DEEP LEARNING-BASED MEDICAL IMAGING: A COMPREHENSIVE APPROACH TO MULTI-LABEL CLASSIFICATION OF THYROID FNAB CYTOLOGICAL IMAGES

Hai Pham-Ngoc*

Data Science Laboratory

Falculty of Mathematics, Mechanics and Informatics

Vietnam National University, Hanoi, Vietnam

Phuong Le-Hong†

Data Science Laboratory

Falculty of Mathematics, Mechanics and Informatics

Vietnam National University, Hanoi, Vietnam

October 23, 2024

ABSTRACT

This study addresses the automation of multi-label thyroid cytopathology classification based on the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC). Our goal is to enhance diagnostic performance while ensuring model interpretability and computational efficiency, targeting three critical labels: benign (B2), suspicious for malignancy (B5), and malignant (B6). We propose a novel deep learning approach incorporating data augmentation and a patient-level diagnosis framework, aimed at improving core feature learning and minimizing noise. Our ThyroidEffi Basic model achieves a macro average AUC of 0.9638, with individual label AUCs of 0.9830 for B2, 0.9499 for B5, and 0.9585 for B6. The macro F1 score reaches 0.8919. In comparison, the upgraded ThyroidEffi Premium model, which processes 12 patches, achieves a higher F1 score of 0.8977, though at the cost of increased computational demand. Additionally, model interpretability is improved using GradCAM to highlight essential image regions contributing to classification. These results demonstrate the potential for deploying our approach in clinical environments, provides efficient and accurate predictions for slide images as well as the potential application of this process with WSI.

***K*eywords** Thyroid Cancer · Medical Image Classification · Diagnosis · Cytopathology · Fine-Needle Aspiration Biopsy · Multi-Label · The Bethesda System · Deep Learning · Machine Learning · Artificial Intelligence

1 Introduction

1.1 Background

Thyroid disorders, particularly **thyroid nodules and cancer**, are among the most common endocrine malignancies worldwide [1]. **Fine needle aspiration biopsy (FNAB)** is a minimally invasive technique and is widely used to diagnose thyroid abnormalities before making important treatment decisions, such as surgery, ... [2]. "**The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC)** is a significant step to standardize the reporting of thyroid fine needle aspiration (FNA). The distribution of cases in various TBSRTC categories is as follows: I—undiagnosed 13.8%, II—benign 75.9%, III—atypical of undetermined significance (AUS)/undetermined follicular lesion level of significance (FLUS) 1.2%, IV—follicular neoplasm (FN)/suspicious for cystic neoplasm (SFN) 3.7%, V—suspicious

*Website: harito.id.vn, Email: harito.work@gmail.com

†Website: phuonglh.com, Email: phuonglh@vnu.edu.vn

for malignant disease (SM) 2.6% and VI—malignant 2.8%". The 6 classification labels in TBSRTC ensure scientific and uniform standards for FNAB diagnosis [3].

Traditional diagnostic methods rely heavily on the expertise of cytopathologists, which often leads to discrepancies in diagnosis [4]. In addition, there is a shortage of cytopathologists in medical facilities. These have placed a great demand on automating the diagnostic process of cytology images from fine needle aspiration biopsy (FNAB). Recently, deep learning has emerged as a powerful tool in medical imaging, demonstrating remarkable success in automating the classification of complex images such as ultrasound images, X-ray images, and CT images [5]. However, its application in cytological imaging, especially for multi-label classification (more than 2: benign - malignant) with TBSRTC, has not been fully explored and requires further research for application [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17].

1.2 Motivation

Current research still has many limitations, specifically:

- Only binary classification of benign or malignant. Or the dataset has many labels according to The Bethesda System, however due to poor multi-label classification results, the main conclusions about the results are still binary classification.
- The data set is small so it is difficult to confirm generality on a large scale.
- Studies only provide some predictions on each image in a subset of images cutting the area of interest, and cannot make predictions for patients. Specifically, for studies using imaging microscopes, they only make predictions at the image level of the region of interest (also called fragments). As for studies using scanners to capture entire slides, they only make predictions at the level of cross-sectional images of the area of interest (also known as slide image).
- Most tend to use heavy modeling approaches; or the process to infer the diagnostic results has too many steps and sub-models, causing complexity and taking a long time to obtain the output.

1.3 Aim and Objectives

Due to the above limitations and the needs of today's physicians, this study aims to improve the performance of automatic multi-label classification at the patient-wide level. with a quick process and ensuring good interpretation of prediction results. Thereby building trust with doctors and being able to apply research results in a real working environment.

To achieve these purposes, we propose deep learning-based approaches for classifying multi-label thyroid FNAB images in The Bethesda System. Our approach aims to:

- improve the performance of automated multi-label thyroid disease diagnosis. Namely 3 labels benign, suspicious for malignancy and malignant (respectively levels 2, 5, 6 in The Bethesda System)
- more reliably (e.g. if binary classification is required, the accuracy and f1 score macro must be 97% or higher)
- accurately diagnose the final result for the patient
- simplifies the model sizing and inference process (model must infer less than 1 second per patient with existing hardware infrastructure)
- ensure the explainability of the model (e.g. automatically clearly identify image regions with high probability of belonging to the predicted label)

1.4 Novel Contributions

Besides overcoming the limitations of previous studies, our new contributions are:

- Established **new data augmentation method** so that the model learns better from core input features, limiting the influence of noisy features
- Established **new training process** so that the model can make good decisions immediately for patient-level images, limiting the impact of data imbalance between labels
- Testing multiple model architectures to find the **balance between classification performance and computational efficiency**

- Learn **apply sequence analysis models** such as Transformer, ... to make the final diagnosis of the patient (discover great application potential for either images from microscope - slide images - or images from automated slide scanners - whole slide images: WSI)

1.5 Scope of Research

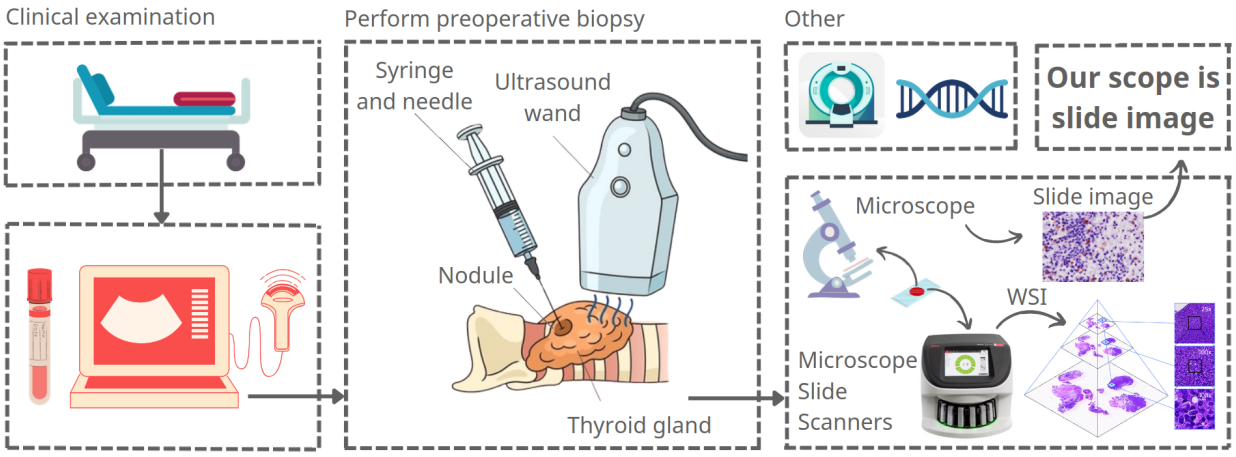


Figure 1: Methods in the thyroid tumor diagnostic process and the scope of our research

Figure 1 is the current steps in diagnosing thyroid tumors. Starting with clinical diagnosis such as palpation, observing external manifestations, analyzing medical history, etc. After undergoing blood hormone testing or ultrasound, it is necessary to a decision to treat such as surgery will require an FNAB. Some other methods such as CT scan, molecular gene analysis,...

Currently, there are two main ways to capture cytological images at the FNAB step on glass slides. The first method uses a microscope with a camera attached to photograph a small part of the biopsy sample, resulting in a "slide image". The second way usesm an digital pathology microscope slide scanners, the result is a WSI consisting of many "slide images" that cover the entire area of the slide. These two ways of taking biopsy samples have their own advantages and disadvantages. While using an digital pathology microscope slide scanners is commonly used in developed countries. The use of microscopes with cameras is common in developing countries. The reason is because the cost of an digital pathology microscope slide scanners is expensive; Complicated procedures for collecting, freezing and cutting biopsy samples; The time to analyze the entire WSI by experts or machines is limited. In addition, because WSIs are essentially composed of many "slide images", research on WSI images still essentially revolves around the problem of "slide images". If the problem at the "slide image" level is solved well, the results will be correspondingly good at WSI.

The scope of our current research is to automate the diagnosis of a corresponding "slide image" of a patient's results. Holding good results in "slide image" will ensure the potential for model development even for WSI.

2 Related Work

In 2018, [18] used CNN to classify PTC:B6 and non-PTC:B2 using 370 slide images. After training a simple CNN on these images, the model achieved an accuracy of 85.06% on the test dataset.

In 2019, [19] presented at a conference the use of 908 whole slide images (WSI) with training:test 799:109. The proposed approach employed two convolutional neural networks (CNNs): the first CNN identified image regions containing thyroid follicular cell clusters, while the second CNN predicted the likelihood of thyroid cancer based on the selected regions. This study is notable for having five output labels, owing to the advantage of a large dataset. However, the results for five-class classification were not satisfactory. See more in figure 2a.

In the same year, [20] used 279 slide images, divided into 159 cases of papillary thyroid carcinoma (PTC: B6) and 120 cases of benign lesions (non-PTC: B2). The researchers trained two CNN models, VGG-16 and Inception-v3. The VGG-16 model achieved an accuracy of 97.66% on fragmented images and 95% accuracy at the patient level. However, the study manually segmented fragments from patient-level images and did not propose a reasonable model to aggregate

fragment results for patient-level prediction. The 95% accuracy was derived from misdiagnosing two patients out of three misclassified fragments during testing.

In 2020, building on the previous work [19], the same research group published [21]. Their approach remained the same, and the classification results are shown in figure 2b.

In 2021, [22] utilized 367 slide images, including 222 cases of PTC and 145 cases of benign lesions (non-PTC: B2). The study implemented a complex preprocessing pipeline involving an automatic segmentation algorithm, followed by manual filtering of the fragments by a team of experts. The images were split into fragments, each containing tissue clusters or regions of interest (ROI), which were used to train and test CNN models. The authors used six CNN models from ResNet, DenseNet, and Inception. The results showed that DenseNet161 achieved the best performance, with an average accuracy of 0.9556. Additionally, combining AdaBoost with fragment results from six CNN models (input vector size 6 * 2 = 12 dimensions) achieved an accuracy of up to 0.9971. However, the major drawback is the inference time due to combining multiple models with numerous parameters. Furthermore, the test set used to evaluate this performance consisted of fragment images, not full FNAC slides. Although fragment usage increased the dataset size and improved classification performance, it limited the model's generalizability to full FNA slide images. The study proposed a method for inferring patient-level predictions based on the average classification of fragments, but the patient-level accuracy was not clearly mentioned.

In 2022, [23] used 360 WSI from thyroid fine-needle aspiration biopsy (FNAB), divided into 222 benign (B2) and 138 malignant (B6) cases. The authors proposed a two-stage system using CNNs. The first stage used YOLO V4 to detect malignant regions and extract fragments, while the second stage employed EfficientNet to classify these regions as benign or malignant. The results showed that the two-stage system achieved an accuracy of 81.84%, 3.16% higher than using YOLO V4 alone. However, the two-stage model required more computational resources, and the paper did not account for multiple diagnostic categories beyond B2 and B6.

In 2023, the research group from [21], aiming to create a compact model for WSI, published [24]. The study used 964 WSI scanned with a Leica AT-2 scanner at 40x magnification. These WSI were split into many tiles to train a MobileNetV2 model, which was later tested on images captured by a smartphone attached to a microscope. The study provided numerous results for AUC, Cl, and P of various methods and concluded that mobile-based machine learning shows promise for thyroid cancer diagnosis, particularly in resource-constrained areas. However, no accuracy or F1-score metrics were provided to evaluate the model's overall effectiveness across labels or specific label analyses.

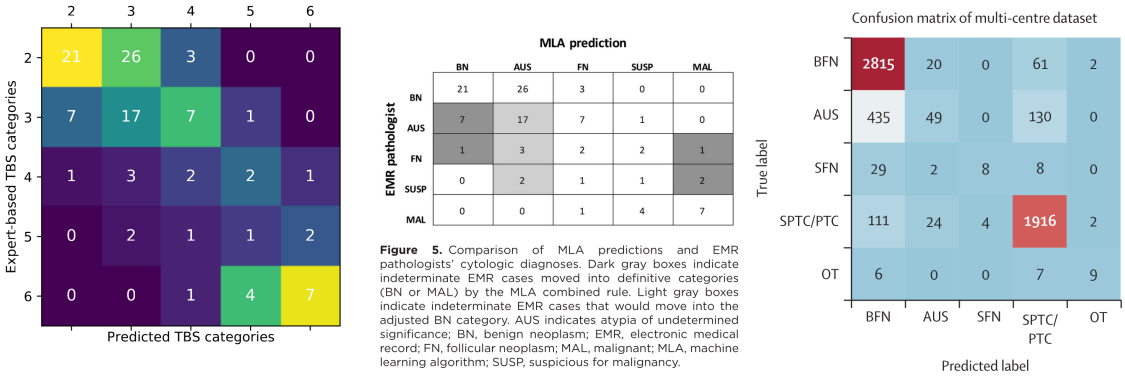
After acquiring additional data, the group from [21] published [25], using 1928 WSI from FNAB. They maintained their two-stage approach, using two VGG11 models for detecting ROI in slide images and then classifying these ROIs. The goal was to classify benign, indeterminate, and malignant cases. The algorithm screened 45.1% of cases as benign or malignant, with risk of malignancy (ROM) rates of 2.7% and 94.7%, respectively. The algorithm also served as an auxiliary test to reduce the number of indeterminate cases, decreasing them by 21.3% with a benign ROM of 1.8%. The inefficiencies of the inference process remain unaddressed.

A new research group, [26], used 577 WSI from frozen thyroid surgical specimens. This study focused on surgical samples rather than preoperative diagnosis, aiming for rapid and accurate intraoperative thyroid nodule diagnosis (IOPD). The researchers combined computer vision techniques with CNN and SVM to automatically detect all types of thyroid diseases. Their system consisted of three main stages: complete slide-level disease classification, benign/malignant classification at the patch level, and subtype classification at the patch level. The proposed method demonstrated accurate thyroid nodule diagnosis during surgery, with 72.65% sensitivity, 100.0% specificity, and an AUC of 86.32% on 191 test slides. For subtype diagnosis, the best AUC was 99.46% for medullary thyroid carcinoma, with an average inference time of 237.6 seconds per slide. However, the study needs to improve the process for faster inference in a surgical environment to reduce waiting time, as the freezing, slicing, and imaging steps already consume significant time. Moreover, the study excluded nodules with uncertain malignancy potential, used a single uniform scanner, was retrospective in nature, and faced cost-related barriers regarding scanning equipment and data storage, potentially limiting the deployment of AI systems in developing countries.

[27] involved 1535 slide images (1128 benign and 407 malignant) from 124 patients. The researchers used correlation optical diffraction tomography (CODT) to obtain papanicolaou-stained images and the three-dimensional refractive index (RI) distribution of FNAB samples. A machine learning algorithm (MLA) was designed to classify benign and malignant cell clusters using color images, RI images, or both. The results showed that the MLA using combined data from both color and RI images achieved 100% accuracy in classifying cell clusters. The study concluded that integrating RI image data with Papanicolaou-stained color images could improve the accuracy of MLAs in diagnosing thyroid cancer from FNAB samples. However, data limitations include the absence of non-diagnostic or indeterminate samples.

In 2024, [28] used 17,966 WSIs from 7,420 patients, divided into a training set, an internal validation set, three external validation sets, and a prospective validation set. The study developed an AI-assisted system named ThyroPower, which utilizes deep learning to diagnose thyroid nodules according to the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC). ThyroPower extracts cell-level features using the PAGIN model, then combines two WSI-level classification models (Random Forest and Top-N Feature) to make the final diagnostic decision. Results showed that ThyroPower achieved high performance in distinguishing benign cases from TBSRTC III+ (AUROC 0.930 for internal validation and 0.944-0.971 for external validation) and TBSRTC V+ (AUROC 0.990 for internal validation and 0.965-0.991 for external validation). This study had the advantage of utilizing a large and multi-center dataset, enhancing the AI model's reliability and generalizability. The AI system developed, ThyroPower, demonstrated high performance in classifying benign and malignant thyroid nodules, with potential to assist less experienced pathologists in making more accurate diagnoses. However, the study also acknowledged some limitations; the AI model still struggled with accurately classifying several types of thyroid nodules, especially SFN. See more in figure 2c.

[29] used human thyroid cell samples stained with methylene blue (MB) and captured through fluorescence polarization imaging (Fpol). The researchers developed a 2D CNN U-Net for automatic cell segmentation from Fpol images, aimed at reducing data analysis time and facilitating the clinical application of MB Fpol technology. Results showed that the U-Net model successfully segmented 15.8% of the cells. The study concluded that the implementation of automatic cell analysis renders quantitative fluorescence polarization diagnosis clinically feasible. Future research directions should focus on providing predictive labels in addition to simply offering cell segmentation labels.



(a) Test results of Study [19] (b) Test results of Study [21] (c) Test results of Study [28]

Figure 2: Test results of Study [19, 21, 28] - Rare studies had data classifying more than 2 labels

3 Methodology

3.1 Data Acquisition

Details about the dataset: The dataset consists of microscopic images of fine-needle aspiration biopsy (FNAB) samples collected by medical experts from 108 Military Central Hospital. The images were captured using an optical microscope, specifically the Olympus BX43, equipped with a digital camera to capture high-resolution images of the biopsy samples on glass slides. A total of 1804 images were collected, each corresponding to a unique patient. The dataset is divided into three categories: 482 images labeled as benign (B2), 541 images labeled as suspected malignant (B5), and 871 images labeled as malignant (B6). This balanced and diverse dataset represents a wide range of thyroid biopsy samples, allowing for robust classification model development.

Dividing data sets: The data from the experiment is randomly divided with a seed of 42 into 3 sets: training - to estimate the parameters of the models (70%), validation - to pause training when the model has reached to the trade-off point and enter the overfitting process (15%), testing - to test the models and choose the best performing model (15%). See details in figure 3.

3.2 Design Experiments

In this study, we pose a few hypothetical questions to achieve the set goals. Those hypothetical questions are:

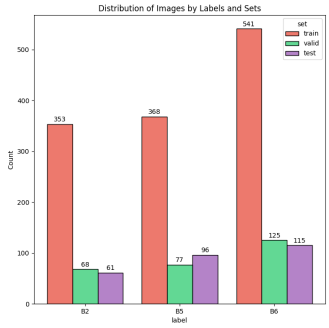
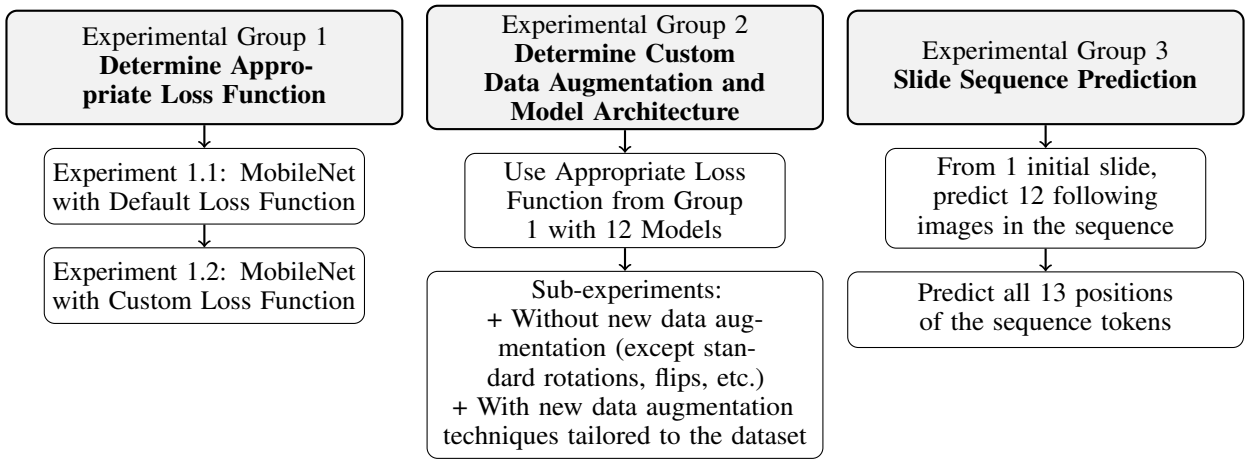


Figure 3: How we divided the datasets for this study

- How much can **change the definition of the loss function** help improve classification performance (especially for classifying minority labels in the data)?
- How does **proper data augmentation** affect how effectively a model can learn from a training data set? **Operation flow and model architecture** should be such that the classification results are good enough while the activity flow is not complicated, the number of calculations is small.
- How much can considering **input image and its 12 cropped grids as tokens into the model** improve classification performance? From there, evaluate the effectiveness of the application potential even with WSI because WSI is a collection of combined biopsy image slides.

To validate the 3 hypothetical questions mentioned above, our research team designed 3 specific groups of experiments at 3.3 - main point is Experimental group 1 setup, 3.4 - main point is Experimental group 2 setup and 3.5 - main point is Experimental group 3 setup. For brevity, below is a visual representation of the three sets of experiments we will perform.



3.3 Loss Function

In the multi-class classification problem, the choice of loss function plays an important role in model training efficiency. This study will design the loss function in two directions to evaluate which approach provides the best accuracy, F1-score, and AUC.

Default loss function for multi-label classification

A popular loss function for multi-label classification is Cross-Entropy Loss. The formula of the Cross-Entropy Loss function in the case of multi-label classification is defined as

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_{ij})$$

In there:

- N is the number of samples in the batch
- C is the number of labels (classes)
- y_{ij} is the actual label value of the ith sample for label j, with $y_{ij} \in \{0, 1\}$
- p_{ij} is the model's prediction probability for the ith sample for label j

Loss function with weights for labels

The Weighted Cross-Entropy Loss function is a function whose weight is calculated based on the number of records belonging to each label. The weights are calculated so that minority labels are given greater weight, so the model will pay more attention to these labels. The Weighted Cross-Entropy Loss formula is adjusted as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j y_{ij}\log(p_{ij})$$

Where w_j is the weight for label j. There are many ways to initialize this weight and below is one way that our research team initializes.

$$w_j = \frac{\text{total samples}}{\text{num classes} \times \text{frequency}(j)}$$

In there:

- total samples: Total number of samples in the data
- num class: Total number of labels (classes)
- information(j): Number of samples belonging to label j

Experimental group 1 setup
The goal of experimental group 1 is to evaluate the impact of the change of the loss function on the evaluation results of interest. The F1 Score macro - the measure we are most interested in in this study. Giving more priority to minority labels can help the model limit bias towards majority labels. However, we still need to seriously consider both ways of initializing the loss function. Because if the data structure has a lot of noise in minority labels while the number of records is not large enough for the noisy features to be not too common when being learned, the model can easily learn the noisy features. The general model architecture is MobileNetv1 [30] with the hyperparameters mentioned in 3.4 except that the data will not be enhanced by x34 times. The images when training can be rotated, flipped, ... to ensure diversity for the samples when being trained.

3.4 Data Augmentation, Model Architecture and Training Strategy

Unlike previous studies with overly complex patient-level diagnostic procedures, models with overly parametric architectures. We aim to use a simple prediction flow, a model with few parameters but still need to ensure good performance results for all 3 classification labels. This is a challenge that has not been resolved by any previous research, placing a lot of pressure on the built model. In the hope of resolving the above pressure, we propose that there should be a difference from the past in terms of:

- **Data augmentation method in training set**. We design a method to automatically identify areas of interest in images, techniques for marking areas of interest such as drawing bounding boxes on the original image, cropping according to the cropped range grid, selecting areas of interest,... From here, the training set is built to strengthen and eliminate many interfering factors. Or it helps the main model automatically focus on these regions (no need to go through the secondary model to identify and cut them out to make predictions. The process is shortened and all tasks are done. by a minimal parameter model. For more details, see Data Augmentation.

- **Inference process and model architecture**. The inference making process needs to be shortened and use a single block of models instead of a combination of multiple models at once or their sequential combination. In Model Architecture, we will discuss in more detail how the 12 models are used and how we choose the model that best suits both goals: simplicity of the inference process and the effectiveness of the results obtained.
- **Initialize hyperparameters** to prepare for training. Also, we know that the order in which the data is augmented in the training set used in an epoch can greatly affect the output. We cover details about pre-training settings at Training Strategy

Data Augmentation

Figure 4 describes how we use YOLOv10 [31] combining grid selection techniques to obtain enhanced image sets A(x1), B(x1), C(x8), D(x12), E(x12). Specifically, we manually labeled (**without the assistance of medical experts**) 120 images, each image includes from 0 to 8 rectangles surrounding the areas of cell cluster concentration care. Because we are not labeling experts, we are only interested in areas with many cell concentrations. Which cancer label that cell area belongs to is not of concern. Training this model helps us obtain a sub-model that supports the data augmentation process.

Because this model is just a sub-model for the data augmentation process and in addition, the manually assigned labels are assigned by us (not by medical experts). Therefore, the measurement evaluations for this YOLOv10 model are unnecessary and have no (or difficult to assess) impact on the final classification results, so they will not be mentioned in detail in this study.

As for the characteristics and properties of image sets A, B, C, D, E. Basically, image E is an image cropped according to a 256x256 grid from the original image A. Image set D is an image cropped according to a grid. 256x256 from image B - the image has identified bounding boxes of clusters with many cells. Image set C includes a set of 8 images cropped from areas with many cell concentrations. When the photo does not cut into 8 pieces. Assuming there is a shortage of i pieces, we will randomly pick i pieces out of 8 pieces cut from image set B (including 6 images cut according to the 512x512 grid and 2 images cut according to the 768x768 grid) - with picture B resized to 1024x768.

The consequence of this is that from an initial set of A images in the initial training set we created a set of A, B, C, D, E images for a total of 34 times.

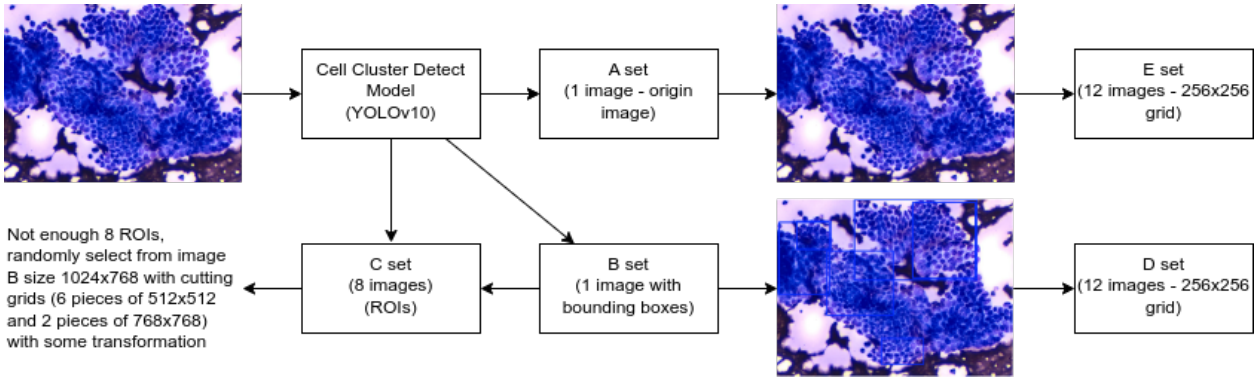


Figure 4: How we enhanced the training data set to obtain 34 times more images than the original

Model Architecture

In this study we will consider 12 models (including the model with the most parameters and the least parameter) from 6 model groups. The goal of this selection is to evaluate and compare models from various architectures, ranging from convolutional neural networks (CNNs) like VGG and ResNet to more recent architectures such as the Transformer-based ViT. By including both the most and least parameterized versions of each group, we aim to investigate the trade-offs between model complexity, computational efficiency, and accuracy in a variety of tasks, especially those involving imbalanced datasets.

In addition to the default versions of these models, we will implement custom versions with loss functions adjusted to address the issue of class imbalance, leveraging techniques such as class weighting. This strategy is crucial to ensure that the model does not become biased towards majority classes, which is a common problem in multi-class classification scenarios with skewed data distributions.

The six model groups are briefly described below:

- **VGG**: A deep convolutional network known for its simplicity and effectiveness in image classification, using a stack of convolutional layers [32].
- **ResNet**: Introduces the concept of residual learning, enabling very deep networks by addressing the vanishing gradient problem [33].
- **MobileNet**: A lightweight architecture designed for mobile and embedded applications, emphasizing efficiency with depthwise separable convolutions [30].
- **DenseNet**: Connects each layer to every other layer in a feed-forward fashion to promote feature reuse and reduce the number of parameters [34].
- **EfficientNet**: Balances network width, depth, and resolution using a compound scaling method, achieving STATE-of-the-art accuracy with fewer parameters [35].
- **Vision Transformer (ViT)**: Applies transformer architecture, previously successful in NLP, to image classification by treating image patches as tokens [36].

The experimentation with these models will provide insights into which model group and configuration are most suitable for tasks where computational resources and accuracy must be balanced.

Training Strategy

The following elements are applied consistently across all experiments. The model's parameters are optimized using the Adam optimizer. The initial learning rate is set to $\alpha = 0.001$ and is gradually reduced by a factor of 0.5 every 10 epochs. Each experiment runs for a maximum of 100 epochs. However, early stopping is employed, where training is halted if the validation loss does not improve for 10 consecutive epochs, thus ensuring that we capture the model with the best performance. The results obtained from testing these models on the test set are detailed in Section Results and Discussion.

The special point that we design during the 1-epoch training process is the order of images in the data augmentation set when training. Specifically, the images in an epoch when training, although randomly selected, are still in the order of set E first, then set E, then sets D, C, B, A, respectively. The main purpose is to ensure that the model learns the characterized at many different scales, from noisy images (set E) to low-noise images (sets D, C, B) and then to noisy images (set A). And finally, the model will learn to match the most suitable features and make the best prediction for image set A - that is, the original image taken from the camera, without undergoing any costly image editing of computational resources. From there, just predicting on this image level is enough to make the final conclusion.

Experimental group 2 setup

Thus, this experimental group number 2 will include 2 implementation directions: no x34 data augmentation and x34 data augmentation. Data augmentation aims to diversify the number of samples. However, non-focused augmentation on the features that determine the output label will cause the model to learn too many noisy features. With testing 12 image classification models, the total number of experiments obtained in experimental group number 2 is 24. Both implementation directions will have training image transformations such as flipping, rotating, ... before training. These are safe augmentations when relatively preserving the features of the original image. **The purpose** of experimental group number 2 is to determine whether the current x34 data augmentation method ensures that the model learns the features that determine the output classification better. At the same time, testing with many of the best current architectures helped our research team identify a model that both optimizes the number of parameters and ensures impressive classification results.

Algorithm 1 Predictive label inference process based on experimental group 1 & 2

Input: A single slide image I.
Output: A three-dimensional vector I' representing classified labels.
Description:
+ The model $f(I)$ may consist of layers from CNNs or Transformers, depending on the architecture used.
+ This model classifies labels directly without preprocessing to remove noisy dimensions.
+ The simplified inference reduces computational time and resource consumption.
+ The output I' is a three-dimensional vector representing three classified labels.

3.5 Sequence of approach and potential use with WSI

With the goal of simplifying the inference process, we decided that this single model block must be able to distinguish between noise features and decision features to output prediction labels. How to augment data in Data Augmentation and how to set up image order in 1 epoch during training are mentioned in detail in Training Strategy which is expected will help achieve the set goals.

However, if we assume we can spend more computational resources to increase the accuracy of prediction, is that feasible? This issue becomes even more important in the case of a device that captures WSI images. WSI images have more capture range and thus make it easier to accurately predict the image's output labels.

Experimental group 3 setup

The goal of Experiment 3 is to evaluate the ability to improve the results when predicting multiple times at different scales on the original image. If the results are improved, we will see the potential to apply the designed procedures in this study even to WSIs captured by other devices.

To train the Transformer model mentioned here, we basically still use the train:valid:test dataset as mentioned in 3.

Accordingly, the input image is decomposed into 13, through a sequence processing model such as Transformer and thereby help predict across multiple scales and take advantage of the diversity in available data. Detail about mathematical formulation for this idea:

Algorithm 2 Predictive label inference process based on the idea of experimental group 3

Input: Given an input image I.
+ The image I is first split into 13 sub-images $I_0, I_1, \ldots, I_{12}$, each representing a different region of the original image. Mathematically, we represent this as: $I \rightarrow \{I_0, I_1, \ldots, I_{12}\}$.
+ Each sub-image I_i is processed by a prediction model to generate a corresponding token T_i, which is a vector of size 3 (since the model outputs 3-dimensional tokens at this stage): $T_i = f(I_i) \quad \text{for} \quad i = 0, 1, \ldots, 12$ where f denotes the function of the prediction model applied to each sub-image I_i - the best model trained in Experimental group 2.
+ Each token T_i (a 3-dimensional vector) is multiplied by a corresponding transformation matrix W_i, which is independent for each token, to convert it into a 9-dimensional vector: $T_i' = W_i \cdot T_i \quad \text{for} \quad i = 0, 1, \ldots, 12$. Here, $W_i \in \mathbb{R}^{9\times 3}$ is the transformation matrix for token T_i. After this transformation, we have 13 new tokens $T_0', T_1', \ldots, T_{12}'$, each of size 9.
+ These 13 transformed tokens are then passed through some Transformer encoder blocks. The encoder processes the sequence of tokens with multi-head self-attention, where the number of attention heads is set to 3: $\{T_0', T_1', \ldots, T_{12}'\} \rightarrow$ Transformer Encoder with 3 heads. The output of the Transformer encoder is a set of processed tokens $\hat{T}_0', \hat{T}_1', \ldots, \hat{T}_{12}'$, each of size 9.
+ From the output tokens, we extract only the token corresponding to T_0' for further processing: $\hat{T}_0' \quad (\text{size } 9)$. This token is then passed through a feed-forward neural network (FFNN) to predict the final labels for the input image. The FFNN outputs 3 labels based on the input token: $y = \text{FFNN}(\hat{T}_0')$ where $y \in \mathbb{R}^3$ is the predicted label for the original image I.
Output: The final output of the model is the prediction y, which represents the label prediction for the entire input image I.

4 Results and Discussion

The evaluation metrics we used in this study can be viewed more clearly at Appendices Evaluation Metrics.

4.1 Experimental group 1

Specifically, the result of using the same weight of the loss function on the labels obtained an F1 score macro of 0.76. Meanwhile, initializing the loss function with the priority of the minority label gave a better result of 0.81. This shows that the loss function with the priority of the minority label is giving better results on average for all 3 labels, at least on a simple model like MobileNetv0. With such results, the second experimental group will be set up with the loss function with the priority of the minority label.

4.2 Experimental group 2

Figure 5 is a comparison of F1 score macros for 24 experiments. Based on these results, our research team draws the following conclusions:

- The x34 data augmentation approach provides a significant improvement in performance for almost all deployed models (except for the ViT-B16 model). This proves that the data augmentation approach has brought diversity to the training data, focusing on the most decisive features for the output prediction label. As a result, noisy features have less impact on the final results, improving the learning efficiency of the model.

- The group of models with too few parameters such as MobileNet; or the group with too many parameters such as DenseNet, ViT shows that models with more parameters will give better classification results (shown in the red lines). In contrast, the group of models with moderate number of parameters such as EfficientNet, VGG or ResNet shows that the model with fewer parameters will give better results (shown in the blue lines).

- The model with the highest F1 score macro is the EfficientNetB0 model with training method using x34-boosted training set. The result is 89.19% F1 score macro. The results mentioned below will be the performance analysis of this model. For easy distinction, we will name this model ThyroidEffi Basic - to easily distinguish it from the ThyroidEffi Premium model built in group of experiments number 3.

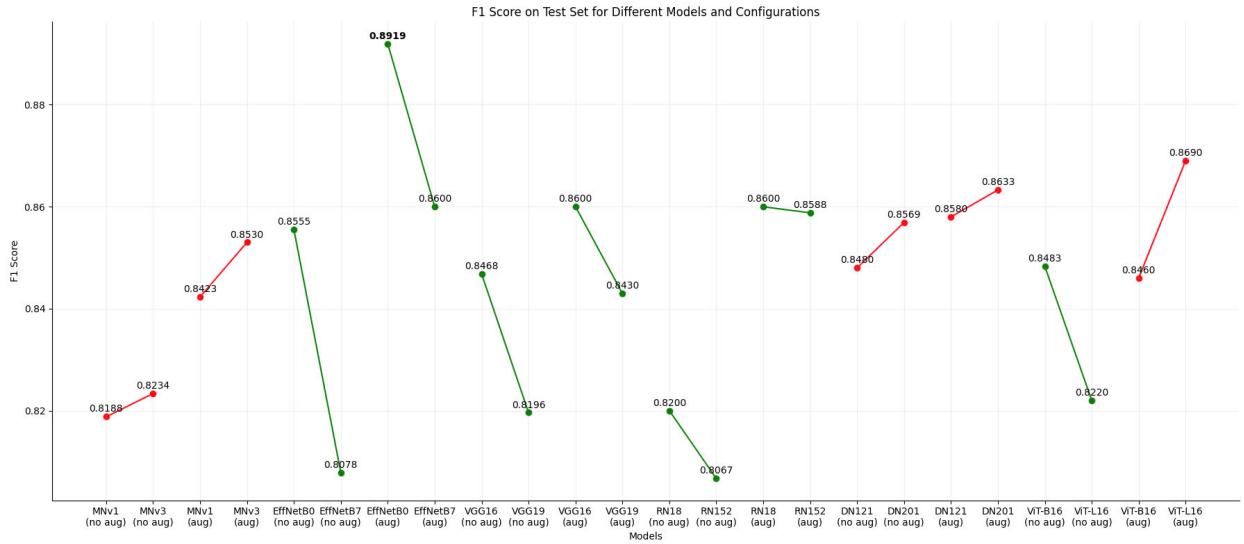


Figure 5: Comparing F1-Score macro for 3 labels B2, B5, B6 between not augmenting the training set and augmenting the training set for 12 models from 6 groups of the best image classification models today

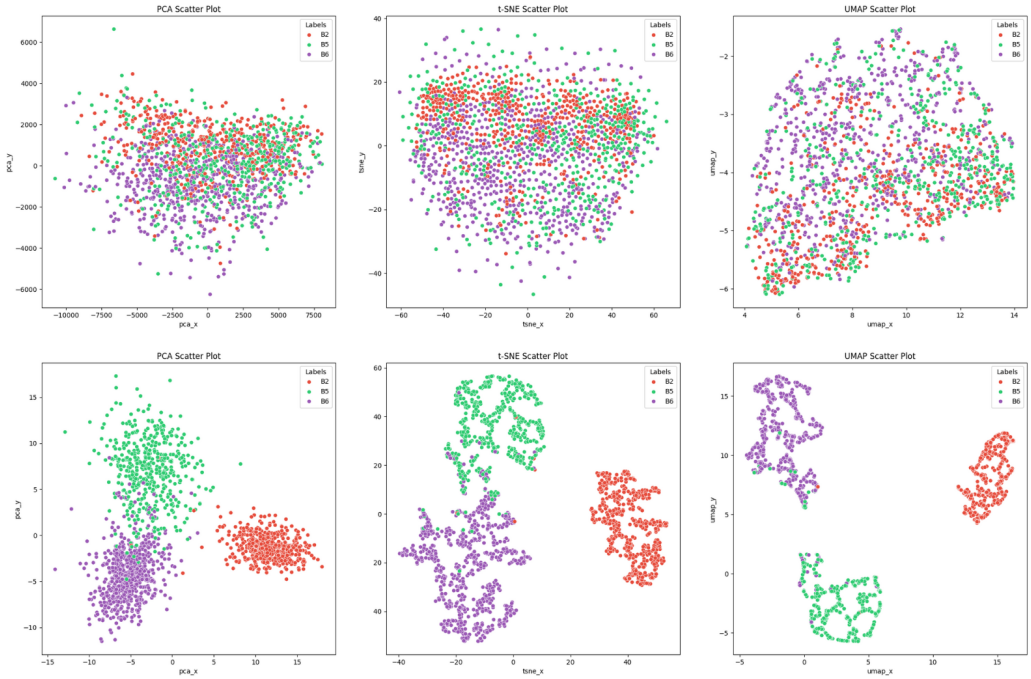
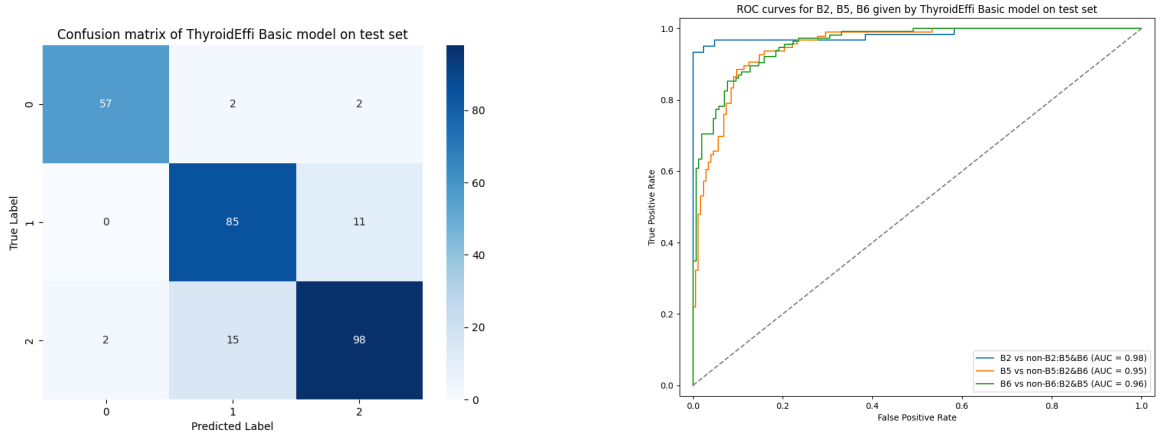


Figure 6: Visualize the distribution of the original data before and after applying ThyroidEffi Basic. The 3 images above are our data set (images are resized to 64x64 = 4096 dimensions) when applying 3 methods of dimensionality reduction to 2-dimensional space. The 3 images below are our data set through our best model (ie 3 dimensions corresponding to the ability information belonging to 3 labels B2, B5, B6) when applying 3 dimensionality reduction methods to 2-dimensional space. The three dimensionality reduction methods are PCA [37] and tSNE, UMAP [38] respectively.

Figure 6 is a visual simulation of how ThyroidEffi transforms complex features into 3-dimensional ones for B2, B5, B6 respectively. The visualization is performed in 2-dimensional space to ensure ease of visualization and evaluation. The ThyroidEffi Basic model with **only 4,011,391 parameters** is one of the most compact models ever implemented for thyroid cytology biopsy image classification. In terms of workflow, the model only needs 1 slide image to produce a predictive label of benign/suspicious malignant/malignant with model evaluation metrics shown in below:



(a) Confusion matrix of ThyroidEffi Basic model on test set

(b) ROC curves of ThyroidEffi Basic model on test set

Figure 7: Performance evaluation of ThyroidEffi Basic model with impressive AUC: B2 (0.9830), B5 (0.9499), B6 (0.9585)

Metric	Macro Average	Per Label
F1 Score	0.8919	B2 (0.9500), B5 (0.8586), B6 (0.8673)
AUC	0.9638	B2 (0.9830), B5 (0.9499), B6 (0.9585)
Accuracy	0.8824	-

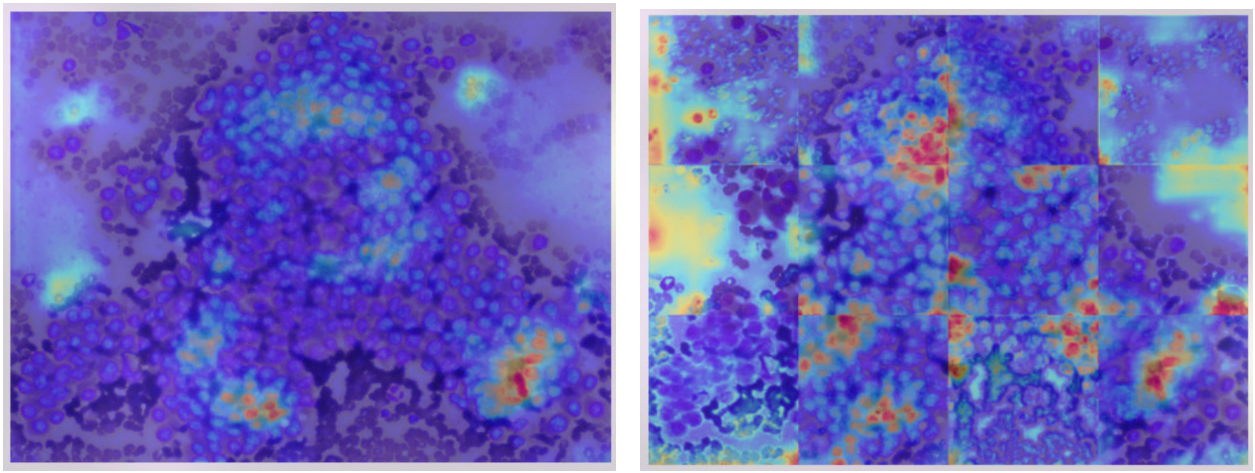
Table 1: Summary of performance metrics of the ThyroidEffi Basic model

4.3 Experimental group 3

The results obtained with the Transformer test have been improved, the resulting model we call ThryroidEffi Premium to distinguish it from Thyroid Base in group 2 of tests. The F1 score macro has been improved by 0.58%, increasing to 89.77%. The improvement is not too much compared to the increased computational cost.

This result shows that the procedure presented in this study is not only effective for slide images, but can even take advantage of the wide range factor of WSI to predict more accurately with this type of data.

To improve the explainability of the model with the predicted output label, we used GradCAM [39]. This is a technique to draw heatmaps assessing the importance of input pixel features. Figure 8 shows that it is very effective in selecting pixel features, focusing on important regions of the image for classification.



(a) CAM of a slide image after pass ThyroidEffi Basic

(b) 12 CAM of 12 patches after pass ThyroidEffi Basic

Figure 8: CAM of a slide image or 12 patches after pass ThyroidEffi Basic

5 Conclusion

It can be seen that, by setting up a suitable loss function, enhancing the data appropriately to increase core features, and removing noisy features, the ThyroidEffi Basic model has really brought outstanding results. With F1 score macro for 3 classification labels being 89.19%, AUC macro for 3 labels being 0.9638; ThyroidEffi Basic shows great potential for practical application with few parameters and good prediction results.

ThyroidEffi Premium version performing inference on multiple crop ranges from 12 patches also gives better results than ThyroidEffi Basic with F1 score macro 89.77%. Although it requires more computation, the results obtained on ThyroidEffi Premium show that the procedure with multiple WSI slide images and then through Transformer layers has the potential to improve the prediction results. The effectiveness of the procedure presented in this study is demonstrated not only with slide images for imaging microscope devices but also with the combination of slide images to make predictions for WSI with modern scanning devices.

Future Work

Future research is aimed at developing a complete CAD system to support physicians. This includes comparative evaluation with expert physicians when using no CAD system and when using it in combination.

Research can also be done on WSI data type when appropriate data provided by scanners are prepared.

Acknowledgments

The authors would like to thank 108 Military Central Hospital for their assistance in providing data for conducting this study.

Contributions

Hai Pham-Ngoc: Conceptualization (equal), Methodology, Data Analysis, Writing - Original Draft. **Phuong Le-Hong**: Conceptualization (equal), Writing - Review & Editing, Supervision, Project Administration.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability and Source Code

The data and implementation code will be made available upon successful contact and agreement for collaboration.

Ethical Statement

This study was conducted in accordance with ethical standards set forth by 108 Military Central Hospital & Hanoi University of Science - VNU. Informed consent was obtained from all participants involved in the study.

References

[1] A. Greco, C. Miranda, M. G. Borrello, and M. A. Pierotti, "Chapter 16 - thyroid cancer," in *Cancer Genomics* (G. Dellaire, J. N. Berman, and R. J. Arceci, eds.), pp. 265–280, Boston: Academic Press, 2014.

[2] A. Aliyev, I. Aliyeva, F. Giammarile, N. Talibova, G. Aliyeva, and F. Novruzov, "Diagnostic accuracy of fine needle aspiration biopsy versus postoperative histopathology for diagnosing thyroid malignancy," *Endocrinology, Diabetes & Metabolism*, vol. 5, no. 6, p. e373, 2022.

[3] B. Anand, A. Ramdas, M. M. Ambroise, and N. P. Kumar, "The bethesda system for reporting thyroid cytopathology: A cytohistological study," *Journal of thyroid research*, vol. 2020, no. 1, p. 8095378, 2020.

[4] E. P. Balogh, B. T. Miller, J. R. Ball, E. National Academies of Sciences, Medicine, *et al.*, "Overview of diagnostic error in health care," in *Improving diagnosis in health care*, National Academies Press (US), 2015.

[5] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Frontiers in Public Health*, vol. 11, p. 1273253, 2023.

[6] L.-R. Li, B. Du, H.-Q. Liu, and C. Chen, "Artificial intelligence for personalized medicine in thyroid cancer: current status and future perspectives," *Frontiers in Oncology*, vol. 10, p. 604051, 2021.

[7] M. Ilyas, H. Malik, M. Adnan, U. Bashir, W. A. Bukhari, M. I. A. Khan, and A. Ahmad, "Deep learning based classification of thyroid cancer using different medical imaging modalities: A systematic review," *VFAST Transactions on Software Engineering*, vol. 9, no. 4, pp. 1–17, 2021.

[8] Y. Lu and B. Zhang, "Application of artificial intelligence based on deep learning in the diagnosis of thyroid cancer," *Int J Radiat Med Nucl Med*, vol. 46, no. 12, pp. 760–764, 2022.

[9] M. Ludwig, B. Ludwig, A. Mikuła, S. Biernat, J. Rudnicki, and K. Kaliszewski, "The use of artificial intelligence in the diagnosis and classification of thyroid nodules: An update," *Cancers*, vol. 15, no. 3, 2023.

[10] V. Fiorentino, C. Pizzimenti, M. Franchina, M. G. Micali, F. Russotto, L. Pepe, G. B. Militi, P. Tralongo, F. Pierconti, A. Ieni, M. Martini, G. Tuccari, E. D. Rossi, and G. Fadda, "The minefield of indeterminate thyroid nodules: could artificial intelligence be a suitable diagnostic tool?," *Diagnostic Histopathology*, vol. 29, no. 8, pp. 396–401, 2023.

[11] Y. Habchi, Y. Himeur, H. Kheddar, A. Boukabou, S. Atalla, A. Chouchane, A. Ouamane, and W. Mansoor, "Ai in thyroid cancer diagnosis: Techniques, trends, and future directions," *Systems*, vol. 11, no. 10, 2023.

[12] C. M. Wong, B. E. Kezlarian, and O. Lin, "Current status of machine learning in thyroid cytopathology," *Journal of Pathology Informatics*, vol. 14, p. 100309, 2023.

[13] G. Slabaugh, L. Beltran, H. Rizvi, P. Deloukas, and E. Marouli, "Applications of machine and deep learning to thyroid cytology and histopathology: a review," *Frontiers in Oncology*, vol. 13, 2023.

[14] P. C. Rizzo, S. Marletta, N. Caldonazzi, A. Nottegar, A. Eccher, F. Pagni, V. L'Imperio, and L. Pantanowitz, "The application of artificial intelligence to thyroid nodule assessment," *Diagnostic Histopathology*, vol. 30, no. 6, pp. 339–343, 2024.

[15] E. David, H. Grazhdani, G. Tattaresu, A. Pittari, P. V. Foti, S. Palmucci, C. Spatola, M. C. Lo Greco, C. Inì, F. Tiralongo, *et al.*, "Thyroid nodule characterization: Overview and state of the art of diagnosis with recent developments, from imaging to molecular diagnosis and artificial intelligence.," *Biomedicines*, vol. 12, no. 8, 2024.

[16] Y. Habchi, H. Kheddar, Y. Himeur, A. Boukabou, A. Chouchane, A. Ouamane, S. Atalla, and W. Mansoor, "Machine learning and vision transformers for thyroid carcinoma diagnosis: A review," *arXiv preprint arXiv:2403.13843*, 2024.

[17] X. Zhang, V. Lee, and F. Liu, "From data to insights: A comprehensive survey on advanced applications in thyroid cancer research," *arXiv preprint arXiv:2401.03722*, 2024.

[18] P. Sanyal, T. Mukherjee, S. Barui, A. Das, and P. Gangopadhyay, "Artificial intelligence in cytopathology: A neural network to identify papillary carcinoma on thyroid fine-needle aspiration cytology smears," *Journal of Pathology Informatics*, vol. 9, no. 1, p. 43, 2018.

[19] D. Dov, S. Z. Kovalsky, J. Cohen, D. E. Range, R. Henao, and L. Carin, "Thyroid cancer malignancy prediction from whole slide cytopathology images," in *Machine Learning for Healthcare Conference*, pp. 553–570, PMLR, 2019.

[20] Q. Guan, Y. Wang, B. Ping, D. Li, J. Du, Y. Qin, H. Lu, X. Wan, and J. Xiang, "Deep convolutional neural network vgg-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study," *Journal of Cancer*, vol. 10, no. 20, p. 4876, 2019.

[21] D. D. Elliott Range, D. Dov, S. Z. Kovalsky, R. Henao, L. Carin, and J. Cohen, "Application of a machine learning algorithm to predict malignancy in thyroid cytopathology," *Cancer Cytopathology*, vol. 128, no. 4, pp. 287–295, 2020.

[22] N. T. Duc, Y.-M. Lee, J. H. Park, and B. Lee, "An ensemble deep learning for automatic prediction of papillary thyroid carcinoma using fine needle aspiration cytology," *Expert Systems with Applications*, vol. 188, p. 115927, 2022.

[23] W. Duan, L. Gao, J. Liu, C. Li, P. Jiang, L. Wang, H. Chen, X. Sun, D. Cao, B. Pang, R. Li, and S. Liu, "Computer-assisted fine-needle aspiration cytology of thyroid using two-stage refined convolutional neural network," *Electronics*, vol. 11, no. 24, 2022.

[24] S. Assaad, D. Dov, R. Davis, S. Kovalsky, W. T. Lee, R. Kahmke, D. Rocke, J. Cohen, R. Henao, L. Carin, and D. E. Range, "Thyroid cytopathology cancer diagnosis from smartphone images using machine learning," *Modern Pathology*, vol. 36, no. 6, p. 100129, 2023.

[25] D. Dov, D. Elliott Range, J. Cohen, J. Bell, D. J. Rocke, R. R. Kahmke, A. Weiss-Meilik, W. T. Lee, R. Henao, L. Carin, and S. Z. Kovalsky, "Deep-learning-based screening and ancillary testing for thyroid cytopathology," *The American Journal of Pathology*, vol. 193, no. 9, pp. 1185–1194, 2023.

[26] Y. Ma, X. Zhang, Z. Yi, L. Ding, B. Cai, Z. Jiang, W. Liu, H. Zou, X. Wang, and G. Fu, "A study of machine learning models for rapid intraoperative diagnosis of thyroid nodules for clinical practice in china," *Cancer Medicine*, vol. 13, no. 3, p. e6854, 2024.

[27] Y. K. Lee, D. Ryu, S. Kim, J. Park, S. Y. Park, D. Ryu, H. Lee, S. Lim, H.-S. Min, Y. Park, *et al.*, "Machine-learning-based diagnosis of thyroid fine-needle aspiration biopsy synergistically by papanicolaou staining and refractive index distribution," *Scientific Reports*, vol. 13, no. 1, p. 9847, 2023.

[28] J. Wang, N. Zheng, H. Wan, Q. Yao, S. Jia, X. Zhang, S. Fu, J. Ruan, G. He, X. Chen, S. Li, R. Chen, B. Lai, J. Wang, Q. Jiang, N. Ouyang, and Y. Zhang, "Deep learning models for thyroid nodules diagnosis of fine-needle aspiration biopsy: a retrospective, prospective, multicentre study in china," *The Lancet Digital Health*, vol. 6, no. 7, pp. e458–e469, 2024.

[29] P. R. Jermain, M. Oswald, T. Langdun, S. Wright, A. Khan, T. Stadelmann, A. Abdulkadir, and A. N. Yaroslavsky, "Deep learning-based cell segmentation for rapid optical cytopathology of thyroid cancer," *Scientific Reports*, vol. 14, no. 1, p. 16389, 2024.

[30] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[31] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[35] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," pp. 6105–6114, 2019.

[36] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] M. Greenacre, P. J. Groenen, T. Hastie, A. I. d'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 100, 2022.

[38] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, no. 201, pp. 1–73, 2021.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[40] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, pp. 1015–1021, Springer, 2006.

Appendices

Evaluation Metrics

Following [40], we use the metrics shown in table below.

Classification Metrics for Label j and Macro-Averaged Scores

Metric	Formula/Definition
Confusion Matrix Values	True Positives (TP_j): Correctly predicted positive cases for label j
	False Positives (FP_j): Incorrectly predicted as positive for label j
	True Negatives (TN_j): Correctly predicted negative cases for label j
	False Negatives (FN_j): Incorrectly predicted as negative for label j
Precision (for label j)	$\text{Precision}_j = \frac{TP_j}{TP_j + FP_j}$
Recall (for label j)	$\text{Recall}_j = \frac{TP_j}{TP_j + FN_j}$
F-beta (for label j)	$\text{F}_\beta = \frac{(1+\beta^2) \cdot \text{Precision}_j \cdot \text{Recall}_j}{(\beta^2 \cdot \text{Precision}_j) + \text{Recall}_j}$
ROC and AUC (for label j)	ROC curve: Plot of True Positive Rate (TPR) vs False Positive Rate (FPR) with $\text{TPR} = \frac{TP_j}{TP_j + FN_j}$ and $\text{FPR} = \frac{FP_j}{FP_j + TN_j}$
	AUC: Area under the ROC curve for label $j \rightarrow AUC_j = \int_{\text{FPR}}^{\text{TPR}} ROC_j$
Macro Precision	$\text{Macro Precision} = \frac{1}{n} \sum_{j=1}^{n} \text{Precision}_j$
Macro Recall	$\text{Macro Recall} = \frac{1}{n} \sum_{j=1}^{n} \text{Recall}_j$
Macro F-beta*	$\text{Macro F}_\beta = \frac{1}{n} \sum_{j=1}^{n} \text{F}_{\beta,j}$
Macro AUC*	$\text{Macro AUC} = \frac{1}{n} \sum_{j=1}^{n} \text{AUC}_j$
Overall Accuracy*	$\text{Accuracy} = \frac{\sum_{j=1}^{n} (TP_j + TN_j)}{\sum_{j=1}^{n} (TP_j + TN_j + FP_j + FN_j)}$