RESEARCH ARTICLE

Artificial Intelligence Role in Subclassifying Cytology of Thyroid Follicular Neoplasm

Mona Mohamed Aly Alabrak[1]*, Mohammad Megahed[2], Asmma Abdulaziz Alkhouly[2], Ammar Mohammed[2], Habiba Elfandy[1], Neveen Tahoun[1], Hoda Abdel-Raouf Ismail[1]

Abstract

Objective: Fine needle aspiration cytology has higher sensitivity and predictive value for diagnosis of thyroid nodules than any other single diagnostic methods. In the Bethesda system for reporting thyroid, the category IV, encompasses both adenoma and carcinoma, but it is not possible to differentiate both lesions in the cytology practice and can be only differentiated after resection. In this work, we aim at exploring the ability of a convolutional neural network (CNN) model to sub-classifying cytological images of Bethesda category IV diagnosis into follicular adenoma and follicular carcinoma. **Methods:** We used a cohort of cytology cases n= 43 with extracted images n= 886 to train CNN model aiming to sub-classify follicular neoplasm (Bethesda category IV) into either follicular adenoma or follicular carcinoma. **Result:** In our study, the model subclassification of follicular neoplasm into follicular adenoma (n = 28/43, images n = 527/886) from follicular carcinoma (n = 15/43, images n= 359/886), has achieved an accuracy of 78%, with a sensitivity of 88.4%, and a specificity of 64% and an area under the curve (AUC) score of 0.87 for each of follicular adenoma and follicular carcinoma. **Conclusion:** Our CNN model has achieved high sensitivity in recognizing follicular adenoma amongest cytology smears of follciualr neoplasms, thus it can be used as an ancillary technique in the subcalssification of Bethesda Iv category cytology smears.

Keywords: Thyroid gland - Cytolopathology- AI- convolutional neural network- Diagnosis

Introduction

Thyroid nodules is a common clinical problem, where the incidentally discovered nodules widely range between 20% and 76% (Ogmen et al., 2020). Thyroid fine needle aspiration cytology is a reliable diagnostic technique which identifies benign thyroid nodules, sparing many patients unnecessary surgery (Ali and Cibas, 2018). The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) is a uniform tiered reporting system for thyroid cytology, which stratifies the thyroid lesions into categories. Each diagnostic category is related to a certain risk of malignancy and accordingly guides towards the proper clinical management. The six categories are as follow: non-diagnostic (category I), benign (category II), atypia of undetermined significance/follicular lesions of undetermined significance (category III), follicular neoplasm/suspicion of follicular neoplasm (category IV), suspicion of malignancy (category V), and malignant (category VI) (Ogmen et al., 2020).

The thyroid nodules diagnosed as follicular neoplasm / suspicious for follicular neoplasm (Bethesda category IV) has a risk of malignancy of 25-40% and is considered one of the cytologically indeterminate nodules, as it encompasses two lesions follicular adenoma and follicular carcinoma (Khan and Zeiger, 2020). The main goal of such category is the identification of potential malignancies (follicular carcinomas), rather than identifying all follicular neoplasm, as follicular adenomas are clinically innocuous, and there is little if any evidence that suggests the progression from follicular adenoma to follicular carcinoma. However, the diagnosis of follicular adenoma or carcinoma can only be guaranteed after surgical excision by histopathological examination and determination of tumor capsular and lympho-vascular state (Ali and Cibas, 2018).

The standard of care for Bethesda IV category diagnosis is not proceeded before considering the clinical and sonographic features. The molecular testing may also be used as an ancillary test as an additional assessment of malignancy risk before surgical excision (Ali and Cibas, 2018). However, molecular testing has shown variable successes, and most were related to papillary thyroid carcinoma rather than other follicular lesions (Savala et al., 2018).

Artificial intelligence (AI) has deeply impacted

[1]Department of Pathology, National Cancer Institute, Cairo University, Egypt. [2]Department of computer science, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt. *For Correspondence: Mona.al-abrak@nci.cu.edu.eg

our lifestyle by imitating human intelligence using the computers, it is to execute tasks more efficiently, rapidly, and at a lower cost in solving complex problems than humans (Kamal and Kumari, 2020)A subset of artificial intelligence (AI) called machine learning includes mathematical algorithms that learn from input data to develop the capacity to generate predictions without being explicitly trained to do so (Landau and Pantanowitz, 2019). Machine learning is categorized into supervised and unsupervised learning based on labelling of the dataset. In supervised learning, the input is the value of the object analyzed and the output is the label representing ground-truth. After training phase, the model is able to classify the inputs into their correct label, for example in pathology image classification by assigning a tumor image to it's correct class or prediction the outcome/prognosis of a disease (Zhou, 2018). On the other hand, unsupervised learning does not use predefined labels or values. These algorithms attempt to group similar inputs together by understanding the associations or identifying data patterns. Unsupervised learning solves clustering problem either, for example a new functional or structural subgroupings of breast cancer based on subtle histologic variations (Harrison et al., 2021).

Until recently, machine learning used hand-crafted input data, where raw data is manipulated into a set of features / values before presenting to the model, for example a squamous cell images as an input data, is changed into set of features as cell size, shape, nucleus, cytoplasm, background and etc. Expert labeling of training data can be challenging, time-consuming, and susceptible to errors, especially on detailed labeling as with pathological images (Harrison et al., 2021). Deep learning technique is a special type of machine learning, which led to significant advances in computer vision. It has allowed skipping the expert labeling of features, and instead the model requires only a data set and it's label for training (learning), for example the image of squamous cell with their corresponding diagnosis (label), then the model "learn" the features by itself. Although deep learning is more accurate, however one of the trade-offs is being a "black box", where the model extract features by a complex interaction and those learned features are not understandable by humans, and it is essentially unknown how it arrives to the output (answer), and with each model run there may be a different result, and therefore any slight change in data quality or quantity may alter the learning process during model training and hence the answer the model on validation (Landau and Pantanowitz, 2019).

One of the most popular learning models for machine learning is an artificial neural network (ANN), it is mathematical model that is inspired by functioning of the brain (Emmert-Streib et al., 2020). The sequential layers of the artificial neural network (ANN) are organized as an input layer, then a hidden layer, and finally an output layer. When the architecture has more than two hidden layers is commonly considered as as a deep learning model (Emmert-Streib et al., 2020).

Convolutional neural network (CNN) is a special type of artificial neural network, that uses deep learning algorithms, and it is widely used in image and video recognition, due to it's high performance in recognizing patterns (Ismael et al., 2020). CNN is a mathematical architecture composed of three types of layers (or building blocks): convolution, pooling, and fully connected layers. The convolution and pooling layers are responsible for feature extraction, whereas fully connected layer maps the extracted features into final output, such as classification (Yamashita et al., 2018)

Early detection and accurate diagnosis are very vital aspect for the improvement of cancer prognosis. Applying artificial intelligence as a tool in these domain has shown promising potential for early detection and diagnosing with high precision, and it is just one of many applications of AI in cancer research (Patil et al., 2020). Several studies were published in thyroid imaging used CNN architecture for classifying being and malignant nodules, however, few studies in thyroid histopathology and cytopathology fields were published. (Wang et al., 2019) trained two CNN models using histological images of thyroid nodules to classify them according to their histologic types, achieving an accuracy of 97.34%. Two studied used CNN to distinguish papillary thyroid carcinomas from non-papillary thyroid carcinoma, one study by (Guan et al., 2019) achieved a 100% sensitivity and 94.91 % specificity, where (Sanyal et al., 2018) achieved a 90.48 % sensitivity and 83.33% specificity. (Elliott Rang et al, 202) have used CNN to classify thyroid nodules into benign vs malignant, achieving 92.0% sensitivity and 90.5% specificity. Other studies have applied CNN in the field of cytopathology as (Garud et al., 2017) who classified malignant vs benign breast lesions achieving accuracy of 80.76%, while (Martin et al., 2018) classified cervical cytology according to Bethesda system categories achieving an overall average accuracy of 60%. (Rahaman et al., 2021) trained several CNNs to classify cervical cytology images into benign and malignant achieving accuracies of 98%.

In this work, we aim at exploring the ability of a CNN model to subclassifying cytological images of Bethesda category IV diagnosis into follicular adenoma and follicular carcinoma. A CNN model is trained using microscopic images of thyroid cytology smears, that was later resected with established histopathological diagnosis as the ground truth.

Materials and Methods

This is a retrospective study on glass smear slides of thyroid nodules cytology, diagnosed at the cytopathology unit, NCI, Cairo university from the year 2015 to 2021, after which a surgical resection was done and diagnosed histopathologically at surgical pathology unit, NCI, Cairo university. These cases were sampled/aspirated with by an experienced radiologist under sonographic guidance mostly, or free-handed by an experienced cytopathologist, for clinically prominent and palpable thyroid nodules.

The study was carried out on the images for the cases, our cohort was 886 images captured from 43 follicular neoplasm cases.

Cases collection
Inclusion criteria

Only the cases diagnosed cytologically as follicular neoplasm or suspicious for follicular neoplasm (Bethesda category IV) and after surgical resection confirmed the diagnosis of follicular adenoma or follicular carcinoma.

Exclusion criteria
Cases with an obscuring element as blood or dense inflammation

Cases with dryness artifact or Diff-Quik stained slides (one slide per case), because the first step of stain is induction of dryness artifact to augment the cytoplasmic size and concurrently nuclear size, thus could be hindering the training process of the model.

Cases cytologically misdiagnosed as Bethesda category II (nodular colloid hyperplasia and thyroiditis), or Bethesda category V (suspicious for papillary carcinoma), or diagnosed as Bethesda III (Follicular lesion of undetermined significance). To avoid training of the model on cases, that represent a challenge to the expert cytopathologists.

Cases diagnosed after resection with double pathology as (follicular adenoma with nodular colloid hyperplasia), to avoid the error of sampling from the adjacent lesion, either free handed by the cytopathologist or by the radiologist under sonar -guided as some both lesions might mimic the other radiologically.

Cases histologically diagnosed as Hürthle cell variant of follicular adenoma, as the abundant eosinophilic cytoplasm can be considered as variation in cellular morphology, thus might hinder the training process of the model.

Smear slides revision

Every PAP-stained smear glass slide for each case was microscopically reviewed to fix any artifact caused by the storage as staining artifact or broken slides cover, example of pre and post fixation artifact in as in Figure 1.

Fixing processes

Slides with air bubble artifacts were fixed either by replacing the slide cover as in Figure 1 or re-staining when needed with caution to avoid losing the material. The smear slides that were unrepairable were excluded from the study.

Selecting the slides

Smear slides with the most appropriate diagnostic fields as region of interest (ROI) in the images was selected and placed in a different box for image acquisition.

Image acquisition

Two different microscopes equipped with a microscopic camera were used for image acquisition: (1) An Olympus BX43 microscope equipped with inserted camera EP50, with image dimensions (2592x1944 pixels), (2) a Nikon (eclipseE600) microscope equipped with an inserted camera with image dimensions (1920X1080 Pixels). Using two different cameras and microscopes provides the required variations in color of the acquired images and the illumination as well, ultimately leading to a different sets of pixel values captured. Thus, the CNN can be trained to classify the lesions in varying conditions of illumination.

Imaging setting were checked to ascertain that the microphotograph is mirroring the conventional microscopic image and by selecting the best setting that achieve the best quality of image e.g., focus, illumination, sharpness, saturation etc. However, some of the images underwent adjustment of saturation or lightening after capture, due to non-perfect staining or bad illumination at time of image acquisition.

The images of the region of interest (ROI) on the slide was captured at microscopic magnifications 10x and 40x, however only 40x was used to simplify the learning process / training of the model. The images were acquired blindly with no prior knowledge of the case diagnosis to avoid any bias during image acquisition. With the retrieval of the captured images, they were then annotated/classified (labelled) according to histopathological diagnosis as follicular adenoma, as in Figure 2 and follicular carcinoma, as in Figure 3.

Images / data cleaning

Some of the images (n= 80 image) underwent adjustment of saturation or lightening after capture, due to non-perfect staining or bad illumination, that could not be adjusted at time of image acquisition, as illustrated in Figure 4. While the images with an artifact as haziness or obscuring agent were excluded.

Data preprocessing
Image color and size

Images used were colored using three-channel. 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. All of the images were normalized into 256x256 pixels.

Data augmentation techniques

Deep learning models have a better performance with large datasets. Since medical images are always limited, data insufficiency represents a problem that hinder the model training. Augmentation is one of the ways to overcome such problem, which increases the number of images through manipulation of these images (as flipping and rotation), creating new images with the same label. Furthermore, augmentation solves class imbalance by enriching the diversity of the training samples and oversampling the minority class in the set, thus bringing balance to the data. Therefore, augmentation is known to allow the CNN to perform better, through enhancing the training of the model, which is the root problem, thus allowing the generalization of the results and avoid overfitting problem of the model. The techniques used for augmentation in this study are rotation_range = 2, zoom_range = 0.1, horizontal_flip = True.

Data Split

The images were split randomly into "training" samples from which the model learns, and "test /

validation" samples that are used for evaluation of its performance. The split was with the ratio of (4:1), where 80% and 20% are training and validation samples, respectively.

Proposed convolutional neural network model

The architecture of the network is of feed-forward, convolutional neural network, illustrated in Figure 5 and detailed as follows:

The proposed convolutional neural network is composed of several layers. The first three layers are convolutional layers, which are responsible for feature extraction from the data/images. Each convolutional layer is followed by max-pooling, and all have ReLU as an activation function, where each has a different filter as follows 32 size 2x2, 64 size 2x2 and 128 size 2x2, first, second and third layers respectively. All max-pooling layer size was 2x2. The following layer is a flattened layer. Then the deep network are three layers of network, which is responsible for classification. the first layer with 128 nodes, the second layer with 64 nodes, and the third layer with 32 nodes. And finally, the fourth and final layer with 2 nodes and activation function SoftMax.

As earlier mentioned, CNN takes a 256x256 color image as input, then convolution and pooling layers extract features of that image, then by the final layer, CNN confers an output for binary classification of follicular adenoma versus follicular carcinoma. All previous steps are summarized in Figure 6.

Evaluation metrics

Instead of using statistical methods, the evaluation of convolutional neural network model performance is measured by different evaluation metrics.

The evaluation metrics used in this study are the accuracy, sensitivity, specificity, precision, F1 score and recall as calculated in equations 1, 2, 3, 4, 5 and 6 respectively. These metrics are derived from the confusion matrix in Figure 7, which is plots the actual class versus the predicted class as follow; true positive (tp) and true negative (tn) denoting the number of positive and negative instances that are correctly classified. Meanwhile, false positive (fp) and false negative (fn) denoting the number of misclassified negative and positive instances, respectively.

$$\text{Accuracy} = \frac{tp+tn}{tp+fp+tf+fn} \quad \text{Equation 1}$$

$$\text{Sensitivity} = \frac{tp}{tp+fn} \quad \text{Equation 2}$$

$$\text{Specificity} = \frac{tn}{tn+fp} \quad \text{Equation 3}$$

$$\text{Precision} = \frac{tp}{tp+fp} \quad \text{Equation 4}$$

$$\text{F1-score} = \frac{2*p*r}{p+r} \quad \text{Equation 5}$$

$$\text{Recall} = \frac{tp}{tp+tn} \quad \text{Equation 6}$$

The other evaluation metrics is area under the curve (AUC) also known as ROC curve., simply it is the percentage of the plot box that lies under the ROC curve. The closer it is to 1 (100%), the better. It is calculated in

$$\text{Area under the curve} = \frac{Sp - np(n+1)/2}{np\,nn} \quad \text{Equation 7}$$

Sp is the sum of all positive examples ranked, while np and nn denote the number of positive and negative examples respectively.

Results

The number of follicular neoplasm cases n = 43 cases including follicular adenoma (n = 28; 65.11%) and carcinomas (n = 15; 34.88%) (Table 1). Number of images / regions of interest (ROI) n = 886 images; including follicular adenoma images (n=527; 59.48%) and follicular carcinoma images (n = 359; 40.51%). A convolutional neural network was trained using training sample (n=708; 80%) and validation samples (n= 178; 20 %) to measures model performance.

Several evaluations metrics were calculated through confusion matrix, illustrated in Figure 7, where class 1 represents follicular adenoma and class 2 represents follicular carcinoma.

Where model classification of follicular adenoma (represented as class 1), showed an accuracy 78.0% illustrated in Figure 8, a sensitivity of 88.4%, and a specificity of 64%, while F1-score of 82.5%, precision of 77.3% and recall of 88.46%. While model classification of follicular carcinoma (represented by class 2), showed an accuracy 78%, a sensitivity of 64%, a specificity of 88.4%, while F1-score of 70.6%, and precision of 79.6% and recall of 63.5%. The area under the curve (AUC) score for follicular adenoma vs follicular carcinoma is 0.87, each, where class 0 represents follicular adenoma and class 1 represents follicular carcinoma illustrated in Figure 9, Table 2.

Table 1. Summary of the Number and Percentages of Cases and Region of Interest

Bethesda category	Number of cases	Percentages of cases	Number of images	Percentages of images
Category IV				
Follicular adenoma	28	65.11%	527	59.48%
Category IV				
Follicular carcinoma	15	34.88%	359	40.51%
Total	43	100%	886	100%

Table 2. Summary of Different Metrics for Evaluation of Model Performance

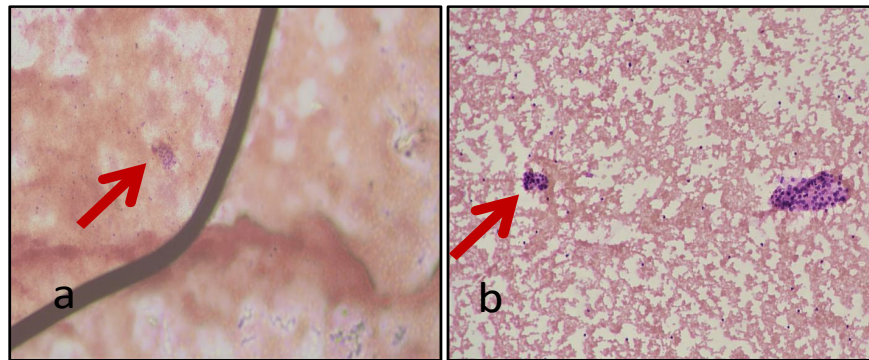| Evaluation metrics	Follicular adenoma	Follicular carcinoma
Accuracy	78.00%	78.00%
Sensitivity	88.4%,	64.00%
Specificity	64.00%	88.40%
Recall	88.4%.	63.50%
Precision	77.30%	79.60%
F1-score	82.50%	70.60%

Figure 1. An Example of Fixation Process. (a), Smear slide with an air bubble artifact mascaraing the follicular group of cells underneath; (b), The smear slide after changing the cover slip (Papanicolaou stained, 10X magnification).
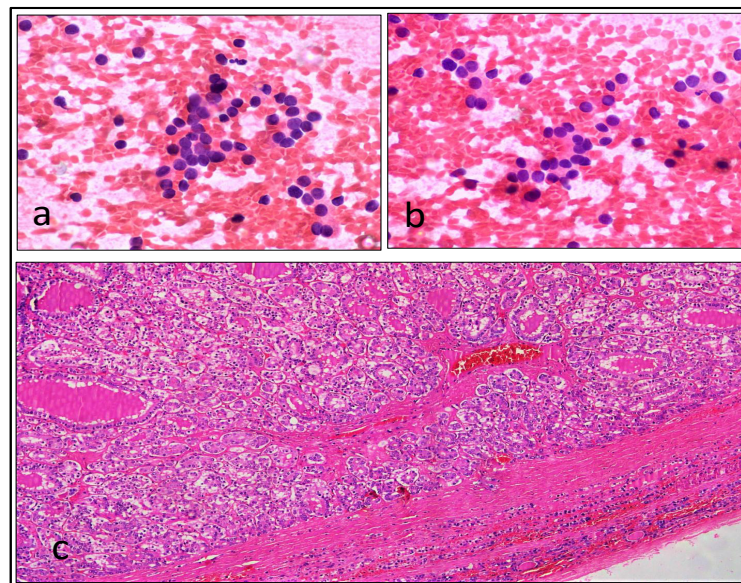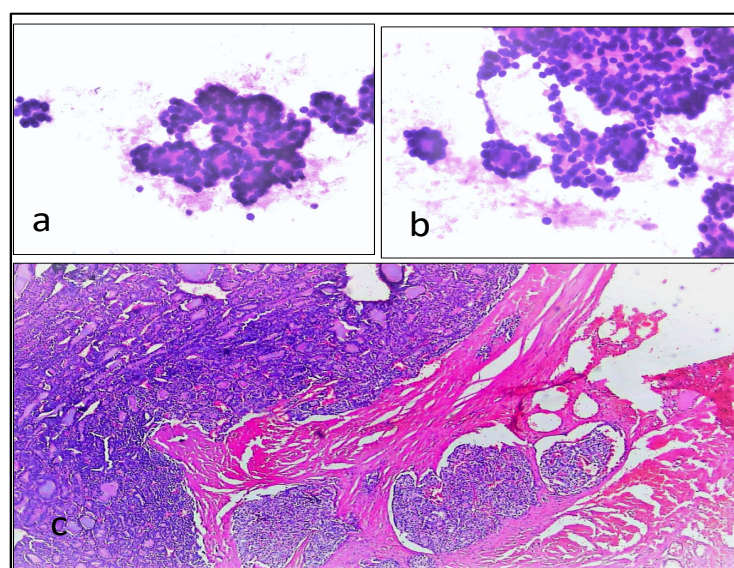

Figure 2. Case of Follicular Adenoma. Cytological diagnosed as follicular neoplasm and diagnosed after resection as follicular adenoma. (a)and(b), Cytological smears shows follicular cell arrangement in mainly in microfollicles pattern within a hemorrhagic background (Papanicolaou stain, 40X magnification); (C), Histological section shows Follicular adenoma with the surrounding intact capsule (HandE stained, 10X magnification).

Figure 3. A Case of Follicular Carcinoma. A case cytologically diagnosed as follicular neoplasm and diagnosed after resection as follicular carcinoma. (a)and(b), Cytological smears are relatively cellular with follicular cell arranged mainly in microfollicular pattern. (Papanicolaou stain, 40X magnification); (C), Histological section shows widely invasive follicular carcinoma with capsular invasion and invasive groups at RED Arrowhead (HandE stained, 10X magnification).
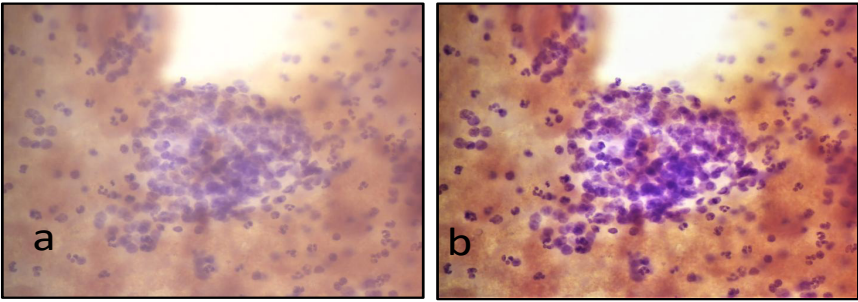
Figure 4. An Example of Image Editing. A case of follicular neoplasm, smears since 2015 (Papanicolaou stain, 40X magnification). (a), Shows hazy follicular groups with fade discoloration; (b), shows the image after edit (color enhancement and increasing the sharpness), which highlighted the follicular group as well .
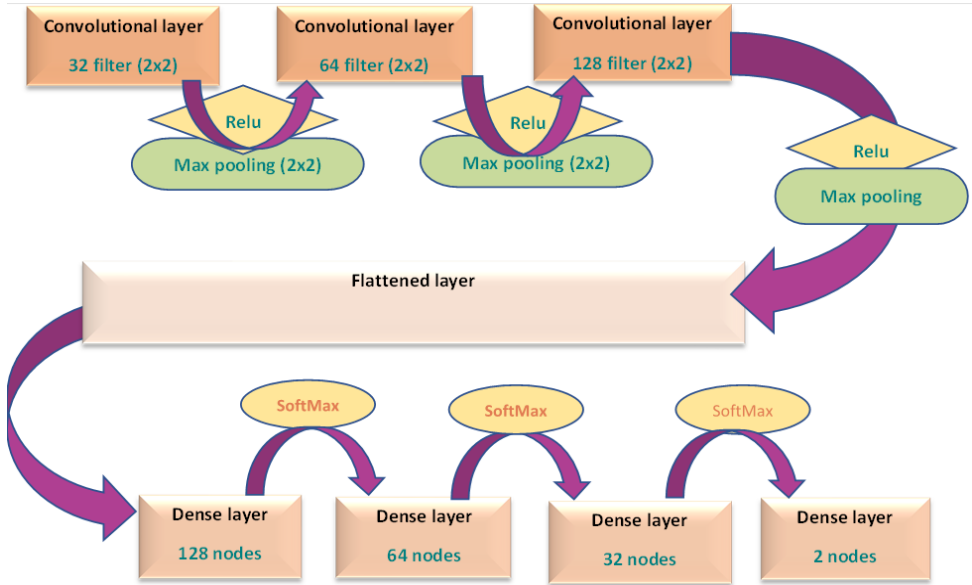

Figure 5. An Illustration of the Proposed Convolution Neural Network. The architecture, showing number of layers and type and specifications of each layer with its activation function.
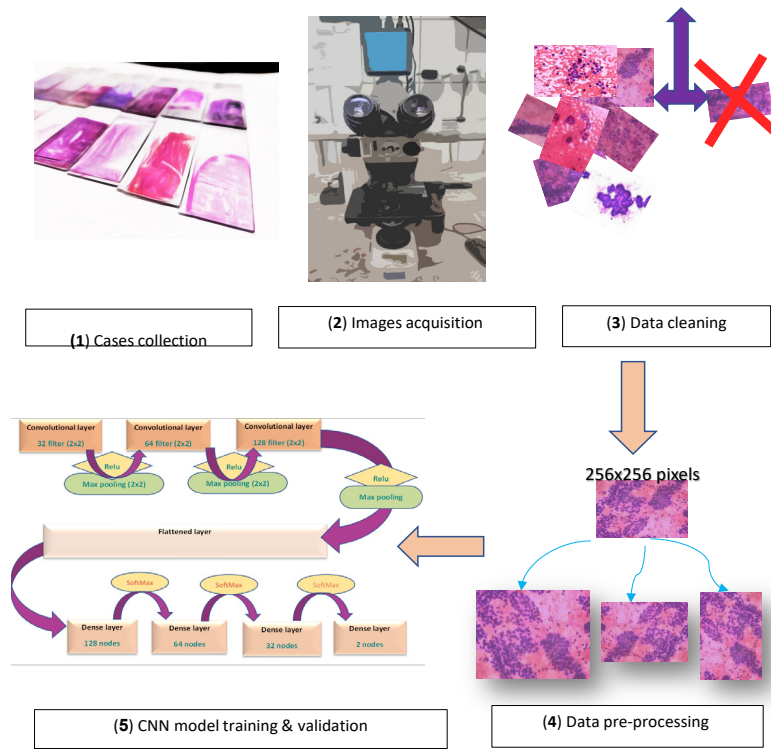
Figure 6. Summary of the Steps Starting with Primary Data Collection Till Training the CNN Model.
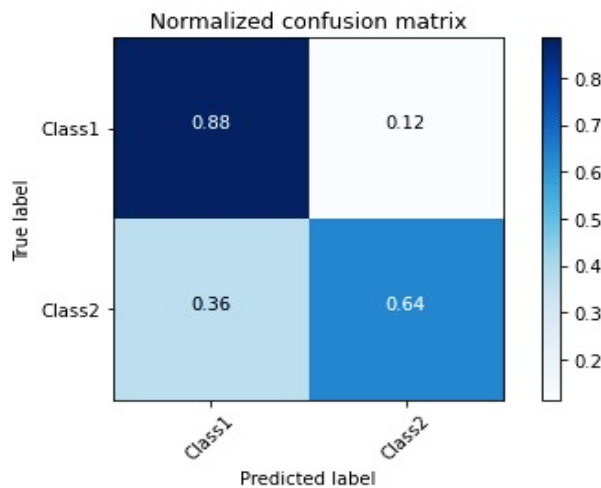
Figure 7. An Illustration of Confusion Matrix where Class 1 Represents Follicular Adenoma and Class 2 Represents Follicular Carcinoma (upper right are the cases of predicted correctly predicted as follicular adenoma, lower right are the cases falsely predicted as follicular carcinoma , upper left are the cases wrongly predicted as follicular carcinoma and lower left are the cases predicted correctly as follicular carcinoma).
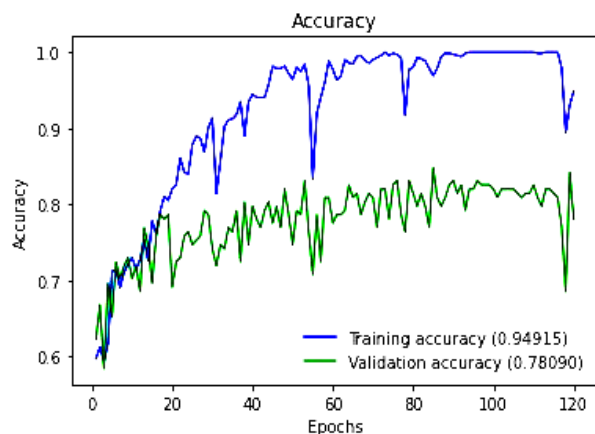

Figure 8. Graph Plotting the Achieved Accuracy during Model Training (represented by blue line) and model validation (represented by green line).

Discussion

Out of all six categories of Bethesda system for reporting thyroid cytopathology, follicular neoplasm / suspicious for follicular neoplasm (Bethesda category IV) is considered one of the indeterminate nodules, as it constitutes both follicular adenoma and carcinoma, where each has different biologic behavior and sometimes requires different management. This study is exploring the utility of applying artificial intelligence represented by a convolutional neural network (CNN) model, for subclassifying thyroid cytopathology images of Bethesda category IV into follicular adenoma and follicular carcinoma.

In our study, differentiating follicular adenoma from follicular carcinoma achieved an accuracy of 78.0%, with sensitivity 88.4%, and specificity 64% and AUC = 0.87. A previous work that classified thyroid cytology into
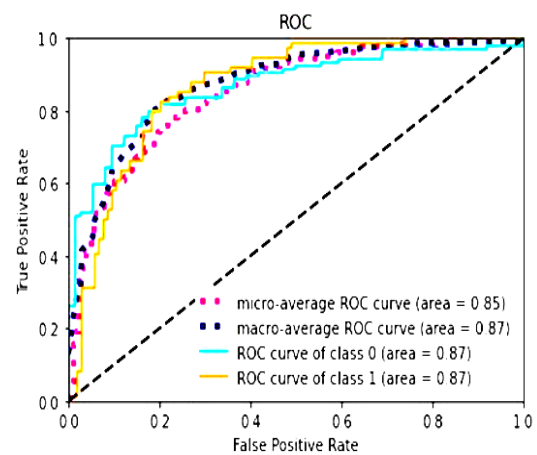

Figure 9. A Graph Plotting Area under the Curve / ROC Curve. The vertical line represents true positive and horizontal line represents false positive rate, class 0 represents follicular adenoma and class 1 represents follicular carcinoma.

follicular adenoma vs follicular carcinoma by applying an artificial neural network, conducted by Savala and his colleague, they achieved an accuracy of 100% and ROC = 1 in the training set (Savala, et al., 2018). Compared to their study, our work stems on convolutional neural network while theirs stems on a backpropagation neural network. The backpropagation neural network, used morphologic cytologic nuclear features (nuclear area, diameter, perimeter, pleomorphism and cellularity .. etc) extracted by using another software, with the integration of clinico-morphometric parameters as age. But in our study, the convolutional and max pooling layers were responsible for feature extraction directly from cytological images only labeled by their own diagnosis.

Few researchers have applied CNN for the purpose of classifying thyroid nodules. Elliott Range and his colleagues used CNN to classify thyroid lesions into benign vs malignant and their work achieved AUC score of 0.93, 92.0% sensitivity, and 90.5% specificity (Range et al., 2020). Compared to our results, they achieved relatively better performances than ours that could be attributed to larger sample size (n=659 comparted to our study n=43). Other studies by Guan and his colleagues in 2019 and Sanyal and his colleagues in 2018, both studies were conducted to differentiate papillary thyroid carcinoma from non-papillary thyroid carcinoma using CNN. (Guan et al., 2019) used two different CNN architectures (VGG-16 and Inception-v3) achieving a better performance by 100% sensitivity and 94.91% specificity by applying VGG-16 model and 98.55% sensitivity and 86.44% specificity with Inception-3 model. (Sanyal et al., 2018) achieved relatively better performance achieving 90.48 % sensitivity of and 83.33% specificity and an accuracy of 85.06% with the use of one objective magnification images for training (either 10X or 40 X). However, the model achieved lower sensitivity of 33.33% with still a better specificity of 98.48% and accuracy 82.76% on using two objective magnifications (x10 and x40). Discrepant results between our study and previous studies of (Elliott Range et al., 2020), (Guan et al., 2019) and (Sanyal et

al., 2018) can be attributed to diverse source of images in our work compared to a single source of images in all the previous works. However, the greatest attribution to the discrepant results is the nature of thyroid lesions. Elliott Range et al., (2020) included malignant category such as medullary carcinoma, anaplastic carcinoma and papillary carcinoma to be compared to benign thyroid nodule that usually does not represent a large burden to the cytopathologist to diagnose. As well papillary thyroid carcinoma has distinctive nuclear morphology with main differential diagnosis is non-invasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP). The studies were designed to distinguish papillary carcinoma from lesions other than NIFTP such as colloid nodular goiter, lymphocytic thyroiditis and follicular neoplasm representing Bethesda II and IV (Sanyal et al., 2018) and benign nodules (Bethesda II) categories as differential lesions (Guan et al., 2019). Consequently, the CNN is expected to achieve high performance, whereas our approach was challenged by a category (Bethesda IV) that is already challenging to the highly experienced cytopathologists on cytology smears.

Our study concludes that convolutional neural network model can be used as an ancillary technique in the diagnosis of thyroid cytopathology. Although the differentiation between follicular adenoma and carcinoma is challenging even to artificial intelligence model and our cohort was not large, however, the model has achieved a high sensitivity in recognizing follicular adenoma. The CNN model can be used as an additional factor in risk classification, as part of pre-surgical intervention assessment for follicular neoplasms. Artificial Intelligence (CNN models) are a long-term investment in the field of pathology, although it requires data (which is not always available), however, with good training of the model it offers an affordable and sustainable ancillary technique, rather than the other costly alternatives.

Our recommendations for future studies for applying artificial intelligence in thyroid FNAC are as follows: (1) A larger dataset is to be used for training the CNN model with better performance. (2) A dataset with higher quality would be better to be used to enhance the performance of the model, even without a larger dataset. (3) Different architectures of a convolutional neural network can be trained on the available dataset, to explore the ability of each architecture to exploit the current dataset of cytology images, regardless of their quality and quantity, to achieve higher performance of classification.

Author Contribution Statement

All the authors have contributed to this study. Data collection, review of the slides and image acquisition were performed by Dr Mona Alabrak, and under supervision of Dr Habiba Elfandy, Prof. Dr Neveen Tahoun and Dr Hoda Ismail. A Pilot CNN model run was done by Asmma A. Alkhouly, to allow better image acquisition of region of interest and the proper magnification to be applied in the study, under supervision of professor Dr Ammar Mohammed. The construction of CNN architecture with the proper hyperparameters as well as CNN model training

was executed by Mohammad Megahed. The validation of CNN model results was done by Professor Dr Ammar Mohammed. The preliminary draft of the manuscript was written by Dr Mona Alabrak and Dr Habiba Elfandy. All authors revised and commented on primary version of the manuscript and approved the final one.

References

Ali SZ, Cibas ES (2018). The Bethesda System for Reporting Thyroid Cytopathology. Definitions, criteria and explanatory notes, In follicular neoplasm/suspicious for follicular neoplasm, Springer international publishing AG, Switzerland, pp 72-4.

Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Front Artif Intell*, **3**, 4.

Garud H, Karri SPK, Sheet D, et al (2017). High-Magnification Multi-views Based Classification of Breast Fine Needle Aspiration Cytology Cell Samples Using Fusion of Decisions from Deep Convolutional Networks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 828–33.

Guan Q, Wang Y, Ping B, et al (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. *J Cancer*, **10**, 4876.

Harrison JH, Gilbertson JR, Hanna MG, et al (2021). Introduction to Artificial Intelligence and Machine Learning for Pathology. *Arch Pathol Lab Med*, **145**, 1228–54.

Ismael SA, Mohammed A, Hefny H (2020). An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artif Intell Med*, **102**, 101779.

Kamal VK, Kumari D (2020). Use of Artificial Intelligence/ Machine Learning in Cancer Research During the COVID-19 Pandemic. *Asian Pac J Cancer Care*, **5**, 251–3.

Khan TM, Zeiger MA (2020). Thyroid Nodule Molecular

Testing: Is It Ready for Prime Time?. *Front Endocrinol (Lausanne)*, **11**, 809.

Landau MS, Pantanowitz L (2019). Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. *J Am Soc Cytopathol*, **8**, 230–41.

Martin V, Kim TH, Kwon M, et al (2018). A More Comprehensive Cervical Cell Classification Using Convolutional Neural Network. *J Am Soc Cytopathol*, **7**, S66.

Ogmen B, Aydin C, Kilinc I, et al. (2020). Can Repeat Biopsies Change the Prognoses of AUS/FLUS Nodule. *Eur Thyroid J*, **9**, 92–8.

Patil S, Moafa IH, Alfaifi MM, et al (2020). Reviewing the Role of Artificial Intelligence in Cancer. *Asian Pac J Cancer Biol*, **5**, 189–99.

Rahaman MM, Li C, Yao Y, et al (2021). DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Comput Biol Med*, **136**, 104649.

Range DD, Dov D, Kovalsky SZ (2020). Application of a machine learning algorithm to predict malignancy in thyroid cytopathology. *Cancer Cytopathol*, **128**, 287–95.

Sanyal P, Mukherjee T, Barui S, Das A, Gangopadhyay P (2018). 'Artificial Intelligence in Cytopathology: A Neural Network to Identify Papillary Carcinoma on Thyroid Fine-Needle Aspiration Cytology Smears'. *J Pathol Inform*, **9**, 43_18.

Savala R, Dey P, Gupta N (2018). Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. *Diagn Cytopatho*, **46**, 244-9.

Wang Y, Guan Q, Lao I, et al (2019). Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: a large-scale pilot study. *Ann Transl Med*, **7**, 45.

Yamashita R, Nishio M, Do RKG, Togashi K (2018). Convolutional neural networks: an overview and application in radiology. *Insights Imaging*, **9**, 611–29.

Zhou ZH (2018). A brief introduction to weakly supervised learning. *Natl Sci Rev*, **5**, 44–53.