Expert Systems With Applications 188 (2022) 115927

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An ensemble deep learning for automatic prediction of papillary thyroid carcinoma using fine needle aspiration cytology

Nguyen Thanh Duc [a,b,c,d,1], Yong-Moon Lee [e,1], Jae Hyun Park [f,*], Boreom Lee [a,*]

[a] Department of Biomedical Science and Engineering (BMSE), Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea
[b] Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill Univ, Montreal, Canada
[c] McConnel Brain Imaging Center, Montreal Neurological Institute, McGill Univ, Montreal, Canada
[d] Ludmer Centre for Neuroinformatics and Mental Health, McGill Univ, Montreal, Canada
[e] Department of Pathology, College of Medicine, Dankook University, South Korea
[f] Department of Surgery, Wonju Severance Christian Hospital, Yonsei University, Wonju College of Medicine, South Korea

ARTICLE INFO

Keywords:
Papillary thyroid carcinoma
Fine needle aspiration cytology
Computer-aided diagnosis
Deep CNN models
Ensemble learning
ThinPrep

ABSTRACT

Accurately cytopathological diagnosis of Papillary Thyroid Carcinoma (PTC) is of importance for reducing costs and increasing efficiency of treatments. In this paper, we pursue that goal by introducing artificial intelligence (AI) for automatic classification of malignant PTC cell clusters from Fine Needle Aspiration Cytology (FNAC) processed by ThinPrep. High-resolution cytological images obtained with a 40 × objective lens digital camera attached to an Olympus microscope were segmented into fragments and then divided into training, validation, and testing subsets. Fragments are non-overlapped patches containing only regions-of-interest that cover informative tissue structures for making proper diagnoses. Deep learning CNN models were pre-trained and fine-tuned on large-scale ImageNet domain before they were re-trained on cytology fragments. Moreover, we proposed a method to compute certainty of the patient-level prediction that undoubtedly provides additional evidence for reliability and confidence of the prediction. Results showed that the best classification performance on digital FNAC images achieved using DenseNet161, obtaining a mean accuracy of 0.9556 ($p < 0.01$), a mean sensitivity of 0.9734, and a mean specificity of 0.9405 on yet-to-be-seen test-set. Ensemble learning findings suggested combinations of AdaBoost classifier with multiple CNN models boosted predictive performances, up to 0.9971 accuracy. Moreover, stain normalization introduced by Reinhard increased the predictive ability, outperforming histogram specification, and Macenko methods. Presented findings demonstrate deep learning can integrate into computer-aided diagnosis systems to support cytopathologists in accurate diagnosis of PTC.

1. Introduction

PAPILLARY Thyroid Carcinoma (PTC) is the most common malignancy of the thyroid gland and has been increasing in incidence worldwide. Fine Needle Aspiration Cytology (FNAC), a minimally invasive technique using the ultrasonography to the thyroid nodule targeted, has been a vital preoperative diagnostic modality in the evaluation and the ThinPrep (Hologic, Marlborough, MA) processed slides have been widely used since 1996, US Food and Drug Administration (FDA) approval. The Bethesda System for Reporting Thyroid Cytopathology (TBS) has been widely adopted worldwide for thyroid FNAC slides, which recognizes 6 diagnostic Categories for estimations of cancer risk as follows (Cibas & Ali, 2017); 1: nondiagnostic/unsatisfactory; 2: benign; 3: atypia of undetermined significance/follicular lesion of undetermined significance (AUS/FLUS); 4: follicular neoplasm/suspicious for a follicular neoplasm; 5: suspicious for malignancy; and 6: malignant. In this pilot study, we investigate the practical possibility of artificial intelligence (AI) supported in clinical preoperative diagnosis using the ThinPrep slide digital images; the Category 2 (benign) and 6 (confirmed malignant) digital images were used.

Digitizing whole-slide images of tissues has brought the arrivals of Computer-Aided Diagnosis (CAD) systems to accelerate the diagnosis process and allow large-scale screening. However, CAD systems using AI in modern digital pathology can adversely be fraught due to the variations in cytology smears obtained from light microscopy due to various problems such as the images taken from different scanners, and various

* Corresponding authors.
E-mail addresses: thanh.duc.nguyen@mcgill.ca (N.T. Duc), vilimoon@daum.net (Y.-M. Lee), jhoney@yonsei.ac.kr (J.H. Park), leebr@gist.ac.kr (B. Lee).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.eswa.2021.115927
Received 5 January 2021; Received in revised form 27 July 2021; Accepted 16 September 2021
Available online 2 October 2021
0957-4174/Crown Copyright © 2021 Published by Elsevier Ltd. All rights reserved.

acquisition protocols, or variations in stained tissue samples, as well as variations in staining procedure (time and concentrations). Thus, prior to automatically quantitative analysis, stain color normalization is an essential practice to reduce color and intensity variations present in stained cytopathology. In the literature, stain normalization has been shown as an effective process to improve predictive accuracies of the CAD systems in digital pathology (Araujo et al., 2017).

In recent advances, deep Convolutional Neural Networks (CNN) have appeared as new angled approaches for analyzing multiple clinical pathology, i.e., breast cancer (Araujo et al., 2017; Aresta et al., 2019), lung cancer (Kanavati et al., 2020; Teramoto et al., 2019), and PTC (Guari et al., 2019; Mukherjee, Sanyal, Barui, Das, & Gangopadhyay, 2018). In a pioneer study presented by Araujo et al., (Araujo et al., 2017), the authors introduced an approach to diagnose hematoxylin-eosin (H&E) breast biopsy images using deep Convolution Neural Network (CNNs). Histological Whole-Slide Images (WSI) were predicted into four categories including normal, benign lesion, in situ carcinoma, and invasive carcinoma, and two categories, carcinoma and non-carcinoma. The CNN model was able to achieve up to 77.8% accuracy for four-class classification and 83.3% for carcinoma/non-carcinoma classification with a great sensitivity of 95.6%. By using a training dataset of 3,554 WSIs, the authors in (Kanavati et al., 2020) developed a CNN deep learning model to differentiate carcinoma from non-neoplastic tissues in histopathological lung cancer images. Optimal results were obtained for identifying lung carcinoma tissues with high Receiver Operator Curve (ROC) Area under the Curves (AUC) on four independent unseen test dataset, AUCs of 0.975, 0.974, 0.988, and 0.981, respectively. The successful results achieved on these studies ensure promising performances for deep CNN network in cytological diagnosis.

Early machine learning approaches using traditional Support Vector Machine (SVM) or Artificial Neural Network (ANN) classifiers combined with various imaging-based features in automatic CAD systems for thyroid nodule cytology have been introduced (Gopinath & Shanthi, 2013a, 2013b, 2015). More recently, a few pilot studies have introduced deep CNN networks for differential diagnosis of PTC in cytological images and the obtained results are promising. Particularly, Guari et al., (Guari et al., 2019) exploited a VGG-16 deep CNN network to recognize PTC tissues from benign thyroid nodules from cytological slides. The authors achieved the accuracy rates of 97.66% and 92.75% with VGG-16 and Inception-v3 models on fragment-wise binary classification. In addition, Mukherjee et al., (Mukherjee et al., 2018) developed an artificial neural network (ANN) for differentiating PTC and Non-PTC on microphotographs using FNAC smears. The results showed good sensitivity of 90.48%, moderate specificity of 83.33%, and a very high negative predictive value of 96.49% and a diagnostic accuracy of 85.06%. However, these studies have several limitations. First, the classification decision was made on the patch-wise approach, not on entire FNAC images. For accurate diagnoses in clinical practices, cytopathologists are advised to carefully make a diagnostic decision on entire FNAC smears. Second, fragments were patch-based images and were generated manually with a fixed size. For proper training of CNN models, patches that do not contain relevant tissue regions for diagnosis and background patches should be excluded. In addition, the automatic ROIs extraction mechanisms are highly recommended to provide full concepts of automatic end-to-end diagnostics in the modern CAD systems. Finally yet importantly, highlighted feature representations to provide informative hierarchical structures of differential tissues are not provided in these studies.

In this present study, we introduced an automatic Computer-Aided Diagnostic system for differentiating the PTC tissues with benign ones from FNAC digital images processed by ThinPrep ® using deep CNN frameworks. In this pilot study, we have selected to train the CNN models only with cytologically confirmed cases of PTC (TBS Category 6) and not to include suspicious cases. We first proposed an automatic approach for preprocessing and stain normalization of high-resolution cytological digital images captured from the Olympus microscopy systems. Next, the fragments for training the deep CNN models are automatically generated using Canny edge detection and contours methods. We then employed a combination of transfer learning and data augmentation to retrain well-established CNN-based ResNet, DenseNet and Inception models. After being trained on a large-scale dataset, the CNN models are expected to show better performances. Our important contributions in this study are summarized as follows:

- We proposed a novel CNN-based deep learning framework that can be implemented in modern automatic CAD systems for accurately identifying the PTC tissues from the benign ones and computing the certainty level of the prediction using FNAC digital images.
- We proposed an ensemble deep learning model, which combined multiple individual deep learning models to significantly boost the predictive performances.
- We introduced an automatic smear extraction (fragments) algorithm to extract only regions of interest of tissues from original cytology FNAC images.
- We implemented the Gradient-weighted Class Activation Mapping (GRAD-CAM) approach to highlight the important regions that are highly dominant on the discrimination decision-making.
- We evaluated the effectiveness of different stain normalization methods on the predictive performance.
- Our proposed ensemble learning is superior to the existing models for automatic PTC diagnosis in current literature.

The rest of this manuscript is divided as follows. The section II is an overview of the materials and methods used in study that describe the acquisition process, the preprocessing phases and automatic generation of fragment images which contain only the relevant regions of interest served for making proper clinical diagnoses. In section II, we also provide details of CNN deep learning architectures, strategies for training the deep CNN models, and the explanations of performance evaluation. Section III provides the results. In section IV, we discuss the results, limitations and some future works that could be done to improve the diagnosis accuracy and to transfer our promising results closer to the concepts of fully automatic and accurate modern CAD systems that can support the pathologists in diagnosis and analysis of the PTC.

2. Materials and methods

2.1. Subjects and cytology image acquisition

This study was conducted with the approval of the ethics committee from the Institutional Review Board (IRB) of Wonju Severance Christian Hospital, Yonsei University, Wonju College of Medicine (IRB approval number: CR320109). The cytological images required to develop and evaluate our method were collected from patients who underwent thyroid nodule FNAC and thyroidectomy from January 1st, 2013 to December 31st, 2019. Before the FNA procedure, written informed consent was obtained from all patients. An FNAC biopsy was performed with 22-gauge needles by an experienced sonographer (member of the Korean Surgical Ultrasound Society, 8 years of experience with thyroid ultrasound examination after obtaining board certification) under ultrasonographic guidance. Thin layer liquid-based cytology (LBC) preparations are superior to conventional preparations with regard to background clarity, monolayer cell preparation, and cell preservation. It is easier and less time consuming to screen and interpret LBC preparations because the cells are limited to smaller areas on clear backgrounds with excellent cellular preservation. Therefore, most of the FNAC samples were transferred to a 10 ml syringe and then prepared with a natural sedimentation-type thin layer LBC system using a BD SurePath liquid-based Pap Test (Beckton Dickinson, Durham, NC, USA). A digital still camera (DP27, Olympus, Tokyo, Japan) with a 40 × objective lens attached to a microscope (BX45, Olympus) was used to take the pictures for the LBC smears. All images were collected and annotated by two

experienced cytopathologists (two co-authors, members of the Korean Society of Pathologists, 5 and 10 years of experience with thyroid cytopathologic examination after obtaining board certification) and saved in JPEG format.

Our dataset contained 367 hematoxylin-eosin (H&E)-stained images. The dataset included 222 cases of PTC and 145 cases of benign lesions (Non-PTC). These slides were digitized at 400 × magnification. All the PTC images had classic features (including high cellularity, papillary fronds with anatomical edges, enlarged oval nuclei with longitudinal intranuclear grooves, nuclear crowding and overlapping, cellular swirls, and chewing gum colloid (Akhtar, Ali, Huq, & Bakry, 1991)) and were diagnozed by our experienced cytopathologists as category VI according to the Bethesda system for reporting thyroid cytopathology (Cibas & Ali, 2017). All the selected patients underwent a thyroidectomy and were given a pathologic diagnosis of classical type PTC. All the images of benign nodules fit the category II description in the Bethesda system and consisted of an adequately cellular specimen composed of varying proportions of colloid and benign follicular cells arranged as macrofollicles and macrofollicle fragments (Cibas & Ali, 2017). All studied subjects in category II undertook a thyroidectomy and were identified a pathologic diagnosis of benign thyroid lesion such as adenomatous hyperplasia or nodular hyperplasia. Table 1 provides detailed information on the original dataset and the number of balanced fragments for training, validation, and testing in an example balanced subsample dataset.

2.2. Preprocessing and automatic generation of fragment images

Cytological images have different backgrounds as they were captured by different microscope systems. As can be seen from Fig. 1, there are six different background images in the PTC class. Therefore, prior to automatic quantitative analysis, stain color normalizations were performed to standardize the image background. In this study, we performed multiple automatic stain normalization methods as proposed by Reinhard et al., (Reinhard, Ashikhmin, Gooch, & Shirley, 2001), Macenko et al. (2009) and histogram specification (Khan, Rajpoot, Treanor, & Magee, 2014). In principle, a simple statistical analysis was used to impose one image's (source) color characteristics on another (reference). Then, the stain normalization on the source image can be achieved by applying its characteristics to the reference image. Fig. 2 visualizes images before (lower) and after (upper panels) normalization using three stain normalization methods.

A set of preprocessing steps was then performed in order to divide the original cytology images of 4800 × 3600 pixels into a number of fragments, in which each fragment contained relevant non-overlapping clusters of tissues or regions of interest (ROIs). According to the experienced cytopathologists (two co-authors, members of the Korean Society of Pathologists, 5 and 10 years of experience with thyroid cytopathologic examination after obtaining board certification), ROIs are the cell clusters, which should contain at least more than five cells to ensure the reliability for making clinical diagnosis. With that regard, we were not only able to increase datasets served for training the deep CNN models but also were able to adapt the practical diagnosis routines into the AI-supported CAD systems.

Translating the criteria from the cytopathologists into the algorithm, in our initial findings, we observed that fragments sized of 200 × 200 pixels should be enough to cover the informative tissue structures. The data quality-controlled process included careful re-annotation of the fragments by our cytopathologist to ensure that more reliable labels should be obtained prior to further analysis. Any fragments, which did not hold similar labels as its original FNAC slide, were discarded from the training dataset. Fig. 3A presents automatic preprocessing procedures, performed by OpenCV library, starting from original FNAC images to fragments while Fig. 3B illustrates the results provided at each preprocessing phase.

2.3. Deep CNN architectures

CNN-based deep learning models have proved their effectiveness on image classifications in various problem domains (Nguyen Thanh Duc & Lee, 2019, 2020; Nguyen Thanh Duc et al., 2020) and thus are hoped to achieve promising predictive accuracy on our clinical diagnosis of PTC cytopathology. In this present study, we employed six recently developed CNN deep learning models, which had demonstrated to provide high predictive performances in a large-scale image dataset for image classification and recognition as compared to all other previously introduced CNN models, for solving our similar predictive problems; and those well-established models are ResNet, DenseNet and Inception described as follows:

i) Deep Residual network (ResNet): Residual network (ResNet) exploited the idea of bypass pathways used in Highway Networks (He, Zhang, Ren, & Sun, 2016). The identity layer or "shortcut" helps to address gradient vanishing by letting the gradient values flow directly to upper layers during back-propagation. ResNet has five versions (ResNet-18, 34, 50, 101, and 152) depending on the depth of the convolution layers. In this study, we employed the ResNet50 architecture, which can be illustrated in Fig. 4 for the classification task. ResNet50 has 50 convolution layers, and its configuration is as follows (Kim et al., 2020). The set consists of a convolution layer, batch normalization, and ReLU (activation function) with 49 layers, and a fully connected last layer. Only the first convolution layer is set to a 7 × 7 kernel with stride 2 and padding 3, and the kernel size of all subsequent convolution layers is 3 × 3. In addition, in this task we implemented other ResNet versions including ResNet101 and RestNet152 that have 101 and 152 convolution layers, respectively.

ii) Densely Connected Convolutional Networks (DenseNet): Similar to ResNet, DenseNet also uses the concept of residual connection (Huang, Liu, Van Der Maaten, & Weinberger, 2017). However, unlike ResNet that preserves information through additive identity mapping on a subset of layers, DenseNet uses cross-layer connectivity in a denser fashion. DenseNet connects all preceding layers to the next coming layer in the feed-forward mechanism. All feature-maps of previous layers have been used on all subsequent layers. Therefore, DenseNet is considered as a very narrow structure and is parametrically expensive by the increase of the number of feature maps. DenseNet has three versions (DenseNet-121, 161, and 201 with 32 channels). In this work, we deployed two DensetNet121 and DenseNet161 architectures that consist 121 and 161 convolution layers, respectively.

iii) Inception-v3: Most CNN models like VGG, DenseNet, or ResNet just stacked the convolution layers deeper and deeper. Inception architecture brought an idea of splitting a network into multiple output branches and developing the network in horizontal dimension. There are various versions such as Inception-v1, Inception-v2, Inception-v3, Inception-v4, and Inception-ResNet. The problems come from choosing a right kernel size for convolution operation when image's sizes vary frequently. Inception-v3 replaced large size filters (5 × 5 and 7 × 7) with small and asymmetric filters and used 1 × 1 convolution

Table 1
Numbers of original FNAC images in each class and numbers of automatically extracted fragments in each randomly selected balanced subsample.

Cytological type	Original images	Originally generated fragments	Randomly selected fragments	Training fragments	Validation fragments	Test fragments
PTC (malignant)	222	866	700	495	140	70
Non-PTC (benign)	145	713	700	495	140	70
All types	**367**	**1579**	**1400**	**980**	**280**	**140**

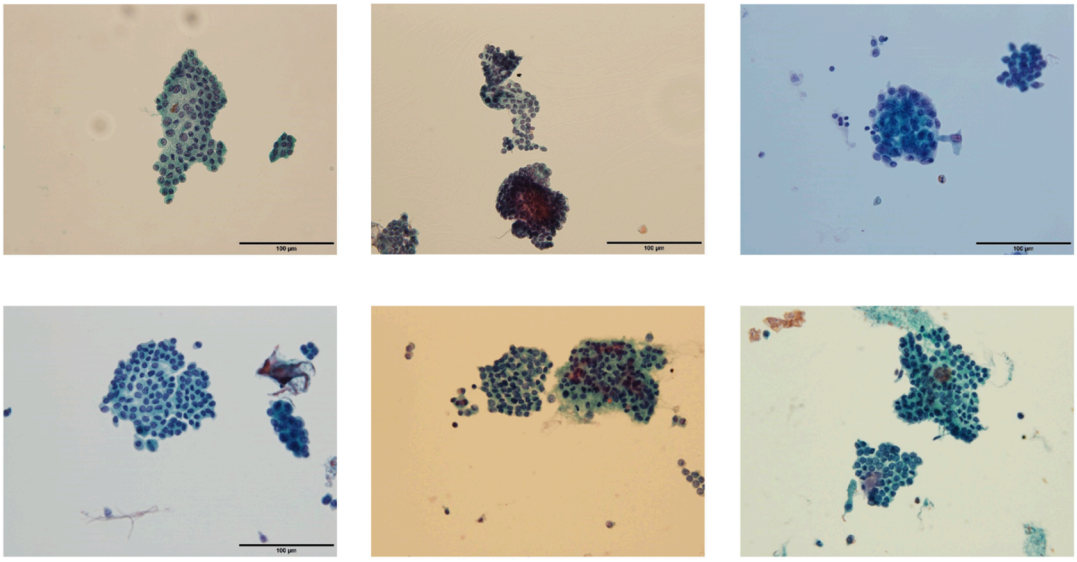

Fig. 1. Various background colours shown in cytological TPC FNAC images captured from multiple microscope systems before implementing of stain normalization methods.

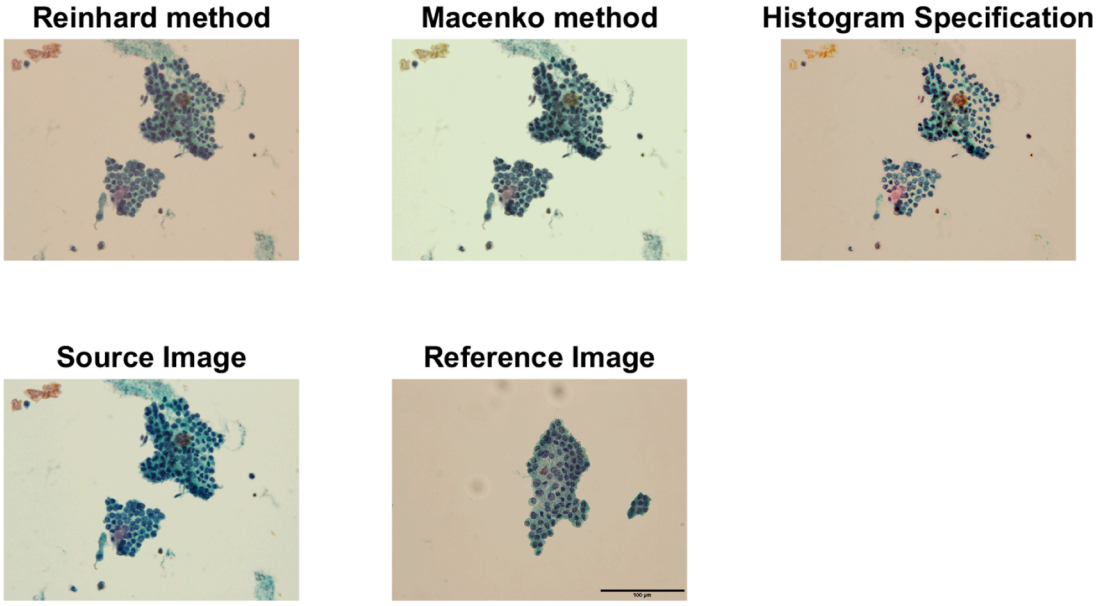

Fig. 2. Illustrations of three stain normalized images (upper panel) implemented in this study. The original FNAC slide source and reference images are also shown (lower panels).

with a large size filter as a bottleneck before rejoining the main branch. Another main point of Inception-v3 is that its input must be a size of 299 × 299 pixel (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016).

2.4. Ensemble learning models for CNN deep learning

CNN deep learning models are non-linear, offer increased flexibility, and their performances are commensurate with the available amount of training data. However, this flexibility has a drawback; that is, the training of CNN models are learned using a stochastic algorithm which means that they are sensitive to the specifics of the training data and, thus, a different set of weights each time they are trained can be achieved, which in turn produce different predictions. In this work, to diminish the variances in predictions and to increase predictive results, we proposed an ensemble deep learning approach, which train multiple CNN deep learning models instead of a single model and combine the prediction results given by these models.

After the training of CNN models reached convergences, we performed ensemble learning with bagging by considering each results of CNN models as a feature using a single machine learning algorithm at the end of inference. In recent advances, ensemble learning has earned reputations as the most effective model, however its mechanism is really burden and costly in tradeoff between accuracy, speed, and computational expenses required high computing infrastructures. In this work, we aggregated the results of six CNN models; each image is fed to six models, which output six different probabilities. Next, we employed several conventional supervised machine learning algorithms (i.e., Support Vector Machine-SVM, K-Nearest Neighbor-KNN, Decision Tree-DT, Random Forest-RF, Neural Net-NN, Adaboost, Naïve Bayes-NB, and Quadratic Discriminant Analysis-QDA) to achieve the final predictive

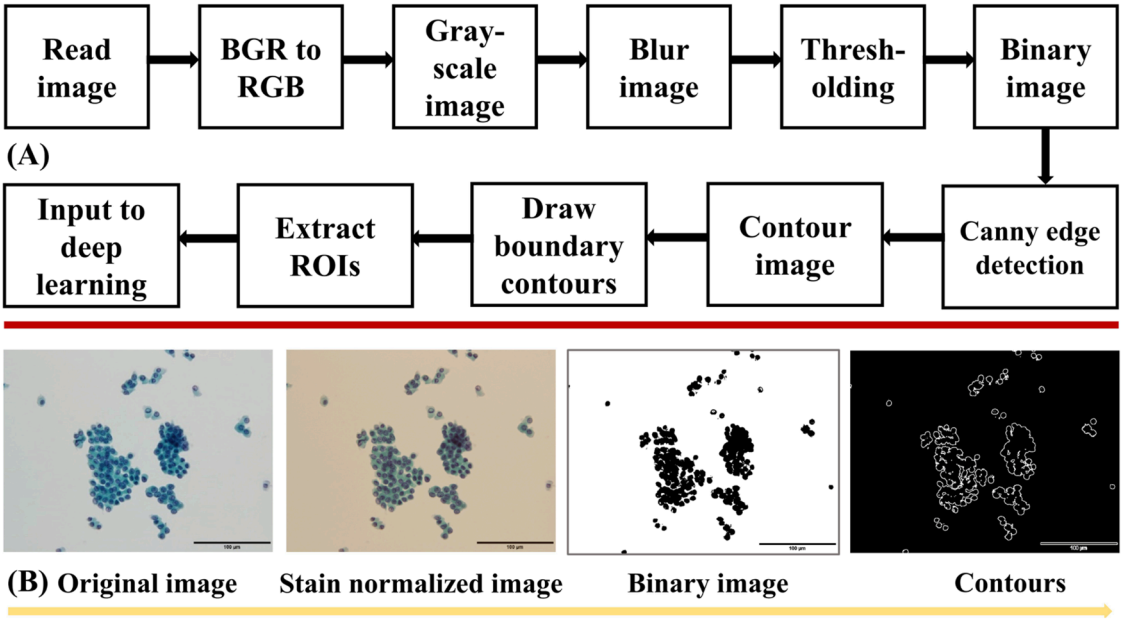

Fig. 3. (A) Multiple preprocessing steps were performed to remove noises and to extract the contours for automatic generation of fragment images and (B) resulted images obtained at individual preprocessing phases.

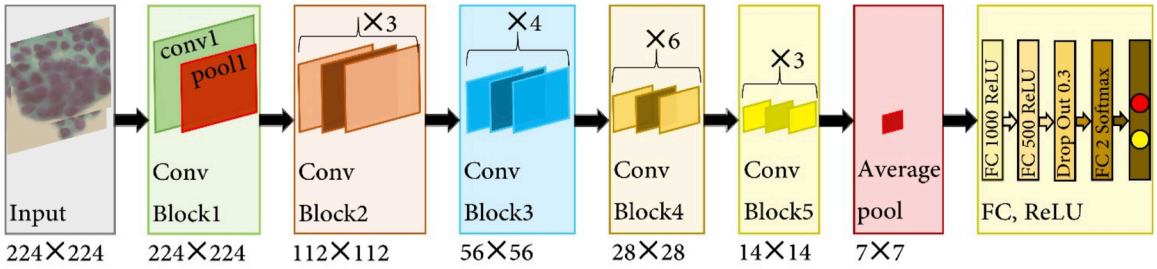

Fig. 4. Illustration of the CNN-based ResNet50 architecture, which has 50 convolution layers. The last Fully-Connected (FC) layer was reconfigured to two for binary classification as in our task. Dropout was set at 0.3.

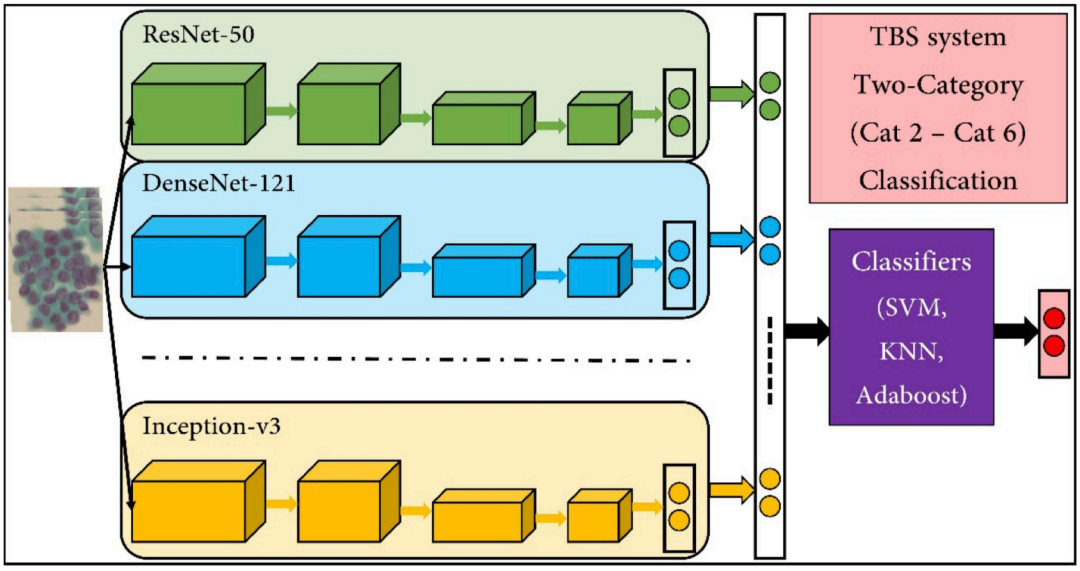

Fig. 5. The architecture for the ensemble deep learning model. Abbreviation: FC: Fully-Connected; SVM: Support Vector Machine; KNN: K-Nearest Neighbors.

output as illustrated in Fig. 5. These conventional machine learning algorithms are implemented using Scikit-learn package (https://scikit-learn.org/) which is the most useful and robust library for machine learning in Python. In this work, for simplicity, the parameter tuning process for selecting the best values for each classifier's parameters was not implemented as we set the default values. We set a threshold 0.5 to indicate the final class in our classification problem. Multiple well-established classifiers are relevant to evaluate the performace comparisons. These traditional supervised classifiers are presented as follows:

(i) Support Vector Machine (SVM) classifier: A SVM is a supervised machine learning model that uses classification algorithms for multi-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new data. A support vector machine takes these data points and outputs the hyperplane (decision boundary) that best separates the categories. In the present study, we utilized a linear SVM kernel with default gamma value of the kernel scale (γ) and the box constraint ($C = 1$).

(ii) K-Nearest Neighbor (KNN) classifier: KNN is a non-parametric, one of the utmost machine learning algorithms used in the variety of applications such as healthcare, handwriting detection, and image recognition. The training examples are vectors in a multidimensional feature space, each with a category label. The training phase involves storing the feature vectors and category labels of the training samples. In the classification phase, a new unlabeled test data is classified by assigning the label which is most frequent among the k training samples nearest to that test data with k is a user-defined parameter. Various commonly used distance metrics in KNN are Euclidean distance, Hamming distance, or correlation coefficients, such as Pearson and Spearman. The classification accuracy of k-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighborhood components analysis. In Scikit-learn package, the default value k was at 5 and weights funtion was set with uniform for KNN classifier implemented in this work.

(iii) Decision Tree (DT) classifier: DT algorithm is one of the easiest and commonly-used supervised learning algorithms that can be used for solving classification and regression problems. The objective of using DTs is to develop a training model that can forecast the category or value of the target variable by learning simple decision rules inferred from training data. Basically, in DTs, for predicting a category label for a record, the algorithm begins from the root node of the tree, followed by comparing the values of the root attribute with the record's attribute to decide the branch corresponding to that value and jump down the tree to some leaf/terminal node.

The key concern when implementing DT algorithm is to determine which attributes should be considered as the root node and each descending level. This process is known as attributes selection. For solving this challenging attribute selection problem, some criteria have been suggested including Entropy, Information Gain, Gini index, Gain Ratio. These criteria will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by descending order, i.e., the attribute with a high value is selected as the root. For DT classifier, we set the maximum depth of the tree at 5 (max_depth = 5) and left other parameters with their default values.

(iv) Random Forest (RF) classifier: RF is an ensemble learning method that builds multiple DT algorithms with bagging method and merges them together to achieve more accurate predictions and reduce the variances of the model, without increasing the bias. The concept of bagging method is constructing a multitude of DTs at training and outputting an outcome that is mean/average prediction of individual trees.

RF can be used to rank the importance degree of the features on the prediction. The feature ranking is operated as follows. The first step is to fit a random forest to the training data. During the fitting process the out-of-bag error for each feature is computed and averaged over the forest. To measure the importance of the k^{th} feature in the dataset after training, the values of this k^{th} feature are permuted, and the out-of-bag error is computed. The importance score for the k^{th} feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The higher importance score presents the more contribution of the corresponding feature in the dataset. By accessing the feature ranking, one will be able to optimize the prediction performance by possibly dropping some features that have low importance scores. For this RF classifier, we set the maximum depth of the tree at 5 (max_depth = 5), the number of trees in the forest at 10 (n_estimators = 10) and all other parameters were left at defaults.

(v) AdaBoost classifier: AdaBoost algorithm, short for Adaptive Boosting, is a boosting technique that is used as an ensemble method in machine learning. AdaBoost combines the outputs of other learning algorithms ("weak learners") with a weighting strategy to produce the final output with a stronger power. It is called adaptive boosting in a sense that weights are re-assigned to each weak learner, with higher weights to misclassified learners. The individual learners can be weak, but the final model can be proven to converge to a strong learner. AdaBoost is also used to reduce bias as well as variance for supervised learning. All of default values for AdaBoost classifier's parameters were configured.

(vi) Neural Network (NN) classifier: A Neural Network, generally referred to as artificial neural networks (ANNs), is the branch of artificial intelligence that works similarly to the biological brain's neural network. An ANN is composed by a collection of connected units or nodes called artificial neurons which artificially model the neurons in a human brain. Like the synapses in a biological brain, each connection can send a signal to other connected neurons. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear activation function of the sum of its inputs. A threshold is applied to each neuron such that a signal is sent only if the aggregate signal exceeds that threshold.

Neurons and connections have weights that can be adjusted by a learning process called backpropagation which determines how changing the weights impact the overall cost function in the neural network. Specifically, the weights that contribute more to the overall "error" (cost function) will have larger derivation values, which means that they will change more. Neurons are grouped into layers to form a neural net. A typical Neural Net is constructed from three type of layers: (1) Input layers: initial data for the neural network, (2) Hidden layers: intermediate layer between input and output layer and place where all the computation is performed, and (3) Output layer: produce the result for given inputs.

(vii) Naive Bayes classifier: Naive Bayes classifier is a straightforward and powerful classification technique based on Bayes' Theorem with an assumption of independence among predictors for simplicity. In other word, presence or absence of a feature does not influence the presence or absence of any other feature. The general idea of Bayes' Theorem is that the probability of an event occurring can be computed given the probability of an event that has already occurred. Naive Bayes classifier predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).

(viii) Quadratic Discriminant Analysis (QDA) classifier: QDA classifier is a statistical classifier that uses a quadratic decision surface to separate measurements of two or more classes of objects or events. It is a more general version of the linear classifier. Similarly, we used default values for QDA classifier.

2.5. Training strategy and experiment setup

The CNN deep learning models were trained, validated and tested by using PyTorch (ver.1.0.1, https://pytorch.org/), a Python deep learning library. In this work, we implemented a transfer learning network,

which had been pre-trained on the well-established large-scale ImageNet database for image classification. In our implementation, the last fully connected layer was re-configured at two classes (output dimension is changed from 1000 to 2 as we have two classes: PTC and Non-PTC). The parameters of the convolution layers in pre-trained networks were used as initial values. The weights in the pre-trained deep learning networks had been learned well through the large amount of images in the large-scale dataset (Russakovsky et al., 2015). Therefore, the transfer learning using the pre-trained weights was expected to learn faster than the scratched networks, avoid overfitting and the better performances are promised. To increase the number of images for training and obtain sufficient robustness, we applied on-the-fly data augmentation in which each training image was flipped to create a new training image. However, the data augmentation was not employed during the validation and testing phases.

We used the binary cross-entropy function as a loss function for binary classification. Adam was selected as an optimizer with a scheduled learning rate for precise training. One epoch is defined as performing backpropagation once for all images in the total training-set (70%) were learned. The deep learning networks were verified for each epoch using the validation-set (20%). Finally, the best models on the validation set were saved for testing on the left-out set (10%). The workflow of training strategy is described in Fig. 6. The large training was performed on a GPU-supported Intel ® Xeon ® Gold 6126 CPU @ 2.60 GHz server with 48 CPU cores, 187 GB RAM and 4 NVidia Titan XP GPUs with 64 GB VRAM. Multiple GPUs support strategic implementations of data parallelism that allows the CNN modules with a huge amount of trainable parameters to be automatically splitted into multiple models and loaded into multiple GPUs for fast training. After each model finishes their job, parallel data for efficient training were collected and merged for the final results.

2.6. Performance evaluation and significant test

i) Performance evaluation: To evaluate the performance of the classifiers, we report the predictive Accuracy (ACC), Sensitivity (SEN), Specificity (SPEC), Positive Predictive Value (PPV), and Negative Predictive Value (NPV) (N. T. Duc, Ryu, Choi, Iqbal Qureshi, & Lee, 2019; Nguyen et al., 2019). TP, TN, FP, and FN indicate the number of true positives, true negatives, false positives, and false negatives, respectively. ACC, SEN, SPEC, PPV, and NPV are computed as follows:

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \quad (1)$$

$$\mathrm{SEN} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP}) \quad (2)$$

$$\mathrm{SPEC} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \quad (3)$$

$$\mathrm{PPV} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP}) \quad (4)$$

$$\mathrm{NPV} = \mathrm{TN} / (\mathrm{TN} + \mathrm{FN}) \quad (5)$$

ii) Significant test: To minimize effects of random weight initializations, effects of random data splitting, and maximize the reliability of the predictive results, we performed the training 20 times. For each training set-up, we randomly selected subsamples of the dataset, retaining a balanced number of fragments in each class. By using a balanced dataset for training, validation, and testing the prediction performance, we increased the reliability of our approach and also minimized sensitivity of the model to specific training data. The means of the classification results were reported. We also performed permutation test (randomization) and the significant p-values for each test were reported. To assess the statistical significance of the classifiers' performance, a permutation test was performed on the classification accuracies, by randomly permuting 1000 times the labels of the test data to get the probability of random successful classification. In general, the lower the p-value of the permuted prediction rate against the prediction rate of the original data labels, the higher the significance of the classifier performance.

2.7. Fragment-level, FNAC Slide-level Classification, and certainty of the prediction

The entire FNAC slide predicted label ($\boldsymbol{L}$) (or patient-level predicted label) of the testing cytology image that has f_N fragments was computed as follows:

$$\boldsymbol{L} = \frac{\sum_{f_1}^{f_N} \boldsymbol{l}_{f_i}}{\boldsymbol{f_N}} \rightarrow \boldsymbol{L} = \boldsymbol{PTC}\ if\ \boldsymbol{L} \in [0.5, 1],\ else\ \boldsymbol{L} = \boldsymbol{Non-PTC} \quad (6)$$

where f_N is the number of fragments segmented from the original FNAC slide and l_{f_i} is the predicted label of the fragment f_i provided as the fragment-level predictive output of the deep learning model. Thus, $\boldsymbol{L}$=PTC$if\boldsymbol{L} \in [0.5, 1], else\boldsymbol{L}$ = Non-PTC. This Eq. (6) allows us to adapt the practical diagnosis from cytopathology in which the malignancy is carefully examined on FNAC slides.

Following the FNAC slide predicted label ($\boldsymbol{L}$), in order to provide

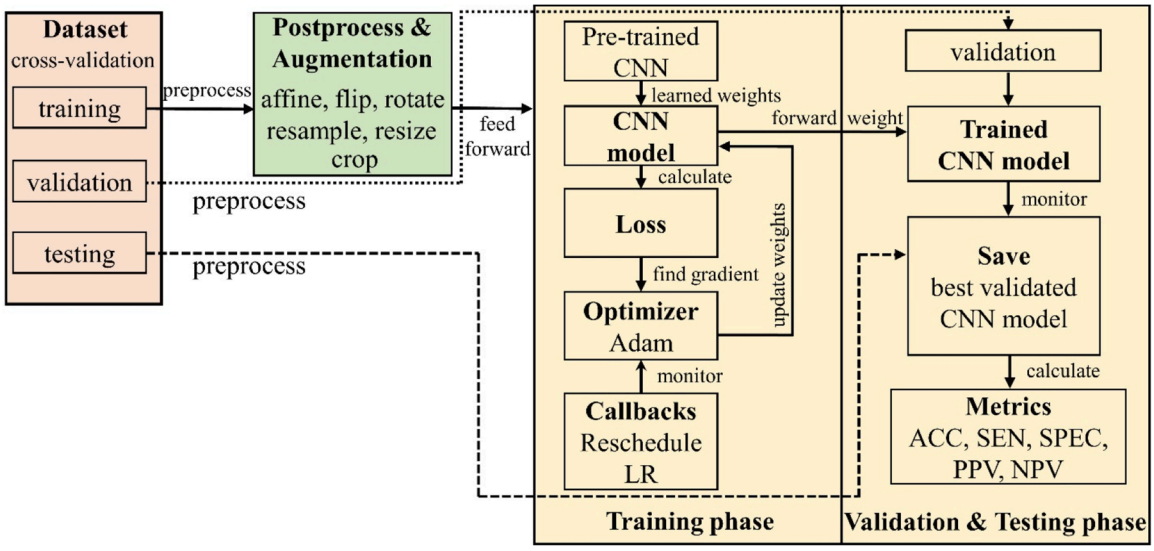

Fig. 6. Illustration of the proposed framework for training, validation and testing of deep CNNs. Abbreviation: CNN: Convolution Neural Network; LR: learning rate: ACC: accuracy; SEN: sensitivity; SPEC: specificity; PPV: positive predictive value; NPV: negative predictive value.

supplemental evidence to the cytopathologist as well as produce higher reliability and confidence of the prediction made by our deep learning models, we also proposed a way how to compute the certainty ($\boldsymbol{C}$) of the prediction as follows:

$$\boldsymbol{C} = \boldsymbol{L}^{*}100 \ if \ \boldsymbol{L} = PTC, \ else \ \boldsymbol{C} = (1 - \boldsymbol{L})^{*}100 \quad (7)$$

3. Results

3.1. Results of automatic smear extraction

Visualizations of fragments extracted from an example original high-resolution (4800 × 3600 pixels) PTC cytology FNAC image is provided in Fig. 7. As shown in Fig. 7, we were able to select the most relevant ROIs that are informative for clinical diagnosis according to the cytopathologists' suggestions, and remove the non-relevant areas (considered as artifacts) which include background and other tissues' types. As suggested by the cytopathologists, the extracted fragments should contain more than five cells to guarantee the quality of the clinical diagnosis. The Fig. 7 shows that all of the fragments have more than five cells and any fragments which do not qualify this criterion were discarded from the dataset. By doing so, we were able to support the cytopathologists in interpretation the results made by AI-supported CAD systems.

The other crucial advantage of the proposed automatic smear axtraction algorithm is that an additional dataset were obtained for sufficiently training the CNN-based deep learning models. Since the original FNAC dataset is quite limited for training the deep learning models, the generated fragments help increase the entire dataset significantly. This process can be widely implemented as a data augmentation technique in any AI studies in digital pathology researches where the chances to collect thousands of patients' samples is extremely low.

3.2. FNAC slide classification results

Table 2 provides FNAC slide classification results on the left-out test dataset, which were not involved in training and validation. We reported the mean ACC, SEN, SPEC, PPV, and NPV of six deep learning models. As can be seen, DenseNet161 achieved the best performance, obtaining a highest mean ACC of 0.9556 ($p < 0.01$), a highest mean SEN of 0.9734, and a mean SPEC of 0.9405, a mean PPV of 0.9697, and a mean NPV of 0.9371. In contrast, Inception-v3 performed worst among all six CNN models. On the other hand, ResNet101 resulted in the best SPEC value of 0.9503 while ResNet152 achieved a maximum PPV value of 0.9711.

In addition, our findings concluded that the training and validation time varied significantly depending on the number of convolution layers and models. Specifically, ResNet152 spent a longest time of around 176 min to finish the training while ResNet50 reached its convergence within only 40 min, a shortest period. Although the training periods were varied with different CNN models, the testing time was identically fast, about<5 s. Note that the results presented here were obtained on the dataset without the use of stain normalization methods. Even though there were no implementations of stain normalization, the obtained classification results look optimistic since predictive accuracies are about 0.95, except for the Inception-v3 model. This result means that the artificial intelligences are providing high performances in automatic predictions of the original FNAC images taken from multiple microscope systems.

3.3. Effects of stain normalization on predictions

In this part, the effects of stain normalization on automatic diagnosis of PTC tissues are analyzed. Here, we report the various evaluation metrics including ACC, SEN, SPEC, PPV, NPV among three stain normalization methods and six deep learning models as shown in Table 3, Table 4, Table 5, Table 6, and Table 7, respectively. Consequently, the stain color normalization method introduced by Reinhard et al. (Reinhard et al., 2001), maximized the predictive ACC, outperforming than histogram specification (Khan et al., 2014), and Macenko (Macenko et al., 2009) method by using all CNN models, except ResNet101 model.

In contrast, according to the results presented in the Tables 3–7, the normalization by histogram specification did not improve the prediction, but decreased them as compared to the results given by no normalization method. Therefore, as shown here, it is not recommended to normalize the original stained images by using histogram specification before the prediction.

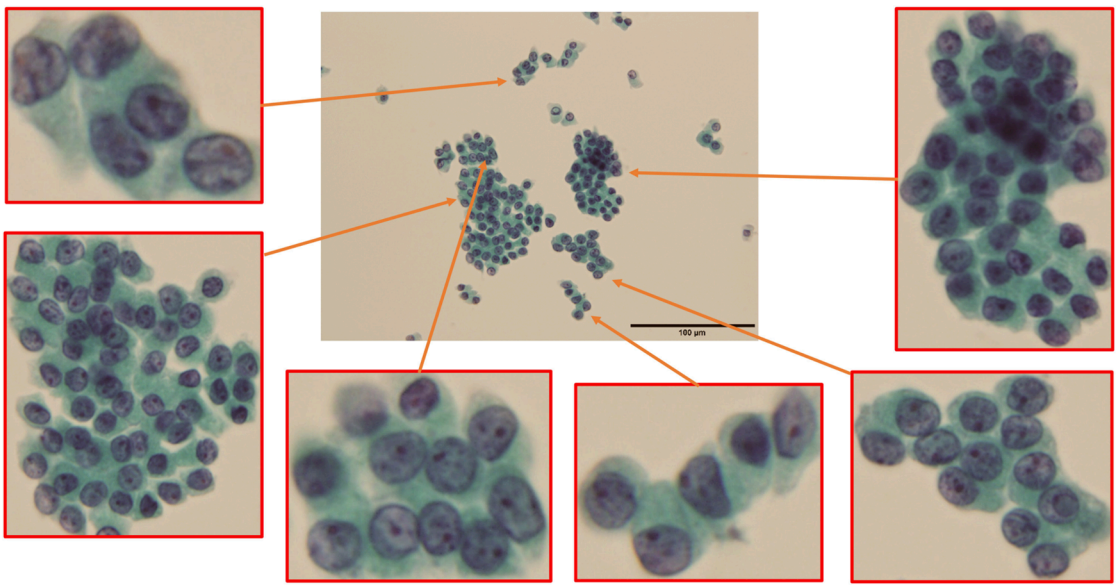

Fig. 7. A cytology FNAC image and its non-overlapped fragments that contain only regions of interest for making proper diagnosis. Fragments containing less than five cells are not sufficient for make pathologic diagnoses, and thus were discarded from the dataset.

Table 2
The FNAC slide classification results on yet-to-be-seen test data of six models presented in five evaluation metrics (ACC, SEN, SPEC, PPV, and NPV). The mean were reported, the p values presents the significance level of permutation test. None of the normalization methods has been implemented in this result. Bold values represent the maximum.

Models	ResNet50	ResNet101	ResNet152	DenseNet121	DenseNet161	Inception-v3
ACC	0.9469 ($p < 0.01$)	0.9512 ($p < 0.01$)	0.9501 ($p < 0.01$)	0.9342	**0.9556** ($p < 0.01$)	0.8939 ($p < 0.03$)
SEN	0.9650	0.9501	0.9500	0.9334	**0.9734**	0.9367
SPEC	0.9152	**0.9503**	0.9502	0.9357	0.9405	0.8128
PPV	0.9524	0.9710	**0.9711**	0.9623	0.9697	0.8979
NPV	0.9361	0.9155	0.9155	0.8888	**0.9371**	0.8797
Training time (minutes)	40	124	176	46	87	55
Testing time (seconds)	< 5	< 5	< 5	< 5	< 5	< 5

Table 3
Effects of stain normalization on the prediction accuracy (ACC).

Stain Normalization	ResNet50	ResNet101	ResNet152	DenseNet121	DenseNet161	Inception-v3
No_Norm	0.9469	0.9512	0.9501	0.9342	0.9556	0.8939
Norm_RBGHist	0.9150	0.9259	0.9376	0.8861	0.9358	0.8716
Norm_Reinhard	0.9519	0.9216	0.9510	0.9493	0.9694	0.9156
Norm_Macenko	0.9493	0.9517	0.9469	0.9482	0.9590	0.8986

NO_NORM = no stain normalization; **NORM_RBGHIST** = Histogram Specification method;
NORM_REINHARD = Reinhard normalization method; **NORM_MACENKO** = Macenko normalization method.

Table 4
Effects of stain normalization on the prediction sensitivity (SEN).

Stain Normalization	ResNet50	ResNet101	ResNet152	DenseNet121	DenseNet161	Inception-v3
No_Norm	0.9650	0.9501	0.9500	0.9334	0.9734	0.9367
Norm_RBGHist	0.9339	0.9422	0.9505	0.8663	0.9505	0.8382
Norm_Reinhard	0.9631	0.9529	0.9821	0.9553	0.9770	0.9196
Norm_Macenko	0.9260	0.9527	0.9327	0.9551	0.9462	0.8969

Table 5
Effects of stain normalization on the prediction specificity (SPEC).

Stain Normalization	ResNet50	ResNet101	ResNet152	DenseNet121	DenseNet161	Inception-v3
No_Norm	0.9152	0.9503	0.9502	0.9357	0.9405	0.8128
Norm_RBGHist	0.8921	0.9060	0.9220	0.9100	0.9185	0.8747
Norm_Reinhard	0.9417	0.8926	0.8694	0.9395	0.9399	0.8901
Norm_Macenko	0.9765	0.9504	0.9634	0.9475	0.9740	0.8695

Table 6
Effects of stain normalization on the positive predictive value (PPV).

Stain Normalization	ResNet50	ResNet101	ResNet152	DenseNet121	DenseNet161	Inception-v3
No_Norm	0.9524	0.9710	0.9711	0.9623	0.9697	0.8979
Norm_RBGHist	0.9129	0.9239	0.9366	0.9211	0.9335	0.8896
Norm_Reinhard	0.9554	0.9198	0.9682	0.9552	0.9697	0.9080
Norm_Macenko	0.9786	0.9572	0.9674	0.9509	0.9769	0.8888

Table 7
Effects of stain normalization on the negative predictive value (NPV).

Stain Normalization	ResNet50	ResNet101	ResNet152	DenseNet121	DenseNet161	Inception-v3
No_Norm	0.9361	0.9155	0.9155	0.8888	0.9371	0.8797
Norm_RBGHist	0.9177	0.9283	0.9389	0.8488	0.9387	0.8168
Norm_Reinhard	0.9468	0.9245	0.8690	0.9310	0.9421	0.8575
Norm_Macenko	0.9189	0.9454	0.9248	0.9475	0.9396	0.8786

3.4. Ensemble deep learning enhances the Predictions' accuracy

Table 8 presents the classification results of ensemble deep learning models in which its architecture was illustrated in Fig. 5. In the ensemble learning models, we evaluated eight different conventional machine learning algorithms; that are, SVM, KNN, Decision Tree, Random Forest, Neural Net, AdaBoost, Naïve Bayes, and QDA, in which one algorithm ensembles the probability outputs of six different CNN deep learning models to make a final decision. Results using three stain normalization methods were also computed and for each ensemble learning configuration, three best models are highlighted in bold.

As can be seen from Table 8, Decision Tree, Random Forest, and

Table 8
The ensemble learning results. Abbreviation: SVM = Support Vector Machine; KNN = K-Nearest Neighbors. QDA = Quadratic Discrimination Analysis. Three best ensemble learning models are shown in bold.

Ensemble learning model	No_Norm	Norm_RBGHist	Norm_Reinhard	Norm_Macenko
SVM	0.9608	0.9494	0.9646	0.9685
KNN	0.9650	0.9621	0.9675	0.9691
Decision Tree	**0.9799**	**0.9711**	**0.9869**	**0.9808**
Random Forest	**0.9746**	**0.9711**	**0.9758**	**0.9749**
Neural Net	0.9565	0.9403	0.9513	0.9627
AdaBoost	**0.9830**	**0.9702**	**0.9971**	**0.9726**
Naïve Bayes	0.9565	0.9312	0.9513	0.9627
QDA	0.9565	0.9439	0.9542	0.9627

Adaboost algorithms always ensure the best performance among eight algorithms. In addition, our findings indicated that ensemble learning models are able to boost the discriminative accuracy, showing the improved classification accuracies (up to 0.9971 by AdaBoost classifier) as compared to individual CNN deep learning models presented in Table 3. Furthermore, stain normalization using Reinhard method proved its efficacy, as its performances are superior to that of stain normalization using Macenko or histogram specification methods.

3.5. FNAC level prediction, certainty and feature representation

Fig. 8 presents prediction results of the four non-overlapped fragments that were automatically generated from a Non-PTC FNAC image using the best-proposed deep DensetNet161 model as given in Table 2. Among four fragments, three fragments (3/4 fragments were predicted as Non-PTC, resulting in $\boldsymbol{L}$ = Non-PTC) were correctly predicted as Non-PTC class and only one fragment was wrongly diagnosed as PTC, thus, according to the Eq. (6) the FNAC slide was correctly classified as Non-PTC class with certainty probability of the prediction made by deep learning-based system is $\boldsymbol{C}$ = 75% certainty based on the Eq. (7).

Moreover, we highlighted the feature representations of the regions of interest (by yellow colour on the right subfigures) using Gradient-weighted Class Activation Mapping (Grad-CAM) method (Selvaraju et al., 2019) which are likely affected on the prediction results. Grad-CAM is a method for visualization of 'visual explanations' for decisions from a large class of CNN models, highlighting the high representations of differential hierarchical features. Basically, the Grad-CAM approach utilizes the gradients of any target concept (say logits for 'PTC' or 'Non-PTC'), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept (Araujo et al., 2017). What presented in Fig. 8 was the Grad-CAM localization given by the last convolution layer of the DenseNet161 model that provided the best discriminative performance.

4. Discussion

4.1. Artificial intelligence for cytology PTC diagnosis and clinical remarks

Papillary Thyroid Carcinoma (PTC) is one of the leading causes of cancer-related surgery in many countries and thus its early diagnoses are of importance to provide optimal treatments. Microscopic cytopathology remains the gold standard in diagnostic surgical pathology; however, the major limitation to morphological diagnosis is diagnostic variability among pathologists (Kanavati et al., 2020). This downside may lead to a negative impact on the quality of the pathological diagnosis, which should be resolved essentially with standard diagnosis approaches to diminish the variations in digital pathology.

In this paper, we presented an AI and computer vision technique for automatic classification of thyroid malignant tissues from the benign ones using FNAC slides obtained from microscope systems. We were able to obtain a high discriminative result. Through this pilot study, we believe that prospective of AI in digital PTC pathology will be grown, approved in national, and worldwide in the near future to greatly support the pathologist in accurate diagnosis. This could be possible due to several following evidences:

i. PTC is the malignancy showing the most typical cytological features in the surgical pathologic field in practice such as oval nuclei, nuclear overlapping, nuclear membrane irregularity, powdery chromatin, chromatin margination, nuclear grooves, and nuclear pseudo inclusions
ii. The diagnosis of PTC is made by FNAC, a rapid, easy, reliable, and well-recognized technique performed with punctuating and

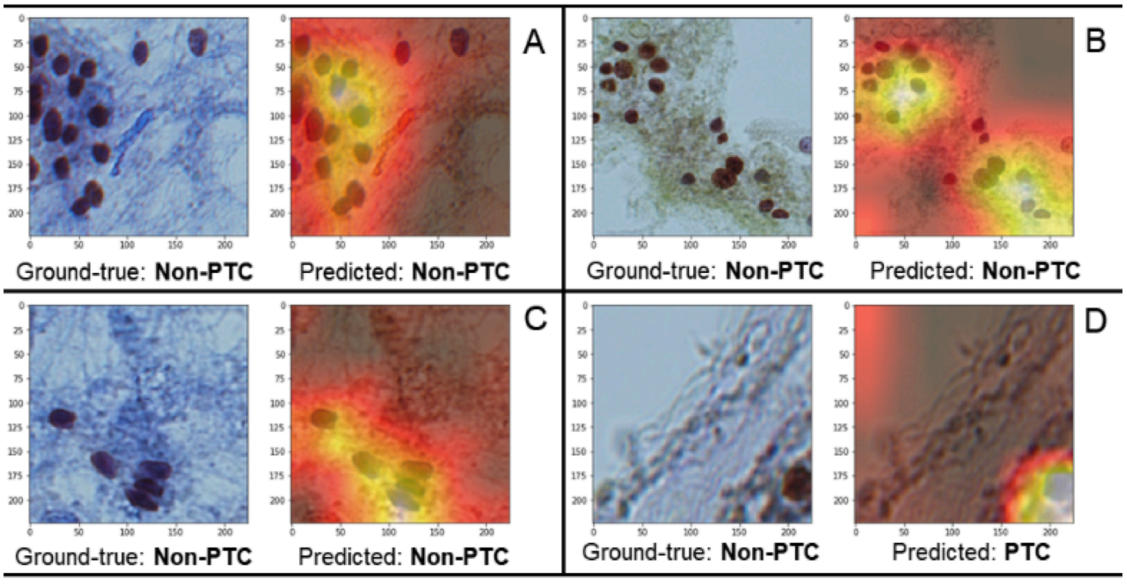

Fig. 8. Fragment-wise and FNAC slide diagnosis on an unseen test Non-PTC class image. The trained CNN model is able to classify the individual fragments (A, B, C, D) independently. Then, the decision for FNAC slide classification was performed based on the Eq. (6). We also implemented GRAD-CAM model to highlight (by heat map) the differential areas that dominantly effect on the diagnostic decision.

aspirating the thyroid nodule in the neck being targeted by bedside ultrasonography safely.

iii. FNAC slides have been standardized worldwide using ThinPrep, combining with advanced computer vison, big data and image processing techniques makes the FNAC images standardized for the diagnosis.

4.2. Automatic smear extraction algorithm and its limitations

In addition to the ensemble learning framework proposed in this work, an automatic smear extraction algorithm which automatically extract the fragments containing only the regions of interest (ROI) of thyroid tissues from original high-resolution cytology FNAC images was also introduced. The concept is that the original high-resolution cytology FNAC images might contain different information including: (1) thyroid tissues which are truly relevant for diagnostic pathology, (2) white-colour background which are not relevant for clinical diagnosis and considered as artifacts, and (3) red blood tissues and other types of fluids; which are NOT relevant for pathologic diagnosis and considered as artifacts.

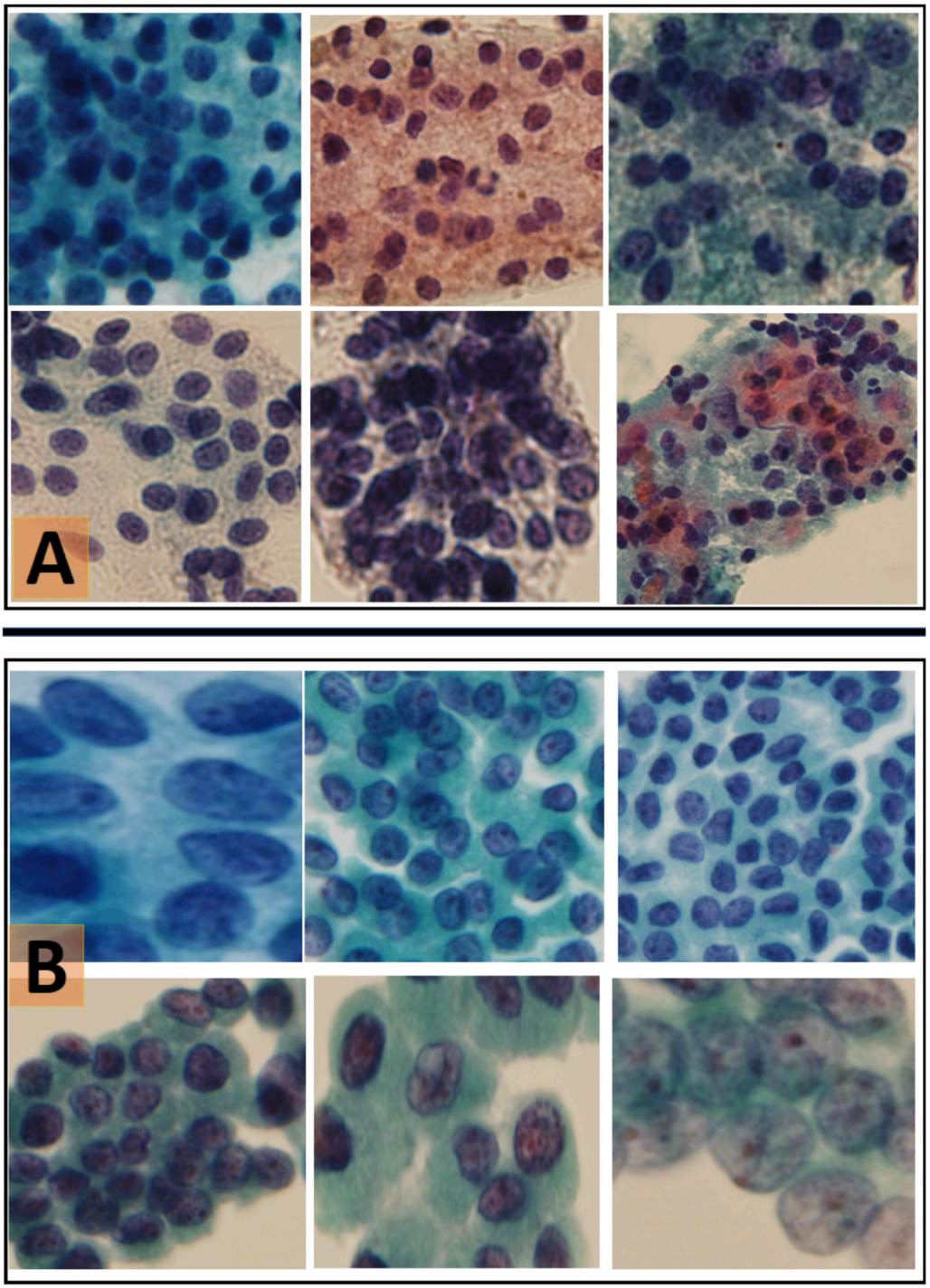

Fig. 9. Examples of ROI-based fragments successfully extracted from original high-resolution FNAC images using automatic smear extraction algorithm. These fragments are used for CNN-based deep learning predictions. The upper panels are fragments from FNAC images of Category II while the lower panels are these of Category VI.

As can be seen in the original FNAC images, there are white-colour background areas which should be discarded for further automatic AI-based analysis, keeping only thyroid cell clusters for proper pathologic diagnosis of PTC. In addition, ROI fragments that contained red blood cells and other fluid artifacts should be discarded because they cause confusions for pathologists in identifications of true PTC cells, as well as they provide noises to the deep learning models.

The automatic smear algorithm is capable to cluster ROIs in FNAC images that contains only thyroid cells. However, our automatic smear algorithm is not always successful in detecting the relevant thyroid cell clusters in all FNAC images. The typical reason for the failures is that in our dataset, a few FNAC images might contain multiple cell types and artifacts fluids, other than the target thyroid cell clusters. Here, we proposed an AI-based pipeline which is the automatic solution for binary classification that performed PTC cell identifications among non-PTC thyroid cells. Therefore, corresponding automatic pre-processing steps are not ready for multiple cell types detections, other than thyroid cells. This is also a limitation in this work that we aim to solve in our future studies to move a step forward to practically clinical deployment.

Therefore, due to these limitations, we have to have a quality-controlled process where the pathologists carefully re-annotate of the generated fragment to ensure that more reliable and accurate labels

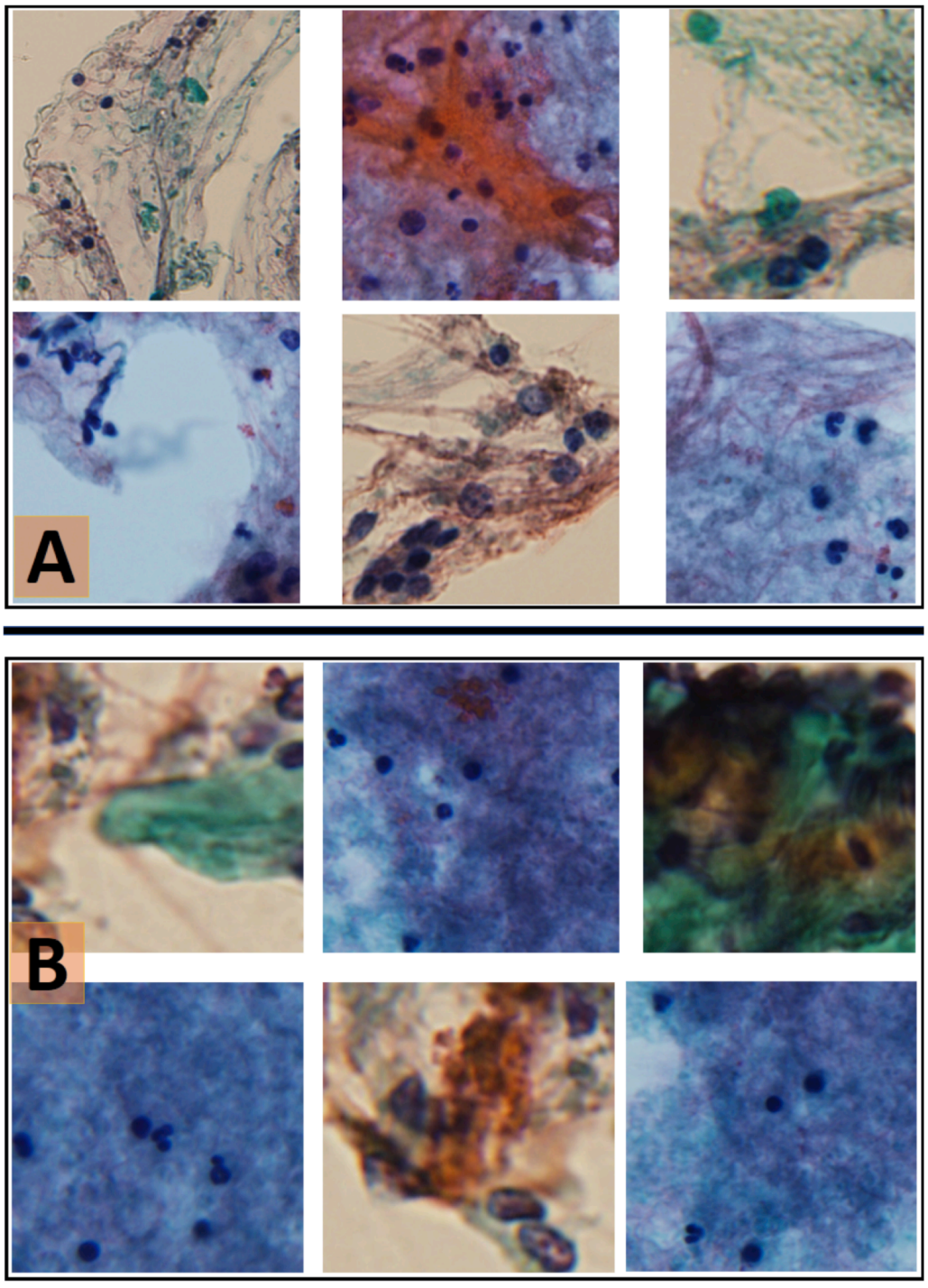

Fig. 10. Examples of ROI-based fragments unsuccessfully extracted from original high-resolution FNAC images using automatic smear extraction algorithm. These fragments are discarded for further predictions. The upper panels are fragments from FNAC images of Category II while the lower panels are these of Category VI.

should be captured before further AI-powered prediction. Six successfully extracted ROI-based fragments and six failed ROIs detected by automatic smear algorithm are presented here in Figs. 9 and 10. In both figures, the upper panels (A) represents the fragments of Category II and the lower panels (B) represents the fragments of FNAC images of Category VI.

Regarding to the Canny edge detection, this is one of the step in the pre-processing pipeline that actually detects the edge/boundary of the cell clusters in each ROIs in the high-resolution FNAC samples. After that, based on the edge clusters of each ROIs, we drew the boundary contours for ROI extractions. Then, from an original FNAC images, multiple ROI images are generated in which each ROI image is a cluster of multiple thyroid cells contoured by Canny edge detection algorithm.

The next step is to discard ROI images that contained less than five cells as we discussed in the criterion strategy for proper pathologic diagnosis. Translating the criteria from the cytopathologists into the algorithm, in our initial findings, we observed that fragments sized of 200×200 pixels should be enough to cover the informative tissue structures.

4.3. Uncertainty Estimations of AI models in Digital Pathology

In our work, to provide a supplemental evidence to the cytopathologist as well as produce higher reliability and confidence of the prediction made by our deep learning models, we also proposed a way how to compute the Certainty (C) level of the prediction. This fragment-driven certainty estimate was recommended by our cytopathologists (coauthors) with the idea that we are developing and validating an AI-supported CAD system for assisting cytopathologists in making clinical diagnosis of PTC patients using the entire FNAC slide, a level of certainty should be provided for the support in decision making to see how many clusters (in percentage) have similar prediction results as the final prediction. In other words, for the practical implementation, the prediction should be made at patient-level, not at fragment-level. Each patient-level FNAC contains hundreds of fragments in which each fragment contains a cluster of cells. In principle, the final prediction certainty will be increased if more fragments are accurately predicted and vice versa. Basically, the certainty level proposed here illustrates the effectiveness of the AI model in predicting the entire relevant clusters in an entire FNAC slide.

In the future works, we are aiming to introduce the uncertainty estimations of the AI-supported model, which is a very challenging task in deep learning theory, in digital pathology. In real clinical application, an AI model for cytologic diagnosis should not only care about the accuracy but also about how certain the prediction is. If the uncertainty is too high, a cytopathologist would take this into account in his decision process. Generally, there are different types of uncertainty in deep learning: structural uncertainty, uncertainty in model parameters, epistemic uncertainty, aleatoric uncertainty, and heteroscedastic uncertainty. Structural uncertainty deals with how AI model structures (i.e. architectures) are used and the accuracy with which they extrapolate information, while uncertainty in model parameters considers the parameters (configuration variables internal to models) chosen to make predictions from a given corpus. Epistemic uncertainty describes what the model does not know because training data was not appropriate. Epistemic uncertainty is due to limited data and knowledge. On the other hand, aleatoric uncertainty is the uncertainty arising from the natural stochasticity of observations. Aleatoric uncertainty cannot be reduced even when more data is provided. Input data-dependent uncertainty is known as heteroscedastic uncertainty.

There have been several advanced proposed methods to deal with model-related uncertainty in deep learning community. The utmost well-established method dealing with uncertainty prediction is Bayesian inference which occupied around 27% of all uncertainty studies. Bayesian modeling allows us to quantify different types of uncertainty. Bayesian deep learning applies the Bayesian framework to deep models and allows estimating epistemic and aleatoric uncertainties as part of the prediction. Such estimates can indicate the likelihood of prediction errors due to the influence factors. Bayesian inference addresses structural uncertainty and uncertainty in parameters while integrating prior knowledge, but it's computationally demanding. Because of lack of suitable experimental data, the effectiveness of Bayesian uncertainty estimation in digital pathology applications with varying levels of influence factors has not yet been systematically studied.

4.4. Comparison with State-of-the-art Results

In recent years, several state-of-the-art studies have been carried out to diagnose the papillary thyroid carcinoma as well as other types of carcinomas at different organs from FNAC slides using either advanced deep learning models (Araujo et al., 2017; Aresta et al., 2019; Gopinath & Shanthi, 2013b; Guari et al., 2019; Kanavati et al., 2020; Mukherjee et al., 2018; Teramoto et al., 2019) or conventional machine learning (Gopinath & Shanthi, 2013a, 2013b, 2015). Studies based on the use of binary classification reported accuracies from about 85.06% to about 98.01%. Table 9 summarizes the predictive performances of recently published studies that discriminate carcinoma malignant tissues from benign ones using FNAC samples and compares these findings with our results. It should be noted that our advanced ensemble learning outperformed the ones proposed in (Araujo et al., 2017; Aresta et al., 2019; Gopinath & Shanthi, 2013b; Guari et al., 2019; Kanavati et al., 2020; Mukherjee et al., 2018; Teramoto et al., 2019). However, direct performance comparison with other studies would not be fair, because of the different datasets, preprocessing pipelines, feature measures, and classifiers.

4.5. Limitations and Future Perspectives

This work has several shortcomings. First, six thyroid cancer Categories diagnosed from cytopathology images according to The Bethesda System (TBS) were not fully considered as our model was currently limited to binary classification problem which aimed at predicting PTC malignant tissues among the Non-PTC benign ones. It would be more valuable for cytopathologists if the CAD system supported deep learning

Table 9
Comparison OF CLASSIFICATION PERFORMANCES. ABBREVIATION: ANN = ARTIFICIAL NEURAL NETWORK; DT = DECISION TREE; ENN = ELMAN NEURAL NETWORK; REF = REFERENCES.

Types	Dataset	Classifier	ACC(%)	Ref
PTC	FNAC smears	ANN	85.06	[7]
Breast carcinoma	Histology images	CNN + SVM	88.90	[2]
Lung Carcinoma	Cytological images	CNN	89.30	[5]
Thyroid carcinoma	FNAC images	DT + KNN	90.00	[10]
Thyroid carcinoma	FNAC images	ENN	93.33	[9]
Thyroid carcinoma	FNAC images	SVM	96.70	[8]
PTC	Cytological images	VGG-16	97.66	[6]
Lung Carcinoma	Whole-Slide Image	EfficientNet-B3 CNN	98.10	[4]
PTC	**Cytological images**	**Ensemble deep learning**	**99.71**	**Our study**

models are able to predict FNAC images into six TBS categories with high and reliable accuracies.

Second, localizations of differential fine-grained regions of interest tissues in six TBS categories were not considered to provide further relevant clinical information to the cytopathologists. Third, the ensemble learning strategy that integrates multiple trained deep CNN classifiers was not implemented for classifications of six TBS categories. Fourth, this study analyzed on the FNAC images, and the analyses using high-resolution whole-slide FNAC images were not considered.

Finally yet importantly, the critical downside of this work is lack of multiple validations evaluated with different external clinical cohorts. Therefore, at this present phase, this study was designed as a proof-of-concept or technical feasibility study without thorough external validation of real-world clinical performance. In subsequent works, those shortcomings would be investigated to move closer to the concept of a fully and highly accurate CAD system that supports cytopathologists in automatic analyses of PTC cancer using cytology images.

5. Conclusion

In this work, the goal for timely and accurate diagnosis of H&E stained cytopathology papillary thyroid carcinoma (PTC) using fine needle aspiration cytology (FNAC) image processed by ThinPrep was extensively addressed by utilizations of artificial intelligence techniques. Our proposed deep learning network is efficient to automatically extract the informative features characterized in PTC smears, neglecting the need of field knowledge that are normally required in current CAD systems. The novel findings also suggest that the proposed ensemble deep learning framework is robust and promising to achieve high predictive ability using cytology dataset which can be integrated into existing CAD systems for automatic PTC diagnosis and are capable for greatly supporting the cytopathologist for clinical cancer diagnosis. Our subsequent studies will be focused on automatic diagnosis of six thyroid cancer TBS categories with high accuracies from whole-slide high-resolution FNAC images. For practical deployment in hospitals for pathology uses, our deep learning models proposed in this work can be integrated into CAD systems for real-time diagnosis.

CRediT authorship contribution statement

Nguyen Thanh Duc: Conceptualization, Methodology, Formal analysis, Software, Writing - original draft, Writing - review & editing. **Yong-Moon Lee:** Conceptualization, Data curation, Writing - original draft, Writing - review & editing. **Jae Hyun Park:** Conceptualization, Data curation, Writing - original draft, Writing - review & editing. **Boreom Lee:** Conceptualization, Writing - review & editing, Funding acquisition, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by “GIST Research Institute(GRI) IIBR” grant funded by the GIST in 2021 and also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, Ministry of Science and ICT) (No. 2020R1G1A110088211).

References

Akhtar, M., Ali, M. A., Huq, M., & Bakry, M. (1991). Fine-needle aspiration biopsy of papillary thyroid carcinoma: Cytologic, histologic, and ultrastructural correlations. *Diagnostic Cytopathology, 7*(4), 373–379. https://doi.org/10.1002/dc.2840070410

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., … Sapino, A. (2017). Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One, 12*(6), e0177544. https://doi.org/10.1371/journal.pone.0177544

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., … Aguiar, P. (2019). BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis, 56*, 122–139. https://doi.org/10.1016/j.media.2019.05.010

Cibas, E. S., & Ali, S. Z. (2017). The 2017 bethesda system for reporting thyroid cytopathology. *Thyroid, 27*(11), 1341–1346. https://doi.org/10.1089/thy.2017.0500

Duc, N. T., & Lee, B. (2019). Microstate functional connectivity in EEG cognitive tasks revealed by a multivariate Gaussian hidden Markov model with phase locking value. *Journal of Neural Engineering, 16*(2), 026033. https://doi.org/10.1088/1741-2552/ab0169

Duc, N. T., & Lee, B. (2020). Decoding brain dynamics in speech perception based on EEG microstates decomposed by multivariate gaussian hidden markov model. *IEEE Access, 1–1*. https://doi.org/10.1109/access.2020.3015292

Duc, N. T., Ryu, S., Choi, M., Iqbal Qureshi, M. N., & Lee, B. (2019). Mild cognitive impairment diagnosis using extreme learning machine combined with multivoxel pattern analysis on multi-biomarker resting-state FMRI. *Conf Proc IEEE Eng Med Biol Soc, 2019*, 882–885. https://doi.org/10.1109/EMBC.2019.8857623

Duc, N. T., Ryu, S., Qureshi, M. N. I., Choi, M., Lee, K. H., & Lee, B. (2020). 3D-deep learning based automatic diagnosis of alzheimer’s disease with joint MMSE prediction using resting-state fMRI. *Neuroinformatics, 18*(1), 71–86.

Gopinath, B., & Shanthi, N. (2013a). Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images. *Australasian Physical & Engineering Sciences in Medicine, 36*(2), 219–230. https://doi.org/10.1007/s13246-013-0199-8

Gopinath, B., & Shanthi, N. (2013b). Support Vector Machine based diagnostic system for thyroid cancer using statistical texture features. *Asian Pacific Journal of Cancer Prevention, 14*(1), 97–102. https://doi.org/10.7314/apjcp.2013.14.1.97

Gopinath, B., & Shanthi, N. (2015). Development of an Automated Medical Diagnosis System for Classifying Thyroid Tumor Cells using Multiple Classifier Fusion. *Technology in Cancer Research & Treatment, 14*(5), 653–662. https://doi.org/10.7785/tcrt.2012.500430

Guan, Q., Wang, Y., Ping, B.o., Li, D., Du, J., Qin, Y.u., … Xiang, J. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: A pilot study. *Journal of Cancer, 10*(20), 4876–4882.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 770-778. 10.1109/cvpr.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. 2261-2269. 10.1109/cvpr.2017.243.

Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., … Tsuneki, M. (2020). Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-66333-x

Khan, A. M., Rajpoot, N., Treanor, D., & Magee, D. (2014). A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. *IEEE Transactions on Biomedical Engineering, 61*(6), 1729–1738. https://doi.org/10.1109/tbme.2014.2303294

Kim, Y. D., Noh, K. J., Byun, S. J., Lee, S., Kim, T., Sunwoo, L., … Park, S. J. (2020). Effects of Hypertension, Diabetes, and Smoking on Age and Sex Prediction from Retinal Fundus Images. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-61519-9

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Xiaojun, G., . . . Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. 1107-1110. 10.1109/isbi.2009.5193250.

Mukherjee, T., Sanyal, P., 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. *Journal of Pathology Informatics, 9*(1), 43. https://doi.org/10.4103/jpi.jpi_43_18

Nguyen, D. T., Ryu, S., Qureshi, M. N. I., Choi, M., Lee, K. H., Lee, B., & Yan, J. (2019). Hybrid multivariate pattern analysis combined with extreme learning machine for Alzheimer’s dementia diagnosis using multi-measure rs-fMRI spatial patterns. *PLoS ONE, 14*(2), e0212582. https://doi.org/10.1371/journal.pone.0212582

Reinhard, E., Ashikhmin, N., Gooch, B., & Shirley, P. (2001). Color transfer between images. *Ieee Computer Graphics and Applications, 21*(5), 34–41.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L.i. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision, 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. 2818-2826. 10.1109/cvpr.2016.308.

Teramoto, A., Yamada, A., Kiriyama, Y., Tsukamoto, T., Yan, K.e., Zhang, L., … Fujita, H. (2019). Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. *Informatics in Medicine Unlocked, 16*, 100205. https://doi.org/10.1016/j.imu.2019.100205