Acta Cytologica

Review

Acta Cytologica 2021;65:324–329
DOI: 10.1159/000512097

Received: June 18, 2020
Accepted: October 6, 2020
Published online: December 16, 2020

Artificial Intelligence in Thyroid Fine Needle Aspiration Biopsies

Brie Kezlarian Oscar Lin

Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Keywords
Artificial intelligence · Machine learning · Thyroid fine needle aspiration biopsy

Abstract
Background: From cell phones to aerospace, artificial intelligence (AI) has wide-reaching influence in the modern age. In this review, we discuss the application of AI solutions to an equally ubiquitous problem in cytopathology – thyroid fine needle aspiration biopsy (FNAB). Thyroid nodules are common in the general population, and FNAB is the sampling modality of choice. The resulting prevalence in the practicing pathologist's daily workload makes thyroid FNAB an appealing target for the application of AI solutions. ***Summary:*** This review summarizes all available literature on the application of AI to thyroid cytopathology. We follow the evolution from morphometric analysis to convolutional neural networks. We explore the application of AI technology to different questions in thyroid cytopathology, including distinguishing papillary carcinoma from benign, distinguishing follicular adenoma from carcinoma and identifying non-invasive follicular thyroid neoplasm with papillary-like nuclear features by key words and phrases. ***Key Messages:*** The current literature shows promise towards the application of AI technology to thyroid fine needle aspiration biopsy. Much work is needed to define how this powerful technology will be of best use to the future of cytopathology practice.

Introduction

Pathology is at the dawn of a major paradigm shift with increased applications of digital and computational pathology. The standard practice of using a light microscope for diagnosis has remained largely unchanged for over a century. This diagnostic model has been transforming since the advent of digital pathology. Glass slides are being digitized into whole slide images (WSIs) more frequently, a process facilitated by the introduction of high-output digital slide scanners. The digital images created from WSIs have allowed the development of the field of computational pathology which uses the massive amount of data generated to facilitate computer-assisted diagnostics [1, 2]. Recent advances in deep learning [3] in solving image classification tasks, such as classification and categorization, have revolutionized the field. Some high-capacity deep neural network models have been reported to be comparable to human performance [4]. Although most current efforts have centered on the evaluation of histologic sections, it is important to remember that one of the earliest commercial applications of computational pathology was a cytology-centered application for cervico-vaginal cytology specimens (Papnet). Although not commercially successful, many other investigators have evaluated the use of computational pathology tools such as artificial intelligence (AI) in different types of cytology specimens. Among the different types of cytology specimens evaluated, thyroid fine needle aspira-

karger@karger.com
www.karger.com/acy

Oscar Lin
Department of Pathology, Memorial Sloan-Kettering Cancer Center
1275 York Avenue
New York NY 10065 (USA)
lino@mskcc.org

tion biopsy (FNAB) cytology specimens are of particular interest due to their potential clinical impact. Thyroid nodules are reported to occur in 19–68% of randomly selected individuals, with higher frequencies in women and the elderly [5, 6], and this incidence has dramatically increased with the advent of high-resolution ultrasound. The main clinical issue is the need to rule out thyroid cancer in all these thyroid nodules, which occurs in 7–15% of cases depending on age, sex, radiation exposure history, family history, and other factors [7]. To address this issue, FNAB of these thyroid nodules is recommended according to their size and radiological appearance [8].

The sensitivity and specificity of thyroid FNAB are reported to be 68–98% and 56–100%, respectively. However, 15–30% of thyroid FNAB are classified as indeterminate. Approximately 25% of the indeterminate cases will be diagnosed as malignant in surgical resections, suggesting that a large proportion of surgeries are unnecessary [9–12]. Although tests based on molecular profiling of thyroid FNAB have been developed in an attempt to refine the diagnosis among indeterminate thyroid FNAB diagnosis, these molecular tests show limitations and demonstrate variable sensitivity and specificity among the several commercially available tests [13, 14]. AI, including machine learning, has the hypothetical capability to further refine the evaluation of thyroid FNAB specimens. Several investigators have studied the use of AI in thyroid FNAB, and their work is summarized below.

Early Studies: Benign versus Malignant

The literature on AI solutions in thyroid cytopathology predates the recent AI boom by over a decade [15]. In 1996, Karakitsos et al. [16] attempted to classify benign and malignant follicular and Hurtle cell lesions using a neural network. Geometric and densometric features of approximately 100 nuclei from each of the 51 patients' Giemsa-stained smears were measured. Using these measurements, 2 different neural networks were trained and tested. The first attempted to classify the cell into one of the 4 categories: benign follicular cell, benign oncocyte, malignant follicular cell, and malignant oncocyte. This performed reasonably well on the training set data (82.24, 76.28, 95.12, and 68.19% of cells correctly classified for follicular cells, benign oncocytes, malignant follicular cells [i.e., papillary carcinoma], and malignant oncocytes, respectively). The test data were less impressive (47.97, 57.25, 71.32, and 26.86% of cells correctly classified for follicular cells, benign oncocytes, malignant follicular cells, and malignant oncocytes, respectively). The second classifier simply attempted to categorize the nucleus as benign or malignant. This performed much better than the four-class classifier at test time, achieving 91.12 and 89.64% for benign and malignant nuclei, respectively, correctly classified for an overall accuracy of 90.61% on test data.

This same group performed a similar study using a learning vector quantizer neural network [16]. This time, the vectorized input were the mean and standard deviations of the measured morphometric data per case. The four-class classifier results were not impressive. The authors postulated that the classifier had difficulty distinguishing oncocytic from follicular lesions. The two-class classifier was able to distinguish benign from malignant lesions with 94.9% sensitivity and 98.9% specificity.

Later studies by this group abandoned the four-class classifier and focused on distinguishing benign from malignant lesions. In their study published in 2006 [17], the authors used morphometric nuclear data extrapolated from May Grunwald-Giemsa-stained smears. Nuclei were categorized into either benign or malignant using 4 independent classifiers: a linear classifier, a two-layer feed-forward neural network, 3 two-layer feed-forward networks combined by the Adaboost algorithm, and a K-nearest neighbour classifier. If the percentage of benign nuclei in a given case exceeded the average between the maximum percentage of nuclei called benign by the classifier in a malignant case and the minimum percentage of nuclei called benign in a benign case, that case was classified as benign. The most successful classifier on test data was the combination of 3 two-layer feed-forward networks generated by the Adaboost algorithm with 5 nodes, though notably this only outperformed the K-nearest neighbour classifier by 0.56%.

Gopinath et al. [18–20] published several studies with a similar approach to classify thyroid tumours into benign versus malignant lesions. In this case, statistical textural features from images of thyroid lesions were input into one or more classifiers, including decision tree, K-nearest neighbour, Elman neural network, and support vector machine. These studies were problematic, however, as the input data were images taken from the Papanicolaou Society of Cytopathology online atlas. Because the classifier was trained by and tested on textbook images, it is unlikely that the results would be widely applicable to real-world applications. This is one example of overfitting. Overfitting occurs when the AI's concept for what constitutes ground truth is too narrow. The issue of overfitting most commonly arises in the context of small sample size or when the training data have insufficient variation to capture the variation seen in practice.

Convolutional Neural Networks to Determine Benign from Malignant

More recent studies utilize convolutional neural networks (CNNs) for analysis. CNNs are ideal for image analysis because unlike other neural network architectures, CNNs retain spatial information while moving from one layer to the next. Few papers have been written applying CNNs to thyroid cytopathology. All narrowed their focus to distinguishing papillary thyroid carcinomas (PTCs) from non-PTCs.

Sanyal et al. [21] published on CNNs to assess thyroid nodules in 2018. They trained their algorithm on 186 images of PTCs and 184 images of goitre, lymphocytic thyroiditis, and nodules cytologically diagnosed as follicular lesions. These images were taken from Romanowsky-stained smears at ×10 and ×40 objective magnifications and then cropped to 512 × 512 pixel images. The evaluation dataset comprised 174 images taken from 10 different smears. The same 21 foci of PTCs and 66 foci of non-PTCs were photographed at both ×10 and ×40 objective magnifications. The images were analyzed separately, and then the data for each focus were compiled. When either the ×10 or the ×40 objective magnification images were required to be called positive to consider the lesion malignant, the algorithm demonstrated a sensitivity and specificity of 90.48 and 83.33%, respectively. If both the ×10 and the ×40 objective magnification images were required to classify a lesion as malignant, the sensitivity (33.33%) suffered significantly at the hands of an admittedly impressive increase in specificity (98.48%).

Guan et al. [22] subsequently published on the use of a CNN experiment using 224 × 224 pixel images of liquid-based preparations (Surepath stained with hematoxylin and eosin) taken at ×40 objective magnifications. Two separate AI architectures, VGG-16 [23] and Inception-v3 [24], were trained on 407 images of PTCs and 352 images of benign nodules and then tested on 69 PTC images and 59 images of benign nodules. All 887 images were taken from 20 Surepath slides and the images randomly assigned into the test and training sets. In an ideal scenario, the test and training cases would be completely separate. The data showed the VGG-16 model was able to predict the correct category with 100% sensitivity and 94.91% specificity. The Inception-v3 model achieved a sensitivity of 98.55%, with a specificity of 86.44%, though it is difficult to ascertain the influence training and testing different images of the same tumour had on the results.

These studies represent a significant step forward as the first use of CNNs to investigate FNAB of thyroid nodules. In distinction from previous studies, the authors use images rather than data extracted from images as input. However, the methodology still requires a significant amount of hands-on preprocessing. Elliott Range et al. [25] published the first paper to use WSIs as input to a CNN architecture without hands-on preprocessing steps. They utilized a semi-supervised method comprised of 2 CNNs. The first identified regions of interest, using pathologist-annotated areas containing follicular cells as ground truth. The second CNN classified these regions of interest into benign or malignant categories, using the surgical resection diagnosis as ground truth. Both CNNs were predicated on the VGG-11 [23] framework. This training methodology produced an algorithm that was able to predict a benign or malignant outcome with 92.0% sensitivity and 90.5% specificity.

They also developed cutoffs to approximate the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) category for each case using an ordinal regression framework. TBSRTC is not a strictly ordinal system, so these categories are not completely analogous, but the results were interesting nonetheless. In the test set comprised of 109 cases, 29 were called benign, with an associated risk of malignancy (ROM) of 0% and 10 were called malignant with a ROM of 100%. Concerning the less definitive categories, 48 were categorized atypia of undetermined significance with a ROM of 4.2%, 14 were identified as follicular nodules with a ROM of 50%, and 8 were called suspicious, with a ROM of 75%.

Non-Graphical Input to Determine Benign from Malignant

While image analysis seems like the most obvious application of AI technology to the analysis of thyroid, several groups have taken a different approach. Instead, Zoulias et al. [26] reviewed 1,886 non-malignant and 150 malignant smears notating the presence or absence of 67 microscopic features. These data were used to train an artificial neural network, support vector machine, and K-nearest neighbour [26]. These were both analyzed separately and using a majority voting classifier. The majority voting classifier, which combined the results of all 3 algorithms, outperformed each algorithm separately, with a sensitivity of 89.1% and a specificity of 99.4%.

Similarly, Ippolito et al. [27] used pathologist-defined microscopic characteristics but also integrated clinical

data and ultrasound characteristics into their algorithm. Their aim was to triage indeterminate and follicular lesions into high- or low-risk categories using a neural network framework. Their algorithm demonstrated a sensitivity of 85.7% and a specificity of 58.8%.

Follicular Adenoma versus Carcinoma

The distinction between follicular adenoma and follicular carcinoma currently requires a surgical resection to evaluate for capsular and vascular invasion. Determination of benign from malignant follicular lesions by FNAB is a problem for which a solution continues to evade cytopathologists. This is exactly the problem for which Shapiro et al. [28] and Savala et al. [29] sought an AI solution. Shapiro et al. [28] trained 3 two-layer automated neural networks on different sets of data. The first was trained on pathologist-identified cytologic features, including the presence of colloid, structure of the tissue fragments, and cytoplasmic and nuclear features. The second used morphometric features as input data. The third neural network used 256 × 256 pixel colour images of Giemsa-stained smears. The cytologic, morphometric, and image algorithms correctly classified 93, 96, and 87% of cases, respectively.

Savala et al. [29] also sought an AI solution to delineate follicular adenoma from follicular carcinoma. The neural network architecture they employed was slightly deeper with 5 hidden nodes. They used a combination of both semiquantitatively evaluated cytologic features and morphometric parameters. The sample size was relatively small, with 26 cases of follicular adenoma and 31 cases of follicular carcinoma in total. The algorithm was able to correctly categorize the 3 follicular adenomas and 6 follicular carcinomas in the testing set.

Non-Invasive Follicular Thyroid Neoplasm with Papillary-Like Nuclear Features versus Papillary Carcinoma

Non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) is a low-risk neoplasm [30] that might not require surgical treatment. As a result, the distinction between NIFTP and PTC is clinically essential. Similar to follicular adenoma and carcinoma, a resection specimen is currently required to make this distinction in order to fully evaluate for presence of papillary formations and invasion. Maleki and colleagues [31] used a novel approach to the application of AI technologies to this important question. They processed key words and phrases from the microscopic descriptions, which were then used to train a support vector machine. There was not a separate test cohort, but the validation showed a sensitivity of 72.6% and a specificity of 81.6%.

Perhaps, the most interesting outcome of this study was the generation of weights for the key words and phrases. The weights reflected how often each key word or phrase was associated with a given outcome (i.e., NIFTP or PTC). The top 2 key words/phrases in favour of NIFTP over classic PTC were "scant colloid" and "microfollicular pattern." The top 2 words/phrases in favour of cPTC were "papillary" and "pale chromatin." While more study is needed, this novel approach could be applied to the generation of diagnostic criteria in the future.

Practical Applications

These studies represent over 2 decades of research applying a variety of AI technologies to evaluate thyroid cytology specimens. Unfortunately, no application has demonstrated to be robust enough for clinical use. Currently, there are still significant technologic hurdles that must be addressed before AI applications to thyroid FNAB are ready for clinical use.

One issue is the development of algorithms that are suitable for thyroid specimens. The majority of studies referenced above are able to perform adequately when the possible interpretations are limited to binary outcomes, but the interpretation of thyroid FNAB is more complex than the current technologies are able to address. Another major technological problem is the time required to scan cytology specimens. Optimal digital images from cytology specimens require scanning of the cytology preparation in multiple layers, so-called "z-stacking," to recreate the three dimensionality usually seen in cytology specimens. Current commercial scanners available on the market require a long time to scan a slide with multiple z-stacking layers, making it very difficult to implement their use in clinical practice. Additionally, these digital scanners are expensive and there is no real perceivable return of investment at this moment, especially in the absence of validated algorithms. Nonetheless, the commercial success of AI applications in automated cervical screening might provide a roadmap towards development of a similar thyroid FNAB solution.

Conclusion

In this article, we have summarized the significant advancements that have been made in the field of AI applications to thyroid cytopathology. While a significant amount of work has been done, we are only at the dawn of a major paradigm shift in pathology towards digital and computational pathology. As more cytopathologists embrace this new technology, we will begin to define how future pathologists can leverage AI technologies to answer essential clinical questions.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The research reported in this publication was supported in part by the Cancer Center Support Grant of the National Institutes of Health/National Cancer Institute under award number P30CA008748. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

B.K. and O.L. each contributed to the conceptualization, investigation, procurement of resources, and writing of this article.

References

1 Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. Comput Med Imaging Graph. 2011 Oct–Dec;35(7–8):515–30.
2 Louis DN, Feldman M, Carter AB, Dighe AS, Pfeifer JD, Bry L, et al. Computational pathology: a path ahead. Arch Pathol Lab Med. 2016 Jan;140(1):41–50.
3 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44.
4 Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med. 2019 Jul;143(7):859–68.
5 Tan GH, Gharib H. Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. Ann Intern Med. 1997 Feb 1;126(3):226–31.
6 Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest. 2009 Aug;39(8):699–706.
7 Hegedus L. Clinical practice. The thyroid nodule. N Engl J Med. 2004 Oct 21;351(17):1764–71.
8 Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, 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. 2016 Jan;26(1):1–133.
9 Ho AS, Sarti EE, Jain KS, Wang H, Nixon IJ, Shaha AR, et al. Malignancy rate in thyroid nodules classified as Bethesda Category III (AUS/FLUS). Thyroid. 2014 May;24(5):832–9.
10 Straccia P, Rossi ED, Bizzarro T, Brunelli C, Cianfrini F, Damiani D, et al. A meta-analytic review of the Bethesda System for Reporting Thyroid Cytopathology: has the rate of malignancy in indeterminate lesions been underestimated? Cancer Cytopathol. 2015 Dec;123(12):713–22.
11 Cibas ES, Ali SZ. The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid. 2017 Nov;27(11):1341–6.
12 Ko YS, Hwang TS, Kim JY, Choi YL, Lee SE, Han HS, et al. Diagnostic limitation of fine-needle aspiration (FNA) on indeterminate thyroid nodules can be partially overcome by preoperative molecular analysis: assessment of RET/PTC1 rearrangement in BRAF and RAS wild-type routine air-dried FNA specimens. Int J Mol Sci. 2017 Apr 12;18(4):806.
13 Zhang M, Lin O. Molecular testing of thyroid nodules: a review of current available tests for fine-needle aspiration specimens. Arch Pathol Lab Med. 2016 Dec;140(12):1338–44.
14 Nishino M, Nikiforova M. Update on molecular testing for cytologically indeterminate thyroid nodules. Arch Pathol Lab Med. 2018 Apr;142(4):446–57.
15 Li F-F, Johnson J, Yeung S. CS231n: convolutional neural networks for visual recognition; 2017.
16 Karakitsos P, Cochand-Priollet B, Pouliakis A, Guillausseau PJ, Ioakim-Liossi A. Learning vector quantizer in the investigation of thyroid lesions. Anal Quant Cytol Histol. 1999 Jun;21(3):201–8.
17 Cochand-Priollet B, Koutroumbas K, Megalopoulou TM, Pouliakis A, Sivolapenko G, Karakitsos P. Discriminating benign from malignant thyroid lesions using artificial intelligence and statistical selection of morphometric features. Oncol Rep. 2006;15(Spec no. 4):1023–6..:
18 Gopinath B, Shanthi N. Support vector machine based diagnostic system for thyroid cancer using statistical texture features. Asian Pac J Cancer Prev. 2013;14(1):97–102.
19 Gopinath B, Shanthi N. Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images. Australas Phys Eng Sci Med. 2013 Jun;36(2):219–30.
20 Gopinath B, Shanthi N. Development of an automated medical diagnosis system for classifying thyroid tumor cells using multiple classifier fusion. Technol Cancer Res Treat. 2015 Oct;14(5):653–62.
21 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. 2018;9:43.
22 Guan Q, Wang Y, Ping B, Li D, Du J, Qin Y, et al. Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. J Cancer. 2019;10(20):4876–82.
23 Simonyan K, Zisserman A. Very deep convolutional networks of large-scale image recognition.

24 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. IEEE conference on computer vision and pattern recognition (CVPR). 2016 Jun 27–30. Las Vegas, NV: IEEE; 2016. p. 2818–26.
25 Elliott Range DD, Dov D, Kovalsky SZ, Henao R, Carin L, Cohen J. Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. Cancer Cytopathol. 2020 Apr;128(4):287–95.
26 Zoulias EA, Asvestas PA, Matsopoulos GK, Tseleni-Balafouta S. A decision support system for assisting fine needle aspiration diagnosis of thyroid malignancy. Anal Quant Cytol Histol. 2011 Aug;33(4):215–22.
27 Ippolito AM, De Laurentiis M, La Rosa GL, Eleuteri A, Tagliaferri R, De Placido S, et al. Neural network analysis for evaluating cancer risk in thyroid nodules with an indeterminate diagnosis at aspiration cytology: identification of a low-risk subgroup. Thyroid. 2004 Dec;14(12):1065–71.
28 Shapiro NA, Poloz TL, Shkurupij VA, Tarkov MS, Poloz VV, Demin AV. Application of artificial neural network for classification of thyroid follicular tumors. Anal Quant Cytol Histol. 2007 Apr;29(2):87–94.
29 Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. Diagn Cytopathol. 2018 Mar;46(3):244–9.
30 LiVolsi VA, Baloch Z. Noninvasive follicular tumor with papillary-like nuclear features: a practice changer in thyroid pathology. Arch Pathol Lab Med. 2020 Mar 30.
31 Maleki S, Zandvakili A, Gera S, Khutti SD, Gersten A, Khader SN. Differentiating noninvasive follicular thyroid neoplasm with papillary-like nuclear features from classic papillary thyroid carcinoma: analysis of cytomorphologic descriptions using a novel machine-learning approach. J Pathol Inform. 2019;10:29.