Artificial Intelligence in Cytopathology: A Neural Network to Identify Papillary Carcinoma on Thyroid Fine-Needle Aspiration Cytology Smears

Parikshit Sanyal[1], Tanushri Mukherjee[2], Sanghita Barui[1], Avinash Das[3], Prabaha Gangopadhyay[4]

Departments of [1]Pathology and [3]Otorhinolaryngology, Military Hospital Jalandhar, Jalandhar, Punjab, [2]Department of Pathology, Command Hospital (WC), Chandimandir, Haryana, [4]Undergraduate Department, Masters of Science Program, Indian Institute of Science, Bangalore, Karnataka, India

Abstract

Introduction: Fine-needle aspiration cytology (FNAC) for identification of papillary carcinoma thyroid is a moderately sensitive and specific modality. The present machine learning tools can correctly classify images into broad categories. Training software for recognition of papillary thyroid carcinoma on FNAC smears will be a decisive step toward automation of cytopathology. **Aim:** The aim of this study is to develop an artificial neural network (ANN) for the purpose of distinguishing papillary carcinoma thyroid and nonpapillary carcinoma thyroid on microphotographs from thyroid FNAC smears. **Subjects and Methods:** An ANN was developed in the Python programming language. In the training phase, 186 microphotographs from Romanowsky/Pap-stained smears of papillary carcinoma and 184 microphotographs from smears of other thyroid lesions (at ×10 and ×40 magnification) were used for training the ANN. After completion of training, performance was evaluated with a set of 174 microphotographs (66 – nonpapillary carcinoma and 21 – papillary carcinoma, each photographed at two magnifications ×10 and ×40). **Results:** The performance characteristics and limitations of the neural network were assessed, assuming FNAC diagnosis as gold standard. Combined results from two magnifications showed good sensitivity (90.48%), moderate specificity (83.33%), and a very high negative predictive value (96.49%) and 85.06% diagnostic accuracy. However, vague papillary formations by benign follicular cells identified wrongly as papillary carcinoma remain a drawback. **Conclusion:** With further training with a diverse dataset and in conjunction with automated microscopy, the ANN has the potential to develop into an accurate image classifier for thyroid FNACs.

Keywords: Artificial intelligence, cytology, fine-needle aspiration cytology, image classification, neural network, thyroid

INTRODUCTION

Artificial intelligence (AI) is the defining characteristic of the new epoch of information technology. AI-powered devices have transformed our daily lives in the form of smartphones, self-driving cars, and intelligent home appliances. One of the domains where AI has made significant advances is image analysis and object recognition.[1] We have attempted to apply an AI model for categorization of thyroid cytology smears. In the present study, we have chosen a neural network paradigm to classify microphotographs of thyroid fine-needle aspiration cytology (FNAC) smears into two categories, papillary thyroid carcinoma (PTCA) and non-PTCA (non-PTCA). A neural network model has been chosen because has proven to be the most successful among machine learning models in image recognition tasks.[2,3] We have chosen to train the network only with cytologically diagnosed

of papillary carcinoma, which were later confirmed on resection and biopsy, and not to include borderline cases for the present study.

Artificial neural networks (ANNs) are a large family of trainable models, where each subfamily of models is optimized for different functions. For the specific task of tumor classification, we chose the ANN subfamily, known as convolutional neural networks (CNNs), the state-of-the-art networks of which

How to cite this article: 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.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2018/9/1/43/246760

are shown to perform image-based object classification.[4] Briefly, CNNs are feed forward neural networks which take a whole image as input and classifies the image in defined categories. The input image is passed through multiple "layers" in a feedforward manner, each layer comprising multiple, independent, and linear convolutional filters.[5,6] The input for each layer is the output of the previous one, with an overlaid nonlinearity. The architecture is based on the hierarchical image classification and object recognition pathway of the primate brain, the ventral visual pathway,[7-9] where the layers represent a particular retinotopic area of the brain, while the filters represent the receptive field of a neuron in that area. The image "features" extracted by the layers are finally fed into a classifier that determines the category the image belongs to. Further details about CNNs, the significance of each of their components, and how they perform image classification have been described by Karpathy *et al*.[10]

In the present study, we have chosen a CNN to classify microphotographs of thyroid FNAC smears into two categories, PTCA and non-PTCA. We have chosen to train the network only with cytologically confirmed cases of papillary carcinoma and not to include borderline cases for the present study. PTCA is the most common malignant neoplasm of the thyroid;[11] in one of the largest series of resections for hypopharyngeal carcinoma, occult papillary carcinoma was found in 2% of all thyroids; an incidence varying between 0.25%–7% has been reported in previous studies.[12] FNAC is usually part of the initial investigations to investigate a solitary thyroid nodule.[11] Papillary carcinoma presents with cytopathologic findings which are easily discernible to the trainee pathologist. However, the features are not unique to papillary carcinoma, and thus, recognition of papillary carcinoma on smears is a nontrivial machine learning problem. The sensitivity of FNAC in detecting papillary carcinoma has been found to be 76.47%–95.2% in various studies, with specificity between 68.4%–94.2%.[13-15] However, small focus of papillary carcinoma is often encountered in smears showing only benign findings elsewhere.[16] It is because of its reasonable sensitivity, specificity, and screening requirements that we have chosen FNAC thyroid for training the neural network.

Subjects and Methods

A retrospective study design was chosen. We collected material from two different tertiary care centers of North India. Archived and well-preserved slides of thyroid FNACs with good material were chosen for the purpose of the study. Only FNACs done within the last 2 years were selected, and any faded slides were discarded. The slides were stained with Romanowsky stain (Leishman Giemsa/May–Grünwald Giemsa) or Papanicolaou stain, to increase the extent of training of the software. Microphotographs were taken at ×10 and ×40 magnification. In keeping with the principle of training a neural network with diverse array of material, a subset of the slides was photographed in a Labomed ATC 2000 microscope, the other in a Nikon DSFi1c. Using two different microscopes

provides the requisite variations in illumination and color of images, ultimately leading to a different set of pixel values captured. Thus, the ANN can be trained to recognize PTCA even in varying conditions of illumination.

The "non-PTCA" category included images from smears of colloid goiter, cytologically diagnosed follicular neoplasms and lymphocytic thyroiditis. In the "PTCA" category, cytologically diagnosed (and histologically confirmed by an oncopathologist, one of the authors) papillary carcinoma smears were photographed. The photos were then segregated into two categories: the "training" category for teaching the neural network and "test" category for concurrent evaluation of its performance. The distributions of the images in the different categories are listed in Table 1. All images were then cropped to a dimension of 512×512 pixels, focusing on the areas of interest in the smear. A total of 370 cropped images (184 non-PTCA, 186 PTCA) photographed from 20 cytology smears (from 20 patients) were used for training of the software. Several microscopic foci from a single smear were photographed.

The CNN was developed in the Python programming language, using the TensorFlow[17] backend and the Keras library[18] (all open source), extending the method by Chollet.[19]

A color image is a three-dimensional array of size width × height × 3, where the values of the red, green, and blue channels are the depth of the image, i.e., 3. A convolutional network applies several arrays of a smaller size with randomly initiated value (i.e., a 5×5 array) and applies it iteratively over the image. Such smaller arrays are called "masks," and the several such masks can be applied over the image. The final output after this operations is the element-wise multiplication of the two arrays and results in an array which is smaller in width and height, but its depth is the same as the number of such masks applied. Thus, the original information in the image is mathematically redistributed into an array of a different shape after a convolution.[10] Convolution is usually followed by pooling operation, which takes slices of an array of specified size and returns the maximum value within the slice, thus reducing the size of the array.

The architecture of the CNN is shown in Figure 1. The CNN takes a 512×512 color image as input, which is an array of dimensions $512 \times 512 \times 3$. This is because each pixel of the image has three color components red, green, and blue, the sum of which is displayed on screen as the final color. A number of convolution and pooling layers then extract features of that image and generate local maxima values from adjacent pixels. The CNN finally confers an output "0" (non-PTCA) or "1" (PTCA).

The CNN was then trained on the set of images in Table 1 over 10 epochs, with a batch size of 16, after which 97.15% accuracy was achieved in the ×10 magnification.

After completion of training, in the performance evaluation stage, the CNN was evaluated against a different set of images. Sixty-six foci of non-PTCA lesions and 21 images of PTCA were photographed from 10 smears. Multiple foci were

Table 1: Distribution of microphotographs in two magnifications and categories (*n*=370) during the training phase

10×				40×			
Training		Test		Training		Test	
Non-PTCA	PTCA	Non-PTCA	PTCA	Non-PTCA	PTCA	Non-PTCA	PTCA
92	89	14	14	68	73	10	10

PTCA: Papillary thyroid carcinoma
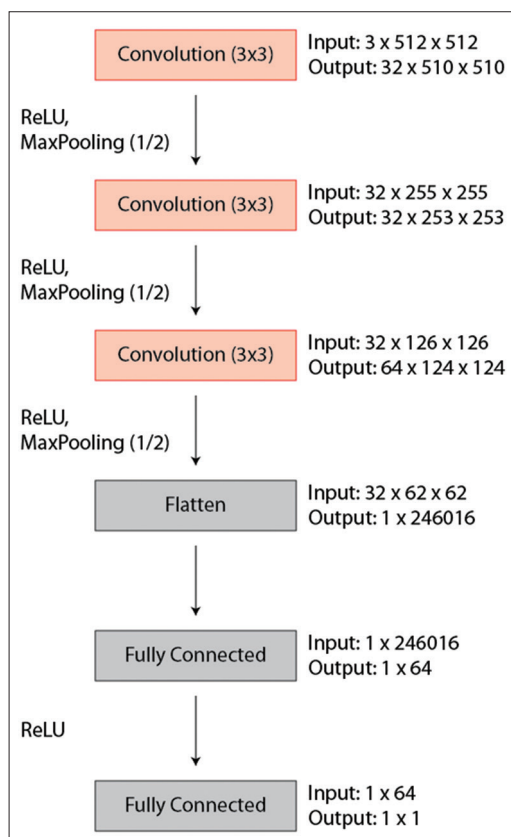
Figure 1: Architecture of the convolutional neural network

photographed from each smear. Every focus was photographed at two magnifications (×10 and ×40), for a total of 174 microphotographs. The cropped images (512 × 512 pixels) were then were examined by the CNN. The sensitivity, specificity, diagnostic accuracy, and positive and negative predictive value of the CNN were then determined from the results, using OpenEpi statistical software (Emory University, Rollins School of Public Health).[20]

RESULTS

The CNN used for smear classification is shown in Figure 1. The network takes in an input image and each subsequent layers extracts the relevant features from the previous layer [Figure 1]; the features extracted by the final layer is used to classify the image to PTCA or non-PTCA. During the training phase, after each epoch of learning, accuracy on the test dataset was measured concurrently. After 10 epochs of training, a contingency table was drawn [Table 2].

A number of false positives were met during the training session; specifically, 2 of the 14 (14.28%) images in the "non-PTCA × 10" set were identified wrongly as PTCA [Figure 2a]. Most other lesions were correctly characterized as in Figure 2b and c. This shows that in the initial phase of the training, the network is prone to false positives, classifying any structure vaguely resembling a papilla as papillary carcinoma. This is attributable to the fact that papillary formations are not unique to papillary carcinoma. Such papillary formations are regarded as "noise" in the image, whereas the nuclear features are the actual "data" to be trained on. Neural networks are prone to recognize noise, a phenomenon known as "overfitting" to the training data.[21]

Once training was over, performance of the CNN was on the evaluation dataset was analyzed. Eighty-seven foci were photographed 10 smears (from 10 patients); multiple foci were photographed from each smear. Each focus was photographed at two magnifications ×10 and ×40, for a total of 174 images. The results were as follows:

a. When using OR-based decision criteria, i.e., a focus is reported by the CNN to be PTCA in either of ×10 and ×40 magnification, the performance data are shown in Table 3, showing 90.48% sensitivity, 83.33% specificity, and 85.06% diagnostic accuracy
b. When using AND-based criteria, i.e., a focus is reported by the CNN to be PTCA in both ×10 and ×40 magnification, the specificity of the CNN showed significant improvement, at the cost of sensitivity [Table 4].

Within the evaluation, dataset images containing thick colloid, macrophages, and stain deposits (objects that are frequently seen on FNAC slides) were also included in the evaluation dataset as true negatives. Except from one focus showing only thick colloid, all of these were reported to be not PTCA by the CNN [Figure 3], i.e., they were correctly classified.

DISCUSSION

The conventional approach to computerized image analysis involves segmentation, blurring, edge detection, and watershed transform to identify geometrical properties in biological images. However, when analyzing cytology smears, the geometrical solution to the problem is inadequate. There are innumerable variations that might be encountered in cytological images, and any fixed program will fail to recognize a large number of them. Instead, we have focused on the "machine learning" approach. In this paradigm, learning is imparted to a model through a didactic manner, i.e., demonstrating a large

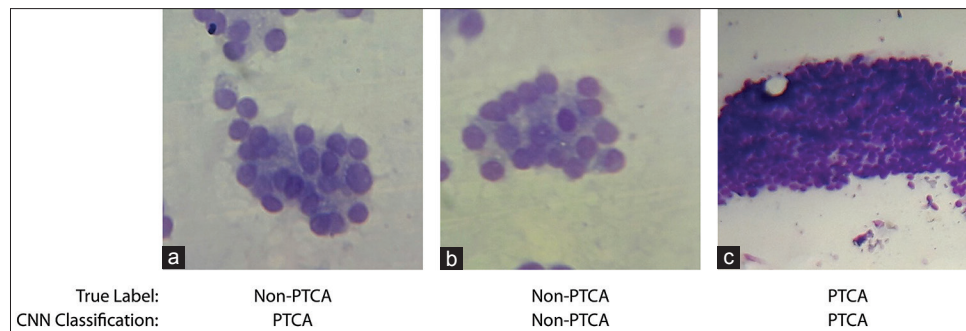|		
True Label:	Non-PTCA	Non-PTCA	PTCA
CNN Classification:	PTCA	Non-PTCA	PTCA

Figure 2: Examples of true and false image classification by the convolutional neural network on the training set. (a) False positive classification by the CNN; normal follicular cell cluster identified as carcinoma. (b) True negative classification by the CNN. (c) True positive classification by the CNN

Table 2: Performance characteristics of the convolutional neural networks on the concurrent test dataset during training ($n=48$)

Actual diagnosis	10×		40×		Total
	CNN detected "non-PTCA"	CNN detected "PTCA"	CNN detected "non-PTCA"	CNN detected "PTCA"	
Non-PTCA	12 (TN)	2 (FP)	10 (TN)	0 (FP)	24
PTCA	0 (FN)	14 (TP)	0 (FN)	10 (TP)	24

CNN: Convolutional neural networks, FP: False positive, TN: True negative, FN: False negative, TP: True positive, PTCA: Papillary thyroid carcinoma

Table 3: Performance characteristics of the convolutional neural networks when using criteria that a focus must be reported papillary thyroid carcinoma in any of ×10 and ×40 magnification to be diagnosed papillary thyroid carcinoma ($n=87$)

	Diagnosis on FNAC by pathologist		Total
	PTCA	Non-PTCA	
Diagnosis by CNN			
Detected PTCA	19 (TP)	11 (FP)	30
Detected not PTCA	2 (FN)	55 (TN)	57
Total	21	66	

Statistics	Value (%)	Lower-upper 95% CI	Method
Sensitivity	90.48	71.90-97.35	Wilson score
Specificity	83.33	72.57-90.43	Wilson score
Positive predictive value	63.33	45.51-78.13	Wilson score
Negative predictive value	96.49	88.08-99.03	Wilson score
Diagnostic accuracy	85.06	76.1-91.05	Wilson score

TP: True positive, FP: False positive, TN: True negative, FN: False negative, CI: Confidence interval, FNAC: Fine-needle aspiration cytology, PTCA: Papillary thyroid carcinoma, CNN: Convolutional neural networks

number of example data. The method is similar to the way the human pathologist is trained to distinguish between cell types, between benign and malignant. Reinforcement learning over several epochs of repetition of example data calibrates a machine learning model to correctly classify images.

Success with one of the modalities of machine learning, namely support vector machines, has already been demonstrated by Gopinath *et al*.[24] CNNs provide another approach to the problem, where a variety of images belonging to a category,

"non-PTCA" and "PTCA," is shown to the machine. The parameters of the model to arrive at the right answer are left for the machine to figure out by trial and error. Over many epochs of training, the machine figures out a set of rules (matrix transformations) which provide the right answer in majority of the cases.

The use of ANN has been examined by Dey who concluded that they are capable of reasoning tasks and can be used for diagnostic difficulties.[22] In a study from PGI Chandigarh, Savala *et al*. applied a neural network to distinguish follicular adenoma from carcinoma in thyroid FNAC smears.[23] They used 39 cases in training set and 9 cases each in validation and test sets. Their model successfully distinguished all 9 cases successfully. A similar tool, based on support vector machines, was used by Gopinath *et al*. to examine 110 thyroid FNAC smears, which achieved a diagnostic accuracy of 96.7% with sensitivity and specificity of 95% and 100%, respectively.[24] As per existing literature, the present study is the only one which employs an ANN to segregate PTCA from nonpapillary lesions. The ANN model has been chosen because of consistently better performance in image analysis than other models, in previous studies.[1-4]

The task, stated in context of papillary carcinoma thyroid, is to identify the following features:
a. At ×10 magnification – syncytial aggregates and sheets with a distinct border, three-dimensional tissue fragments, and papillary tissue fragments
b. At ×40 magnification – nuclear crowding, nuclear overlapping, nuclear grooves, and intranuclear cytoplasmic pseudoinclusions.[25]

Each of these features may be approached in a geometric manner, i.e., an intranuclear pseudoinclusion might be

characterized as a "circle within a circle." But considering the nearly endless variations encountered on cytological smears on this same theme, it is improbable that the "circle within a circle" rule will fit in all the cases [Figure 4].

The CNN has to convert an image to a single number, namely "1" (PTCA) or "0" (not PTCA). This conversion happens in the intermediate layers of the CNN. The layers extract features and shape of the image over successive convolutions and pooling operations, each time reducing and simplifying the information, and passing on to the next layer, until the image array is converted to either "1" or "0" [Figure 5].

It is important to appreciate the limitations of a CNN: the network can operate only on images from thyroid FNACs.
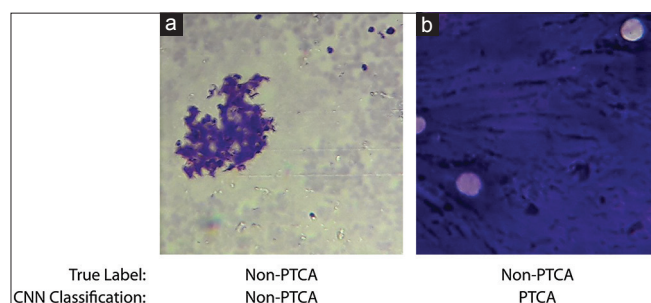

Figure 3: Examples of true and false classification by the convolutional neural network on artifacts. (a) True negative classification by the CNN. (b) False positive classification by the CNN
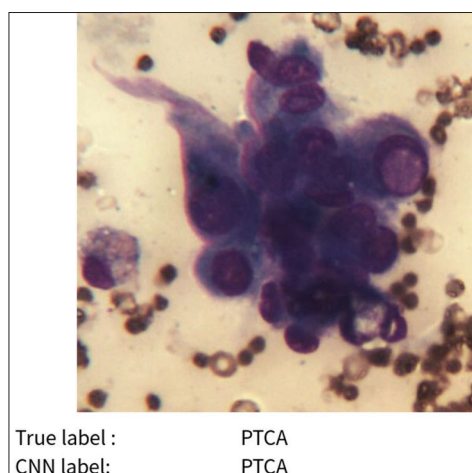

Figure 4: Intranuclear cytoplasmic pseudoinclusion correctly classified by the convolutional neural network

Given any random image of any object, the network will produce a result, either of "PTCA" and "Non-PTCA," i.e., it cannot distinguish between FNAC microphotographs and other photographs. However, implemented in the proper context, especially with an automated microphotography and slide scanning system, the CNN can provide actionable results.

The principal difficulties met during the training of the CNN were to distinguish between vague papillary formations by normal follicular cells [Figure 2a], which in the ×10 magnification was identified wrongly as papillae by the CNN. Furthermore, an area of thick colloid was wrongly identified as papillary carcinoma. This might be attributable to "overfitting" on training data, i.e., recognizing the signal (papillary formations) as well as the noise (overall basophilia of the image) of the training dataset.

When using the criteria that an image must be classified by the CNN as PTCA in both ×10 and ×40 magnification, the sensitivity is 90.49% and specificity only modest, 83.33%. This is lower than reported by Gopinath *et al.*[9] or Savala *et al.*;[8] however, the present CNN has been evaluated on a larger dataset than the aforementioned.

Table 4: Performance characteristics of the convolutional neural networks when using criteria that a focus must be reported papillary thyroid carcinoma in both of ×10 and ×40 magnification to be diagnosed papillary thyroid carcinoma (*n*=87)

	Diagnosis on FNAC by pathologist		Total
	PTCA	**Non-PTCA**	
Diagnosis by ANN			
Detected PTCA	7 (TP)	1 (FP)	8
Detected not PTCA	14 (FN)	65 (TN)	79
Total	21	66	

Statistics	Value (%)	Lower-upper 95% CI	Method
Sensitivity	33.33	17.19-54.63	Wilson score
Specificity	98.48	91.9-99,73	Wilson score
Positive predictive value	87.5	52.91-97.76	Wilson score
Negative predictive value	82.28	72.42-89.14	Wilson score
Diagnostic accuracy	82.76	73.48-89.26	Wilson score

FNAC: Fine-needle aspiration cytology, PTCA: Papillary thyroid carcinoma, ANN: Artificial neural network, TP: True positive, FP: False positive, TN: True negative, FN: False negative, CI: Confidence interval
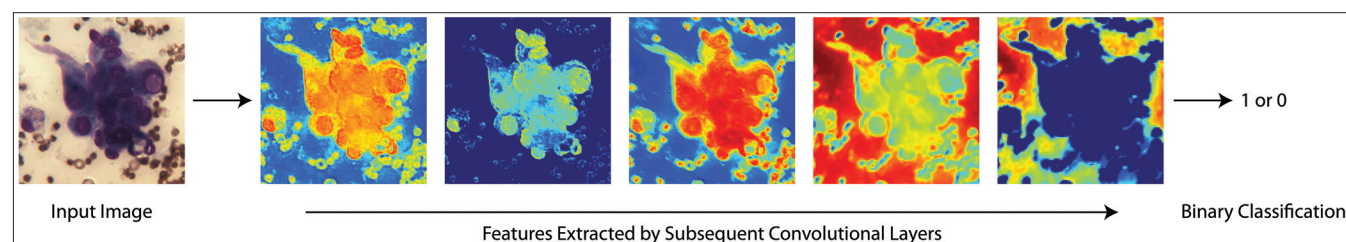

Figure 5: Processing an image by intermediate layers of the convolutional neural network

Unlike the study by Savala *et al.*,[8] false positives were met at both magnifications, which remain a shortcoming. However, it has to be noted here that a typical state-of-the-art deep convolutional network is trained on much larger datasets than the ones available for specific medical diagnostics.[26] This is to introduce sufficient variation in the training dataset and prevent the networks from overfitting to the data. Thus, it is time to develop such diverse repository of images containing large number of diverse pathological images to build a very strong and reliable image classifier.

CONCLUSION

With further training on larger and more diverse datasets, the CNN has potential to develop into an accurate image classifier for thyroid fine needle aspiration cytology.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Redmon J, Ali F. YOLOv3: An Incremental Development. Cornell University Library: ArXiv; 2018. Available from: http://arxiv.org/abs/1804.02767. [Last accessed on 2018 Nov 12].
2. Jarrett K, Kavukcuoglu K, Ranzato MA, LeCun Y. What is the Best Multi-Stage Architecture for Object Recognition? In International Conference on Computer Vision. IEEE; 2009. p. 2146-53.
3. Ciresan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J. High-performance neural networks for visual object classification. Arxiv preprint arXiv: 1102.0183; 2011.
4. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv: 151203385[cs]; 10 December, 2015. Available from: http://www.arxiv.org/abs/1512.03385. [Last accessed on 2018 Sep 17].
5. Lecun Y, Hafner P. Object Recognition with Imaged Based Learning. AT&T labs. Available from: http://www.yann.lecun.com/exdb/publis/pdf/lecun-99.pdf. [Last accessed on 2018 Sep 18].
6. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. Available from: https://www.papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. [Last accessed on 2018 Sep 18].
7. Mishkin M, Ungerleider L. Contribution of Striate Inputs to the Visuospatial Functions of Parieto-Preoccipital Cortex in Monkeys-Science Direct. Available from: https://www.sciencedirect.com/science/article/pii/016643288290081. [Last accessed on 2018 Sep 18].
8. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 2016;19:356-65.
9. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ, *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 2014;111:8619-24.
10. Karpathy A. Stanford University Open Access course. Convolutional Neural Networksk in for Visual Recognition. Available from: http://www.cs231n.github.io/convolutional-networks/. [Last accessed on 2018 Sep 17].
11. Nguyen QT, Lee EJ, Huang MG, Park YI, Khullar A, Plodkowski RA. Diagnosis and treatment of patients with thyroid cancer. Am Health Drug Benefits 2015;8:30-40.
12. Joshi P, Nair S, Nair D, Chaturvedi P. Incidence of occult papillary carcinoma of thyroid in Indian population: Case series and review of literature. J Cancer Res Ther 2014;10:693-5.
13. Hajmanoochehri F, Rabiee E. FNAC accuracy in diagnosis of thyroid neoplasms considering all diagnostic categories of the Bethesda reporting system: A single-institute experience. J Cytol 2015;32:238-43.
14. Sinna EA, Ezzat N. Diagnostic accuracy of fine needle aspiration cytology in thyroid lesions. J Egypt Natl Canc Inst 2012;24:63-70.
15. Cáp J, Ryska A, Rehorková P, Hovorková E, Kerekes Z, Pohnetalová D, *et al.* Sensitivity and specificity of the fine needle aspiration biopsy of the thyroid: Clinical point of view. Clin Endocrinol (Oxf) 1999;51:509-15.
16. Sharma C. An analysis of trends of incidence and cytohistological correlation of papillary carcinoma of the thyroid gland with evaluation of discordant cases. J Cytol 2016;33:192-8.
17. Available from: https://www.tensorflow.org/. [Last accessed on 2018 Oct 26].
18. Available from: https://www.keras.io/. [Last accessed on 2018 Oct 26].
19. Chollet F. Building Powerful Image Classification Models Using Very Little Data. Available from: https://www.blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html. [Last accessed on 2018 Jun 25].
20. Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version. Available from: http://www.OpenEpi.com. [Last updated on 2013 Apr 06; Last accessed on 2018 Oct 26].
21. Noh H, You T, Mun J, Han B. Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization. arXiv: 171005179 [cs]; 14 October, 2017. Available from: http://www.arxiv.org/abs/1710.05179. [Last accessed on 2018 Oct 26].
22. Dey P. Time for evidence-based cytology. Cytojournal 2007;4:1.
23. 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;46:244-9.
24. Gopinath B, Shanthi N. Support vector machine based diagnostic system for thyroid cancer using statistical texture features. Asian Pac J Cancer Prev 2013;14:97-102.
25. Jayaram G, Svante R. Orell thyroid. In: Orell SR, Sterett GF, editors. Orell & Sterrett's Fine Needle Aspiration Cytology. Edinburgh: Churchill Livingstone; 2012. p. 134.
26. Deng J, Dong W, Socher R, Li-Jia L, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Conference Computer Vision and Pattern Recognition; 2009. p. 248-55.