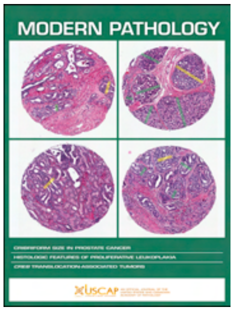


Machine learning streamlines the morphometric characterization and multi-class segmentation of nuclei in different follicular thyroid lesions: everything in a NUTSHELL

Vincenzo L'Imperio, Vasco Coelho, Giorgio Cazzaniga, Daniele M. Papetti, Fabio Del Carro, Giulia Capitoli, Mario Marino, Joranda Ceku, Nicola Fusco, Mariia Ivanova, Andrea Gianatti, Marco S. Nobile, Stefania Galimberti, Daniela Besozzi, Fabio Pagni

PII: S0893-3952(24)00188-1

DOI: https://doi.org/10.1016/j.modpat.2024.100608

Reference: MODPAT 100608

To appear in: *Modern Pathology*

Received Date: 27 March 2024

Revised Date: 16 August 2024

Accepted Date: 29 August 2024

Machine learning streamlines the morphometric characterization and multi-class segmentation of nuclei in different follicular thyroid lesions: everything in a NUTSHELL

Vincenzo L'Imperio[1,2*], Vasco Coelho[3*], Giorgio Cazzaniga[1,2], Daniele M. Papetti[3], Fabio Del Carro[1,2], Giulia Capitoli[1,8], Mario Marino[3], Joranda Ceku[1,2], Nicola Fusco[4,5], Mariia Ivanova[4], Andrea Gianatti[6], Marco S. Nobile[7,8], Stefania Galimberti[1,8,9], Daniela Besozzi[3,8#], Fabio Pagni[1,2#]

[1]School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy

[2]Department of Pathology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy

[3]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

[4]Division of Pathology, European Institute of Oncology IRCCS, Milan, Italy

[5]Department of Oncology & Hemato-Oncology, University of Milan, Milan, Italy

[6]Department of Pathology, ASST Papa Giovanni XXIII, Bergamo, Italy

[7]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy

[8]Bicocca Bioinformatics Biostatistics and Bioimaging Research Centre - B4, University of Milano-Bicocca, Milan, Italy

[9]Biostatistics and Clinical Epidemiology, Fondazione IRCCS San Gerardo Dei Tintori, Monza, Italy

*These authors contributed equally to this work

#Co-corresponding authors

Co-corresponding authors' information:

Prof. Daniela Besozzi, PhD

University of Milano-Bicocca, Department of Informatics, Systems and Communication, Viale Sarca 336, 20126 Milan, Italy

Email: daniela.besozzi@unimib.it

Tel: +39 02 64487874

Prof. Fabio Pagni, MD

University of Milano-Bicocca, Department of Medicine and Surgery, Pathology, IRCCS Fondazione San Gerardo dei Tintori, Via Cadore 48, 20900 Monza, Italy

Email: fabio.pagni@unimib.it

Tel: +39 338 1833651

Conflict of interest

The authors declare there are no competing financial interests in relation to the work described.

Abstract

The diagnostic assessment of thyroid nodules is hampered by the persistence of uncertainty in borderline cases, and further complicated by the inclusion of non-invasive follicular tumor with papillary-like nuclear features (NIFTP) as a less aggressive alternative to papillary thyroid carcinoma (PTC). In this setting, computational methods might facilitate the diagnostic process by unmasking key nuclear characteristics of NIFTPs. The main aims of this work were to (1) identify morphometric features of NIFTP and PTC that are interpretable for the human eye, and (2) develop a deep learning model for multi-class segmentation as a support tool to reduce diagnostic variability. Our findings confirmed that nuclei in NIFTP and PTC share multiple characteristics, setting them apart from hyperplastic nodules (HP). The morphometric analysis identified 15 features that can be translated into nuclear alterations readily understandable by pathologists, such as a remarkable inter-nuclear homogeneity for HP in contrast to a major complexity in the chromatin texture of NIFTP, and to the peculiar pattern of nuclear texture variability of PTC. A few NIFTP cases with available NGS data were also analyzed to initially explore the impact of RAS-related mutations on nuclear morphometry. Finally, a pixel-based deep learning model was trained and tested on whole slide images (WSIs) of NIFTP, PTC, and HP cases. The model, named NUTSHELL (NUclei from Thyroid tumors Segmentation to Highlight Encapsulated Low-malignant Lesions), successfully detected and classified the majority of nuclei in all WSIs' tiles, showing comparable results with already well-established pathology nuclear scores. NUTSHELL provides an immediate overview of NIFTP areas and can be used to detect microfoci of PTC within extensive glandular samples or identify lymph node metastases. NUTSHELL can be run inside WSInfer with an easy rendering in QuPath, thus facilitating the democratization of digital pathology.

Introduction

In the last years, the evolving classification of thyroid tumors has made their cytological and histological assessment more complex, likewise affecting the clinical management of patients, in particular after the recent inclusion of the non-invasive follicular tumor with papillary-like nuclear features (NIFTP) as a less aggressive non-invasive alternative to papillary thyroid carcinoma (PTC)[1–3]. Due to the numerous diagnostic nuclear features that they share, NIFTP and PTC represent a true challenge for pathologists, especially in the preoperative phase[4]. To address at least partly this challenge, ancillary molecular tools were proposed[5] but the diagnostic gap of indeterminate cases is still to be filled[6]. Different nuclear features (size/shape, membrane irregularities and chromatin characteristics) were recognized as components of a nuclear scoring system to distinguish NIFTPs from other follicular patterned thyroid lesions[7]. However, the nuclear score proved to be useful in distinguishing NIFTPs from benign hyperplastic nodules (HP), while having only a limited impact on the NIFTP vs PTC discrimination[8], and showing a certain degree of interobserver variability[9]. In this setting, a better comprehension of the histological features of NIFTPs at the nuclear level and the detection of useful morphometric features associated with the RAS fingerprint[10,11] may potentially support the diagnostic phase. The application of artificial intelligence (AI) approaches to this complex context can overcome the intrinsic limits of the human eye in discriminating nuclei from NIFTP and PTC cases, possibly unveiling human interpretable features and streamlining the pathologists' workflow through the development of attention maps able to point out even small foci of PTC within larger HP/NIFTP nodules.

Previous experiences demonstrated the feasibility of image analysis and AI as integrative and complementary methods to traditional histopathology in distinguishing malignant vs benign thyroid lesions[12,13], with interesting results on histological samples[14,15]. Although promising, these reports still fail to differentiate specific subtypes of thyroid tumors[16]. More generally, a number of deep learning (DL) models were shown to facilitate the analysis of microscopic images in different clinical contexts, despite most models still suffer from reproducibility, reusability, robustness, and replicability problems[17]. In the attempt to address this issue, Kaczmarzyk *et al.* developed WSInfer[18], an open-source collection of software tools that is expected to simplify the sharing and reuse of DL models in digital pathology.

It is from this perspective that, here, we aim at automatically extracting useful nuclear features from HP, NIFTP and PTC lesions and develop a DL tool embedded within WSInfer to reduce the diagnostic variability and offer crucial support to pathologists in identifying these cases. Simultaneously, we aim at providing an in-depth understanding of the nuclear characteristics of NIFTP and explore its heterogeneity with a special focus on molecular features, particularly in relation to RAS-mutated neoplasms.

Materials and Methods

Cases

The enrollment of cases for the present study involved patients from the Ricerca Finalizzata GR-2019-12368592 trial, diagnosed from January 2020 to August 2023. For each case, anagraphic data (i.e. gender and age) and pathological information (i.e. preoperative cytological class, nodule size, surgery type, final diagnosis, histological subtype, invasiveness, oncocytic and high-grade features of PTCs, molecular status) were extracted. Only cases with histological diagnosis of HP, NIFTP or PTC were included in the study, and no additional histotypes (e.g. follicular adenoma/carcinoma) were part of the analysis.

A representative hematoxylin and eosin (H&E) slide of the lesion was selected, retrieved from the pathology departments archives, and anonymized. Only original slides used to render the primary diagnosis were considered, no recuts were employed for the study. Histological slides cut at 2-3 μm thickness and mounted with a film coverslip (Tissue-Tek coverslipping film, Sakura Finetek, Nagano, Japan) were scanned using MIDI II or P1000 platforms (3DHISTECH, Budapest, Hungary) using a magnification of x58 with resolution of 0.1725 µm/pixel and .svs or .mrxs file format, and magnification of x20 with resolution of 0.2426 µm/pixel and .mrxs file format, respectively, to generate whole slide images (WSIs). A total of 55 WSIs from 53 patients were collected. The diagnoses were based on the morphological, immunohistochemical and molecular criteria of the WHO classification[19]. The obtained WSIs were re-evaluated by an expert thyroid pathologist (FP) for the assignment of the nuclear score, as previously defined[8], whose interpretation scheme with iconographic examples is reported in Supplementary Figure 1. For immunohistochemistry (IHC), NRAS Q61R was tested (rabbit monoclonal antibody, clone RST-NRAS, dilution 1:20) on DAKO Omnis (Agilent, Santa Clara, CA, US). Representative formalin-fixed, paraffin embedded (FFPE) tissue blocks were used for next generation sequencing (NGS) analysis (Supplementary Methods). The IHC and NGS analysis were performed on purpose for the present study, to unveil possible correlations with the morphometric analysis results obtained by machine learning (ML) models.

Nuclei segmentation and features extraction

The WSIs were processed with QuPath software to annotate regions of interest (ROIs) by the pathologist, that is, areas within tissue samples containing pathological features[20]. The StarDist extension was utilized to segment nuclei within the ROIs[21], with optimized parameter settings applied by modifying the standard parameters threshold (set at 0.2) and by introducing the intensity and shape features calculation using the model for H&E (he_heavy_augment). An object classifier was then trained for each WSI through annotations and applied to the nuclei to categorize them into one of four distinct classes: “HP”, “NIFTP”, “PTC”, and “Other” (encompassing any nucleus of non-thyroid relevance, such as inflammatory cells, endothelium, ecc.). In order to validate the accuracy of this classification process, the results underwent rigorous examination by two experienced thyroid pathologists (FP and VL).

Morphometric analysis

We used the QuPath software to build a dataset of morphometric features[22] calculated from each classified nuclei appearing in the 55 WSIs. The dataset included shape features (e.g. area, length, circularity, solidity, maximum and minimum diameters, perimeter), and intensity features computed on the optical density sum (OD Sum) color transform at the resolution of 1 μm/px (e.g. nuclear OD Sum mean value, Haralick texture features). The full list of morphometric features is given in Supplementary Table 1. Highly correlated features were discarded using an unsupervised feature selection, whereby we computed the Pearson correlation matrix of the morphometric dataset and fed it to a hierarchical clustering algorithm coupled with Ward's minimum variance linkage method[37]. The linkage matrix produced by the clustering algorithm is shown as a dendrogram in Supplementary Figure 2. We identified a maximum intercluster distance threshold to flatten the dendrogram, obtaining as a result an assignment of the highly correlated features to single clusters. One feature per cluster was randomly selected as the representative.

The reduced set of features were subjected to both unsupervised (principal component analysis, PCA) and supervised (random forest, decision tree) ML methods in order to highlight differences among the "HP", "NIFTP" and "PTC" groups, provide insights into the feature importance, and enable the comparison with morphologic parameters readily understandable to the human eye.

RAS mutant vs wild type NIFTP

An additional dataset containing only NIFTP nuclei was derived from the morphometric dataset, with labels assigned to either known RAS-mutated cases or known WT (wild type) cases. We used both PCA and random forest methods to investigate their capability in readily distinguishing between mutated and non mutated cases, leveraging the same nuclear features used for the morphometric analysis.

Deep learning model

A convolutional neural network (CNN) was used for pixel-based predictions, to identify and classify each cell nucleus in the WSIs as belonging to one of the four considered classes (i.e., “HP”, “NIFTP”, “PTC”, and “Other”). Since the resolution of each WSI is in the order of tens of thousands of pixels and the ROIs size varies among the various images, a tiling procedure was applied on the ROIs to obtain smaller images—of a fixed size—suitable as input for the CNN. Details of the CNN architecture and the tiling procedure are reported in Supplementary Methods. The tiling procedure might generate tiles that are either fully contained or partially contained within a ROI. To account for the sampling distance heterogeneity of the WSIs generated by different scanners, all tiles were downsampled to reach the same pixel size of 0.25 μm/pixel. The set of fully annotated tiles were split into the training and test sets with a 80-20 policy, that is, 80% of the tiles were assigned to the training set and the remaining 20% were assigned to the test set (named “dense test set”). The partially annotated tiles were set aside—as they could deceive the training procedure—and included in another test set (named “partial test set”) to further assess the generalization capabilities of the CNN. The performance of the CNN was first assessed by means of a 5-fold cross-validation (CV) process performed on the training set. The creation of the five folds was done in a stratified fashion in order to keep, within each fold, similar proportions between tiles that present a majority of nuclei belonging to each of the four classes. During the training phase, online data augmentation—consisting in possible mirroring along all axes and random rotations by 90°, 180° or 270°—was applied to increase variability of the image source. The CNN was then trained using the training set according to a 90-10 split policy, and leveraging the same online data augmentation used for CV. The final model was used for nuclei classification on both the dense and the partial test sets. The whole pipeline—from tile generation to final prediction—is shown in Figure 1. Details of the training procedure and performance metrics are provided in Supplementary Methods. Based on the results obtained, a deep learning model named NUTSHELL (NUclei from Thyroid tumors Segmentation to Highlight Encapsulated Low-malignant Lesions) was developed. To compare the final results of NUTSHELL with the nuclear score assigned by the thyroid pathologist, the WSI-level prediction of the CNN model was inferred based on the prevalence of nuclei detected as either HP or NIFTP/PTC for each case. The comparison of WSI-level CNN prediction and pathologist-derived nuclear score was performed through Cohen's kappa measure.

Code availability

NUTSHELL was implemented in Python using the PyTorch 2.2 framework. The source code of NUTSHELL, together with the extension of WSInfer[18], are available under Academic Free License v. 3.0 in the GitHub repository at the following URL: https://github.com/Vsc0/nutshell. The repository also contains an extension of WSInfer to allow the visualization of the predictions made by pixel-based classification models and its easy rendering in QuPath.

Results

Cases

The dataset included 19 HPs (35%), 16 NIFTPs (29%), and 20 PTCs (36%), as indicated in Table 1. In two PTC cases, both the primary site and the respective lymph node metastasis were examined (cases MI_007 and MI_010, see Table 2 for all clinico-pathological characteristics of the PTC series).

Morphometric analysis

A total of 1,002,864 nuclei were extracted and, after the described cleaning phase, the dataset included 318,471 HP nuclei (37%), 248,967 NIFTP nuclei (29%) and 296,211 PTC nuclei (34%). The unsupervised feature selection resulted in a reduction from 21 to 15 features. The 15 features were used to feed the random forest model for the classification of nuclei into the three classes “HP”, “NIFTP” and “PTC”, obtaining an overall accuracy, precision, and recall of 0.70 and F1 score of 0.69 on the test set. The ranking of the most important features obtained from this model is shown in Figure 2a. The application of PCA confirmed the differential distribution of the features among the different diagnostic categories (Figure 2b).

Human interpretable features (HIFs)

A decision tree was applied on the 15 extracted features as an interpretability effort to investigate human interpretable features (HIFs). To help explainability, these features were grouped based on their impact on three different nuclear domains: shape, clarification, and texture (Figure 3). For nuclear shape, HP nuclei were overall smaller but shared similar regular contours with NIFTPs, whose nuclei were on average bigger, rounder and more regular than those of PTC (Figure 3, first row). A surrogate marker for nuclear clarifications, the *“OD sum means”*, demonstrated comparable whitening for NIFTP and PTC nuclei, which were clearer than HP nuclei on average (Figure 3, middle row). The nuclear chromatin texture analysis revealed remarkable inter-nuclear homogeneity of HP cases, even with irregular distribution of chromatin in a finely granular pattern, as opposed to NIFTP nuclei showing higher chromatin texture complexity (*"Measure of Correlation, F11"*) with brightness variation within the same nuclei (*"Inverse Difference Moment, F4"*) (Figure 3, last row). PTC showed intermediate texture characteristics between the other two classes (*"Difference Entropy, F10"*). The resulting decision tree is shown in Supplementary Figure 3.

RAS mutant vs wild-type NIFTP

A subset of 10 NIFTP cases from the initial cohort had available results from genetic testing, allowing a subanalysis on the potential impact of RAS-related mutations on the nuclear morphometry. Of these, 6 were RAS-mutated (4 *NRAS Q16R*, 1 *HRAS A59T,* and 1 *HRAS Q61K*) and 4 were WT (Table 3). The total number of RAS-mutated nuclei was 113,720, while the total number of RAS-WT nuclei was 89,770. An initial analysis through random forest showed good discrimination capabilities of the nuclear features among RAS mutant and WT cases, with an accuracy, precision, recall, and F1 score of 0.72. However, the application of PCA showed a lack of segregation between the two groups of RAS-mutant and RAS-WT nuclei (Supplementary Figure 4), likely due to the batch effects from staining and scanning procedures. These issues should be carefully addressed in future analyses with a larger cohort.

Multi-class segmentation by NUTSHELL

The CNN model was developed and tested as a diagnostic aid for the differentiation of the three classes of this study, further complementing the ML and morphometric approaches. A total of 14,150 tiles were generated from 55 WSIs (average 257.27 ± 256.33 standard deviation, all details on the number of tiles extracted per case/WSI are given in Supplementary Table 2), 12,180 of which were used to generate the training set (9,744 tiles) and the dense test set (2,436 tiles), maintaining a balance among tiles extracted for each class (Supplementary Table 3).

Approximately 14% of the tiles partially exceeded the ROI boundaries and were considered to form the partial test set (1,970 tiles). In all CV iterations, the CNN successfully achieved a score of Intersection over Union (IoU) around 0.91 computed on the validation set (Supplementary Figure 6, left side); the training phases terminated in about 60 epochs, requiring approximately 3 hours on the machine leveraged for this study (Supplementary Methods). Finally, the CNN was trained and validated, achieving a loss function value of approximately 0.13 and a score of IoU around 0.92 (Supplementary Figure 6, right side). The final model was tested on both the dense and partial test sets (Figures 4 and 5), achieving an IoU value of 0.92 and of 0.91, respectively. The comparative analysis of the performance obtained on the tiles extracted from the different scanners did not show significant differences in both the dense test set (IoU of 0.913 on the 352 tiles from MIDI II vs 0.927 on the 2,084 tiles from P1000) and the partial test set (IoU of 0.892 on the 366 tiles from MIDI II vs. 0.911 on the 1,604 tiles from P1000). The confusion matrices evaluated on both the dense and partial test sets are shown in Supplementary Figure 7. The comparison of the results obtained with the nuclear classification using NUTSHELL and the nuclear score given by the pathologist at WSI-level (Supplementary Table 4) demonstrated a good agreement between the two methods—kappa = 0.67 (95% CI, 0.48 to 0.86, p<0.0001)—as shown in Table 4.

Discussion

Discriminating NIFTP lesions from mimickers can be troublesome and challenging, due to the significant overlap of nuclear features with the more aggressive PTC, requiring histology and occasionally ancillary tools to discriminate among these two similar lesions[23]. Although specific pathological aspects are known for being more PTC-correlated (e.g. over clearing, grooves and pseudoinclusions, psammoma bodies and papillae)[24], these are only variably present. The employment of a nuclear scoring system proposed for the assessment of NIFTPs only partially overcame the intrinsic limitations[8]. In this direction, the recent application of high-resolution 3D-structured illumination microscopy revealed fascinating nuclear characteristics of NIFTP, ranging from densely packed DNA and narrower interchromatin spaces[25], as compared to the more striking nuclear pseudoinclusions, marginal micronucleoli, irregular branching sheets, and linear arrangement described for PTC cells, confirming the existence of a certain variability between the two entities[26]. These discoveries, while limited by difficult reproducibility and validation in routine practice, hint at the possibility of exploring new morphometric pathways. The description of morphometric features that can eventually be HIFs, as those obtained through ML in this study, could partly fill the gap in the differential diagnosis of thyroid lesions. The findings from our study confirm that nuclei in NIFTP and PTC share multiple characteristics, setting them apart from HP in various aspects. Despite noticeable overlap, certain subtle features might indicate affiliation with neoplasms displaying either a more indolent or aggressive course. Specifically, PTC nuclei tend to be slightly clearer, while NIFTP nuclei are generally larger, rounder, and less elongated. Moreover, differences in chromatin distribution, though challenging to describe in easily perceptible terms, further distinguish NIFTP nuclei from hyperplastic counterparts.

The introduction of ancillary molecular techniques allowed a morpho-molecular clustering of thyroid lesions, with the so-called *BRAFV600E*-like tumors belonging to the PTC family and RAS-like tumors more on the follicular and NIFTP side[27,28]. This association has already been employed to develop AI models (e.g., conditional generative adversarial networks) able to generate synthetic images depicting the BRAF/RAS-associate morphology spectrum, encompassing nuclear (enlargement, chromatin clearing, membrane irregularities), architectural (elongated follicles, papillae), colloid (darkening, scalloping), and stromal changes (fibrosis, calcification, ossification)[29]. Here, the application of supervised ML approaches demonstrated a moderate ability to distinguish between RAS-mutated and WT nuclei. Despite these intriguing preliminary results, they were not corroborated by unsupervised ML approaches. This discrepancy could stem from the relatively low number of available NIFTP cases, potentially leading to the presence of batch effects within the dataset, or from subtle differences in nuclear morphology between RAS-mutant and RAS-WT NIFTP, suggesting the need for additional investigation of this intriguing aspect on larger cohorts characterized by a higher heterogeneity in terms of pre-analytical and scanning procedures of the slides.

Even if the ML-based morphometric approach to NIFTP vs PTC classification gave promising results, the application of DL for this task is also stimulating, as shown by the few reports available in this setting. AI-based nuclear texture analysis, CNN models, and artificial neural networks improved the accuracy of fine-needle aspiration (FNA) in distinguishing between benign and malignant nodules[14,30,31]. ML methods achieved high concordance with expert cytopathologists, suggesting their potential to reduce the workload and improve the diagnostic accuracy[32,33]. On the histological side, algorithms were also developed to detect tall cells in PTC, a marker of aggressive behavior[15], and DL algorithms detected lymph node metastasis in thyroid cancer patients[34]. Here, the CNN model successfully detected and classified the majority of nuclei in all tiles, with only a few nuclear misclassifications in a small subset of tiles not affecting the proper diagnosis of the WSI's sample. NUTSHELL is also able to generalize the class prediction beyond the ROI boundaries, that is, it detects and classifies nuclei that appear in a tile but are not—or only partially—included in a ROI. Moreover, a good correlation was found between the WSI-level pathologist assessment of nuclear score and the CNN prediction, which further confirms the genuinity of the detections.

The implications of applying an AI tool on WSIs, freely accessible and compatible with QuPath, have numerous practical outcomes. Primarily, it offers a morphological indication to complement the stringent and detailed features defining NIFTP, including criteria like non-invasion of the capsule and absence of papillae[35]. NUTSHELL provides an immediate overview of NIFTP areas within the WSI, simplifying the identification of capsule infiltration zones. Furthermore, the tool's capability to incorporate nuclei from PTC and HP expands its applications. It can be used to detect microfoci of PTC within extensive glandular samples or identify lymph node metastases, potentially prioritizing the oncological cases during the pathologist screening. We also extended WSInfer—a pipeline to run patch-based classification models[18]—in order to easily and quickly render NUTSHELL pixel-based predictions in QuPath. This extension paves the way to the adoption of our DL model—and other pixel-based models—inside WSInfer, helping in the democratization of digital pathology among users, and allowing the use of NUTSHELL directly on the whole WSIs instead of using single tiles of the tissue. Although promising, the application of the DL model on the differentiation of thyroid lesions still has some limitations. The creation of a multicentric cohort of cases and the employment of different scanners helped in obtaining a heterogeneous dataset, but the full generalizability of tools of this kind requires a further validation on external cohorts to tackle any additional variability stemming from the pre-analytical phase (e.g., section thickness, coverslipping, H&E stain) and scanning procedures. Moreover, the current study was mainly focused on the differentiation of HP vs NIFTP vs PTC, and perspective efforts are needed to include the capability of other discriminating follicular patterned neoplasms (e.g., adenoma/carcinoma) from possible mimickers as NIFTPs. In this direction, the development of cytology-based DL-assisted classification systems—also with the aid of recently described methods, such as multiple instance learning[36], and the implementation of more traditional ML tools to provide useful HIFs—can represent the evolution of existing computational pathology instruments serving as computer-aided diagnostics.

In this work, we proposed a user-friendly digital pathology pipeline to perform multi-class segmentation of thyroid nuclei. We are working on the development of another computational pipeline able to automatically generate the most suitable set of filters for the standardization of WSIs; by doing so, we will simplify the classification task when using pre-trained models. We envision that this additional pipeline will partially solve the lack of standards in the generation of WSIs currently existing across different centers, laboratories, scanners, and cohorts. Moreover, we aim to make both NUTSHELL and the filters generator natively executable in QuPath to facilitate their direct use by pathologists in clinical practice.

Acknowledgements

This research was partially supported by Ricerca Finalizzata GR-2019-12368592 and by the Italian Ministry of the University MUR Dipartimenti di Eccellenza 2023-2027 (IMPACT MEDICINE, l. 232/2016, art. 1, commi 314 - 337). DB acknowledges the partial support by the MUR under the grant “Dipartimenti di Eccellenza 2023-2027” of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

Conflict of interest

The authors declare there are no competing financial interests in relation to the work described.

Ethical approval

The study was conducted in accordance with the Declaration of Helsinki and approved by the local ethical committee (GR-2019-12368592).

Author contribution

FP and VL defined the study design; GC and FDC performed the retrospective research and data extraction; NF, MI and AG contributed to the enrollment of cases as part of the multicentric institution network; JC provided assistance in glass slide scanning; GC, FDC, GCap, MM and SG performed the statistical analysis and application of ML approach to the morphometric data; VC, DMP, MSN and DB performed the artificial intelligence and computational analysis; DMP, VL, SG, DB and MSN performed the supervision of the work revising critically the manuscript before the approval by all the authors; DB, SG, FP and VL provided the funding acquisition and administrative support. All authors were involved in writing the paper and had final approval of the submitted and published versions.

Funding

This work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors’ views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Haugen, B. R., Alexander, E. K., Bible, K. C., Doherty, G. M., Mandel, S. J., Nikiforov, Y. E. et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* **26**, 1–133 (2016).
2. Thyroid Cancer - Cancer Stat Facts. *SEER Program:* https://seer.cancer.gov/statfacts/html/thyro.html.
3. Elbalka, S. S., Metwally, I. H., Shetiwy, M., Awny, S., Hamdy, O., Kotb, S. Z. et al. Prevalence and predictors of thyroid cancer among thyroid nodules: a retrospective cohort study of 1,000 patients. *Ann. R. Coll. Surg. Engl.* **103**, 683–689 (2021).
4. Haugen, B. R., Sawka, A. M., Alexander, E. K., Bible, K. C., Caturegli, P., Doherty, G. M. et al. American Thyroid Association Guidelines on the Management of Thyroid Nodules and Differentiated Thyroid Cancer Task Force Review and Recommendation on the Proposed Renaming of Encapsulated Follicular Variant Papillary Thyroid Carcinoma Without Invasion to Noninvasive Follicular Thyroid Neoplasm with Papillary-Like Nuclear Features. *Thyroid* **27**, 481–483 (2017).
5. Seminati, D., Capitoli, G., Leni, D., Fior, D., Vacirca, F., Di Bella, C. et al. Use of Diagnostic Criteria from ACR and EU-TIRADS Systems to Improve the Performance of Cytology in Thyroid Nodule Triage. *Cancers* **13**, (2021).
6. Paja, M., Zafón, C., Iglesias, C., Ugalde, A., Cameselle-Teijeiro, J. M., Rodríguez-Carnero, G. et al. Rate of non-invasive follicular thyroid neoplasms with papillary-like nuclear features depends on pathologist’s criteria: a multicentre retrospective Southern European study with prolonged follow-up. *Endocrine* **73**, 131–140 (2021).
7. Seethala, R. R., Baloch, Z. W., Barletta, J. A., Khanafshar, E., Mete, O., Sadow, P. M. et

al. Noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a review for pathologists. *Mod. Pathol.* **31**, 39–55 (2018).

8. Nikiforov, Y. E., Seethala, R. R., Tallini, G., Baloch, Z. W., Basolo, F., Thompson, L. D. R. et al. Nomenclature Revision for Encapsulated Follicular Variant of Papillary Thyroid Carcinoma: A Paradigm Shift to Reduce Overtreatment of Indolent Tumors. *JAMA Oncol* **2**, 1023–1029 (2016).
9. Thompson, L. D. R., Poller, D. N., Kakudo, K., Burchette, R., Nikiforov, Y. E. & Seethala, R. R. An International Interobserver Variability Reporting of the Nuclear Scoring Criteria to Diagnose Noninvasive Follicular Thyroid Neoplasm with Papillary-Like Nuclear Features: a Validation Study. *Endocr. Pathol.* **29**, 242–249 (2018).
10. Hung, Y. P. & Barletta, J. A. A user’s guide to non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP). *Histopathology* **72**, 53–69 (2018).
11. Rangel-Pozzo, A., Sisdelli, L., Cordioli, M. I. V., Vaisman, F., Caria, P., Mai, S. et al. Genetic Landscape of Papillary Thyroid Carcinoma and Nuclear Architecture: An Overview Comparing Pediatric and Adult Populations. *Cancers* **12**, (2020).
12. Tizhoosh, H. R., Diamandis, P., Campbell, C. J. V., Safarpoor, A., Kalra, S., Maleki, D. et al. Searching Images for Consensus: Can AI Remove Observer Variability in Pathology? *Am. J. Pathol.* **191**, 1702–1708 (2021).
13. Wong, C. M., Kezlarian, B. E. & Lin, O. Current status of machine learning in thyroid cytopathology. *J. Pathol. Inform.* **14**, 100309 (2023).
14. Alabrak, M. M. A., Megahed, M., Alkhouly, A. A., Mohammed, A., Elfandy, H., Tahoun, N. et al. Artificial Intelligence Role in Subclassifying Cytology of Thyroid Follicular Neoplasm. *Asian Pac. J. Cancer Prev.* **24**, 1379–1387 (2023).
15. Stenman, S., Linder, N., Lundin, M., Haglund, C., Arola, J. & Lundin, J. A deep learning-based algorithm for tall cell detection in papillary thyroid carcinoma. *PLoS One*

17, e0272696 (2022).

16. Ludwig, M., Ludwig, B., Mikuła, A., Biernat, S., Rudnicki, J. & Kaliszewski, K. The Use of Artificial Intelligence in the Diagnosis and Classification of Thyroid Nodules: An Update. *Cancers* **15**, (2023).
17. Wagner, S. J., Matek, C., Shetab Boushehri, S., Boxberg, M., Lamm, L., Sadafi, A. et al. Make deep learning algorithms in computational pathology more reproducible and reusable. *Nat. Med.* **28**, 1744–1746 (2022).
18. Kaczmarzyk, J. R., O'Callaghan, A., Inglis, F., Gat, S., Kurc, T., Gupta, R. et al. Open and reusable deep learning for pathology with WSInfer and QuPath. *NPJ Precis Oncol* **8**, 9 (2024).
19. Rindi, G., Mete, O., Uccella, S., Basturk, O., La Rosa, S., Brosens, L. A. A. et al. Overview of the 2022 WHO Classification of Neuroendocrine Neoplasms. *Endocr. Pathol.* **33**, 115–154 (2022).
20. Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D. et al. QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
21. Schmidt, U., Weigert, M., Broaddus, C. & Myers, G. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 265–273 (Springer International Publishing, 2018).
22. Haralick, R. M., Shanmugam, K., & Dinstein, I. H. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **6**, 610-621 (1973).
23. Layfield, L. J., Baloch, Z. W., Esebua, M., Kannuswamy, R. & Schmidt, R. L. Impact of the Reclassification of the Non-Invasive Follicular Variant of Papillary Carcinoma as Benign on the Malignancy Risk of the Bethesda System for Reporting Thyroid Cytopathology: A Meta-Analysis Study. *Acta Cytol.* **61**, 187–193 (2017).

24. Pusztaszeri, M. & Bongiovanni, M. The impact of non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) on the diagnosis of thyroid nodules. *Gland Surg* **8**, S86–S97 (2019).

25. Rangel-Pozzo, A., Dos Santos, F. F., Dettori, T., Giulietti, M., Frau, D. V., Galante, P. A. F. et al. Three-dimensional nuclear architecture distinguishes thyroid cancer histotypes. *Int. J. Cancer* **153**, 1842–1853 (2023).

26. Legesse, T., Parker, L., Heath, J. & Staats, P. N. Distinguishing non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) from classic and invasive follicular-variant papillary thyroid carcinomas based on cytologic features. *J Am Soc Cytopathol* **8**, 11–17 (2019).

27. Sohn, S. Y., Lee, J.-J. & Lee, J. H. Molecular Profile and Clinicopathologic Features of Follicular Variant Papillary Thyroid Carcinoma. *Pathol. Oncol. Res.* **26**, 927–936 (2020).

28. Melo, M., Ventura, M., Cardoso, L., Gaspar da Rocha, A., Paiva, I., Sobrinho-Simões, M. et al. Non-invasive follicular thyroid neoplasm with papillary-like nuclear feature: clinical, pathological, and molecular update 5 years after the nomenclature revision. *Eur. J. Endocrinol.* **188**, (2023).

29. Dolezal, J. M., Wolk, R., Hieromnimon, H. M., Howard, F. M., Srisuwananukorn, A., Karpeyev, D. et al. Deep learning generates synthetic cancer histology for explainability and education. *NPJ Precis Oncol* **7**, 49 (2023).

30. Sanyal, P., Mukherjee, T., Barui, S., Das, A. & Gangopadhyay, P. Artificial Intelligence in Cytopathology: A Neural Network to Identify Papillary Carcinoma on Thyroid Fine-Needle Aspiration Cytology Smears. *J. Pathol. Inform.* **9**, 43 (2018).

31. Hirokawa, M., Niioka, H., Suzuki, A., Abe, M., Arai, Y., Nagahara, H. et al. Application of deep learning as an ancillary diagnostic tool for thyroid FNA cytology. *Cancer*

Cytopathol. **131**, 217–225 (2023).

32. Dov, D., Kovalsky, S., Cohen, J., Range, D., Henao, R. & Carin, L. Thyroid Cancer Malignancy Prediction From Whole Slide Cytopathology Images. *arXiv [cs.CV]* (2019).

33. Dov, D., Kovalsky, S. Z., Feng, Q., Assaad, S., Cohen, J., Bell, J. et al. Use of Machine Learning-Based Software for the Screening of Thyroid Cytopathology Whole Slide Images. *Arch. Pathol. Lab. Med.* **146**, 872–878 (2022).

34. Wu, X., Li, M., Cui, X.-W. & Xu, G. Deep multimodal learning for lymph node metastasis prediction of primary thyroid cancer. *Phys. Med. Biol.* **67**, (2022).

35. Paniza, A. C. de J., Mendes, T. B., Viana, M. D. B., Thomaz, D. M. D., Chiappini, P. B. O., Colozza-Gama, G. A. et al. Revised criteria for diagnosis of NIFTP reveals a better correlation with tumor biological behavior. *Endocr Connect* **8**, 1529–1538 (2019).

36. Gadermayr, M. & Tschuchnig, M. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Comput. Med. Imaging Graph.* **112**, 102337 (2024).

37. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58**, 236–244 (1963).

Figure 1: The NUTSHELL pipeline. The WSIs were processed to generate tiles of the same size, that is, 512 x 512 pixels, corresponding to 128 μm^2 of sample tissue (blue boxes). Partially annotated tiles were set aside for model testing (partial test set). Fully annotated tiles were used for 5-fold Cross Validation (pink box) as well as for CNN training and validation (yellow box); data augmentation was exploited in both cases. Dashed lines depict the partitioning of the tiles into training and validation sets according to the 80-20 and 90-10 policies. The final CNN model (green box) was used to classify the tiles belonging to both the dense and the partial test sets.

Figure 2: In panel (a), the feature ranking shows the impact of morphometric features obtained from the random forest model. In panel (b), the result of PCA shows the differential distribution of the morphometric features among the HP, NIFTP, and PTC classes.

Figure 3: Morphometric characteristics of HP, NIFTP, and PTC nuclei. In the top row: feature crossing of the three shape features, "Circularity", "Area", and "Length". PTC nuclei showed higher "length/area ratio" and lower "circularity" (how closely a shape resembles a perfect circle) as compared to NIFTPs. In the middle row: mean value of "OD Sum", used as a surrogate for nucleus clarification. In the bottom row: feature crossing of the three most relevant features of chromatin texture (F5, F0, and F1), with "Angular Second Momentum, F0" contributing to a consistent and regular appearance of HP nuclei, and high values of "Sum Average, F5" to an irregular distribution of chromatin in a finely granular pattern.

Figure 4: Six tiles belonging to the dense test set (first row), their ground truth (second row), and the corresponding prediction computed by the CNN (third row). Nuclei belonging to the PTC class are depicted in red, to the NIFTP class in blue, and to the HP class in green. In the first column a few nuclei of the PTC class (red, white arrowhead) were classified as hyperplastic (green, black arrowhead), while in the fifth column a few nuclei of the PTC class (red, white arrow) were classified as NIFTP (blue, black arrow). The second and third columns show that the NIFTP class (blue nuclei) was perfectly recognized.

Figure 5: Six tiles belonging to the partial test set (first row), their ground truth (second row), and the corresponding prediction computed by the CNN (third row). Nuclei belonging to the PTC class are depicted in red, to the NIFTP class in blue, and to the HP class in green. In these illustrative examples, the CNN correctly identifies and classifies the nuclei that are outside the ROIs.

Table 1: Description of the characteristics of the analyzed cases.

Diagnosis	n	Center	Age (years) (mean ± sd)	Sex		Size (cm) (mean ± sd)	Cytology (Bethesda class)					
				M	F		II	III	IV	V	VI	NA
HP	5	MB	52 ± 15	1	4	3.1 ± 1.6	4	1	-	-	-	-
NIFTP	10	MB	57 ± 11	2	8	2.9 ± 0.3	2	1	3	1	-	3
PTC	4*	MB	46 ± 20	1	3	1.3 ± 0.6	-	-	2	-	2	-
HP	9	BG	59 ± 8	2	7	3.6 ± 1.6	6	2	1	-	-	-
NIFTP	1	BG	64	0	1	4	-	-	1	-	-	-
PTC	10	BG	53 ± 14	4	6	1.4 ± 1.1	-	1	-	1	8	-
HP	5	MI	49 ± 9	1	4	1.4 ± 1	2	3	-	-	-	-
NIFTP	5	MI	63 ± 14	0	5	0.4 ± 0.1	-	1	3	1	-	-
PTC	4*	MI	58 ± 19	1	3	1 ± 0.7	-	-	-	-	4	-

*Cases in which two different slides were used (one from the thyroid lesion and one from the lymph node metastasis). Legend: HP, Hyperplastic; NIFTP, Non Invasive Follicular Thyroid neoplasm with Papillary-like nuclear features; PTC, Papillary Thyroid Carcinoma.

Table 2: Details on the PTC cases of the cohort, including histological subtype and presence of oncocytic variant, invasiveness, and high grade features.

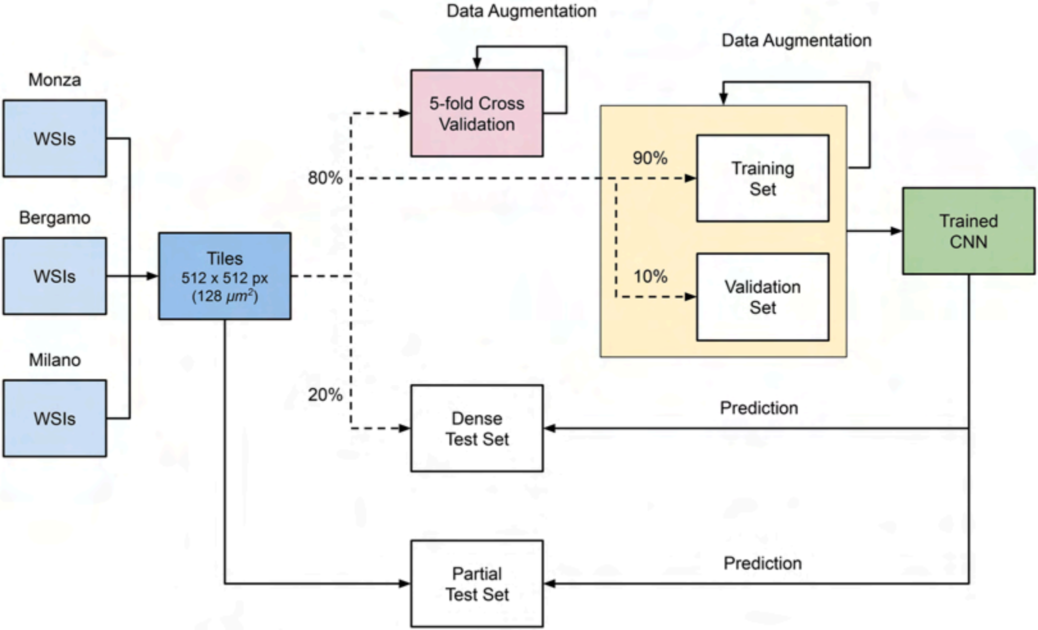
Case ID	Class	Histological subtype	Oncocytic variant (Y/N)	Invasive (Y/N)	High Grade features (Y/N)
BG_001	PTC	Follicular variant	N	Y	N
BG_003	PTC	Conventional	N	Y	N
BG_006	PTC	Conventional	N	Y	N
BG_009	PTC	Follicular variant	N	Y	N
BG_011	PTC	Follicular variant	N	Y	N
BG_012	PTC	Conventional	N	Y	N
BG_015	PTC	Conventional	N	Y	N
BG_016	PTC	Conventional	N	Y	Y
BG_017	PTC	Conventional	N	Y	N
BG_018	PTC	Follicular variant	N	Y	N
MI_002	PTC	Conventional	N	Y	N
MI_005	PTC	Follicular variant	N	Y	N
MI_007	PTC	Lymph node met	N	Y	N
MI_010	PTC	Conventional	N	Y	N
MI_011	PTC	Conventional	N	Y	N
MB_012	PTC	Conventional	N	Y	N
MB_014	PTC	Follicular variant	N	Y	N
MB_015	PTC	Follicular variant	N	Y	N
MB_018	PTC	Conventional	N	Y	N
MB_020	PTC	Conventional	N	Y	N

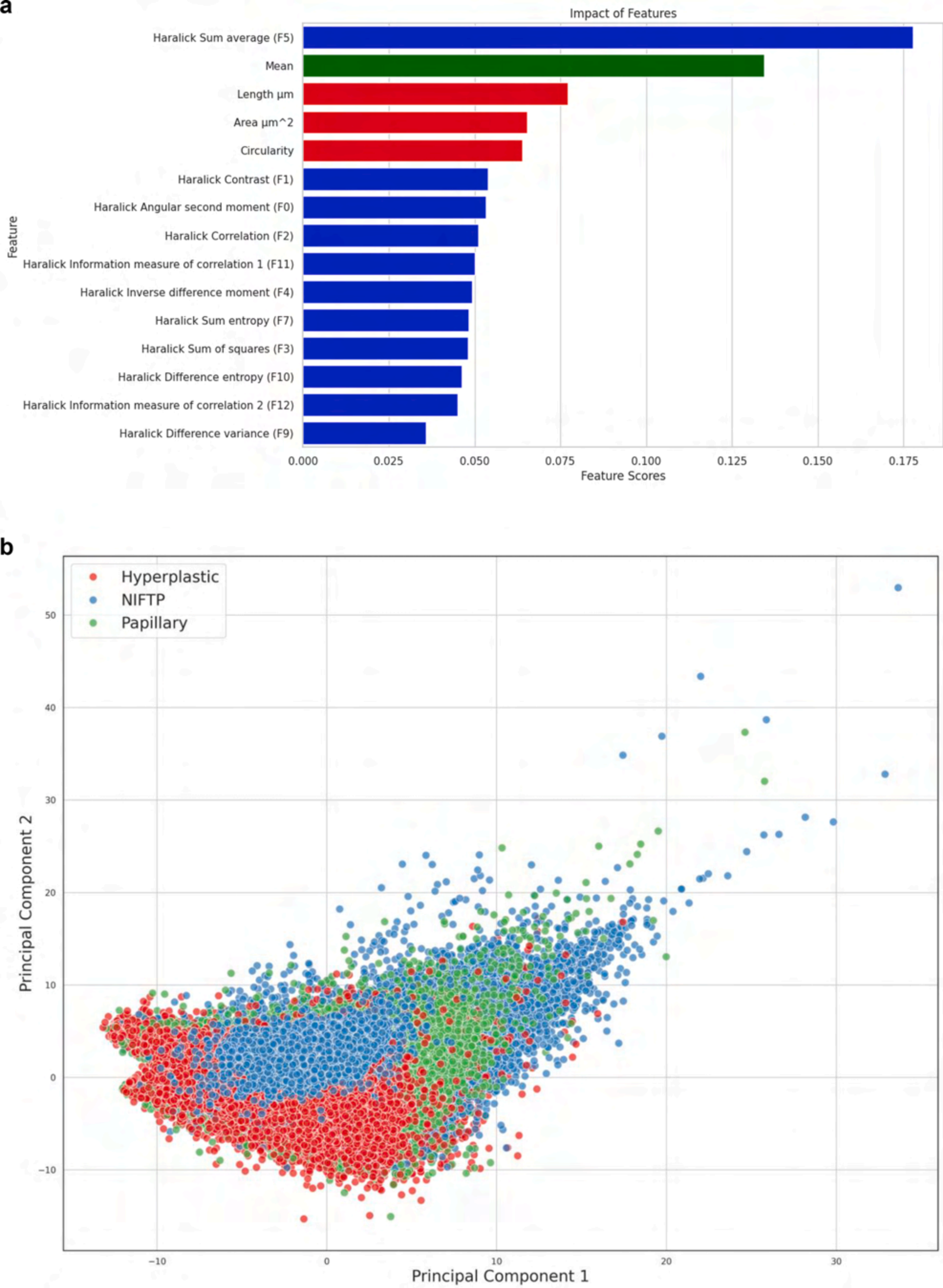
Table 3: Details on the NIFTP cases with available mutational status.

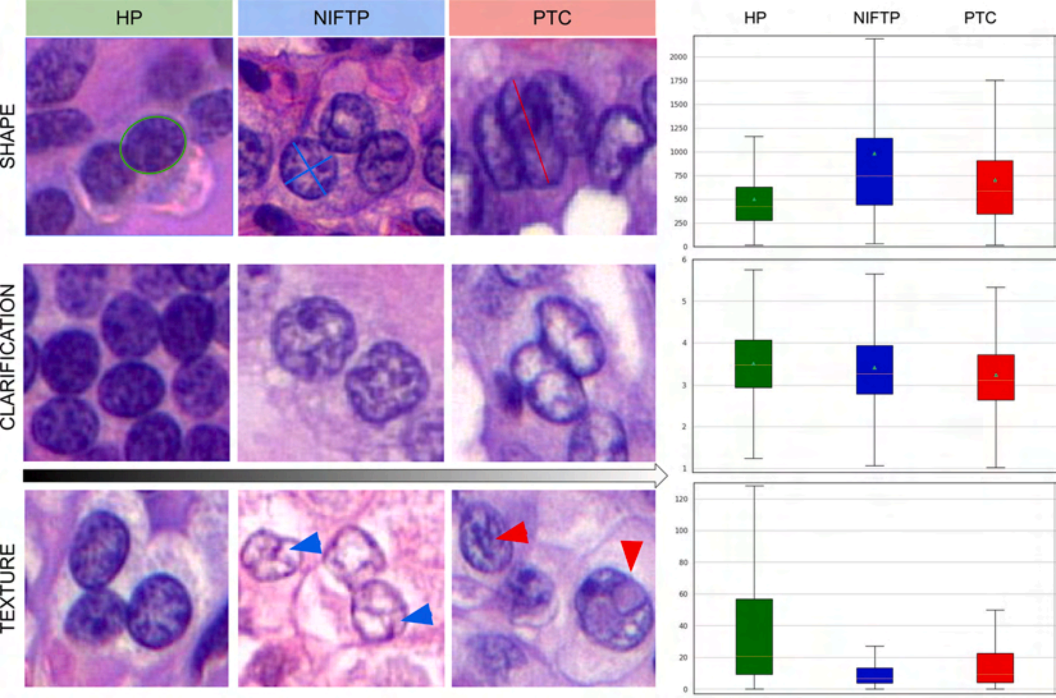
ID	Age (years)	Sex	Size (mm)	Lobe	Mutational status
MB_001	58	F	9	Left	WT
MB_002	56	M	2	Left	HRAS Q61K
MB_003	41	F	5.5	Right	NRAS Q61R
MB_004	56	M	2	Right	WT
MB_005	52	F	0.5	Left	HRAS A59T
MB_006	53	F	1	Right	WT
MB_007	50	F	1.3	Right	NRAS Q61R
MB_008	48	M	3	Right	NRAS Q61R
MB_009	65	F	0.3	Left	NRAS Q61R
MB_010	61	F	1.6	Left	WT

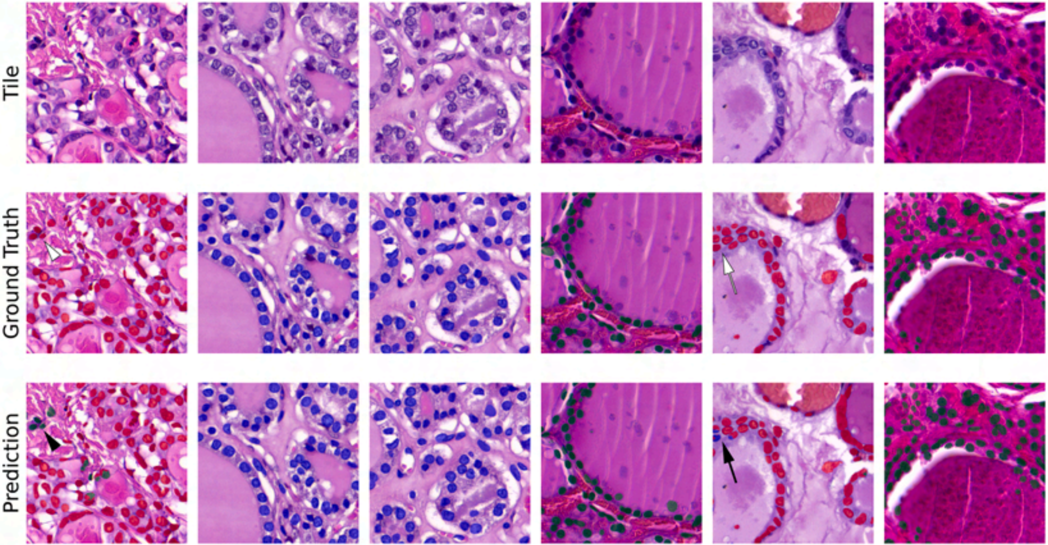
Table 4: Comparison of the disease category assigned to HP and NIFTP/PTC cases based on NUTSHELL with the WSI-level nuclear score assigned by the thyroid pathologist. The number of concordant and discordant cases is reported in the main diagonal and antidiagonal, respectively.

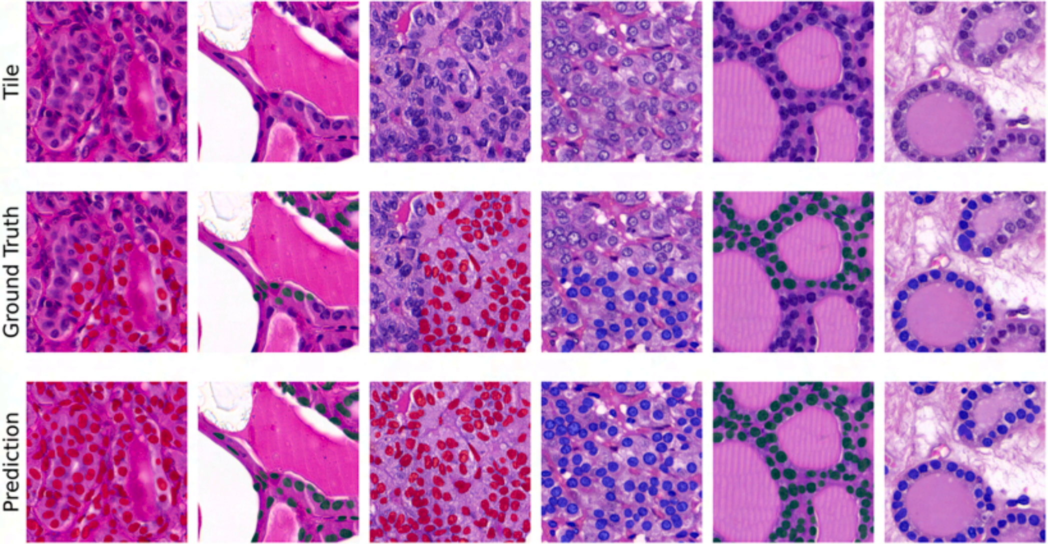
	Nuclear score	
CNN prediction	**score 0-1**	**score 2-3**
HP	18	8
NIFTP/PTC	1	28


Data Augmentation
Data Augmentation
Monza
WSIs
Bergamo
WSIs
Milano
WSIs
Tiles
512 x 512 px
(128 μm^2)
80%
20%
5-fold Cross Validation
90%
Training Set
10%
Validation Set
Trained CNN
Dense Test Set
Prediction
Partial Test Set
Prediction


a
Impact of Features
Feature
Haralick Sum average (F5)
Mean
Length µm
Area µm^2
Circularity
Haralick Contrast (F1)
Haralick Angular second moment (F0)
Haralick Correlation (F2)
Haralick Information measure of correlation 1 (F11)
Haralick Inverse difference moment (F4)
Haralick Sum entropy (F7)
Haralick Sum of squares (F3)
Haralick Difference entropy (F10)
Haralick Information measure of correlation 2 (F12)
Haralick Difference variance (F9)
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature Scores
b
Hyperplastic
NIFTP
Papillary
50
40
30
20
10
0
−10
Principal Component 2
−10
0
10
20
30
Principal Component 1


HP
NIFTP
PTC
SHAPE
CLARIFICATION
TEXTURE
HP
NIFTP
PTC
2000
1750
1500
1250
1000
750
500
250
0
6
5
4
3
2
1
120
100
80
60
40
20
0


Tile
Ground Truth
Prediction


Tile
Ground Truth
Prediction