaou-stained images. Chapters 4 and 5 detail the methodology for developing and evaluating deep learning models for cytopathology applications, encompassing data collection and annotation, model development and training, and evaluation. The evaluation and results of the two models are presented in Sections 4.3 and 5.3, respectively. Chapter 6 concludes the thesis and discusses some directions for future work.

2 BACKGROUND

This chapter provides background information on oral cancer, AgNOR and Papanicolaou staining techniques, and on deep learning methods for medical imaging segmentation. Section 2.1 provides an overview of oral cancer, its prevalence, and the importance of early detection. Section 2.2.1 introduces the AgNOR staining technique, its relevance in cytopathology. Section 2.2.2 presents the Papanicolaou staining technique. Section 2.3 discusses deep learning methodologies for image segmentation.

2.1 Oral Cancer and the Importance of Early Detection

Oral cancer is the sixth most common kind of human cancer worldwide and the most prevalent in the head and neck region (Vigneswaran; Williams, 2014). The incidence of oral cancer has been steadily increasing over the past decade, with an estimated 657,000 new cases and 300,000 deaths annually (OCF, 2024a). Despite advances in surgical and treatment modalities, the five-year survival rate for oral cancer remains below 60% (OCF, 2024b). Late diagnosis is a significant factor contributing to the poor prognosis of oral cancer, as it often leads to advanced-stage disease and limited treatment options (OCF, 2024b). The aggressive nature of oral cancer and its treatment can result in significant functional, cosmetic, and emotional burdens for survivors, affecting their quality of life and overall well-being (OCF, 2024b).

Early detection of oral cancer is crucial for improving patient outcomes and reducing mortality rates. Cytopathology, the study of pathologies that manifest at the cellular level, offers a promising avenue for early diagnosis and intervention. By analyzing cellular abnormalities indicative of malignancy, pathologists can identify potentially malignant lesions in their early stages, enabling timely treatment and improved survival rates (Lingen et al., 2017). Two prominent cytopathology techniques, AgNOR and Papanicolaou staining, have emerged as valuable tools for identifying cellular abnormalities in oral cells, providing insights into the proliferative activity and malignant potential of tissues (Tyagi et al., 2020; Bedell et al., 2019). These staining techniques enhance the visibility of cellular structures, enabling pathologists to identify abnormal nucleolar morphology and cellular changes characteristic of malignant transformations (Rajput; Tupkari, 2010; Shiraz et al., 2020). Leveraging the power of these staining techniques, coupled with advanced image analysis methodologies like the use of CNNs for automatic segmentation of cell

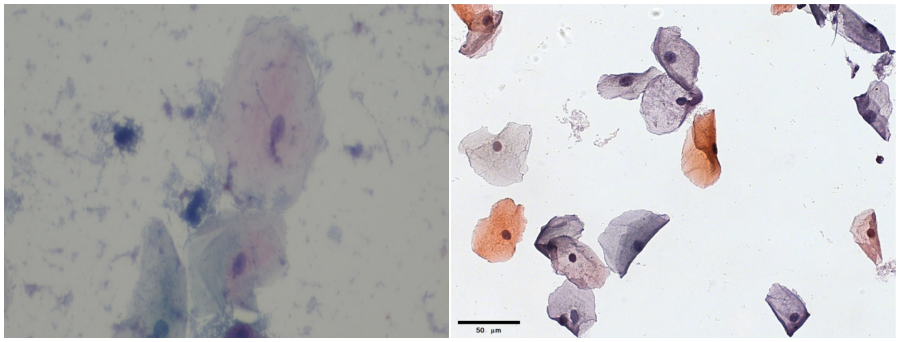
structures, holds immense potential for improving the accuracy and efficiency of detecting the early signs of oral cancer.

Cytopathology is a branch of pathology that focuses on the study of cellular abnormalities and diseases. It plays a crucial role in the early detection, diagnosis, and management of various cancers and other pathological conditions. Cytopathology techniques involve the collection of cells or tissues from the body, their preparation on glass slides, and their staining with dyes to enhance cellular structures and facilitate microscopic examination. Two widely used staining techniques in cytopathology are the AgNOR and Papanicolaou staining techniques, which are instrumental in identifying cellular abnormalities indicative of malignancy. These staining techniques provide valuable insights into the proliferative activity, nuclear morphology, and chromosomal abnormalities of cells, enabling pathologists to make accurate diagnoses and treatment decisions. The following sections provide an overview of the AgNOR and Papanicolaou staining techniques and their relevance in cytopathology.

2.2 Preparation of Microscope Slides

The preparation of microscope slides is a critical step in cytopathology, ensuring the accurate identification and analysis of cellular abnormalities. Traditionally, smearing techniques have been used to prepare slides, where cells collected from tissue samples are spread directly onto a glass slide. This method, although simple, has limitations such as uneven distribution of cells, air-drying artifacts, and the presence of obscuring blood or mucus.

In recent years, liquid-based cytology (LBC) has emerged as a superior alternative to traditional smear techniques. In LBC, cells are collected and suspended in a liquid medium, which is then processed to create a thin, uniform layer of cells on a slide. This method reduces the presence of obscuring material and artifacts, providing clearer, more consistent samples for examination (Strander et al., 2007). See Fig. 2.1 for a comparison of traditional smearing and LBC techniques.



Conventional Smear Liquid-Based Cytology (LBC)

Source: The Authors.

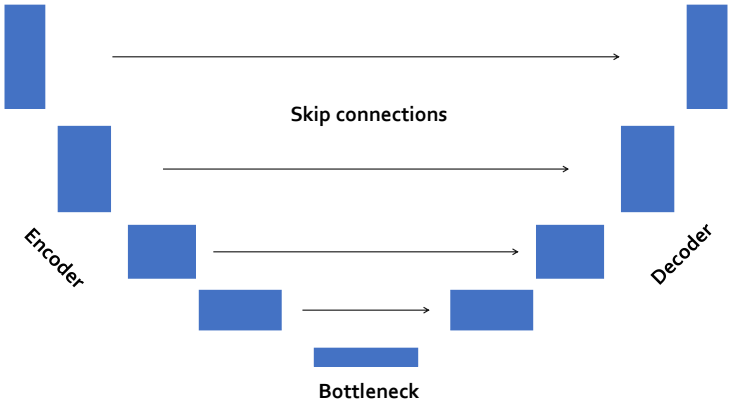
Figure 2.1 – Comparison of traditional smearing and liquid-based cytology (LBC) techniques for preparing microscope slides for Papanicolaou staining. In the conventional smearing technique, cells are spread directly onto a glass slide, leading to uneven distribution, air-drying artifacts, and obscuring material. In liquid-based cytology (LBC), cells are collected and suspended in a liquid medium, which is then processed to create a thin, uniform layer of cells on a slide, reducing artifacts and providing clearer samples for examination. The image on the left corresponds to a sample from our dataset (Rönnau et al., 2024), while the image on the right is from the CRIC dataset (Rezende et al., 2021). They were prepared using conventional smearing and LBC techniques, respectively.

2.2.1 The AgNOR Staining Technique

The AgNOR staining technique is widely used for identifying Nucleolar Organizer Regions (NORs) in cells. NORs are chromosomal regions whose number and size in a cell are indicative of its metabolic activity and proliferative potential, making them valuable markers for assessing cell growth and proliferation (Jajodia et al., 2017; Tyagi et al., 2020). AgNOR staining enhances the visibility of NORs, enabling pathologists to identify cells with abnormal nucleolar morphology, such as increased size, number, or staining intensity, which are characteristic of malignant transformations (Rajput; Tupkari, 2010). The AgNOR staining technique has been instrumental in improving the accuracy of cancer diagnosis and has been shown to be a valuable adjunct to traditional histopathological methods (Caldeira et al., 2011).

2.2.2 The Papanicolaou Staining Technique

The Papanicolaou staining technique, also known as Pap smear, is a widely used method for detecting cervical cancer. Papanicolaou staining enhances cell contrast, facilitating accurate morphological analysis and enabling the identification of cellular abnor-



Source: The Authors.

Figure 2.2 – An overview of the encoder-decoder architecture for semantic image segmentation. The encoder processes the input image and compresses it into a fixed-size context vector. The decoder then takes this context vector and generates the segmented output image. Additionally, skip connections are used to pass high-resolution features from the encoder to the decoder, enhancing the decoder's ability to produce a detailed and accurate segmentation map, identifying and classifying each pixel in the image.

malities, such as changes in nuclear volume, shape, and staining properties (Shiraz et al., 2020). The technique has been instrumental in reducing cervical cancer mortality rates through early detection and intervention (Bedell et al., 2019).

2.3 Deep Learning Methods for Medical Imaging Segmentation

Deep learning has revolutionized the field of medical imaging segmentation, providing automated, accurate, and efficient solutions for various applications (Hesamian et al., 2019). These methodologies leverage the power of convolutional networks (CNNs) to learn complex patterns and features from large datasets of medical images (Wang et al., 2022), enabling them to perform segmentation tasks with high precision. The most commonly used architecture in deep learning for medical image segmentation is the encoder-decoder structure (Long; Shelhamer; Darrell, 2015), which combines a feature extraction network (encoder) with a reconstruction network (decoder) to generate pixel-wise segmentation masks. This architecture, which includes the U-Net (Ronneberger; Fischer; Brox, 2015) and its variations, is particularly effective for tasks that require the segmentation of images with a small number of training samples. The U-Net architecture, for instance, has been widely used to segment various types of medical images, including MRI scans, CT scans, and histopathology images. Fig. 2.2 shows the general structure of

an encoder-decoder network.

In addition to the encoder-decoder architectures, several other popular image segmentation methods have been developed and successfully applied in medical imaging. These include fully convolutional neural networks (FCNs) (Noh; Hong; Han, 2015) that are a type of CNN where the fully connected layers are replaced with convolutional layers, enabling the network to produce spatially dense predictions (Long; Shelhamer; Darrell, 2015). FCNs are capable of learning to segment images of any size and have been effectively used for various medical imaging tasks, including organ and lesion segmentation. Mask R-CNN extends the Faster R-CNN framework by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression (He et al., 2017). This method has shown excellent performance in instance segmentation tasks, making it useful for segmenting individual objects. The DeepLab family of models employs dilated convolutions to capture multi-scale contextual information, along with conditional random fields (CRFs) for precise boundary delineation (Chen et al., 2017). DeepLab has been particularly successful in handling the complex and varied shapes found in medical imaging data. For volumetric medical imaging data, such as 3D MRI or CT scans, 3D convolutional networks are used to capture spatial context in three dimensions (Çiçek et al., 2016). These networks extend the 2D CNN architectures to process 3D inputs, making them ideal for tasks that require the analysis of volumetric data, such as brain tumor segmentation.

Overall, deep learning methods for medical imaging segmentation have been applied to a wide range of tasks, including the segmentation of tumors, organs, and cells. These methods have shown remarkable success, often outperforming traditional image segmentation techniques in terms of both accuracy and efficiency (Wang et al., 2022; Rasheed et al., 2023; Amorim, 2020; Hesamian et al., 2019; Pham; Xu; Prince, 2000). However, the success of deep learning methods in medical imaging segmentation is highly dependent on the quality of the training data. High-quality, annotated datasets are crucial for training robust and accurate models. Furthermore, the specific task at hand also plays a significant role in determining the most suitable deep learning method to use.

2.3.1 Semantic Image Segmentation

In the context of this work, semantic image segmentation is the most appropriate deep learning method due to its ability to classify each pixel in an image into a predefined

category. This method is particularly well-suited for the future objective of calculating the ratios between nuclei and nucleolar organizer regions (NORs) in AgNOR-stained images, and cytoplasm and nuclei in Papanicolaou-stained images.
Advantages of Semantic Image Segmentation:

- Pixel-wise Classification: Semantic segmentation provides a detailed pixel-wise classification of different cellular components, which is essential for accurate quantitative analysis. This allows for precise measurement of the areas of nuclei, NORs, and cytoplasm, facilitating the calculation of ratios between these components;
- Comprehensive Tissue Structure Analysis: By classifying all pixels in an image, semantic segmentation offers a comprehensive view of the tissue structure. This holistic approach is crucial for understanding the spatial relationships between different cellular components, which can provide valuable insights into the pathology of the sample;
- Simplicity and Efficiency: Semantic segmentation models, particularly those based on the encoder-decoder architecture, are relatively simple to implement and train. They can achieve high accuracy even with a limited number of training samples, making them ideal for medical image segmentation tasks;
- Robustness to Variability: Semantic segmentation can handle variability in staining intensity, cell shape, and size, which are common challenges in medical image analysis. This robustness ensures reliable performance across different samples and conditions.

The goal is to accurately segment and classify nuclei, cytoplasm, and NORs to enable the calculation of the nucleus-to-NOR and cytoplasm-to-nucleus ratios. Semantic segmentation is particularly well-suited for this task because it can delineate the boundaries of these components with high precision. The resulting segmentation maps provide the necessary data for detailed morphological and quantitative analysis. In conclusion, compared to other techniques like object detection, semantic image segmentation is the most suitable method for this work due to its ability to provide detailed, pixel-wise classification of cellular components, which is essential for accurate quantitative analysis in AgNOR- and Papanicolaou-stained images. This approach lays a strong foundation for future advancements in automated cytopathological diagnostics.

3 RELATED WORKS

This chapter reviews relevant works on the automatic segmentation and classification of cells in AgNOR- and Papanicolaou-stained images. The chapter is organized into two sections: Section 3.1 presents the related works on AgNOR-stained images and datasets, and Section 3.2 presents the related works on Papanicolaou-stained images and datasets.

3.1 Related Works on AgNOR-Stained Image Segmentation and Datasets

The segmentation and classification of cells' structural components in microscope slide images is an important problem in pattern recognition, with numerous applications in health sciences. Examples of techniques include segmentation of cytoplasm and nucleus (Li et al., 2012), segmentation and classification of cell types (Gençtav; Aksoy; Önder, 2012), cell segmentation in the presence of overlapping boundaries (Zhang et al., 2016), and segmentation of nuclei in fluorescence images (Gharipour; Liew, 2016). Traditional techniques, such as the ones just mentioned, typically use thresholding, energy minimization, or a combination of both. For detailed discussions on deep-learning-based image-segmentation techniques for medical images in general, we refer the reader to the surveys by Hesamian et al. (2019), and by Wang et al. (2022). Pham, Xu and Prince (2000) provide a comprehensive discussion of non-learning-based methods for medical image segmentation. Next, we concentrate on related techniques for segmenting AgNOR-stained images, which is the focus of our work.

3.1.1 Segmentation of AgNOR-Stained Images

Amorim et al. (2020b) performed segmentation of AgNOR-stained images using ResNet-18 (He et al., 2016) for feature extraction and U-Net (Ronneberger; Fischer; Brox, 2015) for image reconstruction. The resulting CNN was trained on a dataset containing 2,540 images of cervical cells obtained from three patients (Amorim et al., 2020a). Compared to Amorim et al.'s CNN, our model exhibits better generalization properties. While it can correctly segment the AgNOR-stained cells from Amorim et al. (2020a)'s dataset (Fig. 4.9), Amorim et al. (2020b)'s model was unable to properly segment many images

from our dataset (Figs. 4.8 and 4.10), even when retrained on it.

Bell et al. (2006) used high dynamic range (HDR) images obtained from multiple exposures to segment AgNOR-stained images exploiting the contrast between nuclei and AgNORs. The technique takes as input multiple pre-segmented images containing a single nucleus (at known position) per image. This limits its use to environments where these strict conditions can be satisfied. In contrast, our method only requires standard images, is automatic, and applied to whole images.

Several researchers used threshold-based AgNOR segmentation for various applications. Ferreira et al. (2011) used color thresholding to segment nuclei and AgNORs in ameloblastoma cells. Their goal was to estimate the mean number of AgNORs per cell. The technique presumes that nuclei, AgNORs, and cytoplasm/image background have distinct colors. While this is often true, images exhibiting low contrast between nuclei and background are common (Fig. 4.8 (b) and (d), Fig. 4.10 (c) and (h)). The resulting segmentation tends to exhibit low-fidelity contours. The pixel-level accuracy of the segmentation process has not been reported, only the result of a qualitative evaluation performed by two observers. Since Ferreira et al.'s code is not publicly available, we cannot compare their results with ours on a common dataset.

Both García-Vielma et al. (2016) and Teresa et al. (2007) used thresholding performed by third-party software to segment nuclei and AgNOR. García-Vielma et al. manually defined the thresholding parameters. The area of the AgNORs in each nucleus was then estimated using the segmentation. Teresa et al. used a two-step thresholding applied first to nuclei and then to AgNORs, with the goal of measuring AgNOR area and AgNOR/nucleus area ratio. Cucer et al. (2007) segmented nuclei and AgNOR by manually tracing their contours, from which they computed AgNOR/nucleus area ratios.

All these thresholding applications require user input specifying one or more threshold values, and are less concerned about the accuracy and time-efficiency of the segmentation process. In turn, our solution can segment AgNOR-stained images automatically with satisfactory accuracy in an efficient way.

3.1.2 AgNOR-Stained Image Datasets

While there are several datasets of Papanicolaou-stained images publicly available, according to a recent survey (Jiang et al., 2022), Amorim et al. (2020a)'s CCAgT dataset is the only previously publicly available AgNOR-stained image dataset, and com-

Table 3.1 – Publicly-Available Annotated AgNOR-stained image datasets.

Dataset	Patients	Nuclei	AgNORs	Cell type	Images	Resolution
Ours (UFRGS AgECOM)	48	3,310	12,337	Oral	1,171	2560×1920
CCAgT	3	4,515	12,196	Cervical	2,540	1600×1200

prised of cervical cells. Our dataset (UFRGS AgECOM) is publicly available (Rönnau et al., 2023c), contributes to fill this gap by providing a diverse dataset of epithelial oral cells. It contains 1,171 images from 48 individuals annotated by specialists. Table 3.1 summarizes the characteristics of these two datasets.

3.2 Related Works on Papanicolaou-Stained Image Segmentation and Datasets

The segmentation and classification of cells in Papanicolaou-stained images is a challenging problem due to the frequent presence of artifacts, such as debris and high background noise, as well as to occurrence of defocused images. The literature on this topic is vast, with traditional methods based on thresholding, clustering, and morphological operations, and more recent methods based on deep learning. This section reviews the most relevant automatic methods and datasets for the analysis of Papanicolaou-stained images. Such methods include both traditional image processing and deep learning techniques.

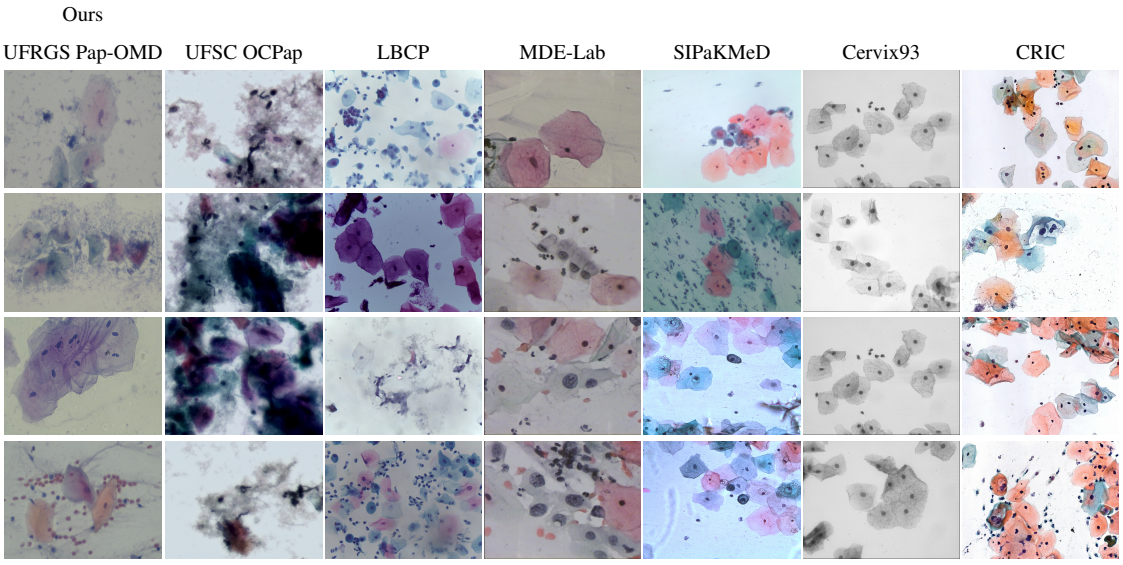
3.2.1 Segmentation of Papanicolaou-Stained Images

Traditional image processing techniques are based on thresholding, clustering, and morphological operations. Boughzala et al. (2016) investigated the impact of color spaces in K-means segmentation of cytoplasms and nuclei of cervical cells using a private dataset. Bandyopadhyay and Nasipuri (2020) used K-means clustering to segment the nuclei from isolated cervical cells in images from the Herlev dataset (Jantzen et al., 2005). Plissiti, Nikou and Charchanti (2010) employed morphological analysis to detect cell nuclei candidates that are refined in a second step using *a priori* knowledge about the shape of the nuclei. The detection technique is only applied to sub-regions of the images defined by binary masks obtained using thresholding. In a subsequent work (Plissiti; Vrigkas; Nikou, 2015), the authors proposed a method based on super-pixels, and more recently (Plissiti et al., 2018), they explored several methods, including Support

Vector Machines (SVM), Multi-layer Perceptron (MLP), and Convolutional Neural Network (CNN) to classify cervical cells from the SIPaKMeD dataset (Plissiti et al., 2018) among the following types: superficial/intermediate, parabasal, koilocytotic, metaplastic, and dyskeratotic. Ragothaman et al. (2016) proposed an unsupervised segmentation method using Gaussian mixture models (GMM) to identify cytoplasms and nuclei in cervical cells from a private dataset. The method employs a shape-based analysis of the nucleus region to deal with false-positive segmentation of nuclei caused by the presence of debris and other artifacts.

Deep learning methods have been increasingly employed to accelerate the diagnostic and improve the accuracy of the results in cytopathology (Jiang et al., 2022; Wang et al., 2022). The encoder-decoder architecture (Long; Shelhamer; Darrell, 2015) is the most commonly used, with the U-Net (Ronneberger; Fischer; Brox, 2015) and its variations being the most popular networks for medical image segmentation, due to its ability to segment images with a small number of training samples. Matias et al. (2021) explored several CNN models pre-trained on the ImageNet dataset (Russakovsky et al., 2015) for segmentation of nuclei in Papanicolaou-stained images. The authors concluded that a U-Net architecture with a ResNet (He et al., 2016) as its encoder performed better than the other tested models. They used the UFSC OCPap dataset of oral cytology images (Matias et al., 2021) (described in Sub-section 3.2.2) to train and evaluate the models. Rasheed et al. (2023) used a variation of the U-Net dubbed C-UNet (Cervical-UNet) to segment cell nuclei from a Papanicolaou-stained cervical image dataset (Zhang et al., 2019). Both of these methods only segment nuclei. Zhao et al. (2022) introduced SPCNet, a star-convex polygon-based CNN for automatic segmentation of cervical cells in Papanicolaou-stained images. SPCNet can segment cytoplasms of adherent cells, but it does not segment nuclei or provide cell type classification.

None of these CNN methods perform simultaneous segmentation and classification of nuclei, cytoplasms, and cell clusters. Moreover, none of them have publicly available pre-trained versions of their models. Our CNN model performs automati0c segmentation of cell nuclei, cytoplasms, and cell clusters in Papanicolaou-stained images. It also automatically classifies cell types among suspicious, superficial, intermediate, anucleate squamous, and binucleate, based on their nucleus types. It classifies cell clusters as suspicious or non-suspicious.



Source: The Authors.

Figure 3.1 – Example of images from different datasets. In all datasets, with the exception of the UFSC OCPap, cell cytoplasms and nuclei are discernible in essentially all cells.

3.2.2 Papanicolaou-stained Datasets

Most available datasets of Papanicolaou-stained images are of cervical cells. A recent survey (Jiang et al., 2023) only reports a single Papanicolaou-stained image dataset of oral mucosa cells, the UFSC OCPap dataset (Matias et al., 2021). This dataset consists of 1,934 whole slide images obtained from two patients. The dataset contains annotations for nuclei only: binary segmentation masks, bounding boxes, and classification as normal or abnormal nuclei. Unfortunately, the images in this dataset contain numerous artifacts and high background noise level that significantly degrade the quality of the images, often resulting in non-discernible cell cytoplasms and nuclei. Fig. 3.1 (second column) shows examples of images from this dataset.

The Liquid-Based Cytology Pap Smear dataset (LBCP) by (Hussain et al., 2020) contains 963 whole slide images of Papanicolaou-stained cervical cells from 460 patients. The images are annotated with four sub-categories of cervical lesions: negative for intraepithelial lesion or malignancy (NILM), low-grade intraepithelial lesions (LSIL), high-grade intraepithelial lesions (HSIL), and squamous cell carcinoma (SCC). The dataset images show well-defined cells and nuclei with high density of objects per image.

Byriel (1999) introduces the cervical cell MDE-Lab dataset containing 500 whole slide images, as well as individual images per cell. The dataset is annotated with the following cell classifications: columnar epithelial, squamous epithelium from the parabasal,

intermediate, and superficial layers, and non-keratinizing mild, moderate, and severe dysplasia. The dataset also includes manually-defined areas for the cells' cytoplasms and nuclei.

Plissiti et al. (2018) introduced the SIPaKMeD dataset containing 966 whole slide images of Papanicolaou-stained cervical cells. The authors do not provide the number of patients or the groups they belong to. The dataset is annotated with five cell classes: superficial/intemediate, parabasal, koilocytitic, metaplastic, and dyskeratotic. The dataset includes manually-defined areas for the cytoplasms and the nuclei of the cells.

Cervix93 (Phoulady; Mouton, 2018) is a dataset containing 93 whole slide images of Papanicolaou-stained cervical cells. It contains nucleus annotations and is divided into three classes: negative, low-grade squamous intraepithelial lesions (LSIL), and high-grade squamous intraepithelial lesions (HSIL). No information about the number of patients or patient groups is provided by the authors.

CRIC (Rezende et al., 2021) is a dataset containing 400 whole slide images of Papanicolaou-stained cervical cells. The dataset is annotated with six classes: negative, atypical squamous cells of undetermined significance (ASC-US), low-grade squamous intraepithelial lesions (LSIL), high-grade squamous intraepithelial lesions (HSIL), atypical squamous cells, and squamous cell carcinoma. The dataset is divided into 11,534 cells and is part of the CRIC Cervix collection.

We evaluated our CNN model on our dataset of oral mucosa cells (UFRGS Pap-OMD), plus on the five public datasets of cervical cells mentioned above: LBCP, MDE-Lab, SIKaKMeD, Cervix93, and CRIC. We choose not to include the UFSC OCPap dataset since its images contain many non-discernible nuclei and cytoplasms (Fig. 3.1).

4 AUTOMATIC SEGMENTATION AND QUANTIFICATION OF NUCLEI AND AGNORS IN AGNOR-STAINED IMAGES

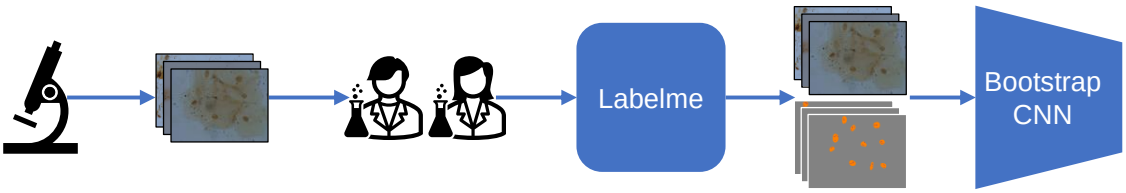
This chapter presents our work on automatic segmentation and quantification of nuclei and AgNORs in AgNOR-stained images. We start by describing our dataset of AgNOR-stained images from the oral mucosa, UFRGS AgECOM (Rönnau et al., 2023c), which was annotated by specialists. We then present our CNN model for image segmentation, which was trained on this dataset. The model performs per-pixel segmentation and classification of AgNOR-stained cell nuclei and AgNORs. We then describe the algorithm to analyze the contours of the segmented objects and reject overlapping nuclei. We show the results of our model on a set of images from our dataset as well as from another publicly avaiable dataset. Finally, we compare the results of our model with the ones produced by three human experts using Intraclass Correlation Coefficient (ICC).

4.1 Building Our AgNOR-stained Cell Dataset

Given our dataset's key role in the CNN architecture selection process, its construction is presented before the CNN model itself, which is described in Section 4.2. The dataset was created from microscope slides containing patients' brushed epithelial cells from oral mucosa. The cells were collected from borders between normal tissue and abnormal wounds, and stained with argyrophilic staining methods (Trere, 2000). They were photographed using an Olympus CX41RF Binocular Microscope, using 100× magnification, with a mounted camera QImaging MicroPublisher 5 RTV. The dataset consists of 1,171 2,560×1,920-pixel RBG images.

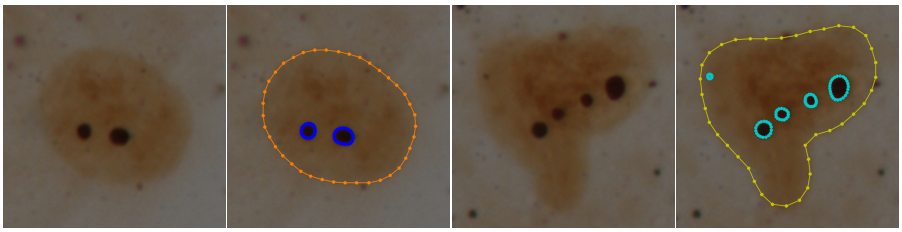
4.1.1 Semi-Automatic Dataset Annotation

The annotation process was performed semi-automatically starting with a *bootstrap CNN* to propose initial segmentation annotations that were then reviewed by specialists. To train and test this bootstrap CNN, 80 images from multiple patients were chosen at random from the images in our dataset and manually annotated using *labelme* (Wada, 2016) (Fig. 4.1). The pixels corresponding to nuclei and AgNORs were delimited and the remaining area was considered as background (Fig. 4.2 left). The annotated images



Source: The Authors.

Figure 4.1 – Creation of a bootstrap CNN for AgNOR-stained image segmentation. A small number of images (80) was manually annotated by specialists using *labelme* (Wada, 2016). The annotated images were used to train a bootstrap CNN to automatically segment AgNOR-stained images.



(a) Suitable nucleus. (b) Unsuitable (overlapping) nuclei.

Source: The Authors.

Figure 4.2 – Images with *labelme* annotation markups for suitable nucleus and AgNORs (left) and for unsuitable nucleus and AgNORs (right).

were divided into three subsets: training (60 images), validation (10 images), and test (10 images). To increase the number of training samples, a pipeline of image augmentation was built using Albumentations (Buslaev et al., 2020). The pipeline consisted of random brightness and contrast changes, vertical and horizontal flips, and two so-called elastic transformations. Each transformation has a 50% chance of being applied, and they are applied sequentially. The order of transformations is randomized for each image. This setup ensures variability in augmentation, ranging from no transformations to all transformations being applied. This data augmentation pipeline was applied six times to each of the 60 images in the initial training set, resulting in a training set with 420 images (60 original + 360 augmented).

The architecture of the bootstrap CNN consisted of ResNet-101 (He et al., 2016) as the encoder and U-Net (Ronneberger; Fischer; Brox, 2015) as the decoder. We chose ResNet-101 because it is used for feature extraction by Mask-RCNN (He et al., 2017), and U-Net because of its success in medical image segmentation. The combination of an encoder and a decoder is known as an encoder-decoder architecture. This kind of

architecture has been widely used in medical imaging segmentation (Minaee et al., 2022; Ghosh et al., 2019; Hesamian et al., 2019). The model was implemented using the Python library *Segmentation Models* (Yakubovskiy, 2019). Despite the fact that the model was trained on an RTX 8000 with 48GB of RAM, the large size of our images would constrain the batch size to only two images. To reduce the stochastic learning behavior caused by training with such small batch size, the images were rescaled to 1280×960 (1/4 of their original resolution). This approach was later disregarded in favor of slicing the images to avoid possible distortions that could affect the segmentation results. The rescaling led to a batch size of 10 images. The model was initialized with weights from the ImageNet dataset (Russakovsky et al., 2015) and fine tuned on our dataset for 100 epochs with a learning rate starting at 10^{-4} and reduced by a factor of 25% when no improvements were obtained for ten consecutive epochs. The loss function used was the sum of the categorical cross entropy (CE) and Dice loss (Yakubovskiy, 2019). The best model, obtained in the 30th epoch, was chosen among all epochs using Dice score as the model selection criterion. The resulting bootstrap CNN achieved a Dice score of 0.81 and intersection over union (IoU) of 0.75 on the test set.

We use Dice score and IoU to evaluate the performance of our model as they are widely used in the literature for image segmentation tasks (Amorim et al., 2020b; Yakubovskiy, 2019; Ronneberger; Fischer; Brox, 2015; Kirillov et al., 2019; Chaurasia; Culurciello, 2017). The Dice score is defined as follows:

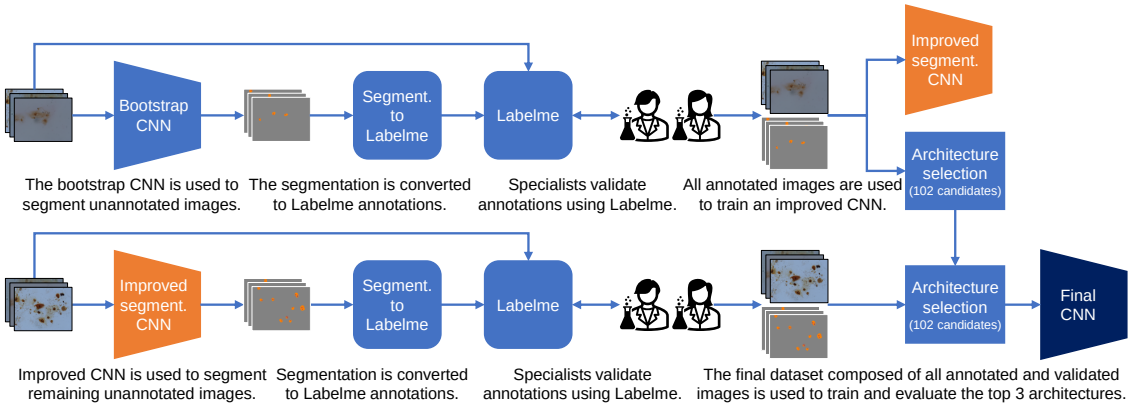
$$Dice\ score = \frac{2TP}{2TP + FP + FN} \tag{4.1}$$

Where TP, FP, and FN are the number of true positive, false positive, and false negative pixels, respectively. The IoU is defined as:

$$IoU = \frac{TP}{TP + FP + FN} \tag{4.2}$$

The Dice score (Equation 4.1) and IoU (Equation 4.2) range from 0 to 1, with higher values indicating better performance. The Dice score is more sensitive to false positives, while the IoU is more sensitive to false negatives.

The bootstrap CNN was then used to propose segmentation annotations for 558 unannotated images (Fig. 4.3 top). The predicted segmentation masks were post-processed using OpenCV (Bradski, 2000) to extract the contours of the nuclei and AgNORs. The



Source: The Authors.

Figure 4.3 – Joint refinement of our AgNOR-stained image segmentation CNN and dataset annotation. (top) A bootstrap CNN was used to segment AgNOR-stained images. Using *labelme* (Wada, 2016), specialists revised the predicted annotations. The revised images were used to train an improved segmentation CNN, as well as to select the best segmentation architecture. (bottom) The improved CNN was used to segment the remaining unannotated images, which were in turn revised by the specialists. The final dataset was used for fine-tuning the three best segmentation architectures, from which we chose the top one as our final AgNOR-stained image segmentation CNN.

extracted contours were analyzed using the algorithm described in Section 4.2.3.1 before being converted into *labelme* annotations. These annotations were validated by specialists, who used *labelme* to adjust or discard incorrect segmentation and include potentially missing ones. This process is summarized in Fig. 4.3 (top). The resulting set of 638 (558 + 80) annotated and validated images were then used to train an improved segmentation CNN (shown in orange in the rightmost part of Fig. 4.3 top).

4.1.2 Improving the Model for Annotation

An *improved segmentation CNN* was trained using the annotated images generated by the previously described process (Fig. 4.3 top). Its training regime was similar to the one used for the bootstrap CNN, except that no data augmentation was used, and the images and masks in the training and validation sets were sliced into four quadrants instead of being rescaled. Such slicing resulted in four new images and masks with 1/4 of the original resolution and no overlapping among the quadrants. Sliced images and masks containing only background pixels were discarded, as they contribute little additional information. We choose to call the slided images and masks as quadrants as opposed to patches given that patches are often used to refer to subimages of any size and shape,

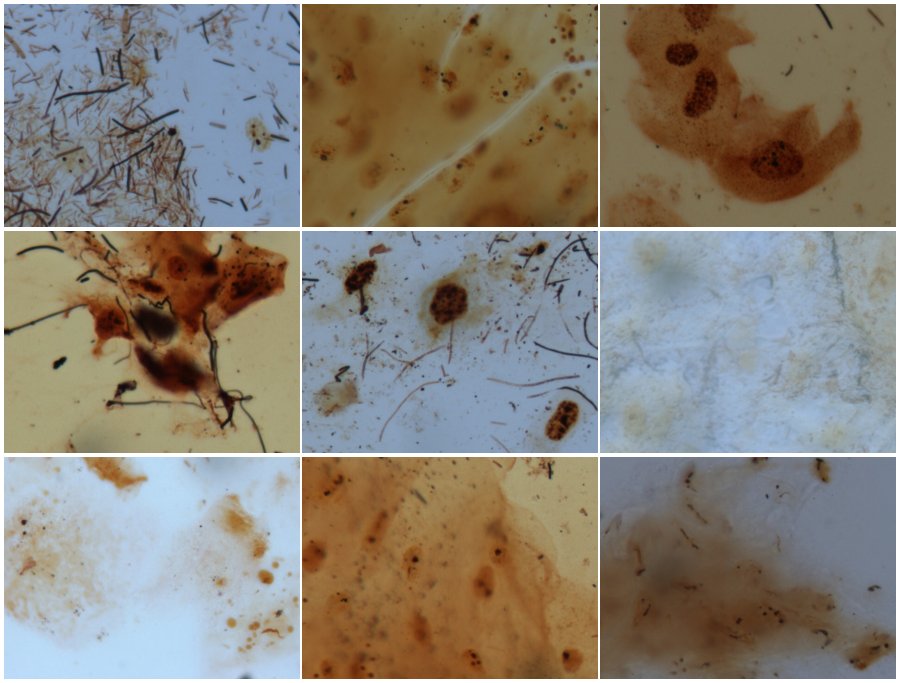
while quadrants are subimages that are always of the same size and shape and are obtained by dividing the original image into four equal parts.

To train the improved segmentation CNN, the image dataset was again divided in three sets: training, validation, and testing. The dataset split occurred in such a way that all cell images of a patient appeared only in one of the devised sets. The training set included 26 patients and 861 quadrant images, while the validation and test sets included 7 patients each. The validation set contained 234 (1280×960) quadrant images. The test set consisted of 103 images kept in their original resolution (2560×1920). Training was performed for 100 epochs using a batch size of 10 images.

When using the improved segmentation CNN for prediction, the input shape of the first layer was changed to match the shape of the original images (2560×1920 with three color channels). The increased number of images and the training with sliced images improved the results compared to the bootstrap CNN. It achieved better performance, with Dice score of 0.86 and IoU of 0.77. The improved segmentation CNN replaced the bootstrap one and was used to propose segmentation annotations for an extra set of 533 images, which were subsequently validated by the specialists (Fig. 4.3 bottom). In total, the specialists produced and/or validated annotations for 1,171 images. Out of these, 169 images were considered by them to contain "unsuitable" nuclei for AgNOR counting purposes. An image was considered unsuitable if it did not contain any discernible nuclei or if it was not possible for one to accurately tell if objects in the nuclei were AgNORs or *foreign* elements. Nevertheless, we kept these so-called unsuitable images in the dataset as negative examples, with their corresponding ground truth masks containing only background pixels. Fig. 4.4 shows examples of unsuitable images. Suitable images examples can be seen in Figs. 4.8, 4.7, 4.9 and 4.10. The final dataset then consists of 1,171 images from 48 patients, with an average of 24 images per patient. It was used for fine-tuning the top three AgNOR-stained image segmentation CNNs (Fig. 4.3 bottom right).

4.2 Our AgNOR Image Segmentation Model

To arrive at the most appropriate CNN model for our application, we explored many combinations of encoder and decoder architectures, as well as different loss functions. We also exploited *transfer learning* for image segmentation (Raghu et al., 2019; Ghosh et al., 2019) to take advantage of large scale pre-trained models. Transfer learning has been largely applied to medical imaging datasets (Hesamian et al., 2019; Wang



Source: The Authors.

Figure 4.4 – Examples of unsuitable images. They have nuclei partially occluded by fungi or materials with high silver precipitation levels. Nuclei and AgNORs with indistinguishable borders are also unsuitable.

et al., 2022; Jiang et al., 2022). Given the numerous options available for encoders and decoders (Huang et al., 2017; Tan; Le, 2019; Szegedy et al., 2016; He et al., 2016; Xie et al., 2017; Simonyan; Zisserman, 2014; Yakubovskiy, 2019), we evaluated 51 network architectures combining 17 encoders with three decoders and two loss functions (Dice loss (Yakubovskiy, 2019) and Focal loss (Lin et al., 2017)). In total, we trained and evaluated 102 models using 638 images from our dataset, distributed as training (413), validation (122), and test (103) images.

We used transfer learning for all trained models. The encoders were initialized with weights from the ImageNet dataset (Russakovsky et al., 2015) and the decoders were initialized with random values. The models were then fine tuned on our dataset for 100 epochs with a learning rate starting at 10^{-4} and reduced by a factor of 25% when no improvements were obtained for ten consecutive epochs. All dataset splits were performed on a patient level (*i.e.*, all images from the same patient were either in the training, in the validation, or in the test dataset). The list of all evaluated encoders, decoders, and loss functions is shown in Table 4.1.

The three best performing architectures were then retrained using our training and validation sets to obtain our final segmentation CNN (Fig. 4.3 bottom right). Next, we

Table 4.1 – Performance of the 102 CNN models trained and evaluated on a subset of our dataset. These models combine 17 encoders, three decoders, and two loss functions. The best results for each decoder and loss function are highlighted in bold.

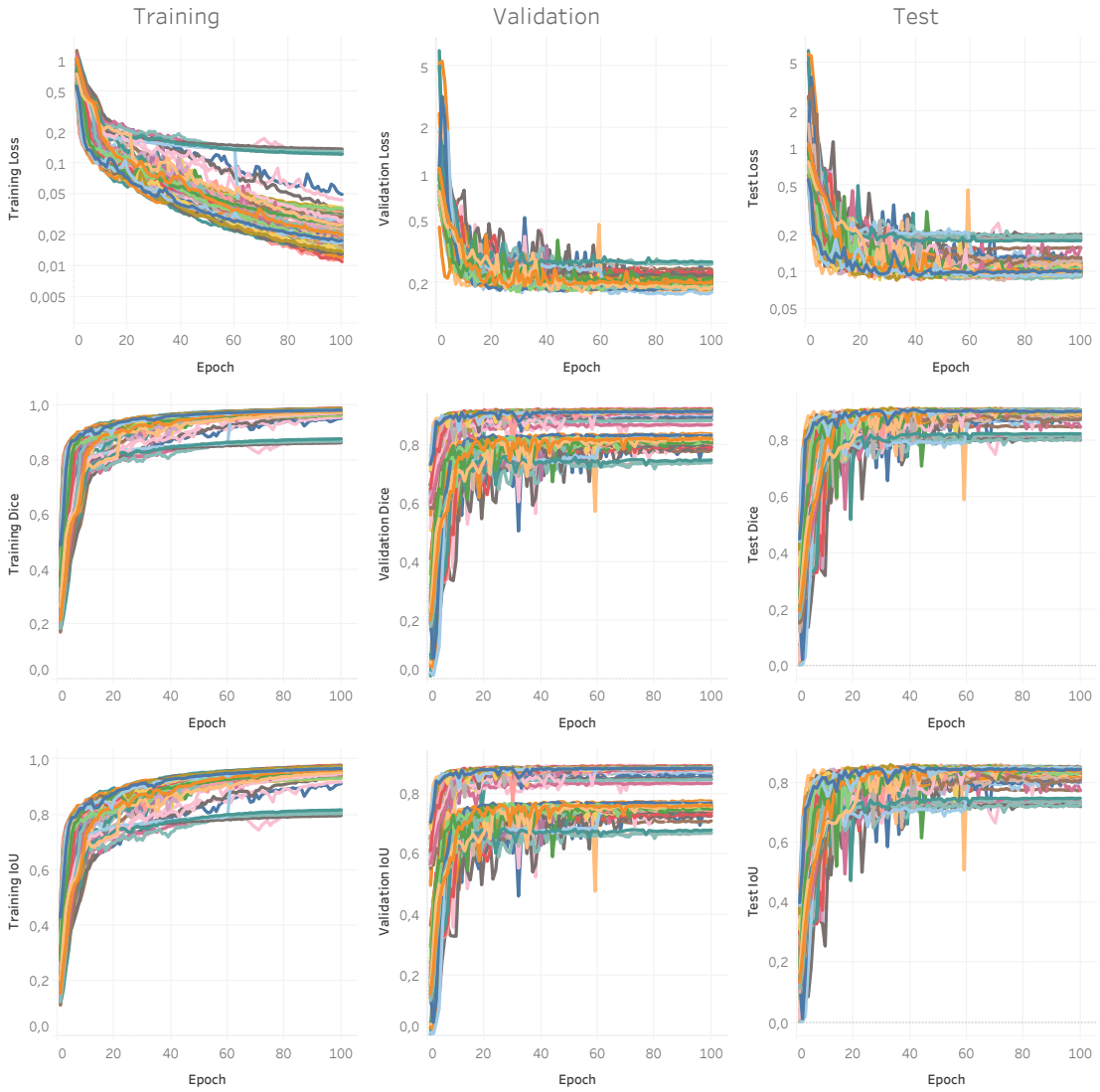
Loss function →	**CE + Dice loss**									**Focal loss**								
Decoder →	**FPN**			**LinkNet**			**U-Net**			**FPN**			**LinkNet**			**U-Net**		
Encoder ↓	Loss	Dice	IoU	Loss	Dice	IoU	Loss	Dice	IoU	Loss	Dice	IoU	Loss	Dice	IoU	Loss	Dice	IoU
DenseNet-121	0.0996	0.9046	0.8444	0.0985	0.9052	0.8453	0.0972	0.9073	0.8477	**0.0902**	**0.9112**	**0.8534**	0.0981	0.9032	0.8430	0.1005	0.9009	0.8400
DenseNet-169	**0.0862**	**0.9169**	**0.8603**	0.1830	0.8218	0.7459	**0.0940**	**0.9096**	**0.8507**	0.0908	0.9102	0.8517	**0.0882**	**0.9129**	**0.8552**	**0.0923**	**0.9088**	**0.8518**
DenseNet-201	0.0932	0.9111	0.8533	**0.0943**	**0.9100**	**0.8522**	0.0961	0.9074	0.8485	0.0911	0.9103	0.8522	0.0973	0.9039	0.8453	0.0974	0.9038	0.8435
EfficientNet-B0	0.1184	0.8877	0.8215	0.1244	0.8802	0.8130	0.1134	0.8913	0.8248	0.1082	0.8932	0.8288	0.1214	0.8804	0.8097	0.1042	0.8971	0.8333
EfficientNet-B1	0.1123	0.8922	0.8273	0.1083	0.8955	0.8307	0.1025	0.9027	0.8410	0.1049	0.8965	0.8324	0.1086	0.8927	0.8271	0.1926	0.8088	0.7291
EfficientNet-B2	0.1026	0.9014	0.8392	0.1998	0.8046	0.7257	0.1033	0.9013	0.8403	0.1040	0.8972	0.8335	0.1958	0.8056	0.7281	0.1026	0.8987	0.8365
EfficientNet-B3	0.1089	0.8961	0.8326	0.1072	0.8972	0.8331	0.1125	0.8923	0.8300	0.1043	0.8971	0.8362	0.1072	0.8943	0.8297	0.1153	0.8868	0.8226
Inception V3	0.1012	0.9042	0.8431	0.0979	0.9064	0.8455	0.0995	0.9039	0.8443	0.0940	0.9070	0.8481	0.1016	0.8996	0.8393	0.0936	0.9077	0.8485
ResNet-18	0.1021	0.9019	0.8395	0.1089	0.8964	0.8321	0.1083	0.8958	0.8328	0.0970	0.9044	0.8439	0.1062	0.8951	0.8331	0.0957	0.9057	0.8453
ResNet-34	0.1057	0.8983	0.8381	0.1062	0.8981	0.8357	0.0992	0.9042	0.8435	0.1088	0.8924	0.8306	0.1006	0.9008	0.8396	0.0930	0.9081	0.8501
ResNet-50	0.1112	0.8939	0.8313	0.1074	0.8968	0.8342	0.1058	0.8988	0.8384	0.1030	0.8986	0.8368	0.1050	0.8965	0.8350	0.0984	0.9027	0.8427
ResNet-101	0.1048	0.9018	0.8418	0.1014	0.9036	0.8425	0.1021	0.9015	0.8398	0.1019	0.8997	0.8384	0.1900	0.8115	0.7342	0.0998	0.9018	0.8413
ResNet-152	0.1011	0.9047	0.8447	0.1966	0.8087	0.7311	0.1015	0.9045	0.8448	0.0959	0.9051	0.8449	0.1915	0.8101	0.7312	0.0938	0.9072	0.8489
ResNeXt-50	0.0982	0.9067	0.8481	0.1032	0.9028	0.8412	0.0984	0.9055	0.8464	0.1015	0.9000	0.8391	0.0912	0.9099	0.8507	0.0958	0.9054	0.8466
ResNeXt-101	0.1010	0.9044	0.8448	0.0990	0.9057	0.8460	0.0977	0.9060	0.8470	0.0983	0.9032	0.8433	0.0991	0.9023	0.8425	0.0938	0.9073	0.8478
VGG16	0.1055	0.8982	0.8334	0.1146	0.8898	0.8238	0.1082	0.8949	0.8312	0.1129	0.8882	0.8218	0.1905	0.8109	0.7330	0.1041	0.8970	0.8343
VGG19	0.1152	0.8902	0.8235	0.1548	0.8510	0.7769	0.1150	0.8887	0.8243	0.1103	0.8909	0.8268	0.1155	0.8859	0.8198	0.1053	0.8958	0.8341

discuss the process of training and evaluating the 102 model candidates, followed by the training and evaluation of the best three performing models to arrive at our final model.

4.2.1 Training and Evaluating 102 Model Candidates

The architectures of the 102 model candidates (Table 4.1) were implemented using the Python library *Segmentation Models* (Yakubovskiy, 2019) and trained in parallel using three Nvidia RTX 8000 GPUs with 48 GB of memory each. The models using FPN (Kirillov et al., 2019), DenseNet-169 (Huang et al., 2017), DenseNet-201 (Huang et al., 2017), ResNet-152 (He et al., 2016), ResNeXt-101 (Xie et al., 2017), EfficientNet-B1 (Tan; Le, 2019), EfficientNet-B2 (Tan; Le, 2019), and EfficientNet-B3 (Tan; Le, 2019) required two GPUs for training. The others were trained on a single GPU. The models were trained and tested under the same regime described in section 4.1.2, except for the batch size reduced to 8 images to ensure the models fit in the GPU memory.

The number of epochs was set to 100 after performing some pre-training experiments to find a number large enough to allow all models to reach a plateau during training. This is illustrated in Fig. 4.5, which shows the training, validation, and test values for the loss, Dice score, and IoU for the 102 evaluated models. For each trained model, weights were saved at the end of each epoch. After training was concluded, the weights from all epochs were individually loaded and tested against the test set. This process is illustrated in the rightmost column of Fig. 4.5. The weights leading to the best Dice score for each model were selected as the final weights for that model. The initial learning rate value



Source: The Authors.

Figure 4.5 – Training, validation, and test values for the loss, Dice score, and IoU for the 102 evaluated models. The different colors represent the various combinations of backbones, decoders, and loss functions used. The first row displays the logarithmic loss of all the models during training. It can be observed that, after 100 epochs, the training loss continued to decrease, while the validation and test losses plateaued around the 50th epoch. The second and third rows show the Dice score and IoU, respectively. The same pattern can be seen, with validation and testing plateauing around the 50th epoch.

Table 4.2 – Comparison of the models trained on the subset dataset and the incremented dataset.

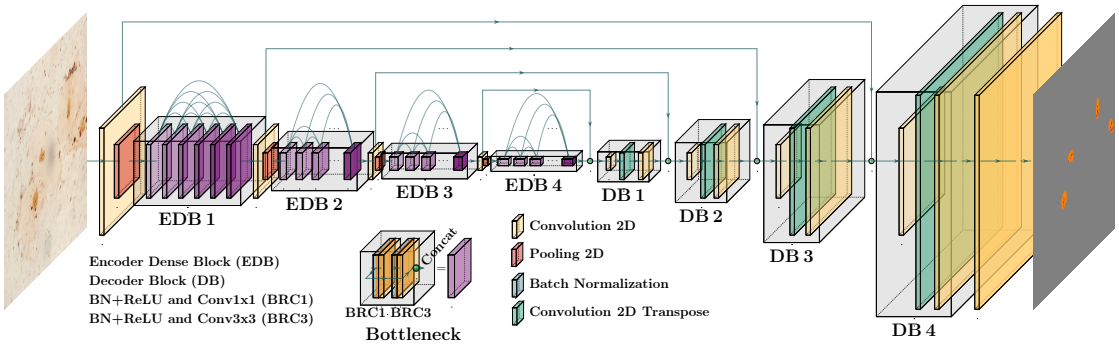
		Test dataset →	Subset			Complete		
Encoder	**Decoder**	**Train dataset ↓**	**Loss**	**Dice**	**IoU**	**Loss**	**Dice**	**IoU**
DenseNet-169	FPN		**0.0862**	**0.9169**	**0.8603**	**0.1182**	**0.8879**	**0.8230**
DenseNet-169	LinkNet	**Subset**	0.0918	0.9129	0.8552	0.1222	0.8863	0.8210
DenseNet-169	U-Net		0.0965	0.9081	0.8501	0.1325	0.8759	0.8100
DenseNet-169	FPN		0.0804	0.9220	0.8686	0.1041	0.9006	0.8391
DenseNet-169	LinkNet	**Complete**	**0.0799**	**0.9239**	**0.8705**	**0.1038**	**0.9025**	**0.8405**
DenseNet-169	U-Net		0.0862	0.9174	0.8619	0.1107	0.8956	0.8324

(10^{-4}) was also set based on pre-training experiments that showed it led to more stable learning.

On average, each model took 6 hours and 23 minutes to train. The total time required to train all models was over 652 hours. The results of the trained models are shown in Table 4.1. For each model, we show the value of the loss function, Dice coefficient, and IoU metrics.

4.2.2 Training the Best Models on the Complete Dataset

In the experiments reported in Table 4.1, the encoder DenseNet-169 achieved the highest Dice score with all three decoders. The best performing model based on the highest Dice score for each decoder (FPN, LinkNet, and U-Net) was selected for further training and evaluation. They were retrained using a complete version of our dataset. The retraining occurred under the same regime used to train the 102 models (Section 4.2.1), except that the selected models were initialized with the weights obtained during the selection phase (*Subset* in Table 4.2), as opposed to with the weights from the ImageNet dataset (Russakovsky et al., 2015). The retraining included 1,002 images containing at least one visible nucleus with AgNORs (i.e., images from the subset of suitable images). They were split into sets of 788 and 214 images for training and testing, respectively. No validation split was used. All images from the same patient were either in the training or in the test set. The images and masks in the training set were sliced into four quadrants, with no overlap among the quadrants. The resulting mask and image quadrants containing only background pixels were not used for training. The results of the retraining are shown in Table 4.2. The training and test sets in *Subset* are contained in the *Complete* training and test sets. The model consisting of *DenseNet-169 + LinkNet with Focal loss*



Source: The Authors.

Figure 4.6 – Our CNN architecture. Its encoder (downsampling portion) consists of the encoding layers of a DenseNet-169. Its decoder uses the upsampling layers of a LinkNet. The encoding layers bypass spatial information to the corresponding decoding layers using skip connections. The resulting architecture exploits the benefits of feature-map concatenation and skip connections: feature propagation reinforcement, feature reuse, and reduction in the number of required parameters.

performed best on the complete dataset. Thus, our selected CNN architecture uses the encoding layers of a DenseNet-169 in combination with the upsampling layers of a LinkNet as its decoder (Fig. 4.6). The resulting architecture exploits the benefits of feature-map concatenation and skip connections, achieving feature propagation reinforcement, feature reuse, and reduction in the number of required parameters.

4.2.3 Quantifying AgNORs per Nucleus

Before counting the number of AgNORs per nucleus, the system discards semantically invalid (segmented) nuclei and AgNORs: nuclei containing no AgNORs, as well as AgNORs not contained by any nucleus. It also removes nuclei and AgNORs whose sizes are outside some specified intervals. For images captured with 100$\times$ magnification, nuclei with sizes bigger than 67,000 pixels or smaller than 1,000 pixels, and AgNORs with sizes bigger than 3,600 pixels or smaller than 6 pixels are eliminated. These thresholds were determined after analyzing 1,002 images from 48 patients whose segmentation have been validated by specialists. For those images, the maximum and minimum nuclei sizes were 66,129 and 1,196 pixels, respectively, with average of 15,783 and standard deviation of 6,670 pixels. Likewise, the maximum and minimum AgNOR sizes were 3,521 and 6 pixels, respectively, with average of 92 and standard deviation of 171 pixels.

Our system uses OpenCV (Bradski, 2000) to extract contours and their hierar-

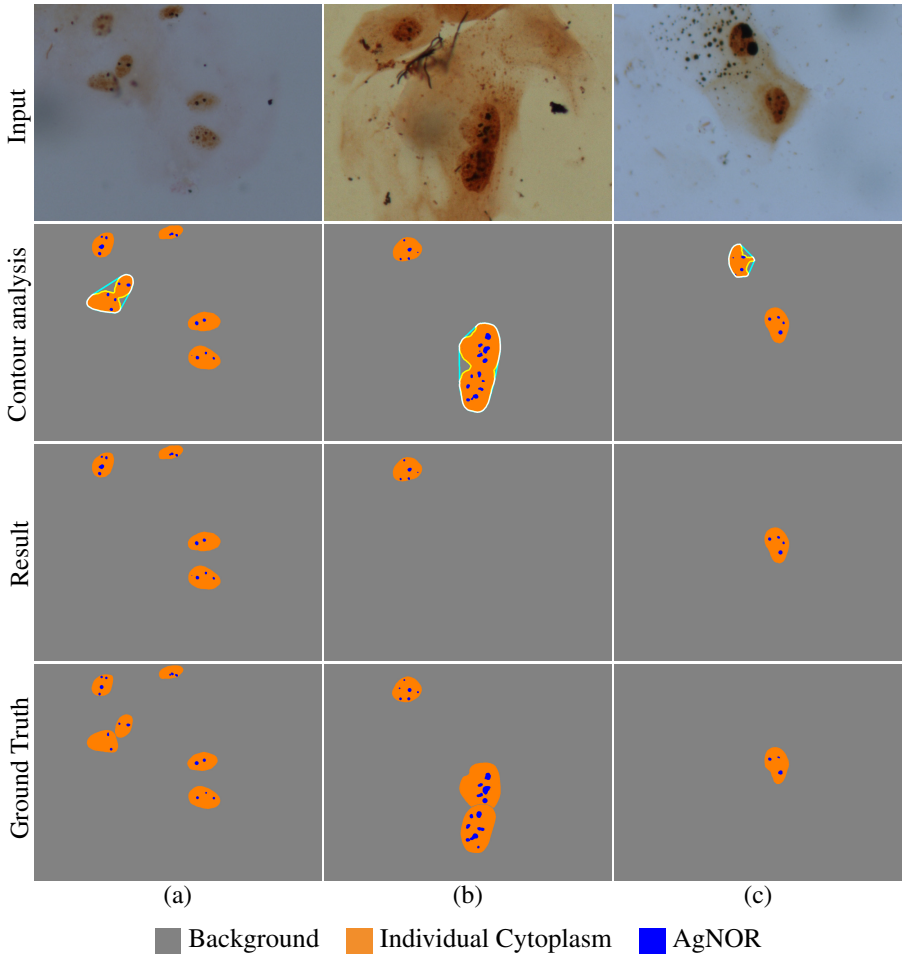
chy from segmented elements, allowing us to identify which AgNORs belong to which nucleus and count them.

4.2.3.1 Discarding Overlapping and Distorted Nuclei

Overlapping nuclei tend to hide AgNORs, thus affecting their true count. We developed a contour analysis algorithm (CAA) to optionally detect and discard potentially overlapping and distorted nuclei. Since cell nuclei tend to define a convex shape, the algorithm works by comparing the percentage difference in the number of pixels contained by each segmented nucleus and by its corresponding convex hull. If the difference exceeds the empirically defined threshold of 5%, the segmented element is discarded since it most likely contains recesses found in overlapping and deformed nuclei (Fig. 4.7). The 5% threshold value was determined based on the contours in the ground truth of over 890 images. Nuclei deformation may result from slice manipulation or improper segmentation (e.g., due to occlusion - Fig. 4.7 c). Another option is to use morphological operations to erode the nucleus and then determine whether this resulted in the nuclei being separated. However, this approach is most effective when nuclei are not too close together, which limits its applicability in our case.

4.2.3.2 Classifying AgNORs Based on Their Relative Sizes

The size and number of AgNORs in nuclei can be an indicator of lesions with malignant potential (Jajodia et al., 2017). Our system can classify AgNORs as “clusters” (big) or “satellites” (small), depending on their relative sizes according to other AgNORs in the same nucleus. Thus, two AgNORs with the same size, but in different nuclei can be classified differently. To address this subtle issue, we trained a decision tree classifier using Scikit-Learn (Pedregosa et al., 2011). Our decision tree model takes as features the ratios of the sizes (in pixels) between the AgNOR itself and three measures: the size of the nucleus to which it belongs to, the size of the biggest AgNOR, and the size of the smallest AgNOR, both in the given the nucleus. The model was trained using annotated data generated by specialists, which included 749 “clusters” and 173 “satellites”. The data was split into training (70%) and test (30%) sets. The resulting model estimates the number of “clusters” and “satellites” in each nucleus, and achieved 0.84 and 0.80 of precision and recall, respectively. We believe that information about size and number of AgNORs may lead to new insights on the dynamics of pre-cancer development.



Source: The Authors.

Figure 4.7 – Examples of use of the contour analysis algorithm. In (a) and (b) the algorithm detected and discarded overlapping nuclei. In (c) it detected and discarded a severely deformed nuclei segmentation.

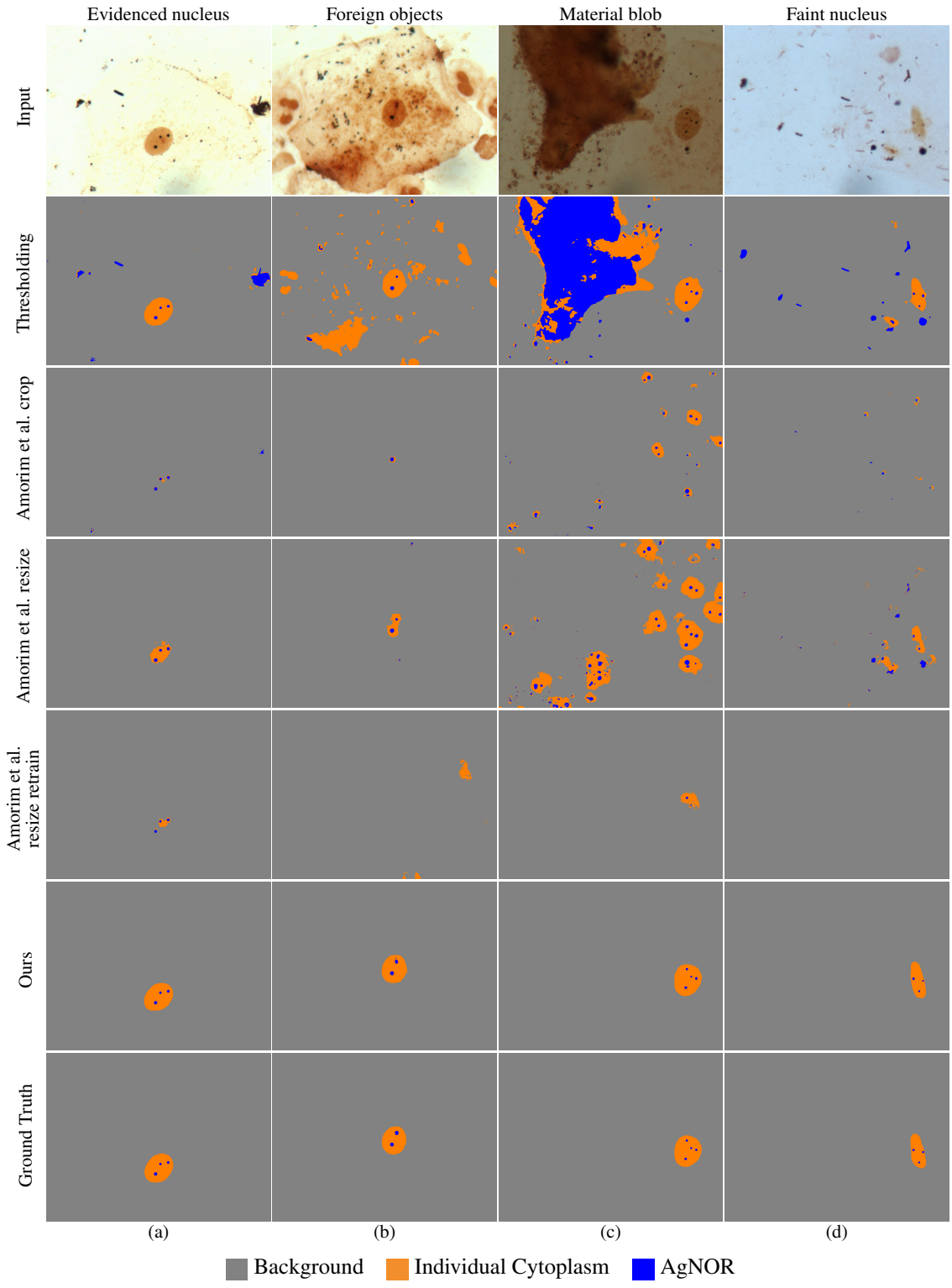
4.3 Results

This section presents some results and compares our method to Amorim et al. (2020b)'s on both our (UFRGS AgECOM (Rönnau et al., 2023c)) and their (CCAgT (Amorim et al., 2020a)) datasets. We also compared our results on the same datasets with the thresholding segmentation (Ferreira et al. (2011), García-Vielma et al. (2016), and Teresa et al. (2007)). Since the details of the thresholding segmentation used in these works are not available (Section 3.1.1), we applied thresholding segmentation using ImageJ (Schneider; Rasband; Eliceiri, 2012) and manually adjusted the threshold values for each image aiming for the best possible results. Color thresholding was performed in four steps: first, we applied a Gaussian blur using $\sigma = 3$ pixels; second, we segmented the pixels corresponding to nuclei by interactively selecting values for hue, saturation, and brightness; third, we segmented the pixels corresponding to AgNORs in a similar fashion; fourth, we combined the segmentation masks and saved the result as an RGB image. Pixels not corresponding to nuclei nor to AgNORs were considered background. Although heavily relying on user interaction and being time-consuming, this process tends to produce poorly segmented and noisy results (see second row of Fig. 4.8). Manually segmenting each input image shown in Fig. 4.8 using color thresholding took an average of 2 minutes and 40 seconds per image. In comparison, our model can segment hundreds of images under one minute.

Segmentation results for typical as well as challenging scenarios in both datasets are shown in Fig. 4.8 (UFRGS AgECOM) and Fig. 4.9 (CCAgT). In both figures, column (a) shows images exhibiting high contrast between one nucleus and the corresponding background; (b) shows one cell sprinkled with foreign objects that can be confused with AgNORs; (c) shows one cell close to a mass of organic material collected during the brushing processes with silver precipitation; and (d) shows nuclei that appear fainted with respect to the background. For the examples shown in Fig. 4.8, our results nicely match the ground truth. In contrast, Amorim et al.'s technique and color thresholding do not produce satisfactory segmentation.

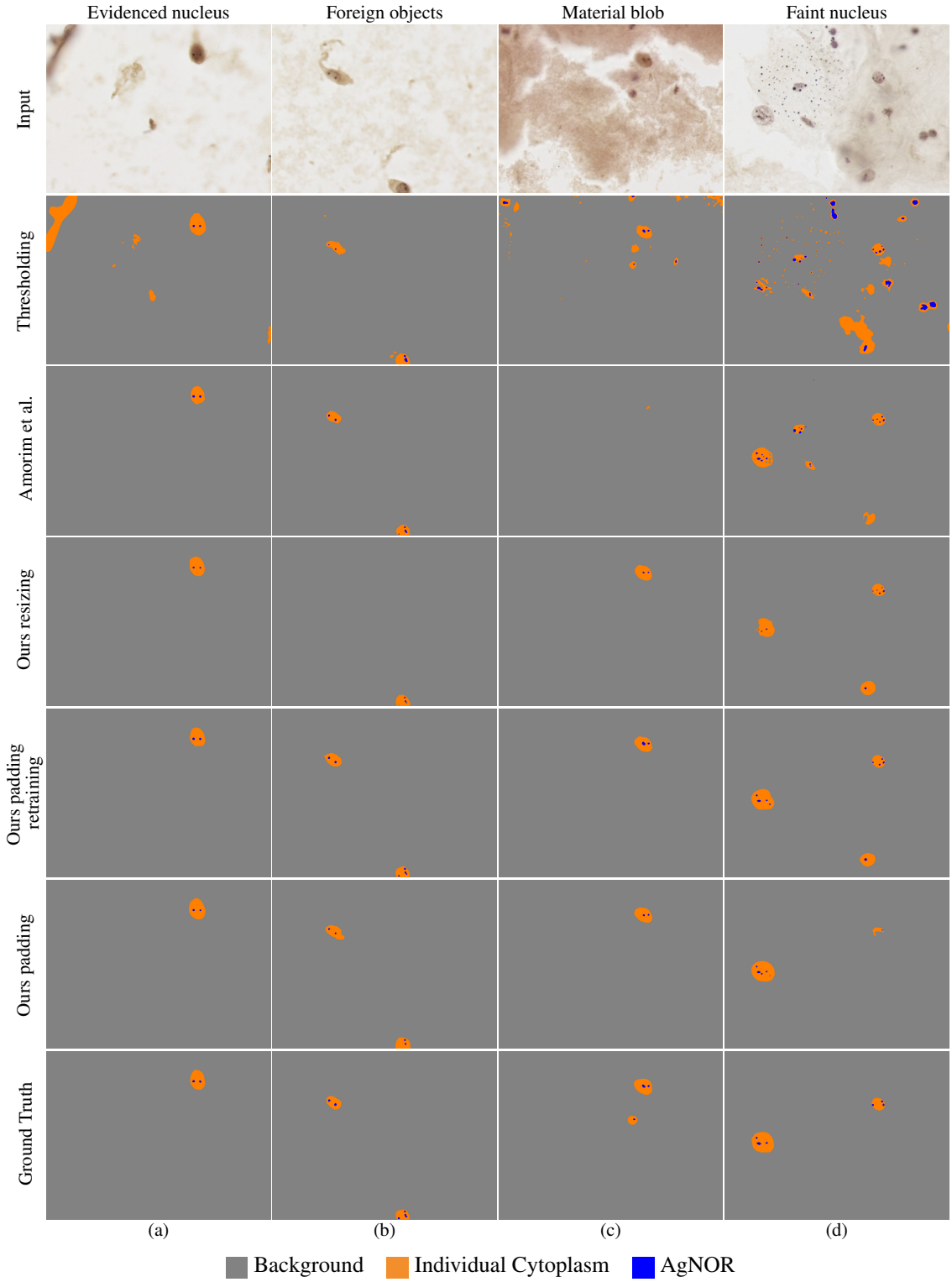
4.3.1 Quantifying AgNORs per Nucleus Results

To evaluate our system's performance, we compare the predicted nuclei and AgNORs to the ground truth masks provided by specialists. The analysis consisted of calcu-



Source: The Authors.

Figure 4.8 – Comparison of segmentation results produced by our model, by Amorim et al.'s, and with color thresholding for typical and challenging images from our dataset. Amorim et al.'s results are shown considering retraining on our dataset (UFRGS AgECOM), and image cropping and resizing to match the image dimensions in their dataset. Threshold segmentation represents the works by Ferreira et al., García-Vielma et al., and Teresa et al.



Source: The Authors.

Figure 4.9 – Segmentation results produced by our method, Amorim et al.'s, and thresholding segmentation on images of the CCAgT dataset. Our method produced segmentation results that better match the ground truth in all tested scenarios.

Table 4.3 – Results of our method for counting nuclei and AgNORs.

Dataset	Metric	Nucleus	AgNOR
Ours	Precision	0.9420	0.8155
	Recall	0.9002	0.7379
CCAgT	Precision	0.8864	0.8823
	Recall	0.7732	0.6799

lating the number of true positive, false positive, and false negative nuclei and AgNORs in the predicted segmentation. A true positive corresponds to a predicted object (nuclei or AgNOR) intersecting at least 50% of the corresponding pixel mask in the ground truth. A false positive corresponds to a predicted object not found in the ground truth. If a prediction misses an object in the ground truth, this is a false negative.

In total, 214 images were used, corresponding to the test set of our dataset. The images contain 541 annotated nuclei and 2,270 AgNORs. Out of those, our method correctly identified (true positives) 487 nuclei and 1,675 AgNORs. There were 30 false positive and 54 false negative nuclei, 379 false positive and 595 false negative AgNORs. Applying the same method on Amorim et al.'s CCAgT dataset, we used our retrained model described in subsection 4.3.3. The test set from CCAgT contains 331 images, 626 nuclei, and 1,896 AgNORs (Amorim et al., 2020a). Our method was capable to correctly identify (true positives) 484 nuclei and 1,289 AgNORs. It produced 62 false positives and 142 false negatives for nuclei, 172 false positives and 607 false negatives for AgNORs.

Table 4.3 summarizes the precision and recall results of our method on both datasets. It performed well on both, despite the fact that the CCAgT dataset consists of cervical cells (Amorim et al., 2020a).

4.3.2 Quantifying AgNORs in User Selected Nuclei

As part of a protocol when manually counting AgNORs, experts tend to restrict themselves to 20 up to 100 nuclei (Rajput; Tupkari, 2010), often selecting the *best* ones in a set of slide images. Thus, we provide our users the ability to indicate, for each image, which nuclei should be considered for counting. This is particularly useful for images with low contrast or containing fungi and other organic materials, and allows users to better reproduce their daily working process. Currently, this is done by interactively specifying a rectangle (bounding box) around the each selected nucleus in the slide image

using *labelme*.

When we applied our method to our test set considering only nuclei within bounding boxes drawn by specialists, the number of nuclei and AgNORs was reduced to 242 and 922, respectively. In this scenario, the values of precision and recall for nuclei increased to 1.0 and 0.99, respectively, and for AgNORs they increased to 0.86 and 0.92, respectively. These numbers represent a significant improvement over the "in-the-wild" results shown in Table 4.3.

4.3.3 Comparison with Other Segmentation Model

The results presented in Figs. 4.8 and 4.9 show that thresholding segmentation, even when specifically adjusted for each image does not produce satisfactory results. Thus, in this section we restrict the comparison of our results with Amorim et al.'s, since they also use a CNN.

Fig. 4.8 shows that our model properly segments nuclei and AgNORs under a variety of conditions. This section compares our results with Amorim et al.'s on both datasets. Since Amorim et al.'s model originally uses 1600×1200-pixel images as input (Amorim et al., 2020b), while the images in our dataset have 2560×1920 pixels, we ensure compatibility in the tests on our dataset applying the following strategies before prediction:

i. Cropping: Crop the central portion of the images and masks in our dataset to match Amorim et al.'s model input size. Cropped images with no nucleus (7 in total) were discarded;

ii. Resizing: Resize the images and masks in our dataset to match Amorim et al.'s model input size;

iii. Retraining: Retrain Amorim et al. (2020b)'s model on the resized versions of the images and masks in our dataset. For this, we used the code provided by the authors (Amorim, 2020). The model was initialized and trained as described in their paper (Amorim et al., 2020b). The retrained model was used for prediction on the resized images.

Fig. 4.8 illustrates the segmentation results produced by Amorim et al.'s model under these three scenarios. In all of them, the results are not consistent with the ground truth, containing a number of false positives and false negatives. It worth noting that the retrained model did not show significant improvement over the other two.

We also tested our model on the CCAgT dataset of cervical cells provided by

Table 4.4 – Comparison of model metrics in the AgNOR datasets.

Dataset	Model	Method	Dice*	IoU*
Ours	Ours	-	**0.9025**	**0.8405**
	Amorim et al.	Crop	0.3889	0.3744
		Resize	0.5118	0.4463
		Retrain	0.5512	0.4970
CCAgT	Ours	Pad	0.6173	0.5631
		Resize	0.7327	0.6619
		Retrain	0.8075	**0.7388**
	Amorim et al.	-	**0.8340**	0.6813

** Values consider the background class.*

Table 4.5 – Performance comparison of our model with human experts on 291 images from 6 new patients.

Patient	Count	Expert 1	Expert 2	Expert 3	Our Solution
A	# Nuclei	79	59	79	76
	# AgNOR	225	179	277	247
B	# Nuclei	58	54	60	60
	# AgNOR	135	132	147	153
C	# Nuclei	60	53	60	59
	# AgNOR	148	124	157	137
D	# Nuclei	52	49	52	52
	# AgNOR	118	117	129	131
E	# Nuclei	44	42	44	43
	# AgNOR	117	118	124	119
F	# Nuclei	81	69	82	71
	# AgNOR	274	238	307	290
Total	# Nuclei	374	326	377	361
	# AgNOR	1,017	908	1,141	1,077
	Time	$\approx$ 2h	$\approx$ 3h	$\approx$ 2h	2m26s

Amorim et al.. Again, to ensure the compatibility in the tests on their dataset, we apply the following strategies before using our model for prediction:

i. Padding: Our model supports images with various sizes, but our CNN architecture requires their dimensions to be multiples of 32. Thus, we zero-padded 16 rows of pixels at the bottom of the images and masks in the dataset. The corresponding rows were removed from the segmentation results;

ii. Resizing: Resize the images and masks in the CCAgT dataset to match the input shape of our model;

iii. Retraining: Retrain our model on the resized images and masks from the CCAgT dataset. The model was initialized with the weights obtained from training on our dataset.

Fig. 4.9 compares the segmentation results produced by our method under these three scenarios and by Amorim et al.'s considering examples from four classes of images.

For these examples, our method outperformed Amorim et al.'s, better matching the ground truth in all scenarios.

Table 4.4 summarizes the performance of the two models on both datasets across the considered scenarios, using Dice score and intersection over union (IoU). Our method achieved significantly higher scores when evaluated on our dataset. For the CCAgT dataset, our retrained model achieved better IoU score and a Dice score very close to the one obtained by Amorim et al.'s. This highlights the robustness of our CNN architecture and its ability to handle different datasets.

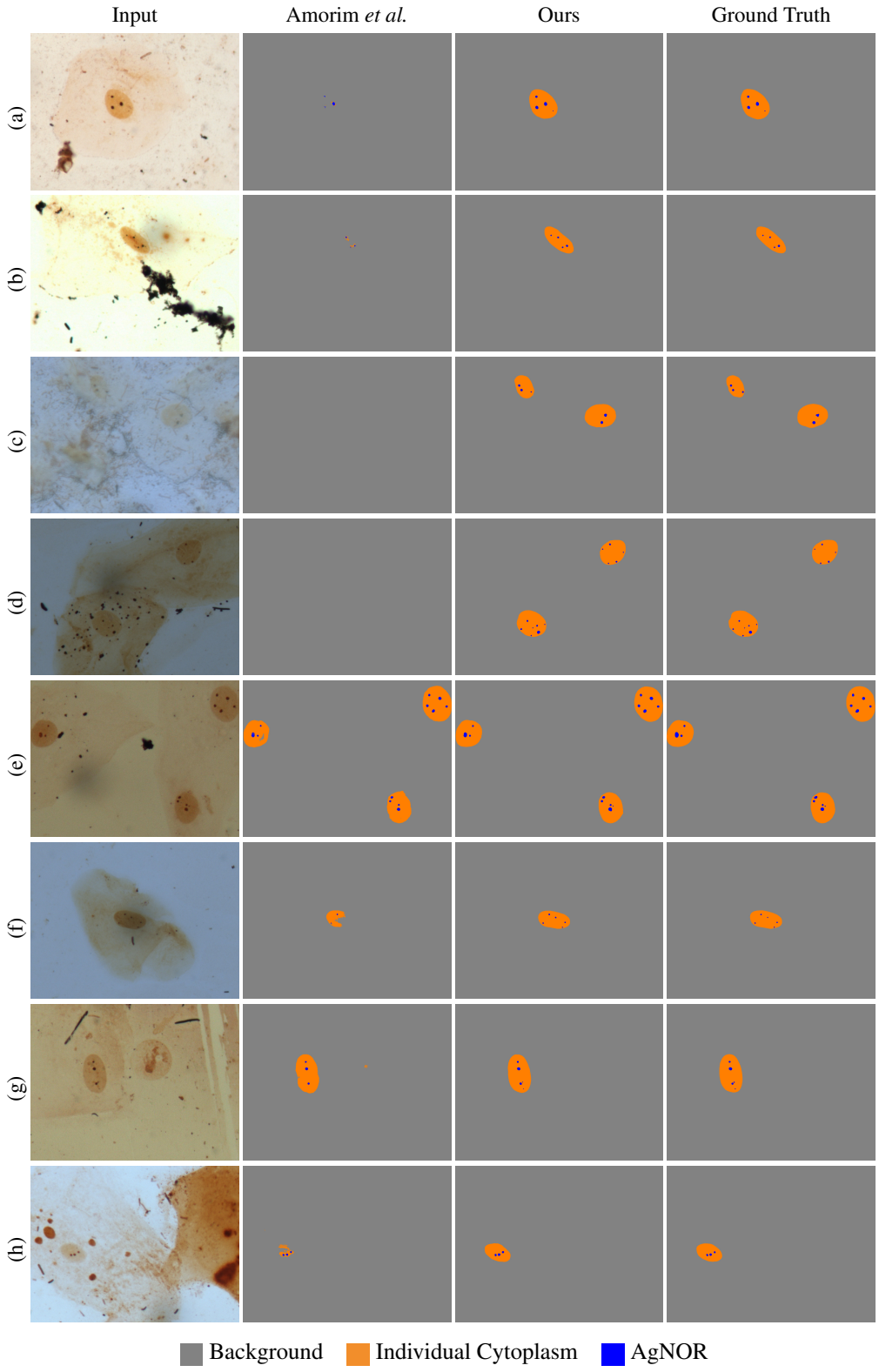
Fig. 4.10 also illustrates the robustness of our method on a variety of challenging images from our test dataset. Note how its results match the ground truth. For comparison, we show the results produced by Amorim et al.'s model retrained and evaluated on resized images of our dataset.

Limitation: Our CNN architecture requires the dimensions of the input images to be multiples of 32, which is due to the convolutional and pooling layers of its encoder. This limitation is easily overcome by padding the input images and discarding the corresponding rows/columns in the segmentation results, as demonstrated in the case of the CCAgT dataset (Fig. 4.9).

4.3.4 Comparing Our Model with Human Experts

We validate the robustness of our CNN model by comparing its performance against conventional counting (*i.e.*, visual inspection) of nuclei and AgNORs performed by three human experts on a selected set of nuclei in 291 AgNOR-stained images from six new patients. These images were captured using a Nikon Eclipse SI microscope with a Nikon Prime CAM 6 camera (different equipment from the one used to capture the training dataset).

The results of this experiment are summarized in Table 4.5, which includes subject-wise comparisons for nuclei and AgNORs. The number of nuclei identified by the experts ranged from 326 to 377, while the number of AgNORs ranged from 908 to 1,141. Our model identified 361 nuclei 1,077 AgNORs. Table 4.5 also compares the amount of time taken to perform the task. Our solution took 2 min and 26 sec on an RTX 3090 GPU. On a laptop with an RTX 2060 GPU, the time was 3 min and 10 sec. These times include loading, predicting, finding and discarding overlapped nuclei, and saving results to disk. The experts took from two to three hours. Since this is a visually tiring process, they



Source: The Authors.

Figure 4.10 – Application of our method on a series of challenging images from our test dataset. (a) and (b) depict nuclei close to foreign objects. (c) depicts a cloudy nuclei. (d) and (e) show examples of silver precipitation resulting in dark spots outside the nuclei that resemble AgNORs. (a), (b), (e), (f), and (g) show highly contrasted nuclei with respect to the cytoplasm. (h) shows a fainted nucleus near a mass of organic material with some silver precipitation on top. The results produced by our model shows that it can robustly segment nuclei and AgNORs under various conditions. The ground truth and the results produced by the version of Amorim et al.'s model retrained and evaluated on resized images are shown for comparison.

performed the task in multiple sessions.

We use the Intraclass Correlation Coefficient (ICC) to assess agreement between experts and our solution, employing the two-way mixed effects model based on single ratings. This analysis was conducted using IBM SPSS Statistics (IBM Corp., 2023). The calculated ICC value for nuclei is 0.91, with a p-value < 0.001 and a 95% confidence interval of [0.89, 0.93]. For AgNORs, the ICC value is 0.81, with a p-value < 0.001 and a 95% confidence interval of [0.77, 0.84]. According to Koo and Li (2015), agreement values above 0.90 represent excellent agreement and values between 0.75 and 0.90 indicate good agreement. Thus, our solution achieved strong agreement with the experts, with the low p-values confirming the statistical significance of these results.

Given its high agreement with human experts, our solution is currently in use in the graduate program of the Faculty of Dentistry at UFRGS, in parallel with the conventional method, in an experimental phase seeking validation for clinical use.

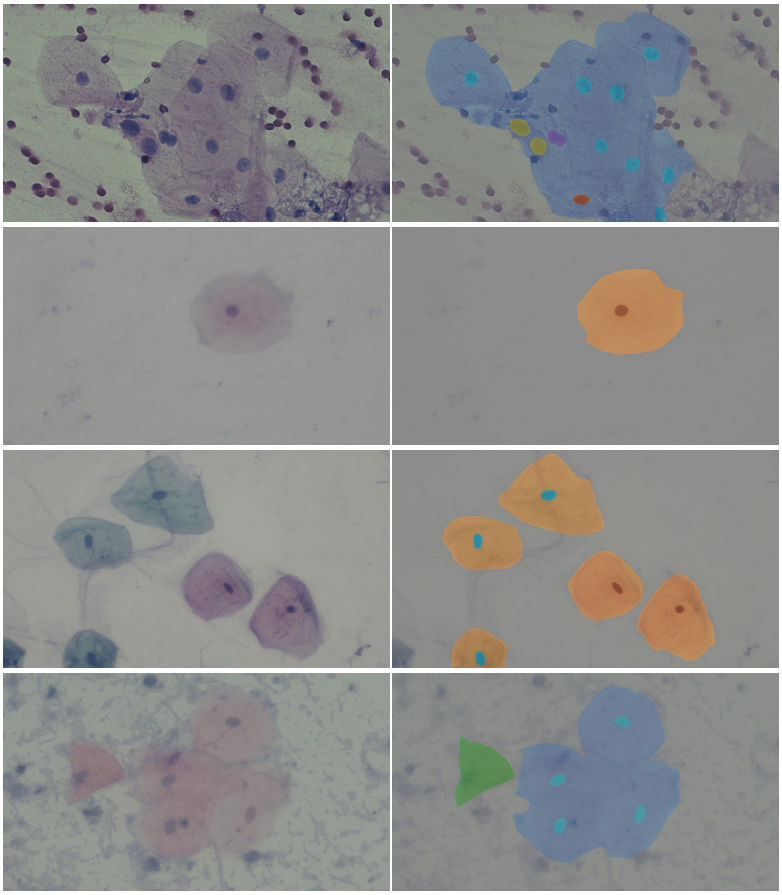
5 AUTOMATIC SEGMENTATION AND CLASSIFICATION OF CELLS IN PAPANICOLAOU-STAINED IMAGES

This chapter presents our work on automatic segmentation and classification of cells in Papanicolaou-stained images. We start by describing our dataset of Papanicolaou-stained images from the oral mucosa, UFRGS Pap-OMD (Rönnau et al., 2024), which was annotated by specialists. We then present our CNN model for image segmentation, which was trained on this dataset. The model performs per-pixel segmentation and classification of Papanicolaou-stained cell nuclei and cytoplasm. We then describe the improvements we made to the model to better segment Papanicolaou-stained images. These improvements include the addition of a new layer and two post-processing steps. We show the results of our model on our dataset as well as on five publicly available datasets of Papanicolaou-stained images from cervix cells. Finally, we compare the results of our model with the ones produced by three human experts using Intraclass Correlation Coefficient (ICC).

5.1 Our Papanicolaou-stained Oral Mucosa Dataset

Our dataset of Papanicolaou-stained images from the oral mucosa, dubbed UFRGS Pap-OMD (Rönnau et al., 2024), consists of 1,563 Papanicolaou-stained images from the oral mucosa of 52 patients. On average, the dataset contains 2.69 cells per image with standard deviation of 3.06. The images were acquired using a Nikon Eclipse SI microscope with a Nikon Prime CAM 6 camera, with resolution of $1,920 \times 1,080$ pixels and three color channels (RGB). These images were annotated by specialists using the software *labelme* (Wada, 2016). The specialists used *labelme* to interactively define polygons delimiting individual elements for the following classes, whose color annotations are shown in parentheses: *individual cytoplasm* (orange), *squamous cell* (green), *superficial cell nucleus* (red), *intermediate cell nucleus* (cyan), *suspicious cell nucleus* (yellow), *binucleate nuclei* (purple), *cytoplasms of cell cluster* (blue). The remaining pixels were considered *background* (gray). Fig. 5.1 shows examples of images from our annotated dataset along with their corresponding color annotations overlaid on them.

For the development and testing of our CNN model, we used 1,163 images from 32 patients. The training, validation, and test sets contained 925, 123, and 115 images,



Source: The Authors.

Figure 5.1 – Examples of annotated images from our dataset. (a) Original images. (b) Expert's annotations overlaid on (a).

respectively. The splits of the dataset were done at the patient level (*i.e.*, all images from any given patient only appear either in the training, validation, or test set).

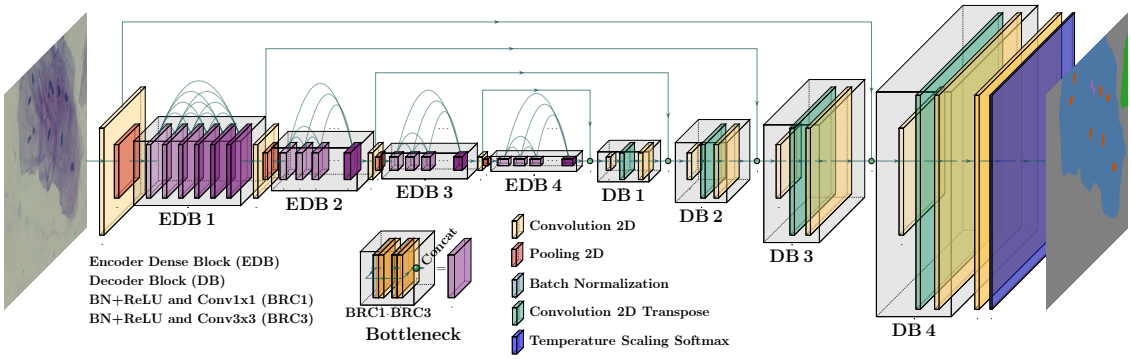
The remaining 400 images from 20 individuals were reserved to statistically compare the results of our model with the ones produced by three human experts (Subsection 5.3.2). To be able to evaluate our model in all expected scenarios, these 20 individuals were chosen from four groups, with five individuals in each group. The groups comprise, respectively: (i) *patients with squamous cell carcinoma* (SCC); (ii) *patients with oral potentially malignant disorders* (OPMD); (iii) *patients exposed to carcinogens* (e.g., *tobacco and alcohol) but without lesions in the oral mucosa* (EXP); and (iv) *control group consisting of healthy patients* (CTL).

5.2 Our Papanicolaou Image Segmentation Model

Our model performs per-pixel segmentation and classification of Papanicolaou-stained cell nuclei and citoplasm. Cell nuclei can be classified as *suspicious* (nuclear atypia), *superficial*, *intermediate*, and *binucleate* based on their morphological and staining properties. Individual cells are classified after their nuclei types, or as anucleate squamous if they have no nucleus. A cell cluster corresponds to a number of cells grouped together and is considered suspicious if at least one cell in the cluster contains a suspicious nucleus. The cytoplasm of individual cells and the aggregated cytoplasms of cell clusters belong to different pixel classes. Squamous cells are anucleate. Background pixels form an additional class of pixels. Fig. 5.1 shows examples of images from our dataset, with the corresponding annotated classes.

We use transfer learning for image segmentation to train our model. Transfer learning is a popular technique to take advantage of pre-trained large-scale models (Raghu et al., 2019; Ghosh et al., 2019) and has been largely applied to medical imaging datasets (Hesamian et al., 2019; Wang et al., 2022; Jiang et al., 2022). There are numerous options available for encoders and decoders (Huang et al., 2017; Tan; Le, 2019; Szegedy et al., 2016; He et al., 2016; Xie et al., 2017; Simonyan; Zisserman, 2014; Yakubovskiy, 2019). We base our model on the architecture we used in our previous work for AgNOR-stained image segmentation (Rönnau et al., 2023a). It is an encoder-decoder architecture consisting of *DenseNet-169* + *LinkNet* (Huang et al., 2017; Chaurasia; Culurciello, 2017), which was selected after a systematic evaluation of 102 alternatives involving multiple encoders, decoders, and loss functions. This architecture was then customized to improve the quality of the segmentation results for Papanicolaou-stained images, and fine-tuned on our own dataset. The customization is discussed in Section 5.2.1, and the resulting architecture is shown in Fig. 5.2.

To implement the customized architecture, we use the *Segmentation Models* library (Yakubovskiy, 2019). The model was initialized with weights from the ImageNet dataset (Russakovsky et al., 2015) and trained using the Adam optimizer (Kingma; Ba, 2014) with a learning rate of 10^{-4} and a batch size of 8 images. The images were resized to 960×544 pixels to fit in the GPU memory. We resize the image height to 544 instead of 540 (half the number of image rows) because the DenseNet-169 model implemented by *Segmentation Models* requires input shapes with dimensions multiple of 32 to prevent shape mismatch errors.



Source: The Authors.

Figure 5.2 – Architecture of our CNN model. The encoding layers are based on DenseNet-169 (Huang et al., 2017), and the decoding layers are based on LinkNet (Chaurasia; Culurciello, 2017). The decoder is modified by replacing the regular softmax layer with a temperature scaling softmax (shown in blue) with a temperature parameter value of 0.1 to increase the confidence of the predictions and avoid the bias towards the prediction of background pixels. The model's prediction is further processed by a semantic reclassification and a segmentation artifact removal steps. See Sections 5.2.1 to 5.2.4 for details about these components.

The model was trained for twenty epochs, each containing ten thousand batches of 8 images. We used image augmentation functions provided by TensorFlow to artificially increase the number of training samples applying brightness, contrast, hue, and saturation changes with 50% of chance of occurrence (brightness delta $\in$ [-0.2,0.2], contrast $\in$ [0.6, 1.6], hue delta $\in$ [-0.2, 0.2], saturation $\in$ [0.6, 1.6]). The model weights were saved to disk at the end of each epoch. The training was performed on a RTX 3090 GPU with 24 GB of memory and took 28 hours. After the training finished, all saved weights were loaded and evaluated on the test set. For the tests, we first predicted the segmentation mask for each image in the test set at 960×544 resolution. Then, we resized the predicted segmentation mask to the original image resolution $(1,920 \times 1,080)$ using nearest neighbor resampling. Finally, we evaluated the resized mask against the ground truth segmentation mask using the Dice score and Intersection over Union (IoU) metrics. The best performing weights on the test set were selected for the final model, which achieved a Dice score of 0.66 and IoU of 0.65.

5.2.1 Improving Segmentation and Generalization

Papanicolaou-stained images typically have a significantly larger number of background pixels compared to nuclei pixels (see Fig. 5.5), which tends to introduce some pixel classification bias towards background. Also, cytoplasm pixels from individual

cells and from clusters of cells look similar, introducing some ambiguity to the model. To address these issues, we extended the base architecture with a new layer and added two "post-processing" steps. Together, these three new components help to improve the model's segmentation and classification results. The first new layer is *temperature scaling softmax* (TSS). It replaces the original softmax layer and helps to avoid the bias towards the prediction of background pixels. The post-processing steps consist of *semantic re-classification* (SR) and *segmentation artifact removal* (SAR). The details involving TSS, SR and SAR are presented in sub-sections 5.2.2 to 5.2.4. Fig. 5.5 shows the application of these components to images from six different datasets (including five of cervical cells), which exhibit significantly different characteristics in terms of colors, cell sizes, background noise, etc. This demonstrates the effectiveness of our solution to generalize to images with highly-distinct spatial features.

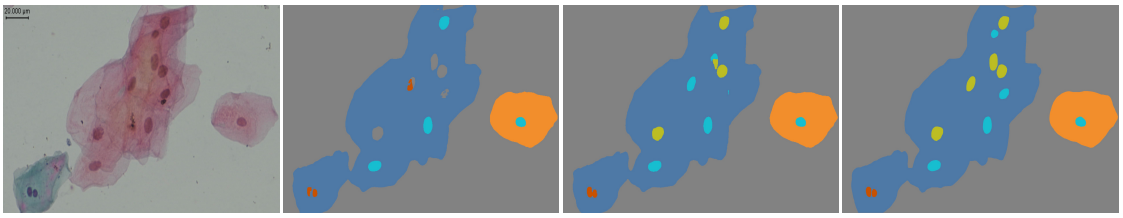
5.2.2 Temperature Scaling Softmax Layer

Temperature Scaling Softmax (TSS) is a technique for calibrating the confidence of a model's predictions. As such, it can improve the reliability of the confidence scores associated with those predictions, and set more appropriate thresholds, potentially improving classification decisions. Calibrated confidence scores can provide more realistic uncertainty estimates, which is essential in applications involving medical images, where handling uncertainty is critical.

The calibrated probabilities p_i are obtained by dividing the model's predicted logits z_i by a "temperature" parameter T before applying the regular softmax function:

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{5.1}$$

According to Equation 5.1, a value of $0 < T < 1$ stresses the differences among the estimated probabilities, indicating higher confidence in the predictions. For choosing the value of T, we trained two models using $T = 0.1$ and $T = 0.5$. Applied to the validation dataset, the model trained with $T = 0.1$ achieved Dice and IoU scores of 0.71 and 0.7, respectively, while the model trained with $T = 0.5$ obtained Dice and IoU scores of 0.69 and 0.68, respectively. When using regular softmax, the obtained Dice and IoU scores were 0.6 and 0.59, which demonstrates the benefits of TSS over regular softmax for our model. Thus, we use a temperature parameter value $T = 0.1$ to make the model's



(a) Input. (b) Prediction without TSS. (c) Prediction with TSS. (d) Ground truth.

Background
Individual Cytoplasm
Cytoplasm of Cell Cluster
Superficial Cell Nucleus
Intermediate Cell Nucleus
Suspicious Cell Nucleus

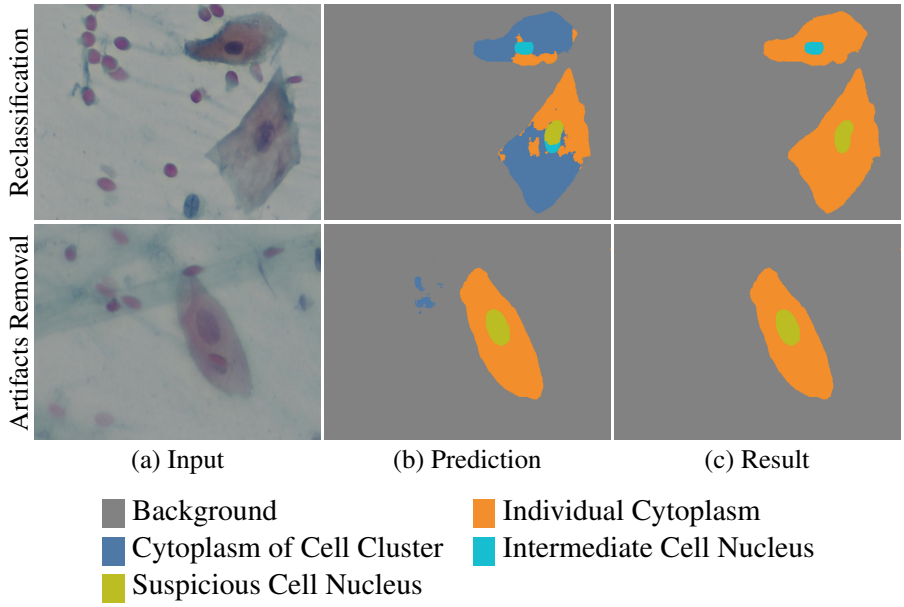
Source: The Authors.

Figure 5.3 – The impact of temperature scaling softmax (TSS) on avoiding the bias towards the prediction of background pixels over nuclei pixels on Papanicolaou-stained images. (a) Input image. (b) Prediction using a model trained with a regular softmax layer. (c) Prediction using our model trained with TSS, but before applying the semantic reclassification and artifact removal post-processing steps (see Sections 5.2.3 and 5.2.4). (d) Ground truth. TSS improves prediction confidence and segmentation.

predictions more confident and avoid the bias towards the prediction of background pixels. Fig. 5.3 illustrates the benefits of TSS in our model. Fig. 5.3 (b) shows the predicted result for the image shown in (a) obtained by replacing our model's TSS layer with a regular softmax layer. Note various nuclei incorrectly classified as background (gray). Fig. 5.3 (c) shows the prediction produced by our model trained with its TSS layer, where the bias towards the background has been fixed. The ground truth is shown in (d).

5.2.3 Semantic Reclassification Step

Since clusters are formed by sets of individual cells touching each other, their cytoplasms may be mistaken with ones of individual cells, and vice-versa (Fig. 5.4b (top row)). To solve this ambiguity, we extract the contours for the union of all pixels from these two classes using OpenCV (Bradski, 2000). Each contour containing a single nucleus is then classified as an individual cell, while the ones containing multiple nuclei are classified as a clusters of cells. Contours with no nucleus are classified as anucleate (squamous cells). We also extract the contours of the union of all pixels from each nucleus. Each nucleus is then classified based on the class with the largest number of pixels (*i.e.*, suspicious, superficial, or intermediate). For example, if a nucleus contour contains more pixels from the intermediate cell nucleus class, it is classified as an intermediate cell nucleus. Fig. 5.4c (top row) illustrates the result of the reclassification process applied to cytoplasm pixels of two cells as well as to the nucleus (yellow and cyan) of one of the cells



Source: The Authors.

Figure 5.4 – Applying semantic reclassification and segmentation artifact removal to the prediction of the model. Input images (a). Prediction of our model (b) to the input image. Result after semantic reclassification (top) and artifact removal (bottom) (c) properly match the ground truth (not shown).

shown in (b). Individual cells are classified based on their nucleus type. Clusters are classified as either *suspicious*, if they contain at least a suspicious nucleus, or *non-suspicious*, otherwise.

5.2.4 Segmentation Artifact Removal Step

After semantic reclassification, we may still face some segmentation artifacts. These consist mostly of small structures with less than 100 pixels, which are then discarded. The number of 100 pixels was empirically defined. Fig. 5.4c (bottom row) shows a result obtained after the removal of segmentation artifacts misclassified as cytoplasm of cell clusters in (b). The final segmentation output is assembled, converted to RGB for visualization, and saved to disk.

Table 5.1 shows the average Intersection over Union (IoU) values for the classification results produced by our system applied to two different test sets: (i) our original test set (OTS) consisting of 115 images; and (ii) our additional test set (ATS) containing 400 images used to compare the performance of our system with human experts (Section 5.3.2). Note the steady increase in the IoU values in both datasets as our solution

Table 5.1 – Progression of the average Intersection over Union (IoU) values for the classification results produced by our system as it goes from model prediction to reclassification and artifact removal, evaluated in two test sets. OTS and ATS stand for Original Test Set and Additional Test Set, respectively.

	Average IoU		
Dataset	Prediction	Reclassification	Artifacts Removal
OTS (115 images)	0.65	0.75	0.77
ATS (400 images)	0.78	0.81	0.82

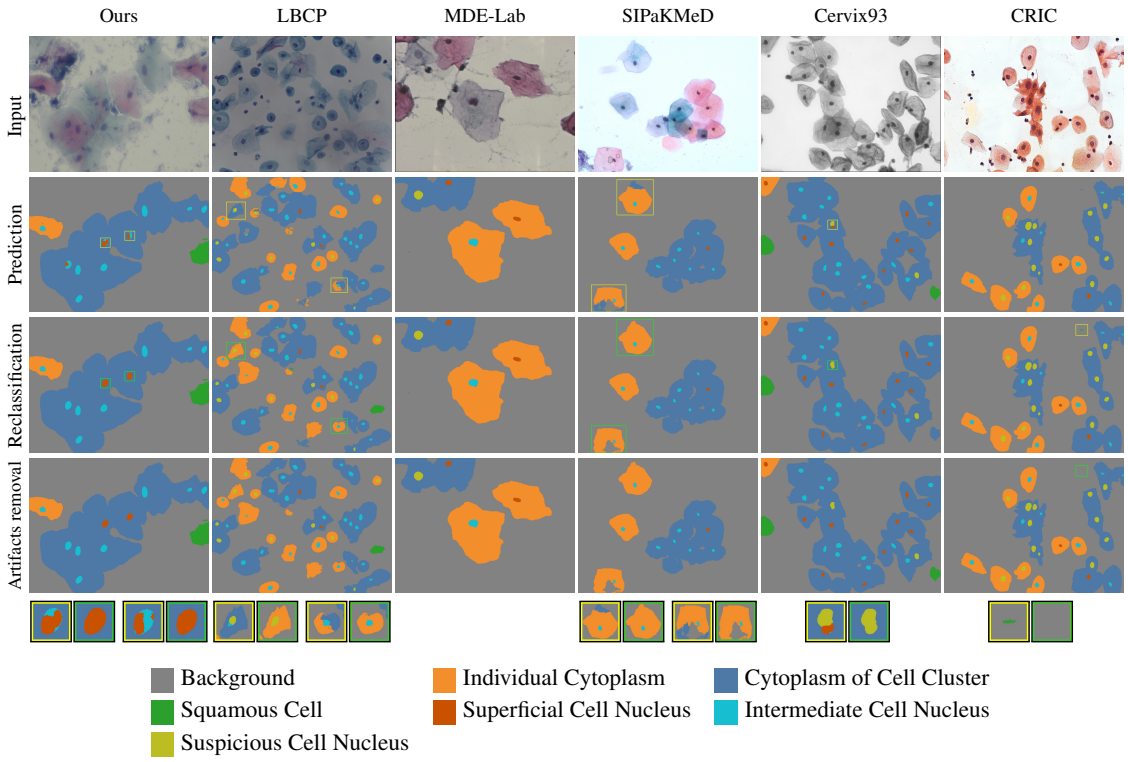
pipeline progresses from model prediction to semantic reclassification and artifact removal, highlighting the improvements introduced by these post-processing steps.

5.3 Results

This section presents the results of applying our model to our dataset of oral mucosa cells and to five public datasets of cervical cells, each presenting distinct features (*e.g.*, color versus gray scale, different cell colors, different background colors and intensity levels, etc.). The use of such diverse datasets demonstrate the ability of our model to generalize to different scenarios, showing that it can be used to segment and classify not only images of the oral mucosa, but cervical images as well. We also compare the results of our model with the ones produced by three human experts on an annotated dataset with 400 images from 20 patients. Our solution can process a Full HD image on an RTX 3090 GPU in 0.63 seconds. Given that none of the techniques discussed in Section 3.2 can segment/classify both cytoplasms and nuclei and no publicly-available pre-trained models or implementations of these techniques are available, we do not include comparisons with them.

5.3.1 Results on Ours and on Five Public Datasets

We evaluate the performance of our model on six datasets of Papanicolaou-stained images, consisting of our own dataset of oral mucosa cells (UFRGS Pap-OMD) plus five public datasets of cervical cells: LBCP (Hussain et al., 2020), MDE-Lab (Byriel, 1999), SIPaKMeD (Plissiti et al., 2018), Cervix93 (Phoulady; Mouton, 2018), and CRIC (Rezende et al., 2021). The images of these datasets were acquired using different micro-



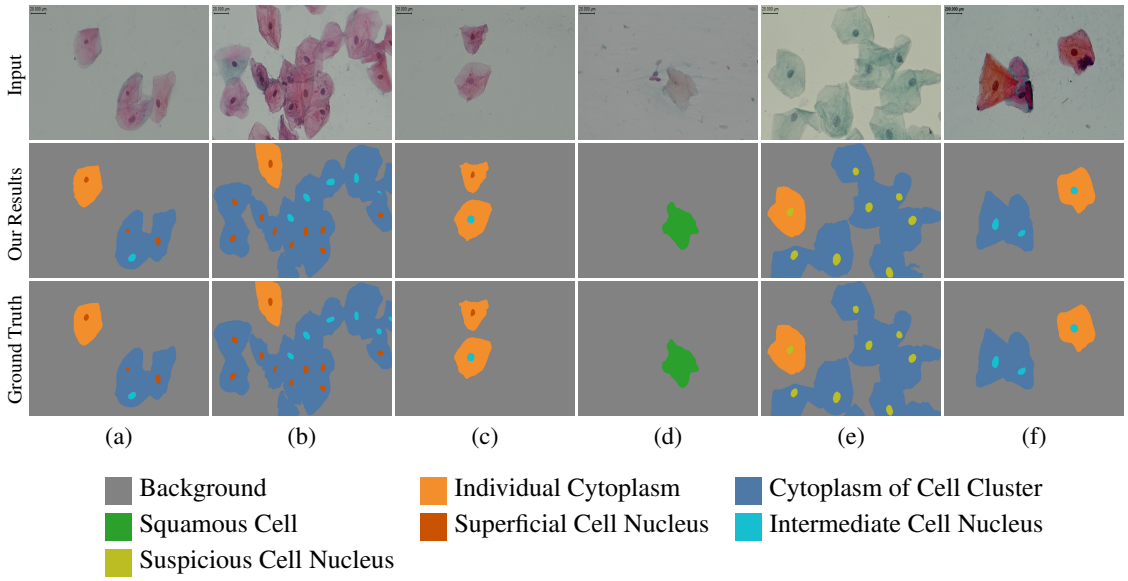
Source: The Authors.

Figure 5.5 – Results of our model applied to images from our dataset and from five public datasets of cervical cells. (first row) Input images. (second row) Our model's predictions before any post-processing. (third row) Results after the semantic reclassification step. (fourth row) Results after the segmentation-artifact removal step. Despite the high variability in the input images, the predictions of our model already correspond to the final results or are very close to them. The reclassification and artifact-removal post-processing steps only make minor changes to the predictions, providing some "final touch". Examples of pixel reclassification and artifact removal, and their corrected values are highlighted by yellow and green outlines, respectively (bottom of the figure).

scopes and cameras, and have different resolutions and color channels. Fig. 5.5 shows the results of applying our model to images from the six datasets, showing the progression of the segmentation and classification process as the input images advance in our pipeline (prediction, semantic reclassification, and artifact removal). The results show that our model generalizes well to the images from these diverse datasets.

The results on the LBCP, MDE-Lab, and CRIC datasets show that our model is robust to background artifacts and successfully classifies such objects as background. The results on the SIPaKMeD and Cervix93 datasets show that our model is able to correctly segment and classify cells in high-density slides, as well as to work with gray-scale images.

Fig. 5.6 illustrates the quality of the segmentation produced by our model on images from our dataset displaying different types of cells, from anucleate, intermediate,



Source: The Authors.

Figure 5.6 – Examples of segmentation produced by our model on images from our dataset (not used in the model's training) displaying different types of cells and clusters. The first row shows the input images. The second row shows the results produced by our model after prediction and the post-processing steps, nicely matching the ground truth shown in the third row.

and superficial, as well as suspicious and non-suspicious clusters. These images were selected among the 400 images used for evaluating the performance of our model against the human experts and, as such, were not in the training set. Note how our model results nicely match the ground truth.

5.3.2 Comparing Our Model with Human Experts

We compare our model's results with the ones produced by three human experts on an additional annotated dataset (ATS) consisting of 400 images from 20 patients. The dataset was annotated by experts using the same procedure described in Section 5.1. The patients in the dataset belong to four groups (five patients in each group): (i) patients with squamous cell carcinoma (SCC); (ii) patients with oral potentially malignant disorders (OPMD); (iii) patients exposed to carcinogens (*e.g.*, tobacco and alcohol) but without lesions in the oral mucosa (EXP); and (iv) control group consisting of healthy patients (CTL). One patient from each group was randomly selected to have a total of 60 annotated images. The remaining patients had each 10 annotated images. We choose to have one patient from each group with a larger number of annotated images to allow more accurate comparison between our model and the experts at a patient level.

Table 5.2 – Intraclass Correlation Coefficient (ICC) values and confidence intervals (95%) considering cells and clusters of various types identified by: three human specialists (Specialists Only); and the same three human specialists and our model (Specialists and Our Model) in a dataset with 400 images from 20 patients. Note the improvement of the ICC values for 4 of the 5 types of cells/clusters when including our model results.

Cell/Cluster Type	ICC	Confidence Interval (95%)	p-value
Cluster	0.974	[0.950, 0.989]	<0.001
Suspicious cluster	0.848	[0.704, 0.932]	<0.001
Superficial Cells	0.906	[0.827, 0.957]	<0.001
Intermediate Cells	0.883	[0.788, 0.946]	<0.001
Suspicious Cells	0.879	[0.771, 0.945]	<0.001

We compared the classification results (for the various types of cells and clusters) produced by our model and by the three human experts. The resulting Intraclass Correlation Coefficients (ICCs) are shown in Table 5.2 for the three experts only (Specialists Only), as well as for the experts and our model together (Specialists and Our Model). The ICC values were computed considering a confidence interval of 95%, using the 400 images from 20 patients. The low p-values (< 0.001) for all classes confirm the results' statistical significance. The improvements in the ICC values for 4 of the 5 types of cell/clusters when including our model demonstrates that it achieves consistent expert-level performance.

We also calculate the ICC for the randomly selected patients who had 60 annotated images. The results in Table 5.3 show the ICC values and confidence intervals (95%) for each patient and object class. According to the criteria described by (Koo; Li, 2015), our results show excellent or good agreement between our model and the experts for all patients and object classes, except for the *suspicious cell* and *suspicious cluster* classes of patient (C). As shown in Table 5.2, the ICC for the suspicious cell and for the suspicious cluster classes in the whole dataset were 0.879 and 0.848, which correspond to excellent agreements. The ICC for the remaining object classes for patient (C) are 0.91 for cluster, 0.82 for superficial cell, and 0.86 for intermediate cell. The zero ICC value for the class of suspicious cells from patient (C) is explained by the existence of only 6 such cells, on which there was no agreement among the three experts: our model classified two cells as suspicious, while experts #1, #2, and #3 reported, respectively, one, zero, and one cell as suspicious, with no agreement among the three, resulting in an ICC value among themselves of -0.11. Although ICC values are typically between 0 and 1, the negative value in this particular case indicates that the experts were highly inconsistent in their answers. The low agreement for the suspicious cluster for patient (C) (ICC value of 0.55)

Table 5.3 – Intraclass Correlation Coefficient (ICC) values and confidence intervals (95%) per object class for four randomly selected patients (with 60 images each) from different groups, identified by our model and by three human specialist.

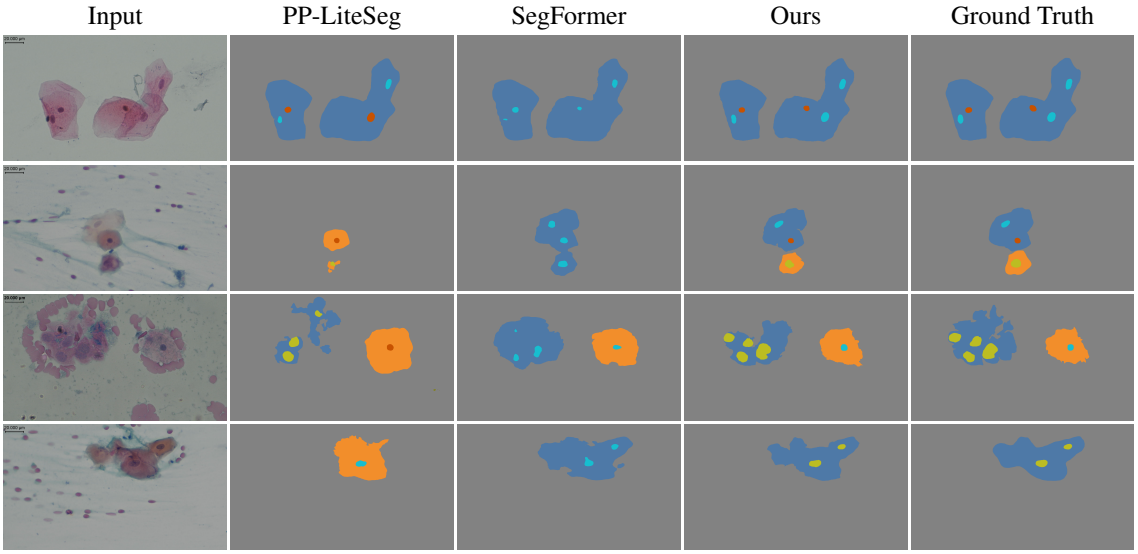
Patient	Group	Cell/Cluster Type	ICC	Confidence Interval (95%)	P-value
		Cluster	0.90	[0.59, 0.79]	<0.001
		Suspicious Cluster	0.92	[0.64, 0.82]	<0.001
(A)	SSC	Superficial Cell	0.73	[0.27, 0.55]	<0.001
		Intermediate Cell	0.79	[0.36, 0.63]	<0.001
		Suspicious Cell	0.76	[0.32, 0.58]	<0.001
		Cluster	0.96	[0.79, 0.90]	<0.001
		Suspicious Cluster	0.80	[0.37, 0.63]	<0.001
(B)	OPMD	Superficial Cell	0.91	[0.61, 0.82]	<0.001
		Intermediate Cell	0.93	[0.68, 0.84]	<0.001
		Suspicious Cell	0.77	[0.32, 0.59]	<0.001
		Cluster	0.91	[0.62, 0.81]	<0.001
		Suspicious Cluster	0.55	[0.11, 0.38]	<0.001
(C)	Exposed	Superficial Cell	0.82	[0.41, 0.66]	<0.001
		Intermediate Cell	0.86	[0.49, 0.72]	<0.001
		Suspicious Cell	0.00	[-0.1, 0.11]	0.62
		Cluster	0.89	[0.56, 0.77]	<0.001
		Suspicious Cluster	0.82	[0.42, 0.66]	<0.001
(D)	Control	Superficial Cell	0.92	[0.65, 0.82]	<0.001
		Intermediate Cell	0.86	[0.48, 0.71]	<0.001
		Suspicious Cell	0.82	[0.41, 0.66]	<0.001

is also due to the existence of only 4 suspicious clusters for this patient. In comparison, patients (A), (B), and (D) have, respectively, 68, 45, and 43 suspicious cells, and 13, 11, and 15, suspicious clusters. Given the small number of suspicious cells/clusters for patient (C), a single missed suspicious cell/cluster, by either our model or by an expert, has a significant impact on the ICC value.

5.3.3 Discussion

We are currently using our system to detect suspicious cells and clusters. Patients with a nucleus-cytoplasm ratio greater than 0.17 in their suspicious cells or presenting suspicious clusters are referred to close monitoring by experts.

Our model was able to correctly segment the important classes for determining the malignancy of oral mucosa cells (suspicious cells and clusters, cytoplasms, superficial cells, and intermediate cells) and was able to generalize well to images from other datasets of Papanicolaou-stained images. However, currently it is not as accurate in segmenting binucleate cells. This is not a limitation of the model itself, but rather results from the low number of examples of this class in the training dataset. Even though anucleate cells were not included in our evaluation against the experts, our model is capable of segmenting them correctly as shown in Fig. 5.6. While the other classes in our dataset (suspicious



Source: The Authors.

Figure 5.7 – Comparison of results produced by our model and by other segmentation architectures (PP-LiteSeg and SegFormer) on four images from our dataset.

cell, cytoplasm, cluster, superficial cell, and intermediate cell) have over 400 examples each, the anucleate and binucleate classes have only 121 and 37 examples, respectively. This shortcoming is not critical for the purpose of our model, (spotting suspicious oral mucosa cells), and it can be addressed by collecting more images with examples of these classes and fine-tuning the model.

5.3.3.1 Experiments with Different Architectures

During the development of our work, new architectures were proposed in the literature. We tested a few of them, including PP-LiteSeg (Peng et al., 2022) and Segformer (Xie et al., 2021). We fine tuned these models on our dataset and compared their segmentation with the ones produced by our model. The results showed that our model outperformed the predictions by PP-LiteSeg and Segformer. Fig. 5.7 compares the results of applying PP-LiteSeg, SegFormer, and our model on four images from our dataset. The results show that our model is able to segment and classify the cells more accurately than PP-LiteSeg and Segformer. Our model correctly classifies cell nuclei classes, and more precisely segments the cytoplasm outlines.

We also explored Segment Anything (Kirillov et al., 2023), a model for instance segmentation. However, it does not perform semantic segmentation, which is required for

our work in order to categorize cell nuclei, and cytoplasm/clusters of cells. As such, we decided not to pursue further experiments with it at that time. All the tested architectures hold promises for future work and may be considered for further experiments.

6 CONCLUSIONS

We presented two CNN-based methods for automatic segmentation and classification of oral mucosa cells. Our first method is an efficient solution for automatically segmenting and counting the number of AgNORs in cytological images. Our method can discard overlapping and distorted nuclei, and further classify the detected AgNORs based on their relative sizes. Users can specify the nuclei for AgNOR counting and classification by annotating the images with rectangles delimiting the regions of interest. Our segmentation CNN can process 100 high-resolution images under one minute on a laptop. We also introduced an annotated AgNOR-stained image dataset of epithelial cells from the oral mucosa containing 1,171 images from 48 patients (Rönnau et al., 2023c). To the best of our knowledge, this is the most diverse annotated AgNOR dataset available. We demonstrated the effectiveness and robustness of our solution on many challenging configurations on two datasets. On our dataset, our method achieved Dice and IoU scores of 0.90 and 0.84, respectively, indicating very good agreement with the ground truth. On a third-party dataset of cervical images, Dice and IoU scores were 0.80 and 0.74, respectively. Our solution achieved a performance similar to human experts on a set of 291 images from 6 new patients, while significantly reducing the time required to quantify the number of AgNORs per nuclei. The results of this experiment show high ICC values and low p-values, confirming their statistical significance and agreement with human experts.

Our second method is a CNN-based solution for automatic segmentation and classification of Papanicolaou-stained oral mucosa cells. Individual cells are classified as either suspicious, superficial, intermediate, anucleate, or bi-nucleate. Clusters of cells are classified as suspicious or non-suspicious. To the best of our knowledge, ours is the first technique that simultaneously performs segmentation and classification of Papanicolaou-stained cells. Our model achieved expert-level performance in an experiment comparing its results with the ones of three human experts on a set of 400 images of the oral mucosa from 20 patients. The results of this experiment show high ICC values and low p-values, confirming their statistical significance. We also presented a Papanicolaou-stained image dataset of oral mucosa cells containing 1,563 Full HD images from 52 patients, annotated by specialists. This is the most diverse oral mucosa cell dataset in terms of number of patients, containing a balanced number of images from four classes of patients: with squamous cell carcinoma; with oral potentially malignant disorders; exposed to carcinogens, but without lesions in the oral mucosa; and healthy. We evaluated the performance of our

model on our dataset and on five public datasets of cervical cells. The results show that despite being trained on images of oral mucosa, our model generalizes well to images from different datasets, with different characteristics (*e.g.*, captured with different microscopes and cameras, and having different resolutions, colors, background intensities, noise levels, and Papanicolaou staining methods). The results on these datasets exhibit high-quality segmentation and plausible classification (no ground truth is available). This suggests that our model can be successfully used, especially after some fine-tuning, for segmentation and classification of other types of Papanicolaou-stained images, helping in the detection of other types of cancer.

Our methods can be used to assist pathologists in detecting the first signs of oral cancer, especially in resource-limited settings, where AgNOR and Papnicolaou staining technique are still widely used. Our models and datasets are publicly available (Rönnau et al., 2023a; Rönnau et al., 2023c; Rönnau et al., 2024) and we hope they can help practitioners and stimulate new research in early oral cancer detection.

6.1 Future Work

There are several directions for future work that could further improve the performance and usability of our methods. The development of an end-to-end model for AgNOR-stained datasets that not only segments the images but also outputs the count of AgNORs directly is a promising direction. The development of a similar end-to-end model for Papanicolaou-stained images that outputs not only the segmentation of the images but also the number of cells in each class and the overall nuclei/cytoplasm ratio per cell or cell cluster is another promising future work. Training our Papanicolaou CNN with additional examples of anucleate and binucleate cells would improve the model's performance on these types of cells. Fine-tuning our model for other types of Papanicolaou-stained images is also a promising direction for future exploration. We also envision the development of a unified software tool that integrates both methods, allowing users to analyze both AgNOR-stained and Papanicolaou-stained images in a single environment. This tool could also include a user-friendly interface for annotating images, training new models, and evaluating the performance of the models. The development of a web-based version of this tool would make it accessible to a broader audience.

REFERENCES

AMORIM, J. G. A. **Segmentation of AgNOR-Stained Cytology Images Using Deep Learning**. [S.l.]: GitLab, 2020. <https://codigos.ufsc.br/lapix/segmentation-of-agnor-stained-cytology-images>.

AMORIM, J. G. A. et al. **Cytology Dataset - CCAgT: Images of Cervical Cells with AgNOR Stain Technique**. [S.l.]: Lapix, 2020. <https://lapix.ufsc.br/agnor-dataset/>. [Online; accessed April 2022].

AMORIM, J. G. A. et al. A novel approach on segmentation of agnor-stained cytology images using deep learning. In: **2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)**. IEEE, 2020. Available from Internet: <https://doi.org/10.1109/cbms49503.2020.00110>.

BANDYOPADHYAY, H.; NASIPURI, M. Segmentation of pap smear images for cervical cancer detection. In: IEEE. **2020 IEEE Calcutta Conference (CALCON)**. [S.l.], 2020. p. 30–33.

BEDELL, S. L. et al. Cervical Cancer Screening: Past, Present, and Future. **Sexual Medicine Reviews**, v. 8, n. 1, p. 28–37, 11 2019. ISSN 2050-0513.

BELL, A. A. et al. High dynamic range images as a basis for detection of argyrophilic nucleolar organizer regions under varying stain intensities. In: IEEE. **2006 Intl. Conference on Image Processing**. [S.l.], 2006. p. 2541–2544.

BOUGHZALA, O. et al. Automatic segmentation of cervical cells in pap smear images. Václav Skala-UNION Agency, 2016.

BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, 2000.

BUSLAEV, A. et al. Albumentations: Fast and flexible image augmentations. **Information**, v. 11, n. 2, 2020. ISSN 2078-2489. Available from Internet: <https://www.mdpi.com/2078-2489/11/2/125>.

BYRIEL, J. Neuro-fuzzy classification of cells in cervical smears. **Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation**, 1999.

CALDEIRA, P. C. et al. Oral leukoplakias with different degrees of dysplasia: comparative study of hmlh1, p53, and agnor. **Journal of oral pathology & medicine**, Wiley Online Library, v. 40, n. 4, p. 305–311, 2011.

CHAURASIA, A.; CULURCIELLO, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: IEEE. **2017 IEEE Visual Communications and Image Processing (VCIP)**. [S.l.], 2017. p. 1–4.

CHEN, L.-C. et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 40, n. 4, p. 834–848, 2017.

ÇIÇEK, Ö. et al. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: SPRINGER. **Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19**. [S.l.], 2016. p. 424–432.

CUCER, N. et al. Two-dimensional agnor evaluation as a prognostic variable in urinary bladder carcinoma: A different approach via total agnor area/nucleus area per cell. **Micron**, Elsevier, v. 38, n. 6, p. 674–679, 2007.

FERREIRA, A. A. et al. An image processing software applied to oral pathology. **Pathology-Research and Practice**, Elsevier, v. 207, n. 4, p. 232–235, 2011.

GARCÍA-VIELMA, C. et al. Digital image analysis of agnors in cervical smears of women with premalignant and malignant lesions of the uterine cervix. **Biotechnic & Histochemistry**, Taylor & Francis, v. 91, n. 2, p. 102–107, 2016.

GENçTAV, A.; AKSOY, S.; ÖNDER, S. Unsupervised segmentation and classification of cervical cell images. **Pattern Recognition**, v. 45, n. 12, p. 4151–4168, 2012. ISSN 0031-3203. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0031320312002191>.

GHARIPOUR, A.; LIEW, A. W.-C. Segmentation of cell nuclei in fluorescence microscopy images: An integrated framework using level set segmentation and touching-cell splitting. **Pattern Recognition**, v. 58, p. 1–11, 2016. ISSN 0031-3203. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0031320316300280>.

GHOSH, S. et al. Understanding deep learning techniques for image segmentation. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 52, n. 4, p. 1–35, 2019.

HE, K. et al. Mask r-cnn. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2961–2969.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HESAMIAN, M. H. et al. Deep learning techniques for medical image segmentation: achievements and challenges. **Journal of digital imaging**, Springer, v. 32, n. 4, p. 582–596, 2019.

HUANG, G. et al. Densely connected convolutional networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 4700–4708.

HUSSAIN, E. et al. Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. **Data in brief**, Elsevier, v. 30, p. 105589, 2020.

IBM Corp. **IBM SPSS Statistics**. 2023. Available from Internet: <https://www.ibm.com/analytics/spss/statistics>.

JAJODIA, E. et al. Brush cytology and agnor in the diagnosis of oral squamous cell carcinoma. **Acta cytologica**, Karger Publishers, v. 61, n. 1, p. 62–70, 2017.

JANTZEN, J. et al. Pap-smear benchmark data for pattern classification. **Nature inspired smart information systems (NiSIS 2005)**, p. 1–9, 2005.

JIANG, H. et al. Deep learning for computational cytology: A survey. **Medical Image Analysis**, Elsevier, p. 102691, 2022.

JIANG, H. et al. Deep learning for computational cytology: A survey. **Medical Image Analysis**, Elsevier, v. 84, p. 102691, 2023.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

KIRILLOV, A. et al. Panoptic feature pyramid networks. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2019. p. 6399–6408.

KIRILLOV, A. et al. Segment anything. **arXiv:2304.02643**, 2023.

KOO, T. K.; LI, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. **J Chiropr Med.**, v. 15, n. 2, p. 155–163, 2015.

LI, K. et al. Cytoplasm and nucleus segmentation in cervical smear images using radiating gvf snake. **Pattern Recognition**, v. 45, n. 4, p. 1255–1264, 2012. ISSN 0031-3203. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0031320311003979>.

LIN, T.-Y. et al. Focal loss for dense object detection. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2017. p. 2980–2988.

LINGEN, M. W. et al. Adjuncts for the evaluation of potentially malignant disorders in the oral cavity: diagnostic test accuracy systematic review and meta-analysis—a report of the american dental association. **The Journal of the American Dental Association**, Elsevier, v. 148, n. 11, p. 797–813, 2017.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 3431–3440.

MATIAS, A. V. et al. Segmentation, detection, and classification of cell nuclei on oral cytology samples stained with papanicolaou. **SN Computer Science**, Springer, v. 2, n. 4, p. 285, 2021.

MINAEE, S. et al. Image segmentation using deep learning: A survey. **IEEE Trans Pattern Anal Mach Intell.**, v. 44, n. 7, p. 3523–3542, 2022.

NOH, H.; HONG, S.; HAN, B. Learning deconvolution network for semantic segmentation. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2015. p. 1520–1528.

OCF. **The Oral Cancer Foundation**. 2024. <https://oralcancerfoundation.org/>. [Online; accessed April July 2024].

OCF. **The Oral Cancer Foundation, Cancer Screening Protocols**. 2024. <https://oralcancerfoundation.org/discovery-diagnosis/cancer-screening-protocols/>. [Online; accessed April 2024].

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PENG, J. et al. Pp-liteseg: A superior real-time semantic segmentation model. **arXiv preprint arXiv:2204.02681**, 2022.

PHAM, D. L.; XU, C.; PRINCE, J. L. A survey of current methods in medical image segmentation. **Annual review of biomedical engineering**, v. 2, n. 3, p. 315–337, 2000.

PHOULADY, H. A.; MOUTON, P. R. A new cervical cytology dataset for nucleus detection and image classification (cervix93) and methods for cervical nucleus detection. **arXiv preprint arXiv:1811.09651**, 2018.

PLISSITI, M. E. et al. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: IEEE. **2018 25th IEEE International Conference on Image Processing (ICIP)**. [S.l.], 2018. p. 3144–3148.

PLISSITI, M. E.; NIKOU, C.; CHARCHANTI, A. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. **IEEE Transactions on information technology in biomedicine**, IEEE, v. 15, n. 2, p. 233–241, 2010.

PLISSITI, M. E.; VRIGKAS, M.; NIKOU, C. Segmentation of cell clusters in pap smear images using intensity variation between superpixels. In: IEEE. **2015 International Conference on Systems, Signals and Image Processing (IWSSIP)**. [S.l.], 2015. p. 184–187.

RAGHU, M. et al. Transfusion: Understanding transfer learning for medical imaging. **Advances in neural information processing systems**, v. 32, 2019.

RAGOTHAMAN, S. et al. Unsupervised segmentation of cervical cell images using gaussian mixture model. In: **Proceedings of the IEEE conference on computer vision and pattern recognition workshops**. [S.l.: s.n.], 2016. p. 70–75.

RAJPUT, D. V.; TUPKARI, J. V. Early detection of oral cancer: Pap and agnor staining in brush biopsies. **Journal of oral and maxillofacial pathology: JOMFP**, Wolters Kluwer–Medknow Publications, v. 14, n. 2, p. 52, 2010.

RASHEED, A. et al. Cervical cell's nucleus segmentation through an improved unet architecture. **Plos one**, Public Library of Science San Francisco, CA USA, v. 18, n. 10, p. e0283568, 2023.

REZENDE, M. T. et al. Cric searchable image database as a public platform for conventional pap smear cytology data. **Scientific Data**, Nature Publishing Group, v. 8, n. 1, p. 1–8, 2021.

RÖNNAU, M. M. et al. A cnn-based approach for joint segmentation and quantification of nuclei and nors in agnor-stained images. **Computer Methods and Programs in Biomedicine**, p. 107788, 2023. ISSN 0169-2607.

RÖNNAU, M. M. et al. **Medical Image Segmentation: pre-trained models and source code**. [S.l.]: GitHub, 2023. <https://github.com/maikelroennau/medical-image-segmentation>.

RÖNNAU, M. M. et al. **UFRGS AgNOR-stained image dataset of epithelial cells from oral mucosa (UFRGS AgECOM)**. [S.l.]: GitHub, 2023. <https://github.com/maikelroennau/AgECOM>.

RÖNNAU, M. M. et al. **UFRGS Papanicolaou Oral Mucosa Dataset (UFRGS Pap-OMD)**. [S.l.]: GitHub, 2024. <https://github.com/maikelroennau/UFRGS-Pap-OMD>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: **Intl. Conf. on Medical image computing and computer-assisted intervention**. [S.l.: s.n.], 2015. p. 234–241.

RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. **International journal of computer vision**, Springer, v. 115, n. 3, p. 211–252, 2015.

SCHNEIDER, C. A.; RASBAND, W. S.; ELICEIRI, K. W. Nih image to imagej: 25 years of image analysis. **Nature methods**, Nature Publishing Group, v. 9, n. 7, p. 671–675, 2012.

SHIRAZ, A. et al. The early detection of cervical cancer. the current and changing landscape of cervical disease detection. **Cytopathology**, v. 31, n. 4, p. 258–270, 2020.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

STRANDER, B. et al. Liquid-based cytology versus conventional papanicolaou smear in an organized screening program: a prospective randomized study. **Cancer cytopathology**, Wiley Online Library, v. 111, n. 5, p. 285–291, 2007.

SZEGEDY, C. et al. Rethinking the inception architecture for computer vision. In: **Proc. IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 2818–2826.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2019. p. 6105–6114.

TERESA, D. B. et al. Computer-assisted analysis of cell proliferation markers in oral lesions. **Acta histochemica**, Elsevier, v. 109, n. 5, p. 377–387, 2007.

TRERE, D. Agnor staining and quantification. **Micron**, Elsevier, v. 31, n. 2, p. 127–131, 2000.

TYAGI, K. K. et al. Agnor as an effective diagnostic tool for determining the proliferative nature of different types of odontogenic cysts. **Journal of Family Medicine and Primary Care**, Medknow, v. 9, n. 1, p. 125–130, 2020.

VIGNESWARAN, N.; WILLIAMS, M. D. Epidemiologic trends in head and neck cancer and aids in diagnosis. **Oral and Maxillofacial Surgery Clinics**, Elsevier, v. 26, n. 2, p. 123–141, 2014.

WADA, K. **labelme: Image Polygonal Annotation with Python**. 2016. <https://github.com/wkentaro/labelme>.

WANG, R. et al. Medical image segmentation using deep learning: A survey. **IET Image Processing**, v. 16, n. 5, p. 1243–1267, 2022.

XIE, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. **Advances in neural information processing systems**, v. 34, p. 12077–12090, 2021.

XIE, S. et al. Aggregated residual transformations for deep neural networks. In: **Proc. IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 1492–1500.

YAKUBOVSKIY, P. **Segmentation Models**. [S.l.]: GitHub, 2019. <https://github.com/qubvel/segmentation_models>.

ZHANG, J. et al. Segmentation of overlapping cells in cervical smears based on spatial relationship and overlapping translucency light transmission model. **Pattern Recognition**, v. 60, p. 286–295, 2016. ISSN 0031-3203. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0031320316300802>.

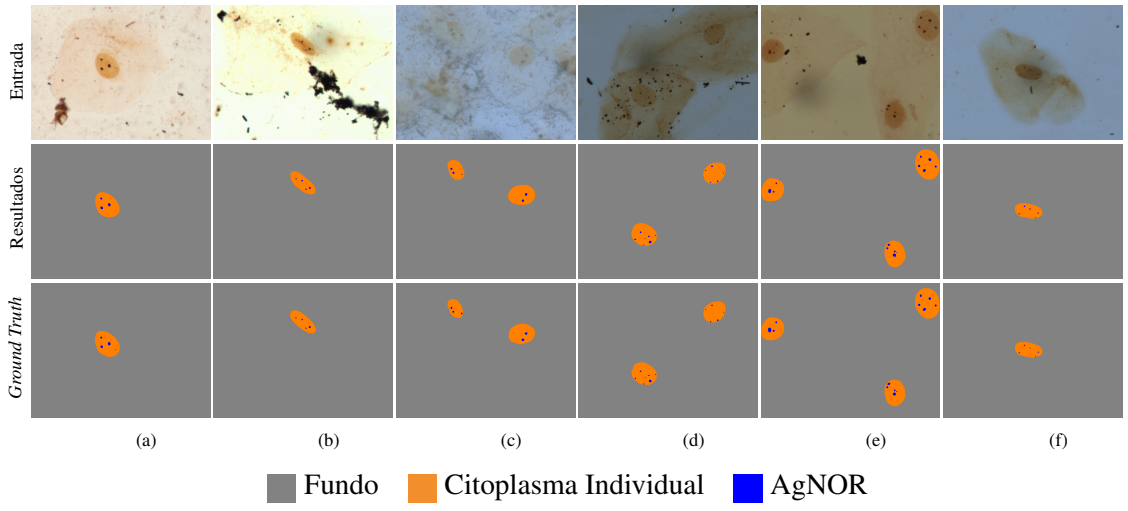
ZHANG, J. et al. Binary tree-like network with two-path fusion attention feature for cervical cell nucleus segmentation. **Computers in biology and medicine**, Elsevier, v. 108, p. 223–233, 2019.

ZHAO, Y. et al. Automatic segmentation of cervical cells based on star-convex polygons in pap smear images. **Bioengineering**, MDPI, v. 10, n. 1, p. 47, 2022.

APPENDIX A — RESUMO EXPANDIDO

Esta dissertação apresenta dois métodos baseados em redes neurais convolucionais (CNN) para segmentação e contagem de núcleos e AgNORs em imagens coradas pelo processo AgNOR, além da segmentação e classificação de células em imagens coradas pelo processo Papanicolaou. Para desenvolver e avaliar os métodos propostos, foram construídos dois conjuntos de imagens de células da mucosa oral coradas com AgNOR e Papanicolaou, respectivamente, anotadas por especialistas.

O conjunto de imagens de células coradas pelo processo AgNOR é composto por 1.171 imagens de 48 pacientes. Este conjunto é o mais diversificado disponível em termos de número de pacientes, sendo o primeiro de células da mucosa oral. O conjunto de imagens de células coradas pelo processo Papanicolaou é composto por 1.563 imagens de 52 pacientes, sendo o mais diversificado em número de pacientes para células da mucosa oral coradas este processo. Ambos os conjuntos foram anotados por especialistas e estão disponíveis publicamente.



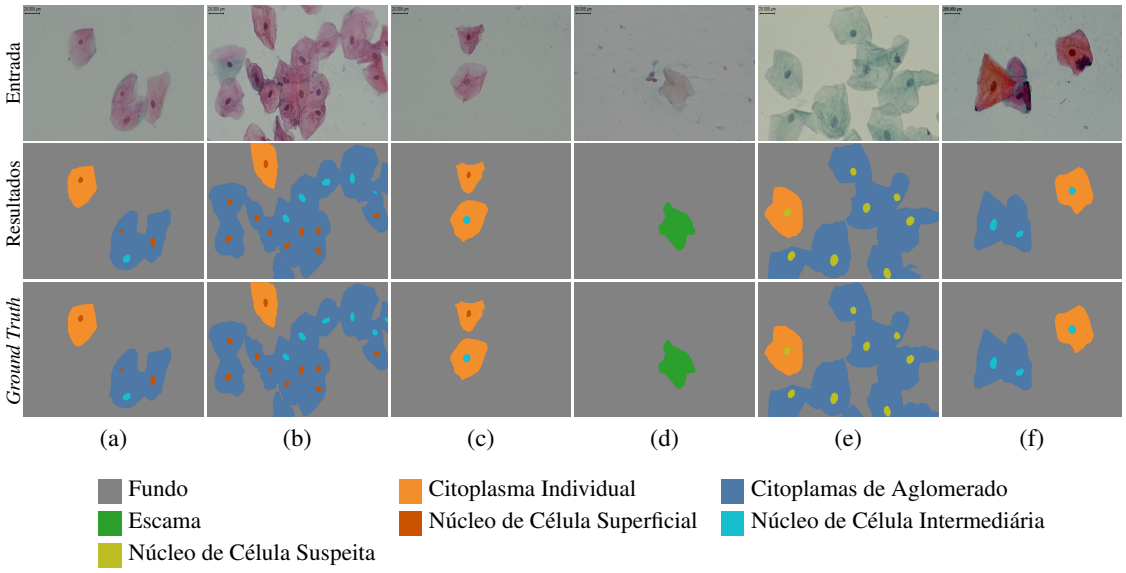
Fonte: Os Autores.

Figure A.1 – Aplicação do nosso método em uma série de imagens desafiadoras do nosso conjunto de dados de teste. (a) e (b) mostram núcleos próximos a objetos estranhos. (c) mostra um núcleo nublado. (d) e (e) mostram exemplos de precipitação de prata resultando em manchas escuras fora dos núcleos que se assemelham a AgNORs. (a), (b), (e) e (f) mostram núcleos altamente contrastados em relação ao citoplasma. Os resultados produzidos pelo nosso modelo comparados ao *padrão ouro* mostram que ele pode segmentar robustamente núcleos e AgNORs sob várias condições.

Nosso modelo para segmentação e contagem de núcleos e AgNORs foi avaliado em um conjunto de 291 imagens de células coradas pelo processo AgNOR anotadas por especialistas. O modelo obteve resultados superiores aos métodos de segmentação de

núcleos e AgNORs da literatura. Em uma comparação com especialistas humanos, o modelo proposto alcançou um Coeficiente de Correlação Intraclasse (ICC) de 0,91 para núcleos e 0,81 para AgNORs, com *p-value* < 0.001, indicando forte concordância com os especialistas. A Fig. A.1 mostra exemplos de segmentação produzidos pelo modelo proposto.

Nosso modelo para segmentação e classificação de células coradas pelo processo Papanicolaou foi avaliado em um conjunto de 400 imagens anotadas por especialistas. O modelo demonstrou capacidade de generalização para diferentes conjuntos de imagens. Em uma comparação com especialistas humanos, o modelo proposto alcançou ICCs acima de 0,84 para todos os tipos de células, mostrando excelente ou boa concordância para a maioria dos tipos de células. A Fig. A.2 mostra exemplos de segmentação e classificação produzidos pelo modelo proposto.



Fonte: Os Autores.

Figure A.2 – Exemplos de segmentação produzidos pelo nosso modelo em imagens do nosso conjunto de dados (não utilizadas no treinamento do modelo) exibindo diferentes tipos de células e aglomerados. A primeira linha mostra as imagens de entrada. A segunda linha mostra os resultados produzidos pelo nosso modelo após a predição e as etapas de pós-processamento, correspondendo bem ao *padrão ouro* mostrada na terceira linha.

Nossos modelos atingiram níveis de precisão e concordância comparáveis ao de especilistas humanos e atendem ao requisito de escalabilidade para o uso difundido dos testes de AgNOR e Papanicolaou, auxiliando profissionais de saúde na detecção precoce de câncer bucal. Nossos modelos são capazes de processar centenas de imagens de alta resolução em cerca de 1 minuto, sendo significativamente mais rápidos do que a análise manual. Os modelos treinados, o código e os conjuntos de dados estão disponíveis no

GitHub e podem estimular novas pesquisas na detecção precoce do câncer oral (Rönnau et al., 2023a; Rönnau et al., 2023b; Rönnau et al., 2023c; Rönnau et al., 2024).