# Overcoming Sample Scarcity and Label Ambiguity in Cell Segmentation and Classification of Thyroid Cancer: A Kernel-Based Approach on Top of CellSAM

Livia Lüscher
*Centre for Artificial Intelligence*
*Zurich University of Applied Sciences*
Zurich, Switzerland
email: luescliv@students.zhaw.ch

***Abstract*—This study addresses the challenge of automated thyroid cell analysis in the presence of incomplete expert annotation. It builds on the CellSAM model used for segmentation and adds a kernel-based approach for classification. Recognizing the limitations of conventional methods due to sample scarcity and label ambiguity, this research integrates clustering algorithms to overcome label ambiguity and improve diagnostic accuracy. As a consequence of the discrepancy in the diagnostic evaluation of pertinent cells resulting from the manual validation of the segmented mask, a metric was devised to capture the attributes of cancerous cells. Relevant features of cancer cells were identified and used together with the prior knowledge of annotated cells to validate the results of the clustering algorithms. Cancerous cells can be detected using a specific dye that binds to the cells, enabling the monitoring of cell growth and the analysis of fluorescence intensity values to identify these cells. The intensity values, which are critical for identifying cancerous cells, were first normalized. These normalized values were then transformed using a linear kernel and subsequently used as input for the clustering algorithms. Semi-supervised learning with clustering algorithm gave too much weight to already annotated cells and did not improve the quality of the dataset. Spectral clustering showed the best result based on the defined metric used in this work.**

## I. Introduction

The incidence rate of thyroid cancer is one of the highest among cancer types in the United States [45]. In the past decade, there has been a significant advancement in the understanding of the molecular mechanisms underlying thyroid tumour formation. The identification of distinct genetic alterations and the discovery of novel oncogenic pathways have greatly enhanced our understanding of the molecular biology involved in tumorigenesis and malignant transformation in the thyroid gland [49]. In a recent study fluorescence polarization imaging with methylthioninium chloride has been used to detect diagnostically relevant cells. Methylthioninium chloride (MB) is a method to stain cells [15] and is closely aligned with clinical applications. Fluorescence polarization (Fpol) which uses MB is an optical technique that is able to detect cancer cells more accurately. It measures how light, emitted from methylthioninium chloride, behaves when it interacts with cells. The feature attaches better to cancerous cells, affecting how the light is polarized. This change in polarization can be measured and visualized in an image. Higher values indicate the presence of cancer [15]. The instrumentation of this unique technique is patent protected [47]. In the same study, a 2D U-Net model was trained on an ambiguous dataset provided by experts to segment and classify diagnostically relevant thyroid cells [15]. Whilst on average it exceeded good results, there were some challenges that will be addressed in this paper.

In the field of artificial intelligence, three categories of AI can be distinguished: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Super Intelligence (ASI). [21].The category ”Artificial Narrow Intelligence” (ANI) is used to describe AI systems that have been designed to perform specific tasks or solve particular problems [10] such as the classification of thyroid cells. Medical experts utilise a range of imaging techniques to diagnose patients underscores the importance of prompt disease detection to improve survival rates. The application of artificial intelligence has yielded promising results in the field of diagnostics, with notable improvements in diagnostic procedures [27]. A review of the literature in the medical fields revealed that the division of the classification and segmentation tasks in the context of medical images is an effective approach [19].

## II. Problem Description

The utilisation of machine learning in a practical context presents a distinct set of challenges compared to its deployment in research environments, particularly with regard to data quality and quantity. In industrial settings, the collection of data samples and labels is often challenging due to the costs and complexities associated with data gathering. This results in incomplete or missing labels within the resulting datasets [40].

- The manually annotated diagnostically relevant cells used in this thesis resulted in masks that did not accurately capture the cell outlines, leading to imprecise Fpol values and inaccurate mask representations.

- Cells that were not annotated but potentially diagnostically relevant contributed to an inaccurate ground truth, resulting in the detection of unannotated yet relevant cells or failed to detect the annotated cell entirely. Consequently, the loss function did not adequately represent the problem, failing to capture all diagnostically relevant cells within an image.
- The dual representation in the dataset of the classification and segmentation task lead to a more complex learning.

The objective of this paper is to enhance the segmentation and classification of diagnostically relevant cells using a kernel-based approach that could be adapted to different applications. As manual improvement and enhancement of the dataset is costly, we propose utilising the foundation model CellSAM to imrpove the segmentation of thyroid cells [42]. In order to enhance the detection of diagnostically relevant thyroid cells and to expand the data set with supplementary annotations, we employed a linear kernel as the input for binary clustering to improve the ground truth.

## III. Related Work

As of 2023, AI has achieved levels of performance that surpass human capabilities across a range of tasks [25]. The emergence of foundation models, including BERT, DALL-E and GPT-3, has brought an advance in the field of AI. Trained on vast and diverse datasets, these models have demonstrated remarkable capabilities in performing a multitude of downstream tasks, as evidenced in recent research [3]. In 2023, Meta researchers introduced the model Segment Anything (SAM) which has been trained on a dataset with 1 billion masks [17]. The model has demonstrated an enhancement in the efficiency of human-performed segmentation tasks. The proficiency of the foundation model SAM highlights the critical importance of leveraging a comprehensive and robust dataset to effectively support human tasks [25]. The downside for SAM in the application of cellular image segmentation is the default uniform grid prompting strategy which leads to inaccurate cell segmentation, due to varying cell densities. In 2024, CellSAM was introduced as a model for cell segmentation. A comprehensive dataset encompassing various archetypes was used for training. CellSAM integrates SAM's Vision Transformer (ViT) for feature extraction and uses CellFinder with a Anchor DETR Framework to generate bounding boxes as prompts for SAM to enumerate masks. [14]

The utilisation of a pretrained semantic segmentation model requires in many instances much less data to adapt the model to a specific task or data set compared to training from scratch. The segregation of the semantic and classification tasks is expected to diminish the complexity of the model. The separation also allows for the subsequent utilisation of extracted cells as a template for manual annotations, thereby reducing the time required for manual annotation and facilitating the enhancement of the dataset.

In the preceding work [15], a 2D U-Net was trained on the entire image. The encoder-decoder architecture comprises a pixel-wise softmax function, which converts feature maps into probability distributions over classes for each pixel [36]. The ambiguity inherent to the labels within the data set made it challenging to achieve an accurate prediction. In order to address this issue, we designed a kernel-based approach with the aim of increasing the quality of the ground truth.

### A. Segmentation

Semantic segmentation involves identifying objects within an image at a pixel level, distinguishing between individual instances of the same class while preserving spatial information [7]. Classification on the on the other hand focuses on the global context aiming to classify all the extracted pixels on a global or local image representation [35]. This hierarchical approach is analogous to natural language processing (NLP), where the initial step of tokenization divides text into words, which are then processed using techniques such as named entity recognition (NER) to derive their meaning [24]. In semantic segmenation local and global context has been used to identify a specific object [29]. While segmentation tasks in computer vision are similarly structured to classification tasks, the underlying problem to be solved is distinct. Segmentation addresses the question of where to search for a specific object (local context), whereas classification uses this knowledge to determine what to search for (global context). It has also been shown that the aforementioned approach to segmentation and classification, which entails the splitting of these tasks, explains the more accurate and stable learning [8].

Transfer learning has demonstrated that the freezing of features and the utilisation of the output of the previous task for a new one results in expeditious training and a reduction in computational complexity and more accuracy. This approach is based on a hierarchical training model, which allows for the separation of segmentation and classification tasks [28]. Although this approach is capable of distinguishing between two tasks, it is unable to address the issue of improving the frozen layer. In light to separate the global and local context, we splitted the segmentation and classification task. We propose a simple approach to enhance the quality of the dataset reflecting to reduce the computational complexity [3].

### B. Feature Extraction

Highly aggressive thyroid cancer exhibits a degree of homogeneity, it can be assumed that other cancer cells demonstrate also elatively uniform characteristics, although not entirely homogeneous [46]. Under this assumption we allowed to focus on prediction per cells instead of the prediction per image. The maximum number of cells in an image predicted with the CellSAM model [14] was 266. Consequently, the training set could be augmented by a significant number. The utilisation of morphological features in other medical application like endometrial cytology enhances the diagnostic accuracy and prognostic prediction [30]. The fluorence polarization method, which enables the capture of the dye retained for an extended period of time in cancer cells, was conducted using a fine-needle aspiration [15]. To improve pathological feature extraction of medical images Gaussian shaped probability models

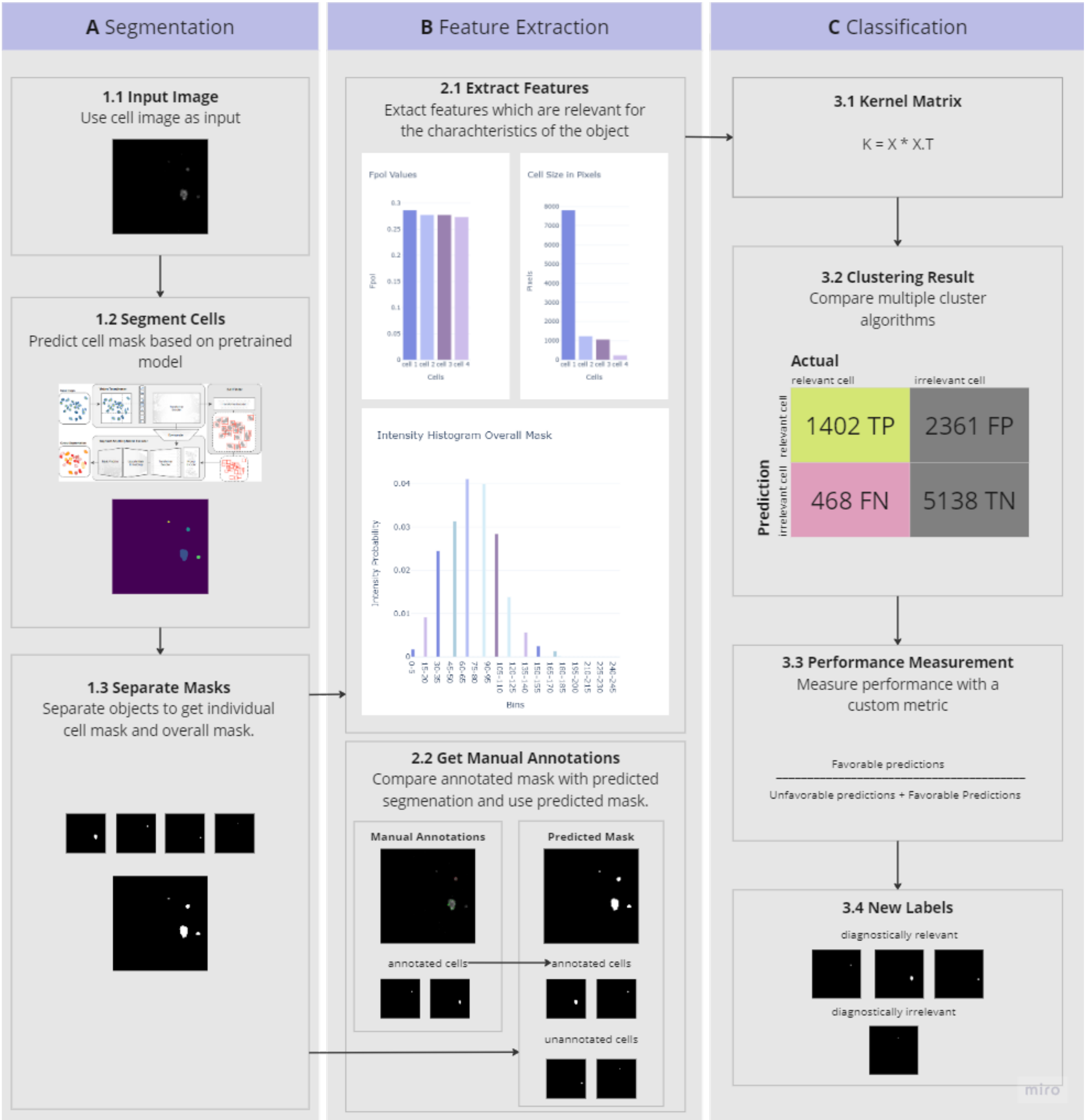


Fig. 1. A comprehensive overview of the cell segmentation and classification pipeline, which has been developed with the objective of facilitating the detection of diagnostically relevant thyroid cells, is provided herein.

can help to enhance the quality of images and therefore lead to better diagnosis [20]. The captured intensity values can be normalised by utilising the histogram of the intensity values of the diagnostically relevant cell in conjunction with a probability distribution function (PDF). A PDF describes the distribution of values of a random variable [41]. By normalizing the intensity values, the distribution's mean and variance are standardized to 0 and 1 [4]. In this application, the intensity changes that are characteristic of cancerous cells can be employed to standardise the differentiation between multiple cells.

## C. *Classification*

The previously used 2D U-Net model, which had been trained on an ambiguous ground truth dataset, was subjected to a data selection process to reduce noise and potentially improve the prediction. Only high-quality cells were included in the selected dataset [15]. The 2D U-Net result was validated by experts and rated as ”very useful”. However, due to the limitations of the dataset, which consisted of only 90 available images, the model's generalizability could not be demonstrated [44]. This result confirmed the importance of enhancing the quality of the dataset prior to prediction [15]. As discussed in the section Feature Extraction III-B the intensity value histograms for each cell were normalised into probability density functions, which were employed as a feature for each cell in the present study.

Kernels are functions that are utilized for the purpose of computing a similarity measurement between data points. In order for a kernel to be classified as symmethodological, it must be attained through the mechanism of multiplication. A further property of kernels is that the resulting inner product in a high-dimensional space is a valid, positive semi-definite kernel. Consequently, it is necessary that the kernel satisfy Mercer's theorem [26].

The efficacy of kernel methodologies has been demonstrated in scenarios that necessitate the enhancement of generalization capabilities and in situations where the availability of estimation is uncertain. This renders them a more suitable option for datasets that are limited in quality and size, particularly when utilized for deep learning methods [37]. The strength of kernel-based approaches also lies in the invariance of transformation [33]. This indicates that a transformation is unlikely to affect the diagnostic relevance of a cell, thereby supporting the selection of the kernel transformation invariant option.

### D. *Classification*

Data clustering is a technique that is employed in a variety of fields for the purpose of organising objects into groups, or clusters, based on a measure of similarity between data points. The application of clustering can assist in the comprehension of the intrinsic structure of the data, the identification of patterns, and the streamlining of the data for subsequent analysis [34]. It can thus be posited that this methodology may be employed for the discovery of analogous, unclassified cell clusters within a given dataset.

*1) **Agglomerative Clustering**:* The Agglomerative Clustering algorithm is used to generate clusters from a single object. It merges the two most similar clusters and therefore builds hierarchical clustering by successive merging or splitting [11].The approach is capable of handling clusters of varying shapes and sizes, which was assumed to be suitable for the binary prediction of thyroid cells.

*2) **BIRCH**:* The Balanced Iterative Reducing and Clustering (BIRCH) algorithm uses a clustering feature tree (CF-tree) which maintains a compact summary of the dataset. The algorithm is used to process larger datasets due to its efficient and scalable nature [50]. Although the dataset employed in this study is not particularly extensive, the hierarchical structure may offer insights into the formation of subgroups within the data.

*3) **Spectral Clustering**:* Spectral clustering is a data reduction technique that employs eigenvalues derived from a similarity matrix to facilitate subsequent cluster analysis. It is adequate for handling non-convex [43] and convex [38] clusters, and is capable of capturing the intrinsic data structure in a way that allows for effective cluster identification [43]. The shape of a cell can be convex or non-convex, when the form of the cancer cell is irregular and uneven [12]. Because of these properties, the algorithm may be suitable for the detection of diagnostically relevant cells.

*4) **KMeans**:* The K-means algorithm used to cluster a set of data points was introduced in 1967. Each data point belongs to the cluster with the nearest mean [23]. K-Means is still used as an effective clustering methodology in the medical field [2] and could therefore be suitable for cell images.

*5) **Mini Batch KMeans**:* The computational challenges of KMeans have been addressed with the Mini Batch KMeans algorithm. The algorithm can handle large datasets efficiently using random and small batches.The random sampling helps to explore different parts of the data that might avoid solution in local minima [39].

## IV. Methodology

Data cleaning is a fundamental issue to be solved for researcher in the medical field to avoid incorrect clinical measurements [32]. Accordingly, we devised an approach that is conducive to improving data quality and can be tailored to categorize the thyroid cells that require analysis. This section also emphasises an approach which distinguishing between the segmentation and classification tasks. The following section illustrates the methodology used to improve the accuracy of diagnostically relevant cell detection.

### A. *Segmentation*

The segmentation of cancerous cells in microscopy images is a common step in quantitative tissue analysis for both research and diagnostic purposes [6].

The manual annotations can be used to train machine learning algorithms to automate this process. However, if the annotations are incomplete and doesn't represent the task which needs to be solved, the training potentially leads to inaccurate results. Since the manual correction of the previously annotated images through expert is time consuming and also costly, the problem was solved computationally. CellSAM tends to segment cells liberally. As a result, the result is sensitive, meaning that most cells are found, but some segmented cells may not be diagnostically relevant. The quality of the dataset could be enhanced by improving the representation of cells that are relevant for diagnostic purposes. Each image used in this project is recorded as an 8-bit grayscale image with a resolution of 1000x1000 pixels in the TIFF file format. The overall cell mask for each image was split into individual masks after using CellSAM for segmentation. This process increased the dataset from 139 images to 9369 statistical representations of each cell, of which 1870 were manually annotated as diagnostically relevant cells.

### B. *Feature Extraction*

In the context of manual analysis, the Fpol value plays a role in determining whether a cell should be classified as diagnostically relevant [15]. The Fpol value incorporates the intensity values of a given cell. In this project, a kernel was employed to represent the intensity changes within an image. Following the segmentation process, the quality of the clustering algorithm should have been assessed using a small dataset comprising 20 fully annotated images. In order to evaluate the efficacy of the assessment, a single image was selected for analysis by two experts. In light of the considerable divergence of expert opinions on this single image, it was deemed necessary to adopt an alternative approach to the validation of the results. The efficacy of the utilized clustering algorithms for binary classification was substantiated through a process of validation that involved the integration of prior knowledge regarding the annotated diagnostically relevant cells with statistical data that reflected the diverse characteristics of a potential cancer cell.

*1) **Kernel Feature**:* We processed the co-polarized and cross-polarized images which after masking represented an image for every cell. The intensity values of these images range from 0 to 255. To construct the histogram, we divided the intensity range into bins of size 5 units and calculated the probability density function. which is calculated as follows

$$f(x) = \frac{d}{dx}F(x)$$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

The function $f(x) = \frac{d}{dx}F(x)$ represents the rate of change of the accumulated quantity, while $F(x) = \int_{-\infty}^{x} f(t)\,dt$ denotes the integral of $f(t)$ from negative infinity to $x$. [18]

The created histograms represents the distribution of pixel intensities in the images for every cell. By applying it to every cell separately, we were able to reduce certain noise by employing. After this normalization a linear kernel matrix $K$ was computed by multiplying the concatenated feature matrix $\mathbf{X}$ with its transpose $\mathbf{X}^T$ :

$$K = \mathbf{X}\mathbf{X}^T$$

[13]

The process was applied to the co-polarized and cross-polarized images separately and in combination, with the objective of evaluating the performance of the kernels at a later stage.

*2) **Feature for Metric Calculation**:* Cancer cells release exosomes into the surrounding environment. Exosomes are small vesicles that facilitate the transportation of various biomolecules, including proteins. The quantity and type of released proteins can be quantified. The intensity of cellular fluorescence, which can be measured using a fluorescent dye, is a reliable indicator of cancerous cells. This is because cancerous cells often secrete a greater number of exosomes and contain specific proteins that can be visualized using the fluorescent dye. This characteristic renders them valuable for the diagnosis and monitoring of cancer [48]. Accordingly, the Fpol values, which quantifies the intensity levels of the co-polarized and cross-polarized images, were selected for assessment of the efficacy of the diagnotically relevant cell detection.

$$\text{Fpol} = \frac{\overline{I}_{\text{co}} - G \cdot \overline{I}_{\text{cross}}}{\overline{I}_{\text{co}} + G \cdot \overline{I}_{\text{cross}}}$$

The calculation includes the mean co-polarized ($I_{\text{co}}$) and cross-polarized ($I_{\text{cross}}$) image for each cell. In the preceding study, the fluorescence polarization value, denoted as $F_{\text{pol}}$, was determined using a calibration factor, $G$. This factor was set to a value of 0.75, consistent with the methodology applied in the current analysis [15].

The width-to-height ratio of cells, which compares the minor length (width) to the major length (height), is a parameter utilized in the classification of cancerous cells [6].

$$\text{Ratio Width Height} = \frac{\text{minor length}}{\text{major length}}$$

In this context, the term "major length" refers to the length of the major axis of the ellipse, while "minor length" denotes the minor axis of the ellipse.

The measurement of circularity can be employed as a means of detecting cancer cells, given that cells with greater metastatic potential are more susceptible to deformation [16].

$$\text{Circularity} = \frac{4\pi \times \text{Area}}{\text{Perimeter}^2}$$

The area represented the number of non-zero pixels, while the perimeter indicated the total length of the contour [51].

An additional morphological feature that can be utilized to illustrate the intricacy of the shape is the outline length of a cell.

$$\text{Outline length} = \text{n*l}$$

In this equation, the variable "n" represents the number of sides with the same length, while the variable "l" represents the length of the sights.

In the context of a malignant diagnostic procedure, the size of a cell appears to be a significant factor in the classification process [22].

$$\text{Size (µm)} = \sum_{i,j} \text{M(i,j)} * \text{pixelsize}$$

In this calculation, $i$ represents the index of the row and $j$ the index of the column in an image. The pixel size utilized in this study was 0.205, as specified by the experts who provided the images.

The above features tried to capture the characteristics of the cancer cells. The information was used for the performance measurements of the clustering results.

## V. Model Development

We used multiple clustering algorithm which are suitable for binary clustering problems to predict the new ground truth. We tried multiple linear kernels of the co-polarized and cross-polarized images as input values for the clustering algorithms, with the objective of predicting the new label. Three distinct kernel configurations were employed for the analysis of intensity values in cell-masked images obtained from co-polarized and cross-polarized microscopy. The initial kernel contained the histogram of intensity values derived from the cell mask of the co-polarized image, the second contained the histogram from the cell mask of cross-polarized image, and the third was a composite kernel integrating both the co-polarized and cross-polarized intensity values into a unified histogram for comprehensive analysis.

In the initial stage of the clustering algorithm, a basic configuration was employed. The number of clusters was set to two for the following algorithms: Agglomerative Clustering, BIRCH, Spectral Clustering, K-Means, and Mini Batch K-Means. This setting permitted the differentiation of diagnostically relevant and diagnostically irrelevant cells. A minimal threshold was selected for the BIRCH clustering to merge smaller clusters into larger ones and therefore potentially enhance the cluster quality. The clustering models were employed with and without label propagation, with the objective of potentially leveraging prior knowledge.

*1) **Evaluation Metric**:* In order to evaluate the results, we integrated the existing knowledge about the annotated cells with the metrics that define their characteristics. To establish a metric, it was essential to define the criteria for a favorable and an unfavorable prediction. Because of the absence of prior knowledge, the cell characteristics weren't weighted against each other. We utilized a straightforward calculation to utilize the characteristics as a metric for evaluating the efficacy of the predictions. In scenarios where the data exhibits minimal variation, the geometric mean is more closely aligned with the unbiased outcomes therefore used in this work [9] .

Every feature used for the calculation of the clustering metric (Performance Label Ambiguity) was scaled to a range between 0 and 1 to ensure that each feature contributes equally to the analysis.

$$Scaling_i = \frac{x_i - \bar{x}}{s}$$

[1]

In the calculation above $Scaling_i$ represents the z-score, $x_i$ the $i$th value, $\bar{x}$ the mean and $s$ the standard deviation.

After scaling the features, they were multiplied together.

$$\text{Features}_i = \text{Fpol}_i \times \text{Size (µm)}_i \times \text{Circularity}_i \times \text{Ratio Width Height}_i \times \text{Outline Length}_i$$

The characteristic value for each feature, $\text{Characteristics}_i$, is then calculated as the geometric mean of the features:

$$\text{Characteristics}_i = \text{Features}_i^{\frac{1}{\text{Number of features}_i}}$$

**Favorable Prediction** The detection of a cell as a diagnostically relevant cell, subsequently annotated by experts, was considered a positive prediction. Where $y_i$ represents the ground truth labels and $\hat{y}_i$ the predicted labels.

The formula for calculating True Positives (TP) is given by:

$$\text{TP} = \sum_{i=1}^{n} (y_i = \hat{y}_i \land y_i = 1)$$

In instances where a cell was not manually annotated, but the calculation of its characteristics fell within the range of the experts decision, the cell was assumed to be correctly recognised by the algorithm, despite the experts having failed to label it. $\min(\text{Characteristics}_i)$ represents the lower bound and $\max(\text{Characteristics}_i)$ the upper bound of the range where the experts decision were laying.

$$\text{Correct Recognized Missed Cells} = \sum_{i=1}^{n} (\min(\text{Characteristics}_i) \leq \hat{y}_i \leq \max(\text{Characteristics}_i))$$

**Unfavorable Prediction** In the event of the clustering algorithm failing to detect the annotated cells, this should be regarded as an unsuccessful outcome.

$$\text{FN} = \sum_{i=1}^{n} (y_i \neq \hat{y}_i \land \hat{y}_i = 0)$$

If the $\text{Characteristics}_i$ of the predicted cell didn't lay within the range of the experts decision and in absence of manual annotation, the cell was deemed to have been incorrectly identified by the algorithm.

$$\text{Wrong Recognized Cells} = \sum_{i=1}^{n} (\min(\text{Characteristics}_i) \geq \hat{y}_i \geq \max(\text{Characteristics}_i))$$

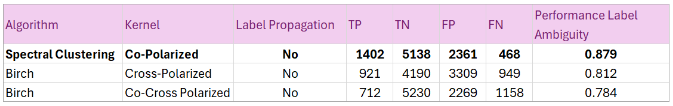
**Ambiguity Metric** In order to create a metric that fulfils all of the aforementioned purposes, namely the favourable and unfavourable factors, all the elements were combined to construct a metric that addresses the issue of label ambiguity.

$$\text{Performance Label Ambiguity} = \left( \frac{\text{TP} + \text{Correct Recognized Missed Cells}}{\text{FN} + \text{Wrong Recognized Cells}} \right)$$

The metric named "Performance Label Ambiguity" was employed to identify the most effective clustering algorithm to detect labeled and missed labeled diagnostically relevant cells.

## VI. Results

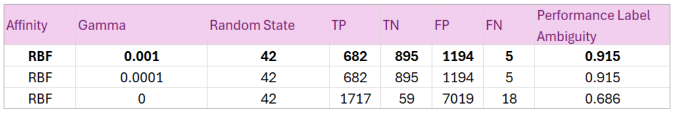
As mentioned in the section on model development, we used the standard settings provided by scikit-learn [31] for nearly every clustering algorithm. The algorithm that demonstrated the highest performance was then employed to train on multiple hyperparameters.



| Algorithm | Kernel | Label Propagation | TP | TN | FP | FN | Performance Label Ambiguity |
|---|---|---|---|---|---|---|---|
| **Spectral Clustering** | **Co-Polarized** | **No** | **1402** | **5138** | **2361** | **468** | **0.879** |
| Birch | Cross-Polarized | No | 921 | 4190 | 3309 | 949 | 0.812 |
| Birch | Co-Cross Polarized | No | 712 | 5230 | 2269 | 1158 | 0.784 |

Fig. 2. Top 3 Results Cluster Algorithms

Since spectral clustering gave good performance compared to other clustering algorithms, we used this algorithm for hyperparameter tuning. The three best results are shown in the figure 3. The findings revealed the importance of assigning greater significance to features that are more effective in detecting cancerous cells, to reflect in the performance metric. The preliminary results of the comparative analysis of multiple algorithms demonstrated better accuracy, as evidenced by a higher number of correctly identified cells that had also been manually annotated.



| Affinity | Gamma | Random State | TP | TN | FP | FN | Performance Label Ambiguity |
|---|---|---|---|---|---|---|---|
| **RBF** | **0.001** | **42** | **682** | **895** | **1194** | **5** | **0.915** |
| RBF | 0.0001 | 42 | 682 | 895 | 1194 | 5 | 0.915 |
| RBF | 0 | 42 | 1717 | 59 | 7019 | 18 | 0.686 |

Fig. 3. Top 3 Results Hyperparametertuning Spectral Cluster

## VII. Discussion and Future Work

This study necessitated a considerable amount of time for the data preparation procedures. Although the manual annotations and matching mask were available for consultation, the co-polarized and cross-polarized images were only accessible in multiple PowerPoint presentations. Accordingly, the images were prepared for utilization in this study. Furthermore, the manual and CellSAM masks had to be separated into individual cell masks. In order to validate the segmented result of CellSAM, it was necessary to identify the matching cell mask from the manual annotation. Additionally, the individual cell masks extracted from the segmented result of CellSAM were applied to the entire image, as each image was assigned a unique cell mask number for identification purposes. As a result, the extensive preparatory work limited the scope of our analysis, preventing a comprehensive evaluation of various clustering algorithms. In the future, testing additional hyperparameters on multiple clustering algorithms could further enhance the robustness and accuracy of our findings.

The validation of one segmented image by experts resulted in divergent answers. Therefore, the uncertainty in the manual annotations is still present. The predicted result of the clustering algorithm should be validated by experts, despite the potential for uncertainty in manual annotations due to the practical nature of the application. If the prediction is found to be insufficient, the quality of the dataset may need to be manually improved or the performance metric need to be adjusted.

The influence of the utilized features in the metric, designated as Performance Label Ambiguity, on the identification of diagnostically pertinent cells remains ambiguous. In the event that particular cell attributes exert a pronounced impact on the outcomes, it would be advantageous to ascribe weights to these attributes. This refinement would facilitate a more comprehensive comprehension of the results.

The labeled dataset and the clustering algorithm employed in this project facilitate the selection of appropriate cell masks from CellSAM. These cell masks can be integrated into a unified mask for each co-polarized and cross-polarized image. Subsequently, the consolidated masks can be applied to the 2D U-Net, which was trained in previous work. This application enables a comparison of the performance outputs between this project and earlier efforts.

In the medical field, any judgment can have bad consequences for a patient. Based on examination, a doctor will form a reasonable explanation [5]. Although the clustering algorithm employed in this study may enhance the ground truth, it is deficient in terms of elucidating the rationale behind the decisions made. In future research, incorporating additional metrics to explain the classification could facilitate not only prediction but also the understanding of the reasoning behind the prediction.

In order to make optimal use of the heterogeneity present in the data, future models should not only focus on individual cells but also consider the entire image, thus preventing the loss of valuable global contextual information. Implementing this approach could potentially incorporate the comprehensive features of the image into the analysis, thereby enhancing the model's diagnostic capabilities.

In this study, we implemented a hierarchical approach, utilizing segmented masks for classification to mitigate noise. To fully leverage the benefits of this hierarchical method, which incorporates both global and local contexts for classifying diagnostically relevant cells, an analysis comparing single and multi-head networks would be beneficial. Such an investigation could help determine whether a dual-headed approach enhances the segmentation and classification tasks. Additionally, exploring whether a hierarchical strategy—initial segmentation followed by classification—outperforms a sequential approach could provide insightful findings.

A further area of interest is the computational cost and data requirements of the various models under consideration. In particular, it would be interesting to evaluate the efficacy of a unified model that is capable of both classification and segmentation, in comparison to two distinct models that have been developed for each of these tasks. This analysis will facilitate an understanding of the trade-offs inherent to model complexity versus performance.

## VIII. Conclusion

The research presented a practical approach to potentially enhance the quality of the dataset provided by experts in the medical field. The separation of the segmentation and classification processes to extract pertinent features helped to reduce noise within the image.

The foundation model CellSAM was employed for the purpose of segmenting thyroid cells. Thereafter, a linear kernel was applied to the cells extracted from the aforementioned segmentation, which were then utilized as input for the classification of diagnostically relevant thyroid cells.

In this study we have addressed the significant challenges with sample scarcity and label ambiguity, that are inherent to manual annotations. The proposed approach has the potential to improve the quality of the dataset. Although the initial result obtained with the basic setting appeared promising, it is imperative to seek expert validation before drawing any definitive conclusions. Following the validation process, it may be necessary to make adjustments to the manual annotations and metric calculation in order to evaluate the results in a more robust manner. Subsequently, this approach could be extended to other applications where an improvement in data quality is necessary and the use of prior knowledge to assess results would be advantageous.

With slight adaptation, extracting features from images and using a machine learning approach for relevant characteristics could also be implemented in various other fields. Overall, the approach we have used shows potential for enhancing diagnostic processes in medical contexts, potentially paving the way for more effective and timely cancer treatments.

# References

[1] Edward I Altman, Małgorzata Iwanicz-Drozdowska, Erkki K Laitinen, and Arto Suvas. Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. *Journal of international financial management & accounting*, 28(2):131–171, 2017.

[2] Tina Babu and Rekha R Nair. Colon cancer prediction with transfer learning and k-means clustering. In *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*, pages 191–200. Springer, 2023.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Kelsy Cabello-Solorzano, Isabela Ortigosa de Araujo, Marco Peña, Luís Correia, and Antonio J. Tallón-Ballesteros. The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis. In *International Conference on Soft Computing Models in Industrial and Environmental Applications*, pages 344–353. Springer, 2023.

[5] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634, 2023.

[6] Tulasi Gayatri Devi, Nagamma Patil, Sharada Rai, and Cheryl Philipose Sarah. Real-time microscopy image-based segmentation and classification models for cancer cell detection. *Multimedia Tools and Applications*, 82(23):35969–35994, 2023.

[7] Tanuj Dhamija, Ashish Gupta, Sanjeev Gupta, et al. Semantic segmentation in medical images through transfused convolution and transformer networks. *Applied Intelligence*, 53(2):1132–1148, 2023.

[8] John Doe and Jane Smith. Robust clinical applicable cnn and u-net based algorithm for mri classification and segmentation for brain tumor. *Expert Systems with Applications*, 200:122347, 2023.

[9] Alex Kane Eric Jacquier and Alan J. Marcus. Geometric or arithmetic mean: A reconsideration. *Financial Analysts Journal*, 59(6):46–53, 2003.

[10] Ragnar Fjelland. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7(1):1–9, 2020.

[11] Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024.

[12] Pablo Gottheil, Jürgen Lippoldt, Steffen Grosser, Frédéric Renner, Mohamad Saibah, Dimitrij Tschodu, Anne-Kathrin Poßögel, Anne-Sophie Wegscheider, Bernhard Ulm, Kay Friedrichs, et al. State of cell unjamming correlates with distant metastasis in cancer patients. *Physical Review X*, 13(3):031003, 2023.

[13] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. 2008.

[14] Uriah Israel, Markus Marks, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Shenyi Li, Morgan Schwartz, Elora Pradhan, Ada Ates, et al. A foundation model for cell segmentation. *bioRxiv*, 2023.

[15] Peter R Jermain, Martin Oswald, Tenzin Langdun, Santana Wright, Ashraf Khan, Thilo Stadelmann, Ahmed Abdulkadir, and Anna N Yaroslavsky. Deep learning-based cell segmentation for rapid optical cytopathology of thyroid cancer. *Scientific Reports*, 14(1):16389, 2024.

[16] Siwei Ju, Cong Chen, Jiahang Zhang, Lin Xu, Xun Zhang, Zhaoqing Li, Yongxia Chen, Jichun Zhou, Feiyang Ji, and Linbo Wang. Detection of circulating tumor cells: opportunities and challenges. *Biomarker research*, 10(1):58, 2022.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[18] Arnold Kirsch. The fundamental theorem of calculus: visually? *ZDM*, 46:691–695, 2014.

[19] Hui Kong, Metin Gurcan, and Kamel Belkacem-Boussaid. Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting. *IEEE transactions on medical imaging*, 30(9):1661–1677, 2011.

[20] Chung-Feng Jeffrey Kuo and Han-Cheng Wu. Gaussian probability bi-histogram equalization for enhancement of the pathological features in medical images. *International Journal of Imaging Systems and Technology*, 33(4):123–130, 2023.

[21] Jiewu Leng, Xiaofeng Zhu, Zhiqiang Huang, Xingyu Li, Pai Zheng, Xueliang Zhou, Dimitris Mourtzis, Baicun Wang, Qinglin Qi, Haidong Shao, et al. Unlocking the power of industrial artificial intelligence towards industry 5.0: Insights, pathways, and challenges. *Journal of Manufacturing Systems*, 73:349–363, 2024.

[22] Nana Lyu, Amin Hassanzadeh-Barforoushi, Laura M Rey Gomez, Wei Zhang, and Yuling Wang. Sers biosensors for liquid biopsy towards cancer diagnosis by detection of various circulating biomarkers: current progress and perspectives. *Nano Convergence*, 11(1):22, 2024.

[23] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[24] Ivan Malashin, Igor Masich, Vadim Tynchenko, Andrei Gantimurov, Vladimir Nelyub, and Aleksei Borodulin. Image text extraction and natural language processing of unstructured data from medical reports. *Machine Learning and Knowledge Extraction*, 6(2):1361–1377, 2024.

[25] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The ai index 2024 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

[26] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

[27] Omar Abdullah Murshed, Farhan Alnaggar, Basavaraj N. Jagadale, and Hasib Daowd Esmail Al-Ariki. Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis. *Artificial Intelligence Review*, 221, July 2024.

[28] Sidra Naz, Abdur Ashraf, and Anam Zaib. Transfer learning using freeze features for alzheimer neurological disorder detection using adni dataset. *Multimedia Systems*, 28:85–94, 2022.

[29] Yitong Ni, Jie Liu, Wei Chi, Xin Wang, and Dapeng Li. Cgglnet: Semantic segmentation network for remote sensing images based on category-guided global–local feature interaction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024.

[30] Toshimichi Onuma, Akiko Shinagawa, Tetsuji Kurokawa, Makoto Orisaka, and Yoshio Yoshida. Fractal dimension, circularity, and solidity of cell clusters in liquid-based endometrial cytology are potentially useful for endometrial cancer detection and prognosis prediction. *Cancers*, 16(13):2469, 2024.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Spectralclustering, 2024. Accessed: 2024-07-31.

[32] Hang TT Phan, Florina Borca, David Cable, James Batchelor, Justin H Davies, and Sarah Ennis. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. *Scientific reports*, 10(1):10164, 2020.

[33] Adrien Poulenard, Marie-Julie Rakotosaona, Yann Ponty, and Maks Ovsjanikov. Effective rotation-invariant point cnn with spherical harmonics kernels. In *2019 International Conference on 3D Vision (3DV)*, pages 47–56. IEEE, 2019.

[34] Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang, and Zhenyu Lu. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, 56(8):8219–8264, 2023.

[35] Shangzhen Ren, Nan Zhao, Qi Wen, Guanglu Han, and Shiming He. Unifying global-local representations in salient object detection with transformers. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[37] Carlos Ruiz, Carlos M Alaíz, and José R Dorronsoro. A survey on kernel-based multi-task learning. *Neurocomputing*, 577:127255, 2024.

[38] scikit-learn developers. Clustering. https://scikit-learn.org/stable/modules/clustering.htmlspectral-clustering, 2023. Accessed: 2023-07-31.

[39] David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178, 2010.

[40] Niclas Simmler, Pascal Sager, Philipp Andermatt, Ricardo Chavarriaga, Frank-Peter Schilling, Matthias Rosenthal, and Thilo Stadelmann. A survey of un-, weakly-, and semi-supervised learning methods for noisy, missing and partial labels in industrial vision applications. In *2021 8th Swiss Conference on Data Science (SDS)*, pages 26–31. IEEE, 2021.

[41] Xiaofei Sun, Lin Shi, Yishan Luo, Wei Yang, Hongpeng Li, Peipeng Liang, Kuncheng Li, Vincent CT Mok, Winnie CW Chu, and Defeng Wang. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomedical engineering online*, 14:1–17, 2015.

[42] Joecelyn Kirani Tan, Wireko Andrew Awuah, Sakshi Roy, Tomas Ferreira, Arjun Ahluwalia, Saibaba Guggilapu, Mahnoor Javed, Muhammad Mikail Athif Zhafir Asyura, Favour Tope Adebusoye, Krishna Ramamoorthy, et al. Exploring the advances of single-cell rna sequencing in thyroid cancer: a narrative review. *Medical Oncology*, 41(1):27, 2023.

[43] Chang Tang, Zhenglai Li, Jun Wang, Xinwang Liu, Wei Zhang, and En Zhu. Unified one-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6449–6460, 2022.

[44] Martin Oswald Tenzin Samdrup Langdun. *Deep-Learning-Based Cell Segmentation for the Detection of Thyroid Cancer in Single Cells*. PhD thesis, Bachelor Thesis, 2023.

[45] The American Cancer Society. Cancer facts & figures, 2024. In *Cancer Facts & Figures*. The American Cancer Society, 2024. Accessed: 2024-05-05.

[46] George J Xu, Matthew A Loberg, Jean-Nicolas Gallant, Quanhu Sheng, Sheau-Chiann Chen, Brian D Lehmann, Sophia M Shaddy, Megan L Tigue, Courtney J Phifer, Li Wang, et al. Molecular signature incorporating the immune microenvironment enhances thyroid cancer outcome prediction. *Cell Genomics*, 3(10), 2023.

[47] Anna N. Yaroslavsky and Richard R. Anderson. Fluorescence polarization imaging device and method, March 2012. US Patent 8,139,211.

[48] Dan Yu, Yixin Li, Maoye Wang, Jianmei Gu, Wenrong Xu, Hui Cai, Xinjian Fang, and Xu Zhang. Exosomes as a new frontier of cancer liquid biopsy. *Molecular cancer*, 21(1):56, 2022.

[49] Ping Zhang, Hui Zuo, Takashi Ozaki, Nami Nakagomi, and Kennichi Kakudo. Cancer stem cell hypothesis in thyroid cancer. *Pathology international*, 56(9):485–489, 2006.

[50] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1:141–182, 1997.

[51] Joviša Žunić and Kaoru Hirota. Measuring shape circularity. In *Progress in Pattern Recognition, Image Analysis and Applications: 13th Iberoamerican Congress on Pattern Recognition, CIARP 2008, Havana, Cuba, September 9-12, 2008. Proceedings 13*, pages 94–101. Springer, 2008.

## Appendix A
## Semi-automatic annotation and label ambiguity

We hypothesized, that a semi-automatic approach for cell annotation could be employed. Instead of manually delineating diagnostically relevant cells, the clinicians would only flag individual cells as being diagnostically relevant, being in focus, and having the approximately correct delineation. We prepared a subset of the available 139 frames for ease of access by numbering each cell and providing an Excel file with drop-down selection cells for each category and cell ID to maximize user experience. Two domain experts, who were trained by the developer of the annotation protocol used in [15], performed the task. Processing a single frame with 33 cells took them 33 and 40 minutes, respectively. After a single frame, we rejected the approach because it was too time-consuming. The results, however, indicate the very high degree of ambiguity of the manual annotation, further advocating for data-driven approaches.