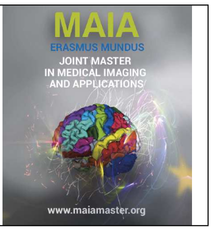
Medical Imaging and Applications

Master Thesis, June 2022



Classification of malignant nodules from 2D ultrasound thyroid images using Deep Convolutional Neural Networks

Tewele Weletnsea Tareke, Alain Lalande (PhD), Sarah Leclerc (PhD)

ImViA Laboratory, Université Bourgogne Franche-Comté, Dijon, France

Abstract

Thyroid nodule is a type of disease that affects the thyroid gland, a small gland at the base of the neck that produces hormones. In clinical routine, thyroid nodules are usually detected manually by expert physician. Manual classification of nodules by physician has several drawbacks: it is time consuming, inaccurate and tends to expose patients to unnecessary fine needle aspirations (FNA). It also suffers from inter-and intra-observer variabilities. Manual classification also is time consuming, less accurate and exposes to unnecessary fine needle aspiration (FNA) biopsies, which brings a lot of stress to patients. Thus, it is of considerable interest to develop an automatic and accurate thyroid nodule classification system. However, automatic methods struggle in presence of noise, artifact and low contrast, all characteristics of ultrasound imaging. In this paper, we propose an automatic Computer Aided Diagnosis System (CAD) for the classification of thyroid nodules using a fine-tuned deep learning model based on Densenet121 architecture in which an attention module is incorporated (Densenet-Attention). This CAD was developed using 595 thyroid nodule images that were fully annotated based on the Bethesda scores established from the biopsy. Out of these samples, 222 images were annotated as positive and 373 were annotated as negative. Out of them 51 images are used as test set to validate our proposed method. Several image enhancement methods were applied, such as histogram equalization, artifacts removal, and range scaling. We augmented the dataset with synthetic images obtained with label-preserving transformations and added a convolutional block appended with an attention module to extract global feature maps and forward them as inputs to the decision layers. Two attention modules (Channel and Spatial) were integrated in this proposed architecture, aiming to help the network focus on the most important feature maps and locations. Our method used a focal loss to encourage prediction accuracy by penalizing the misclassified examples. Moreover, we demonstrated explainability of the decision using gradient-weighted class activation maps(Grad-cam) to identify the most substantial region of the images. The proposed method is evaluated on datasets acquired from Hospitals in Bastia and Dijon. On the test set, our best approach achieved an average accuracy of **90.70%**, F1-score of **92.16%** and sensitivity of **96.42%**, which compares favorably to the state-of-the-art. The proposed method also outperformed similar methods and demonstrated that integrating attention modules improves the classification result.

Keywords: Thyroid nodule, EU-TIRADS, Ultrasound Image, Interoperability, Classification, Deep Learning,

1. Introduction

Thyroid nodules are irregular overgrowth of tissues in the thyroid gland. Most thyroid nodules are not consequential and do not cause symptoms. Some people have one nodule, while others have many. According to the American Cancer Society, there are still numerous deaths due to malignant thyroid nodules. While a moderate percentage of thyroid nodules are cancerous (between 3 and 7% (Hambly et al., 2011)), the death rate for thyroid cancer tends to increase: it augmented by 0.6% per year between 2009 to 2018 according to the Key Statistics for Thyroid Cancer. Autopsy studies have reported incidental thyroid nodules subjectivity up to 50%.

In consequence, early detection and treatment of malignant nodules are very significant. Ultrasound (US) imaging techniques have become an important diagnostic tool in the assessment of thyroid nodules. Though

US images are faster to acquire and effective to analyse thyroid nodules, computed tomography(CT) and Magnetic resonance imaging(MRI) can also be used as imaging tools (Peng et al., 2017). Thyroid ultrasonography has the advantages of being a noninvasive, low-priced procedure widely used to detect and evaluate thyroid nodules risk of being malignant. It plays an important role in providing information such as the nodule positions, dimensions, orientation, and pathologic changes. All in all, it is a highly tactful and core modality for the detection of malignant nodules, though its diagnostic value varies from study to study. Identification of malignancy level is dependent on the quality of the exam, that in turn depends on the physician. Therefore, inter-observer variability exists for the assessment of thyroid nodules. The experience of the sonographer to properly acquire and label the image is substantial, bacause an inaccurate US capture of a nodule might result in unnecessary fine-needle aspiration (biopsy). Hence, an accurate automated diagnosis system is required to avoid unnecessary punctures.

Table 1: European Thyroid Imaging Reporting and Data System

Category	Score Tirads	Eu-TIRADS
0	**–**	**–**
1	**1**	**EU-TIRADS1**
2	**2**	**EU-TIRADS2**
3	**3**	**EU-TIRADS3**
4	**4-6**	**EU-TIRADS4**
5	**7 & more**	**EU-TIRADS5**

Several ultrasound features have been found to be associated to an increased risk of thyroid nodule cancer, the main ones being a cystic composition, a predominantly solid composition, hypo-echogenicity, size, shape (taller-than-wide), margin and the presence of micro-calcification (echogenic foci). Each features are assigned points ranging from 0-3 and the summation of these features' points determine its risk level. In order to standardize the ultrasound report that describes and evaluates thyroid lesions, an agreement which is called an European-Thyroid Imaging Reporting and Data System (EU-TIRADS) has recently established, look (table 1 for more details. From this 1), one can study how nodules range according to the European-Thyroid Imaging Report And System-1 (benign) to European-Thyroid Imaging Report And System-5 (highly suspicious to be cancerous). A high score implies strong suspicion an the need for FNA (Tessler et al., 2018). Usually, thyroid nodules are heterogeneous, composed of various internal echo patterns that are confusing even to experts. Eu-Tirads is a precondition for the Bethesda score system, a reporting system of thyroid cytopathology, which categorizes the nodules as benign, probably benign and malignant based on biopsy features, as shown in (table 2).

The TIRADS score determines the risk level from US images, and helps making the decision on whether to perform a fine needle aspiration on the nodule or not. In clinical routine, if the TIRADS score is above the risk threshold, a fine needle aspiration process is taken on the thyroid nodule, and the Bethesda score, which is the most influential criteria in making the decision to perform surgery on the nodule or not, is calculated from the biopsy features. Stratification and estimating the Bethesda score manually is a tiresome and prone to variability task. Hence, we propose to build a Computer Aided nodule diagnosis system based on Bethesda scores to level the risk and avoid unnecessary surgical process on the patient.

Knowing the orientation of thyroid nodule ultrasound images is one of the important phases in 2D echography analysis. The orientation of a growing nodule is categorized as parallel (when the anteroposterior diameter of a nodule is equal to or less than its transverse or longitudinal diameter) or non-parallel (when the anteroposterior diameter of a nodule is longer than its axial or sagittal diameter). The orientation is categorized according to the relationship between the long axis of a nodule and the long axis of the thyroid gland, regardless of the nodule shape (Shin et al., 2016). For our local ultrasound images, we have two thyroid nodule orientation since two orthogonal views per case are acquired (sagittal and axial). Moreover, we made sure each image acquired contains only one nodule (Fig.1). In this work, we proposed to develop a deep learning algorithm that uses thyroid nodule US images to decide whether a thyroid nodule should undergo a biopsy and to compare the performance of the algorithm with the performance of physicians who adhere to the European-Thyroid Imaging Reporting and Data System (TI-RADS). It is classically composed of four main stages, (I) Pre-processing, (II) Data augmentation, (III) Feature extraction and (IV) automatic classification of benign or malignant. The CAD system should ultimately eliminate the weaknesses of expert dependency, effort, time spend on investigation of nodule and lack of accuracy.

Our work has the following main contributions: 1) We proposed several pre-processing methods that help to enhance the image quality. The pre-processing steps that were implemented are noise removal, cropping, resizing, histogram equalization, and removing artifacts from the images. 2) We demonstrated that generating synthetic images can improve the detection result. 3) We integrated attention modules to the Densenet deep learning architecture, which brought substantial improvement to the classification results. The incorporated attention module specifically helps the network to

Table 2: The Bethesda System for reporting thyroid Cytopathology

Bethesda Category	**Description**	**Risk of malignancy%**	**Managements**
0	**Benign**	**0**	Normal
I	**Undetermined**	**1 - 2**	**Repeat FNA**
II	**Benign**	**3 - 5**	**Follow-up**
III	**Follicular lesion**	**5 - 15**	**Follow-up**
IV	**Follicular neoplasm**	**15 - 30**	**follow-up**
V	**Suspicious malign**	**60 - 75**	**Surgical lobectomy**
VI	**Malignant**	**97 - 100**	**Total Thyroidectomy**

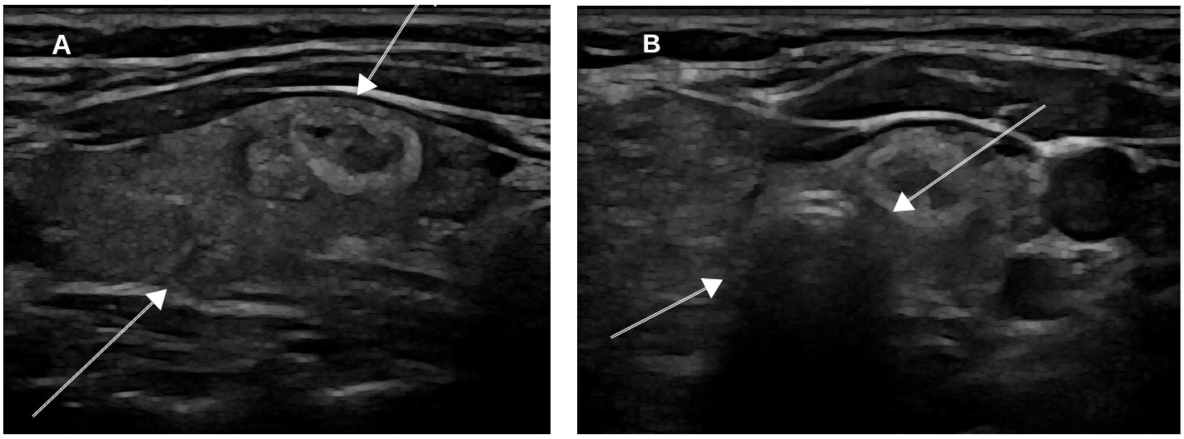


Figure 1: Illustrative example showing the two orientations of thyroids nodules in ultrasound images. The white arrow points at flat shape that indicates in (A) a longitudinal view, while in (B) they point at round structures which are associated to a transversal view nodules respectively.

focus on the strong features while estimating the malignancy. 4) We showed that computing focal loss with automatic assignment of loss weights based on the sample distribution of classes enables to overcome the data imbalance problem and improves the model performance with little rise for the computational cost. 5) We have compared several deep learning methods to overcome our dataset's main limitations which are its low image quality and its small size. 6) We illustrated interpretability of the classification of benign and malignant task using heat maps derived from Grad-CAM.

2. State of the art

A significant number of studies were carried out on this thematic area scientifically (Frates et al., 2005). Nodule detection studies can be classified into two main categories: non-machine learning based and machine learning based techniques. The non-machine learning are usually standard image processing approaches with semi-automatic methods. It is mainly focused on thresholding the risk level by physicians. Most Machine Learning CADs are aimed to outperform experts' assessment accuracy. Basically, the studies involve the comparison of Computer Aided Diagnosis systems with the manual classification of the nodules by experts. We discuss hereafter the state-of-the-art for methods related to our work.

2.1. Classical machine learning algorithms for the detection of malignant nodules

In recent years, few machine learning methods have been proposed to diagnose the malignancy risk of nodules. In (Peng et al., 2017), the authors investigated the feasibility of applying the first order texture features to diagnose thyroid nodules in Computed tomography image (CT). A total of 284 thyroid CT images from 113 patients were used in this study. Their method involved the following steps: first, regions of interest (ROIs) were extracted manually by a physician. Second, some standard filters like median filtering were applied to reduce photon noise before feature extraction. Third, a support vector machine (SVM) algorithm was applied to predict the classification task. The results of this paper work were measured using accuracy and sensitivity scores of 0.880, 0.821 respectively. Chi et al (Chi et al., 2017) presented a CAD system to identify as many malignant nodules as possible. The images used in this research work were from the following two datasets: **Database 1** is a publicly available thyroid ultrasound image database proposed by (Pedraza et al., 2015), consisting of 428 thyroid ultrasound images [1]. **Database 2** is a private database, consisting of 164 thyroid ultrasound images. A pre-trained GoogLeNet deep learning approach was used for feature extraction and a Cost-sensitive Random Forest as classifier to identify the malign nodule.

2.2. *Deep learning algorithm for detection of malignant nodules*

Deep learning design has showed a visible improvement in diagnosis of malign cancer from nodules. Most of the methods that have been implemented in this area use B-mode ultrasound images [2], as we do. (Buda et al., 2019) tackled the classification problem using their local dataset and a deep learning algorithm to provide management recommendations for thyroid nodules observed on ultrasound images, and compared its performance with physicians. They used 1278 nodules for training, and 99 nodules for testing. Their method used three main stages to accomplish the task using a Faster R-CNN network: First, they extracted the region of Interest (ROI) based on caliper markers localization. Secondly, they predicted the risk of malignancy using a multi-task CNN. Lastly, they built a stratification into risk level using the model. They showed that the performance of the algorithm was similar to that of the consensus of three expert readers. On the test set, deep learning achieved an Area under the curve (AUC) of 0.87 (95% Confidence Interval (CI): 0.76, 0.95), which is close to that of expert consensus (0.91; 95% Confidence Interval(CI): 0.82, 0.97). (Wu et al., 2016)'s study consists of 970 radiographical proven thyroid nodules from 970 patients. In this related work, a radial basis function (RBF)–neural network (NN) method was used as classifier. The deep learning method under-performed with respect to the experienced experts. Identification of malignancy by the experienced experts achieved the highest predictive accuracy of 88.66% with a specificity of 85.33%. whereas the radial basis function (RBF)–neural network (NN) achieved the accuracy of 84.74% with specificity of 76%. (Koh et al., 2020)'s research diagnoses thyroid nodules from ultrasound images by ensemble of convolutional neural networks (CNNs). They collected datasets from multiple center, which amount to 15,375 US images of thyroid nodules. CNNs demonstrated higher area under the curves (AUCs) to diagnose malignant thyroid nodules (0.898–0.937 for the the internal test set and 0.821–0.885 for the external test sets) than the physician. AUC was significantly higher for CNNE2 than the one from physician decisions on their test set (0.932 vs.0.840). Recently, a few studies have been proposed to better classify nodules by involving an unsupervised learning method, called Generative adversarial deep learning network (Hang, 2021). This research work diagnoses thyroid nodules using images by the fusion of conventional features and residual-generative adversarial network (Res-GAN) features. Training sets come from an open-source thyroid nodule image dataset named ”database of thyroid ultrasound images” (TDID). Most GANs nowadays are based on the Deep Convolutional Generative Adversarial Networks (DCGANs) architecture. The method which involves the combination of the deep features with the conventional features gives a promising performance in the model.

Focal Loss is a loss function that addresses class imbalance during training in tasks like image classification. It applies a modulating term to the cross entropy loss in order to focus learning on hard to classify examples (Lin et al., 2017). It is a dynamically scaled cross entropy loss, where the scaling factor decays to zero as confidence in the correct class increases. It has two hyper-parameters which are called alpha-α and gamma-γ. The focal loss introduces one new hyper-parameter, the focusing parameterγ, that controls the strength of the modulating term. When $\gamma = 0$, the loss is equivalent to the cross entropy(CE) loss. We define the focal loss in Eq. 1:

$$FL(p_t) = -(1 - p_t)^{\gamma} log(p_t) \tag{1}$$

Where FL is the focal loss, and hyper-parameter γ ranges from 0 to 5.

Attention Mechanism: It is well known that attention plays an important role in human perception, and so does it in artificial neural networks. The main purpose of the attention module is to automatically choose the most important intermediate features, and to carefully refine the best feature maps through the network. There are two well-known convolutional attention modules. The two sequential sub-modules are called channel and spatial attention. **Channel Attention** utilizes the inter-channel relationship of features maps. Every channel of a feature map is considered as a feature detector (Zeiler and Fergus, 2014). Channel attention multiplies the output after max-pooling or average-pooling with a shared network coefficients to scale feature maps. **Spatial Attention** creates a spatial attention map by ex-

[1] https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images

[2] http://cimalab.intec.co/?lang=en&mod=project&id=31/

ploiting the inter-spatial relationship of features on the input images. The difference with channel attention is that spatial attention focuses on the locations where lot of information found, rather than on pondering whole feature maps.

3. Material and methods

3.1. Objective

Our objective was to develop an automatic thyroid ultrasound image classification system to prevent unnecessary fine needle aspiration (FNA). Benign-malignant nodule classification at early stage is a crucial step to prolong patient survival. The aim of this study is to propose a method for predicting nodule malignancy based on deep biopsy features. We came to achieve this general objective by tackling three main challenges. **First**, we have a small dataset to carry out the process of building a computer aided detection system. It is a challenge for classification due to the fact that a diagnosis task is very sensitive and usually needs plenty of datasets to train on. **Second**, the Image format is Joint Photographic Experts Group(JPEG). Since it applies lossy compression to images, this can result in a significant reduction of quality on the images. Also we do not have access to the image resolution as we would with Nifti or DICOM images. Hence, the dataset needs to be pre-processed in order to get important features from the images. **Third**, the target classes show uneven distribution of observation, as the negative (benign) class has more observation than the positive (malign) label.

3.2. Dataset

In this proposal, we used 595 US images of thyroid nodules in Joint Photographic Experts Group format coming from two sources. **Private Dataset:** It contains a set of thyroid Ultrasound images that includes a complete annotation and diagnostic description of thyroid lesions, using the Bethesda score (biopsy features) interpretation criteria. The images are labeled by experts from the Hospitals of Bastia and Dijon. Hence, the annotation criteria might be affected by inter-observer variation. The private database consists of **534** thyroid ultrasound images. All Images from the aixplorer vendor have a size of 1440 × 1080, while from CANON vendor have image size of 1280 × 960 as they come from the CANON vendor. 161 images in the database are labeled positive (with Bethesda score III to VI), while 373 images are labelled as negative (with Bethesda score = 0 or II),including the test set.

Public Dataset is a publicly available and 61 thyroid ultrasound images has been used in our research from the public link mentioned on the footer. [3].The images are in Joint Photographic Experts Group(jpg) format with different dimensions. All the cases are labeled as malignant (positive) and it is an open access resource for the scientific community projects.

We split the validation set from the training set randomly. Out of these images, 473, 71 and 51 images are used as a training set, validation set and test set, respectively. The test set are chosen carefully from both vendors. For more details, see table 3 below.

3.3. Pre-processing

Images are collected from different ultrasound machines, leading to imbalance in exposure, size and other parameters. Before undertaking feature engineering, we should pre-process the raw images by including noise reduction and image enhancement in order to feed the model with better quality of images. Image enhancement is the procedure of improving the quality and information content of original digital images data before processing. We introduce several commonly used image enhancement techniques for our experiment, which are cropping, resizing, interpolation, histogram equalization, adding variability, normalization, removing artifact from images, and Gamma correction.

3.3.1. Normalization

Feature scaling is one of the most important data preprocessing step, the intensity of every patient image is normalized to have zero-mean and unit-variance. Algorithms that compute the distance between features are biased towards numerically larger values if the data is not scaled, so calibrating the details of images from different sources to the same scale is consequential.

3.3.2. Cropping and resizing

The images were firstly cropped and resized to have the same resolution which is the physical space represented by each pixel in the image, as shown in figure 2.

3.3.3. Histogram Equalization (HE)

HE usually increases the global contrast of many images, especially when the image is represented by a narrow range of intensity values. Through this adjustment, intensities can be better distributed on the histogram, utilizing the full range evenly. If most pixels are concentrated in the low gray area, the image will appear completely dark, but if they are concentrated in the high gray area, it will appear bright. Histogram equalization is therefore applied to elevate the contrast of the image, thus improving the visual effect of the image (Patel et al., 2013).

[3] https://www.kaggle.com/datasets/dasmehdixtr/ddti-thyroid-ultrasound-images

Table 3: The distribution samples in training, validating and testing groups of the dataset (I=Images), positive-samples =Bethesda -Score: 3, 4,5 and 6: negative-Samples =Bethesda-Score =0 and 2

Dataset	**Samples (Bastia and Dijon)**	**Positive-Samples**	**Negative-Samples**
Training	544 **I**	201 **I**	343 **I**
Val-Split	71	**Random**	**Random**
Testing	51 **I**	21 **I**	30 **I**

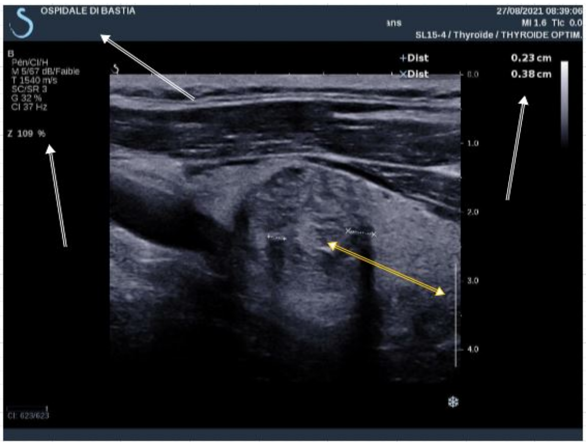


Figure 2: Ultrasound Image with noise and artifacts covering the textures.**White arrow:** indicates noises that appear on the images.**Yellow arrow:** indicates artifacts on the images

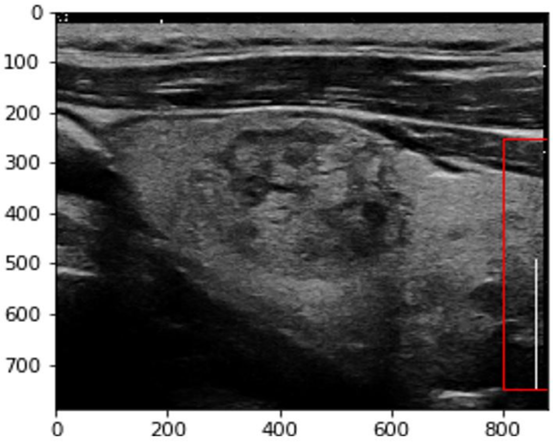


Figure 3: Pre-processed thyroid Ultrasound image and details of how bounded by the red rectangle to remove the artifact

3.3.4. Removing Artifacts

In this step, we implemented an opening morphological operation to discard artifacts on the images. Artifacts are characteristics which appears in an image and which are not present in the original imaged object. They usually appear at the center or corners of medical images. A rectangle method was used to make bounding boxes around artifacts via an anchor point xy and its width, height and 4-connectivity method. We used a bounding box in order focus the kernel in some part of the images, otherwise we may loose essential information if this method is applied for the entire image. Following this, morphological opening was applied to remove small white thin lines from an image while preserving the shape and size of larger objects in the image. We used a small structuring element to maintain the texture information, as shown in figure 4.

3.4. Data Augmentation

As mentioned above, one of the big challenge for this thesis work is to overcome the limitations of the dataset. We employed on-the-fly data augmentation, which allows transformed images to be produced from the original images with very little computation as the transformed images are not stored on disk. We used the TensorFlow ImageDataGenerator class to augment the images. Each generated image is randomly different from the original in certain aspects depending on the augmentation techniques. We do this by extracting random 800 × 600 patches from the various size of the images, and train our network on these extracted patches. At every iteration, batch size of transformed images are generated with different parameters like shifting, rotating, flipping, etc. Such image augmentation techniques not only expand the size of the dataset, but also incorporate a level of variation in the dataset which allows the model to generalize better on unseen data. To observe the produced images, see Fig. 5.

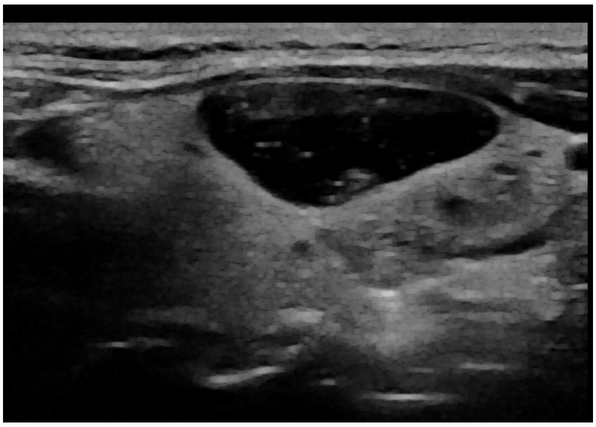


Figure 4: Ultrasound Image after pre-processing

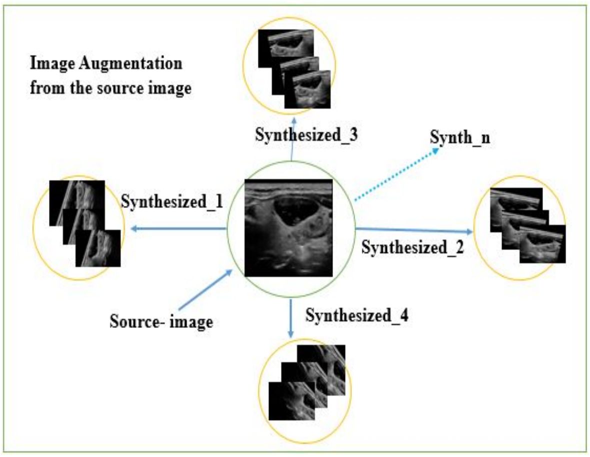


Figure 5: data augmentation: A source image is used to synthesize images like synthesized_1(rotation-range), synthesized_2(shift-range), synthesized_3(shear-range and shift-range),etc, on US image respectively

3.5. *System and Running*

Recently, boosting the training to a satisfactory extent was achieved by using Graphics processing unit (GPU)(Chen et al., 2014). This enhancement allowed us to efficiently utilize computational resources from the available GPUs and other software package tools in Imagerie et Vision Artificielle (ImViA). We used persistence-m nvidia type of GPU with 12GB size of memory and CUDA version of 11.6 to launch the training for 320.40 min. We used Ubuntu 20.04.4 LTS as operating system and a virtual environment with python 3.9.

3.6. *Proposed Pipeline*

The proposed pipeline consists of US image as inputs, pre-processing, data augmentation, deep learning based feature extraction, interpertability and thyroid nodule classification, as illustrated in (Fig.6). We compared several deep learning networks using this pipeline to different learning schemes. In our approach, the pipeline consists of an additional module. The module(attention and conv block) is incorporated with Densenet to extract features and classify malignant nodules, as can be seen from the figure see(Fig.6). The attenton module are integrated between convoutoinal layers. The conv block are appended at the attention module in the end for providing the output of the classification task.

3.6.1. *Network Architecture*

Convolutional neural networks have become the dominant machine learning approach for object classification (LeCun et al., 1989). We implemented three different deep learning techniques to tackle ultrasound thyroid image classification. First, we built a simple convolutional neural networks, and evaluated it on our dataset. Second, we implemented fine-tuned deep convolutional neural networks like Resnet-18, Efficient-netB0, and Densenet121. Lastly, we employed the proposed deep learning architecture that incorporate an attention-conv module and Densenet. inside of it. The different architectures details are discussed as follow afterwards.

3.6.2. *Convolutional Neural Network models*

We first built a simple CNN model with twenty five (25) layers and let the neural network learn from scratch. This section briefly discusses the role of some components in CNN architectures. **Convolutional layers** are composed of a set of convolutional filters where each neuron acts as a feature detector and extracts feature pattern. **Pooling layer:** Once features are extracted, its exact location becomes less important as long as its approximate position relative to others is preserved. Pooling or down-sampling is an operation that sums up similar information and outputs the dominant response within this local region in order to compress information spatially. **LeakyRelu Activation Function:** It is how the weighted sum of the input is transformed into an output from a node or nodes in a layer of the network.It adds non-linearity to the transformed outputs of layers. **Batch normalization** is used to address the issues related to the internal covariance shift within feature maps. The internal covariance shift is a change in the distribution of hidden units values which slows down the convergence(Ioffe and Szegedy, 2015). Data are scaled not only before entering training, but continues to stay scaled while it is training. **Dropout** introduces regularization to the network, which ultimately improves generalization by randomly skipping some units or connections with a certain probability. **Fully connected layers** are mostly used before the output for the final decision. It is a global operation and takes input from feature extraction stages to globally analyses the output of all the preceding layers. **Softmax Activation function** is used as the activation function in the output layer of neural network models in order to predict a binomial probability distribution.

3.6.3. *Deep Convolutional Neural Network*

DCNNs are a type of Neural Networks, which have deep layers and have shown exemplary performance on several competitions related to computer vision. We have used fine-tuned deep CNN architectures where layers are added to the trained model to adapt it for our task. We have tried different deep learning architectures that have been already trained on the ImageNet database(Deng et al., 2009). Hence, we got an opportunity to compare the architectures performance based on the result with our proposed methods. We have also confirmed that fine-tuned models work better than the model that we built and learn from scratch,See section (5) for more details. The following pre-trained deep

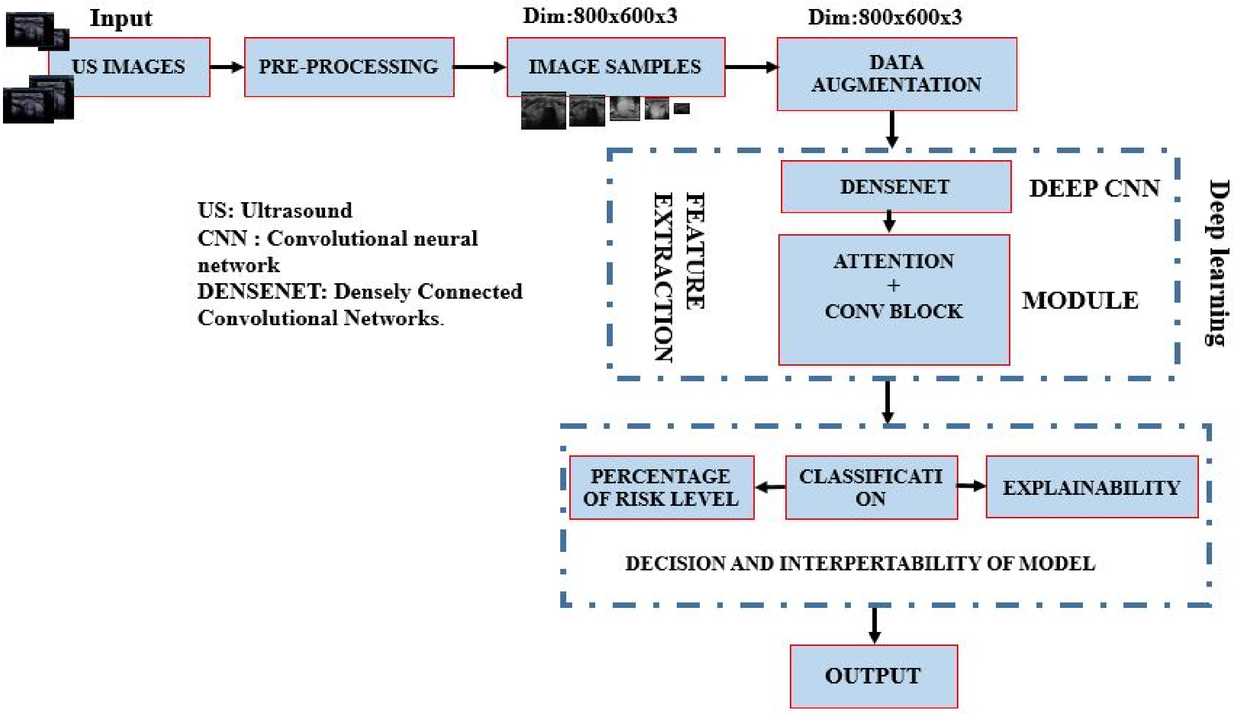


Figure 6: **Proposed pipeline**:Deep convolutional neural networks model for image classification, **800x600x3**: The dimension of image samples and its channel, **US**:Ultrasound image given as input for the model

Convolutional neural network models have been used to perform classification in our task. **Residual Networks, or ResNet-18** is a convolutional neural network that is 18 layers deep. we have implemented the ResNet-18 architecture in two steps in the Tensoflow framework(Pang et al., 2020). we first discarded layers after the 18^{th} in ResNet-50, then we added a block that we have designed to fit our problem and train the model. The block is composed two layers(ReLU activation function and GlobalMaxPooling2D) and one classifier function. ResNet-18 helps to overcame the vanishing gradient problem issue by introducing a so-called skip connections-that leaps over one or more layers(He et al., 2016). Some Layers were frozen to prevents the weights from being modified and random seed method is used to controlling random initialization of weights. **Efficient-netB0** is a deep convolutional neural network architecture with 237 layers. It uses a scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice that arbitrary scales these factors, the Efficient-NetB0 scaling method uniformly scales network width, depth,and resolution with a set of fixed scaling coefficients. EfficientNetB0 is a well known neural network architecture for compound model scaling methods though it does not work well for our task due to overfitting. In other words, scaling every dimensions balance all dimensions of the network depth, width, resolution and improves model performance (Tan and Le, 2019). We have set drop-connect-rate =0.4 during our demonstration.We happended two layers and one classifier function to this architecture to adapt the model to our task.

Densenet121: It consists of 427 layers with 120 Convolutions and 4 average pooling layers. This network was designed to address the problem of vanishing gradient by directly connecting each layer to every other layer in a feed-forward fashion. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers and further exploits the effects of shortcut connections. Unlike residual neural networks (ResNets), the feature maps received from previous layers are concatenated not summed. Other than tackling the vanishing gradients problem, Densenet has a strong feature propagation, feature reuse and reduced number of parameters(Huang et al., 2017). And also, a DenseNet network has translation layers between adjacent block and uses to update the size of feature-map through convolution and pooling layers. By knowing these all features of Densenet pre-trained model, We fine-tuned this model to our specific task and outperform other models that we mentioned them in this section. Hence, we propose an approach that integrated Densenet and module for the betterment of the result.

3.6.4. *Proposed Architecture*

We implemented several deep learning architecture to detect the malignant nodule. Densenets are efficient for classification tasks, because they have skip connections and better transmission of features across the network. Hence, we propose a method that incorporates Densesnet121 and module(attention+convolution block) for further improvement of the result. we chose Densenet121 for further improvement for two reasons: First, Densenet architecture gave the the highest result comparing with the other models. Second, this network does not suffer with the problem of vanishing gradient. One simple interpretation of this is that the output of the identity mapping was added to the next block, which might impede information flow if the feature maps of two layers have very different distributions(Simonyan and Zisserman, 2014). We proposed an approach that incorporates a module(attention+conv blok) to Densenet121's architecture, intending to enhance the performance of the networks. There are two types of attention module 2, and we have arranged them in a sequential manner(Woo et al., 2018). We have used channel attention first and then spatial attention,as illustrated in (Fig.7), because of this arrangement gave the better result, than the spatial-first chain. The attention modules are integrated between convolutional layers to refine the most important feature maps, as illustrated in figure(Fig.9). Channel attention helped the proposed model to concentrate the substantial information of the input US image. Spatial attention brings focus to specific parts of spatial information, pondering feature maps to enhance regions of interest on the images. The attention module down-sampled feature maps using average and max-pooling operations. Therefore, the attention mechanism played a great role in guiding the networks to concentrate on the most important feature maps.

We have also add a block that consists the succession of layers: Separable convolution, batch Normalization, GlobalMaxpooling2D, rectified linear activation unit, dropout, a fully connected layer and finally a Softmax classifier function, as can been seen in this figure (Fig.8). **Separable convolution** is a kernel in which a single convolution can be divided into two or more convolutions to produce the same output with a much lesser computation cost. **GlobalMaxpooling2D** used to reduce the dimensionality of the feature maps and give one maximum value for a whole region to strongly compress information. The rest layers have been explained in this(Sec 3.6.2) section. Basically, We appended the block to the proposed network to minimize the covariance shift problem and let the network learn representation pattern our dataset. Because, the block contains batch normalization, separable conv,etc. Therefore, the proposed method includes: Densenet121, attention module and convolutiona blocks and a classifier function.

4. Optimizer

We used the adaptive moment estimation (ADAM) optimizer to monitor the training and optimize the convergence when training the model.

4.0.1. Loss Function

We used loss function to optimize the parameter values in the proposed neural networks. Basically, It is a method of evaluating how well our model get well with the the given input data. As an objective function, we have used two different loss functions to evaluate our model by computing the error. **Categorical Cross Entropy Loss** is the probability value among the given classes for a classification task. Cross-Entropy calculates the average difference between the predicted and actual probabilities, as explained in this (Eq. 2) equation.

$$\mathbf{L} = -\sum_{i=1}^{N} y_i log\hat{y}_i \tag{2}$$

where $\mathbf{L}$ is loss, N is output-size the output size, $\hat{y}_i$ is the i-th scalar value of the model output, y_i is the corresponding target value, and the output size is the number of scalar values in the model output. Categorical Cross Entropy Loss is not recommended to be used as error optimizer for imbalanced dataset. Hence, we proposed another loss function which is called focal loss, for more detailed, (Sec 2) section. Focal loss is the modification of cross entropy loss and have two hyper-parameters. while α balances the importance of positive/negative examples, γ tries to penalize the misclassified examples. As γincreases, the shape of the loss changes so that easy examples with low loss get further discounted. We tried γ values from 1 to 5 and observed that model's classification accuracy increases with γ values. However, It became almost constant after γ reached 3.5 in our case. According to our experiment, Focal loss works well and helps diminishing the impact of data imbalance compared to Categorical Cross Entropy Loss, see (table 4) for more detail.

4.1. Performance Evaluation Metrics

After applying the deep learning algorithms, evaluating the model is very important to know how the system behaves on unseen data. A tool which is called a metric is introduced to measure the accuracy of the models. In this paper, we use several common metrics for classification problems to obtain valuable information about the performance of algorithms and to run a comparative analysis. These metrics are accuracy, f1-score, confusion matrix and classification report. We rarely used sensitivity and specificity in the evaluation mechanism as confusion matrices are easier to interpret.

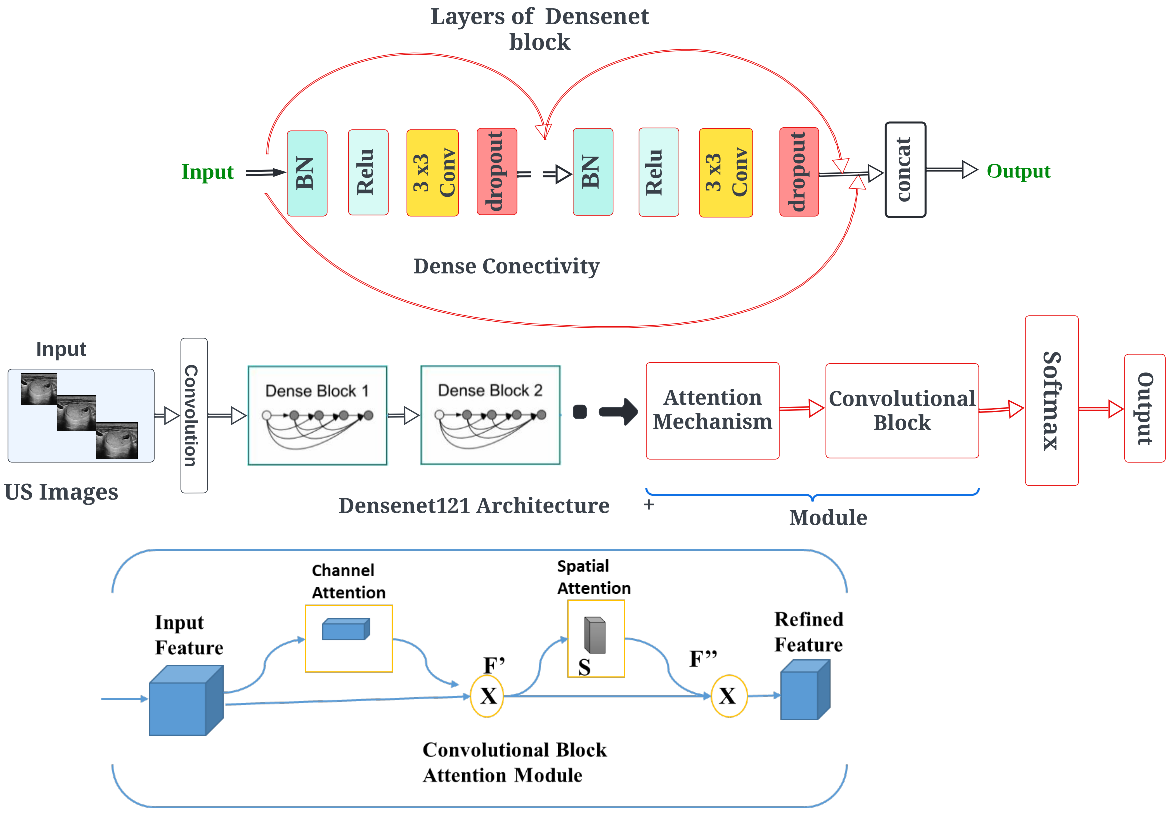


Figure 7: Proposed DenseNet Architecture with concatenated attention module and block

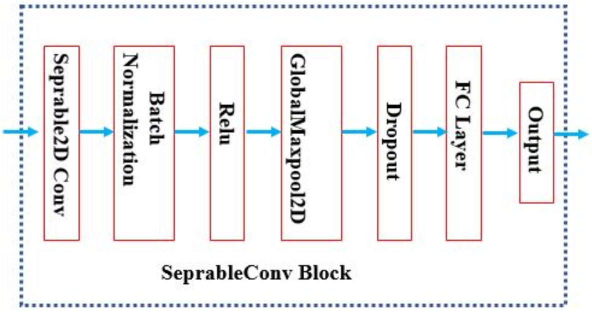


Figure 8: Convolutional block that consists of six top layers

4.1.1. Accuracy

It is the most used and maybe the first choice for evaluating an algorithm performance in classification problems. It can be defined as the ratio of accurately classified data items to the total number of observations, see (Eq. 3). Despite the widespread usability, accuracy is not the most appropriate performance metric in some situations, especially in the cases where target variable classes in the dataset are unbalanced (Vakili et al., 2020).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

TP, TN, FN and FP represent the True Positive, True Negative, False Negative and False Positive of predicted image respectively. Basically,It is the summation of TP and TN which are correctly classified over the total datasets.

4.1.2. F1-score

This metric, which is also known as f-score or f-measure, takes both precision and recall into consideration in order to calculate the performance of an algorithm(Goutte and Gaussier, 2005). Mathematically, it is the harmonic mean of precision, see (Eq. 4a) and recall, see (Eq. 4b) formulated as follows (Eq. 5):

$$precision = \frac{TP}{TP + FP} \tag{4a}$$

$$Recall = \frac{TP}{TP + FN} \tag{4b}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

We observed that accuracy metric does not work well with data imbalance condition, Since it does not distinguish between the numbers of correctly classified images of different classes. F1-score is a proper measure when working on classification tasks in which the data points are imbalanced.

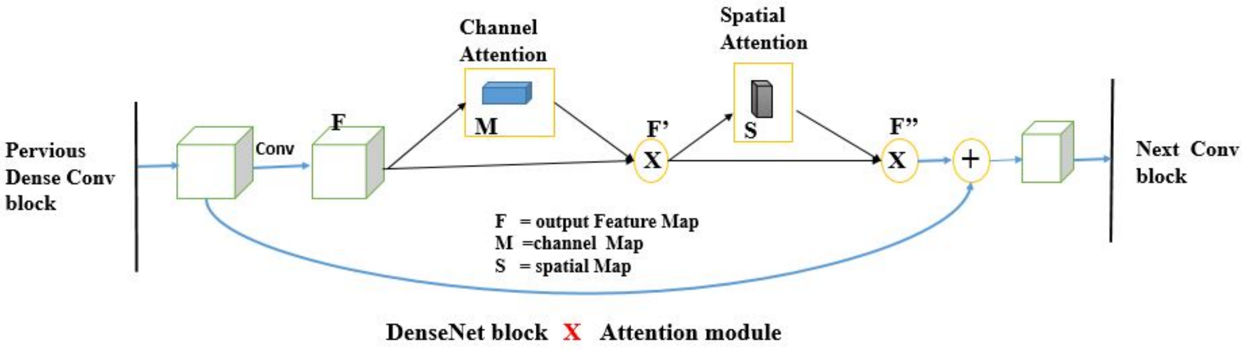


Figure 9: Attention module integrated with DenseBlock in Densenet architecture, **F** indicates the refined outputs feature map

4.1.3. Confusion Matrix

This matrix is one of the most intuitive and descriptive metrics used to find the accuracy and correctness of a machine learning algorithm. Its main usage is in classification problems where the output can contain two or more types of classes (Townsend, 1971). We can compute sensitivity focusing on the True positive rate and specificity focusing on the false positive rate from the confusion matrix.

4.1.4. Classification Report

Classification report is an evaluation metric in deep learning machine learning. It is used to display precision, recall, F1-score and support for the trained model as package. Support is the exact number of occurrences of each class in the specified testing dataset.

4.2. Thyroid Nodule Risk Level Assessment

The estimation of the risk level from ultrasound imaging of thyroid nodules is extremely difficult. In this thesis work, we provided a probability expressed in percentage(0% to 100%) to represent the malignancy risk level. This would be very helpful for physician to take decision in the need for fine needle aspiration. For instance, there are some images that obtain a score between 40% and 60%, which we call "gray zone" associated to an uncertain prediction, and which would need strict follow up of treatment. If the image get a score above 61% of benignity, it is in normal status. Otherwise, it might be needed to take a serious measurement or follow up at the patient.

4.3. Interpretability

Deep neural networks have been widely-known for their magnificent performance in playing with different machine learning tasks. However, because of their exceeding-parameterized "black-box" nature, it is usually back-breaking to understand the prediction results of deep models(Dong et al., 2017). Interpretability(Explainability) is the degree to which a human can realize the cause of a decision and outputs can be described in the way that make sense to deal with deep understanding how a model makes prediction. It also helped us to debug the network. In our proposed method, we have used the gradient-weighted class activation map (Grad-CAM) to give insights on the decision making (Selvaraju et al., 2017). Grad-CAM uses gradients to give a coarse localization map highlighting the most substantial regions in the input image when predicting the result. The class activation map simply indicates the discriminative region in the image which the CNN uses to classify that image in a particular category. We can identify the importance of the image regions by projecting back the weights of the output layer on to the convolutional feature maps. A graphical representation which is called heatmap method is responsible for highlighting the discriminative region used by the model.

4.4. Training

Weights were initialized using *He normal* initialization method (He et al., 2015). It draws samples from a truncated normal distribution centered on 0. The optimization of the weights are done using Adam as the optimizer with learning rate of 0.0001. The mini-batch size was 8, because a small batch size is recommended for small datasets. The models were trained until convergence for various numbers of epochs depending on the type of the model. We empirically selected a weighting factor of 0.50 for α, and 3.50 for γ, the hyperparameter in of the focal loss. For above 200 epochs, we were using α =0.50 and γ =2.0, as the fact γ hyperparameter has reciprocal relationship with number of epochs. We used the programming language Python and the library Tensorflow to implement the deep learning models. We fixed the random seed to 42 to set the integer starting value used in generating random numbers. Setting random seed to fixed value is very important so as to get stable or gives reproducible result with TensorFlow framework.

In order to avoid over-fitting, we adopted three techniques: dropout, early stopping and data augmentation.

Dropout is a regularization technique where randomly selected neurons are dropped during training. The ignored neurons will not have contribution during a forward and backward propagation. Dropout reduces overfitting by preventing complex co-evolution on the training data. In our all experiments, we used dropout with a probability of 0.25.

During training, the training and validation losses decrease, usually in the staring the training and validation loss decreases. As the number of epochs increases, the training loss will continue to decrease but the validation loss will slowly diverge over time. This phenomenon indicates overfitting, which does not generalize well to unseen data. To monitor this, we used early stopping techniques, which triggers when the validation loss starts to increase. The training immediately stops after certain number of epochs, to give the possibility to the validation loss to decrease again, in case the training curves are noisy. In our experiments, the patience parameter for the early terminating of the training process was 20 epochs.

The last techniques that we employed to handle overfitting is data augmentation. A lot of similar images were synthesized synthesized applying transformations such as shearing, rotating,zca-whitening,etc. This helps to artificially increase the dataset size, which helps avoiding over-fitting. The reason for it is that, as we generate more data,the model can not learn by heart the training data and is forced to learn generalizable features and give good performance on unseen data.

5. Experiment and results

We conducted experiments with the models, and training schemes previously mentioned in the methodology section,(Sec (3). The proposed network outperform other networks. To evaluate the classification results of thyroid nodule, we used accuracy, F1-score, confusion matrix, specificity and sensitively metrics. To evaluate our models, we split the dataset between train, validation and test sets, as done traditionally. The validation set is drawn from the training data, it is kept aside the optimization to set hyper-parameters and to detect overfitting. Train-validation-test evaluation methods as well as a validation set that are separate section from the training dataset to get evidence how well the model is performing on images that are not being used in training.

To evaluate the effect of pre-processing in our method, we compared the results with and without the mentioned pre-processing steps. when we say unprocessed image, images are given as raw data and not scaled for the proposed method. This direct classification of malignancy from the full-sized unpre-possessed thyroid ultrasound images(Fig.2). our method yielded accuracy and F1-Scores of 0.772 and 0.813 respectively. The model suffer from overfitting problem due to the noisy and unrepresentative training data. In the sense that model is learning a detail of noise in the training data to the extent of it negatively impacts the performance of the model on raw data. However, when we employed pre-processing, our method achieved an improved accuracy and F1-scores of 0.9007 and 0.9216 respectively.

We can observe the effect of changing the loss function by looking at the diagram in (Fig13), and (table 4). We can see that the focal loss helps to deal with a limited and imbalanced dataset. Especially, the loss curves are less noisy with focal loss than with cross entropy loss function. Focal loss is used often with hyper-paramaters of $\alpha = 0.50$ and $\gamma = 2.0$. but we used other values after tuning which ranges from $\gamma = 1$ to $\gamma = 5$.

Table 4: Quantitative comparison of loss functions using accuracy and F1-score with proposed approach, ±:shows small variation of the value per training

Loss Function	Accuracy	F1-Score
Cross-entropy	0.850 ± 0.030	0.840 ± 0.045
Focal Loss	**0.8700±0.0250**	**0.9005±0.0216**

We also demonstrated that adding synthesized images improves the performance of the model effectively. The results of the proposed model with and without data augmentation, and with a batch size of 8, are shown in this table (5).

Table 5: Classification results of the proposed architecture with data augmentation effect, ±: shows small variation of the value per each training

Metrics	Proposed w/o augmentation	Proposed with augmentation
Accuracy	0.701 ± 0.041	0.870 ± 0.021
F1-Score	0.750 ± 0.037	0.900 ± 0.021
Specificity	0.694 ± 0.036	0.850 ± 0.022
Sensitivity	0.802 ± 0.028	0.9300±0.0342

We compared our proposed approach with four different networks regarding training time and performance. We can see that the proposed architecture works better than the other architectures on a number of the same experiments, see(table 6). Hence, Densenet incorporated with our attention module outperforms other neural networks when considering all metrics.

For the 51 test nodules, the proposed deep learning algorithm outperform the reported results of the previous related research works, as can be seen in this table

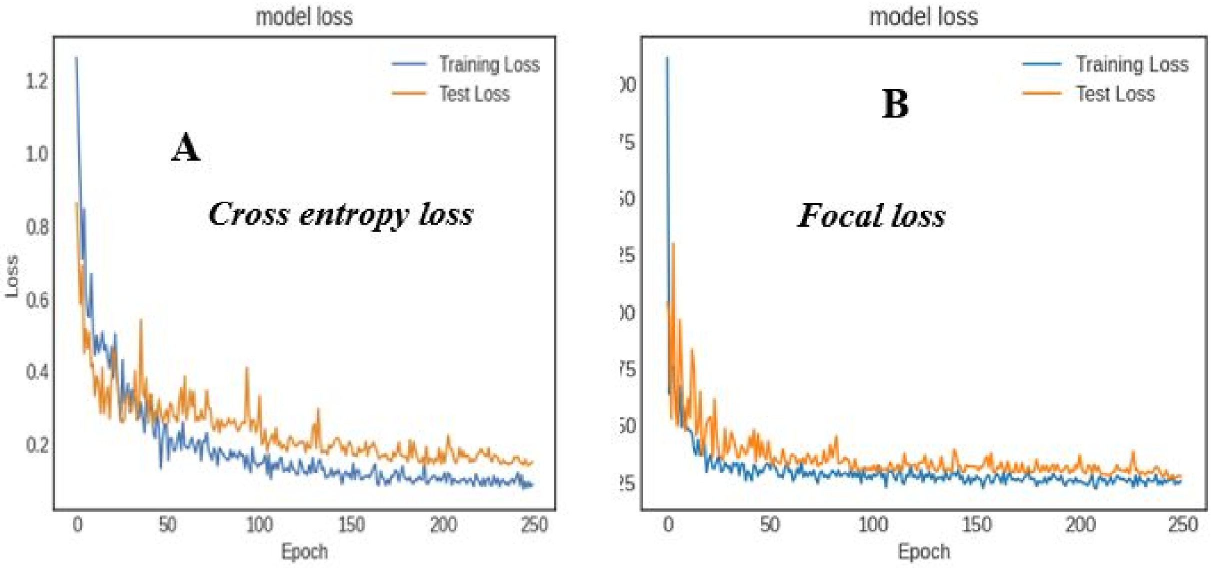


Figure 10: Qualitative comparison of loss functions on monitoring the proposed model during training. A) Cross-Entropy loss, B) Focal loss

(Table 7). Taking into account the inter-observer variation of manual identification of malignancy, the overall accuracy of the proposed deep learning method for thyroid nodule biopsy recommendations is better than the experts, as shown in (Sec 2). These results, however promising, were obtained on a small dataset, and can not be compared directly to other studies.

We compared the proposed method with fine-tuned Densenet121 network, and three other networks which employed in the same pipeline except the modules. The module(attention-conv block) has only integrated with the proposed approach. The hyper-parameters were set to fixed value during the comparison. But, we got different results from each approach due to variation in size, behaviour and structure of the network architecture. We observed that significant changed can be achieved in the result by modifying the networks. The new method is proposed on a little modification of Densenet architecture, which is consolidating a module within it. The proposed method outperform others in all experiments. This is due to the fact that attention method and conblock are playing good role in extracting the most important features. The comparison has been done both quantitatively and qualitatively, as shown in (table 6) and (Fig. 11) respectively. EfficientNet did not perform well on both classes comparing with the other models due to the degradation problem. We obseved that EfficientNet model tried to ends up memorizing the data patterns and put up with random fluctuations. We can say that this model is suffering with gradient vanishing problem when it trained with our dataset. Hence,it has low performance on the test set with average F1-score of 0.7418. The model which is built and learnt from scratch has performed well, but a bit less than ResNet-18. Because, the model sufferers with model complexity due to the huge number of parameters. ResNe-18 has good performance on the on the classification task due to its size and special structure for handling gradient vanishing issue, but it is biased to positive samples. Well, its overall result is not promising like Densenet, and the proposed method. Densenet is very efficient in handling and reusing features maps with dense connections. And also, it has a translation layers that helps to update the size of feature-maps through layers. It performs well next to the the proposed method. We added very important module to the Densent that consists attention and conv block. The main task of this incorporated module, is to guide the architecture to focus on the most substantial features. The output of feature maps from dense convolutional layer is given as input for these module to downscale and forward as output for the next convolutinal layers, as illustrated in (Fig 9). The proposed method showed high performance with accuracy of 0.9007 and F1-score of 0.9216. We made qualitative comparison between the fine-tuned Densenet and proposed technique, illustrated here(Fig 13).

As shown in (table 6), we also compared the number of trainable parameters of the architectures. ResNet-18 has the lowest number of trainable parameters which is 0.25 million. The proposed method have has 0.75 million parameters. While the CNN has the highest number of parameters with 15.4 millions, which can be explained with the high numbers of parameters for fully connected layers. On top of that, the proposed method can be used to estimate the malignancy risk as illustrated on (Fig 16) by exploiting the non-binarized output. As we can see from the (Fig: 17), the confidence interval is 91% percentage. This indicates that the nodule is in normal condition (not cancerous). This estimation gives an important information for the physician to interpret

Table 6: Accuracy and F1-Score comparison of various methods for thyroid nodule classification

Methods	Accuracy	F1-Score	No of Parameters
CNN	0.750 ± 0.045	0.770 ± 0.064	15.4 million
EfficientNetB0	0.760 ± 0.014	0.770 ± 0.018	4.7 million
ResNet18	0.840 ± 0.020	0.835 ± 0.034	0.25 million
Densenet121	0.880 ± 0.002	0.860 ± 0.039	7.1 million
Proposed Method	**0.870 ± 0.051**	**0.900 ± 0.031**	**0.75 million**

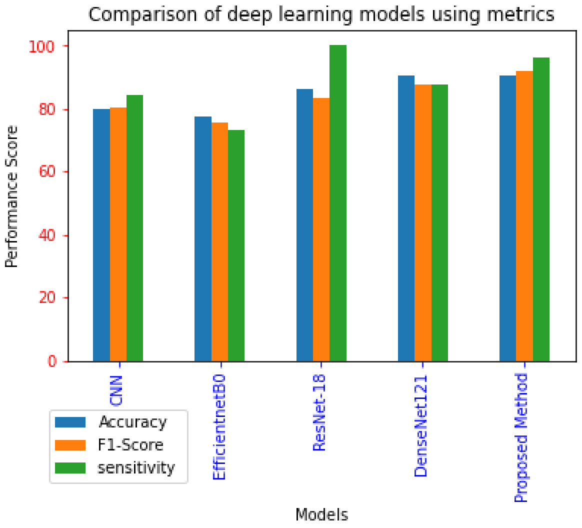


Figure 11: Comparison of models using barplots with maximum value scored by the models

the automatic prediction and to take the required treatment on the patient. This is done based on the probability distribution of being cancerous from 0% to 100% using outputted by the softmax activation classifier, as this score can be interpreted as the model's certainty in its prediction.

The model is classifies the images by looking at some parts of the image. For some images, it is able to look at the centre part of the images, except in a few benign images. In this thesis work, We illustrated a Visual explanations of deep Networks using gradient weighted class activation maps. We observed that the model made classification by extracting the information from center region of the image as might be expected, as can be seen the visualization in (Fig.18). By the using gradient-weighted class activation maps(Grad-Cam), we are able to interpret the reason behind the misclassified images, which is really wonderful and gives a substantial hints for further amendment of the network. The localization of instance has done from the final convolutional layer. The model is not looking at the right part of the the images, See(Fig.19). The reason may be because the nodules are a lot bigger in these images than on the usual ones. This may further be improved by adding more cases such as these.

6. Discussion

In this thesis work, we evaluated our proposed pipeline and proposed method network for classification task on compressed ultrasound image which has two orientations per case. The dataset is very small (595 images) and has poor quality (contains a lot of noise and artifacts), which however corresponds to what experts use.

We found in our experiments that the distribution of image cases (malignant vs benign) in the training group was imbalanced. Therefore, we fused two databases (Private and public) to have enough training samples for both classes. To overcome these limitations, we proposed a pipeline that consists of pre-processing, augmentation, feature extraction, and classification task. Image pre-processing proved to be effective in improving the proposed method: (1) Cropping and resizing the acquired images in order to remove the different noise within the images. (2) Discarding of the artifacts uses to keep away the network from learning meaningless information and recuperate the textures overlapped by the markers made by the experts or physician. This is done using morphological operation using 3 x 3 kernel. (3) Histogram Equalization is contrast adjustments techniques that effectively spreading out the most frequent intensity values throughout the image. (4) the image normalization that adjust the details of images from different sources imaging techniques to the same scale. And also, we have added a contrast variability to the images. Then, the images are well cleaned and have good quality for further process.

As mentioned in Section (3), one of the main problem was the size of the dataset. In spite of getting training examples with higher quality of extracted features, the millions of parameters available 6 to be tuned for the network still need a huge number of examples to prevent the over-fitting. We came to solve it with image data generation techniques. The input image size was 800 x

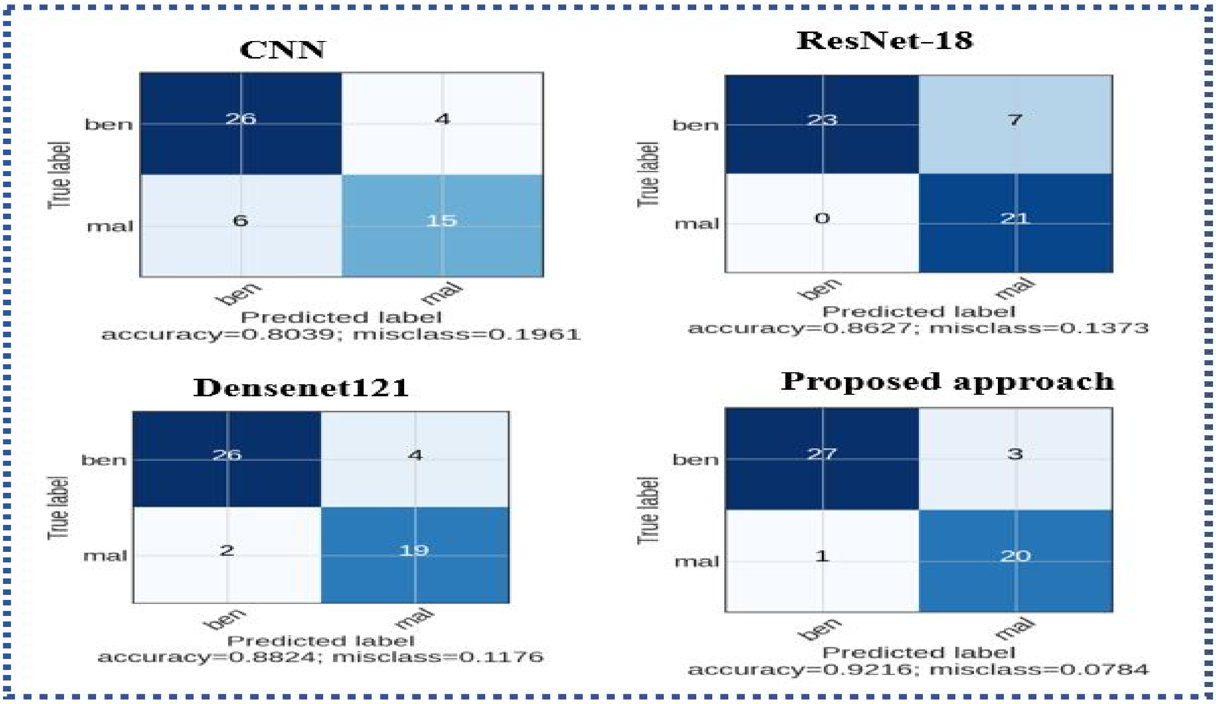


Figure 12: Qualitative comparison of different models on classification of thyroid nodules. **A**: Confusion matrix of CNN with 10 misclassified examples, **B**:Confusion matrix of Resnet-18 with 7 misclassifications, **C**: Confusion matrix of Densenet121 with 6 misclassifications, **D**: Confusion matrix of the proposed method with 4 misclassified examples

and 600, and we augmented the image to have enough number of training samples. The performance of the model might be affected due to cropping technique of the images.

From our experiments, the proposed method proves to have the advantage of needing less training samples to generate a CAD system based on deep learning networks. The system can help in minimizing the time, effort of physician and avoids unnecessary fine needle aspiration on patients. Regarding the use of loss function, the pipeline achieved better result with focal loss in terms of Accuracy and F1-Score comparing with categorical cross entropy loss. This indicates that how focal loss can improve the result by reducing the false positives, false negative and mitigating the class imbalance problem between training samples. Focal loss is efficient to classify the hard examples using penalized learning method. This the most recommend loss function for a research work that involves with data imbalance problem. Automatic assignment of weight to each class play good role in handling of class imbalance as well. We have tried two way of class weight assignment methods. 1) manually assignment high weights for the minority class. 2) automatic assignment of weights based on the distribution of dataset. According our demonstration, the second method works well and can be an hypothesis to tackle unbalanced data. Our research work was restricted in only classifying of thyroid nodules from Ultrasound images (US) into two classes of probably malignant and benign. We do not yet have enough examples with all the 7 different Bethesda scores(0-VI) to attempt a model that can make prediction per each class. The dataset we have used are labeled by experts. Having this in mind, our classification task is still highly dependant on the experience of the experts and their subjectivity in interpreting ultrasound images of thyroid nodules. Our method could have a significant benefit in helping experts during the annotation process.

The performance of the proposed method based on deep learning networks is much improved in compare with state of the art and other deep learning architectures in classification of thyroid nodule task. We compared the performance of the proposed method with the state-of-the-arts in thyroid nodule classifications, see (table7) for more details. and to our CAD systems with the proposed model in classifying benign and malignant thyroid nodule images. In this (table 6), we did a comparison of our approach with other deep learning schemes. EfficinetNetB0 had the worst classification performance. The main reason is that it became overfitted the on the small dataset very fast. The proposed method that uses the added module to focus on the most relevant feature maps achieved better its results. We also compared the number of trainable parameters. Resnet-18 has the lowest number of parameters due to the reason that is has fewest layers than others. Densnet encourages feature reuse which substantially reduces the number of parameters, but still has high parameters due

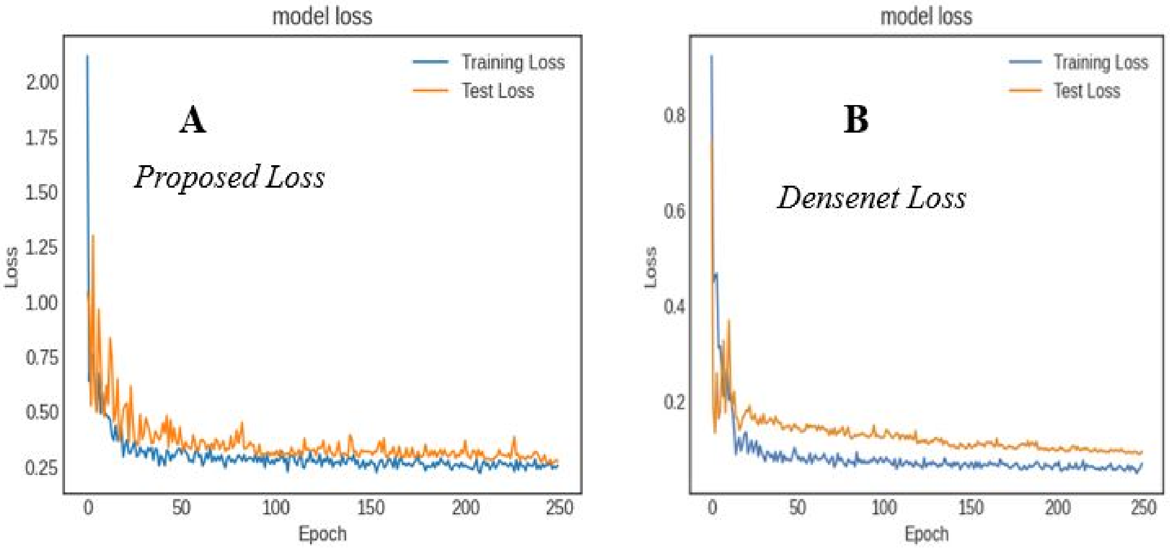


Figure 13: Qualitative comparison of loss functions fine-tuned Densenet and Proposed method monitoring the model during training:A) Loss function of Proposed model, B) Loss function of Densenet model

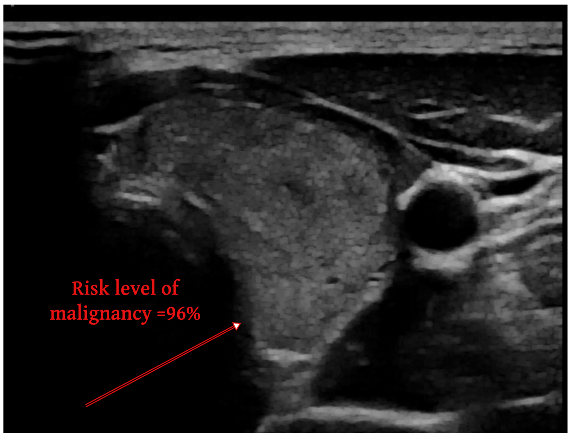


Figure 14: Prediction of malignancy with 96% confidence

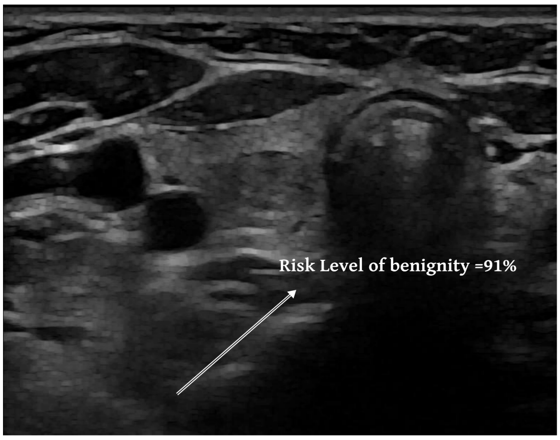


Figure 15: Prediction of benignity with 91% confidence

to its huge structure size. From this research work, We can suggest that our computer aided diagnosis tool can be used in Integrated Healthcare system(IHS) to help physician for early and accurate classification of nodule. we had some challenges and can be hypothesized in some way for further investigation. We do not know where the nodules are exactly on the labeled images. So, the annotation of ground truth can be done directly on the images with help of experts. This could give a comfortable environments to extract the region of interest(ROI) from the images. This could improve the performance of the our CAD system. And also, semi-supervised techniques can be used to annotate unlabeled US images. In spite of data acquisition for this classification task is in ingrowing, auxiliary classifier Generative Adversarial Network(acGAN) can be used to generate synthetic training samples as it required. Future improvement that involves providing of the thyroid nodules which are transversal and longitudinal to the deep learning algorithm as it could provide additional gains in performance and classification results. This task could be incorporated with Generative adversarial network(GAN) to have enough two view training samples.

7. Conclusion

In this work, we proposed a deep learning based Computer aided diagnosis system for automatic classification of thyroid nodule disease from ultrasound images. We demonstrated several deep learning approaches and able to compare them in our method in same dataset. Our method uses incorporated module within the DenseNet architecture,and we showed that

Table 7: Comparison of different CAD system in classifying benign and malignant thyroid nodule images

Methods	**Learning Algorithm**	**Accuracy**	**Sensitivity**
(Buda et al., 2019)	Faster R-CNN	83.00%	87.00%
(Wu et al., 2016)	(RBF)–neural network	84.74%	92.31%
(Peng et al., 2017)	SVM(Kernel=RBF)	88.00%	82.10%
(Koh et al., 2020)	InceptionResNetv2	85.00%	91.80%
Proposed model	**Attention-Densenet121**	**92.15%**	**96.42%**

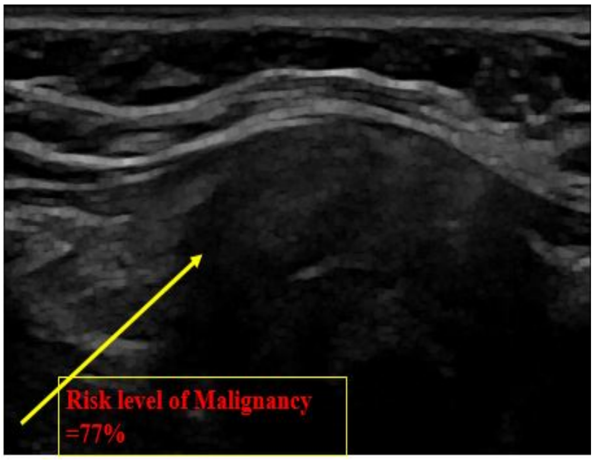


Figure 16: Prediction of malignancy with 77% confidence

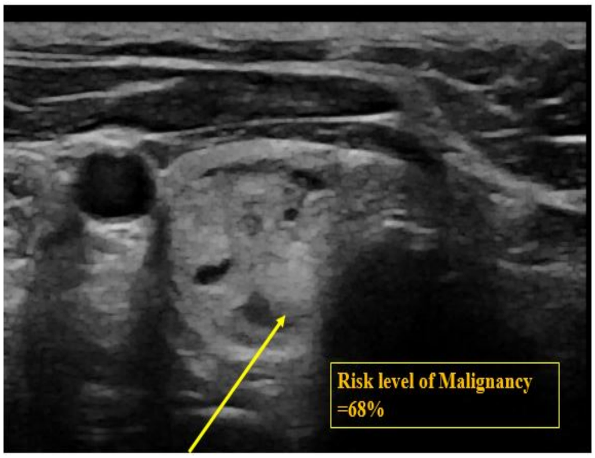


Figure 17: Prediction of benignity with 68% confidence and needs a strict follow up

adding this module to the fine-tuned Densenet121 substantially improves the classification result. We have shown that pre-processing and augmenting effectively improved the the performance of our proposed model. Despite having a small size, heterogeneity, unbalanced, low image quality, our approach obtained overall good result on the test set achieving an accuracy 0.9007 and a F1-score of 0.9216 for detection of nodules, which is higher than the performance reached on recent studies on this thematic area. This method could be used to predict nodule malignancy in clinical practice two reasons where it could bring the following benefits. First, it can eradicate the substantial inter-reader variability, and subjectivity that have been noticed for this task even when a standard Interpretation criteria is used. Second,the proposed approach could reduce the time and effort that is required for analyzing thyroid nodules, which would be of great help for clinical experts.

Furthermore, We have used Gradient-weighted class activation maps(Grad-cam) method to provide an explainable heat map of the primary regions of interest used by the model of the proposed technique. It helps to visualize how the model make decision during the prediction.

8. Acknowledgments

Tewele W.Tareke has awarded an Erasmus+ scholarship from the Erasmus Mundus Joint Master Degree in Medical Imaging and Applications(MAIA), a program which is funded by the Erasmus+ program of the European Union.

Tewele W.Tareke would like to express his sincere gratitude to Dr.Alain Lalande and Dr.Sarah Leclerc,for being his project supervisors and constant source of help and inspiration throughout the work.Their timely advice and guidelines have assisted him to get through a lot of challenges during the master thesis and nice situations.

References

Buda, M., Wildman-Tobriner, B., Hoang, J.K., Thayer, D., Tessler, F.N., Middleton, W.D., Mazurowski, M.A., 2019. Management of thyroid nodules seen on us images: deep learning may match performance of radiologists. Radiology 292, 695–701.

Chen, Z., Wang, J., He, H., Huang, X., 2014. A fast deep learning system using gpu, in: 2014 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE. pp. 1552–1555.

Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M., 2017. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. Journal of digital imaging 30, 477–486.

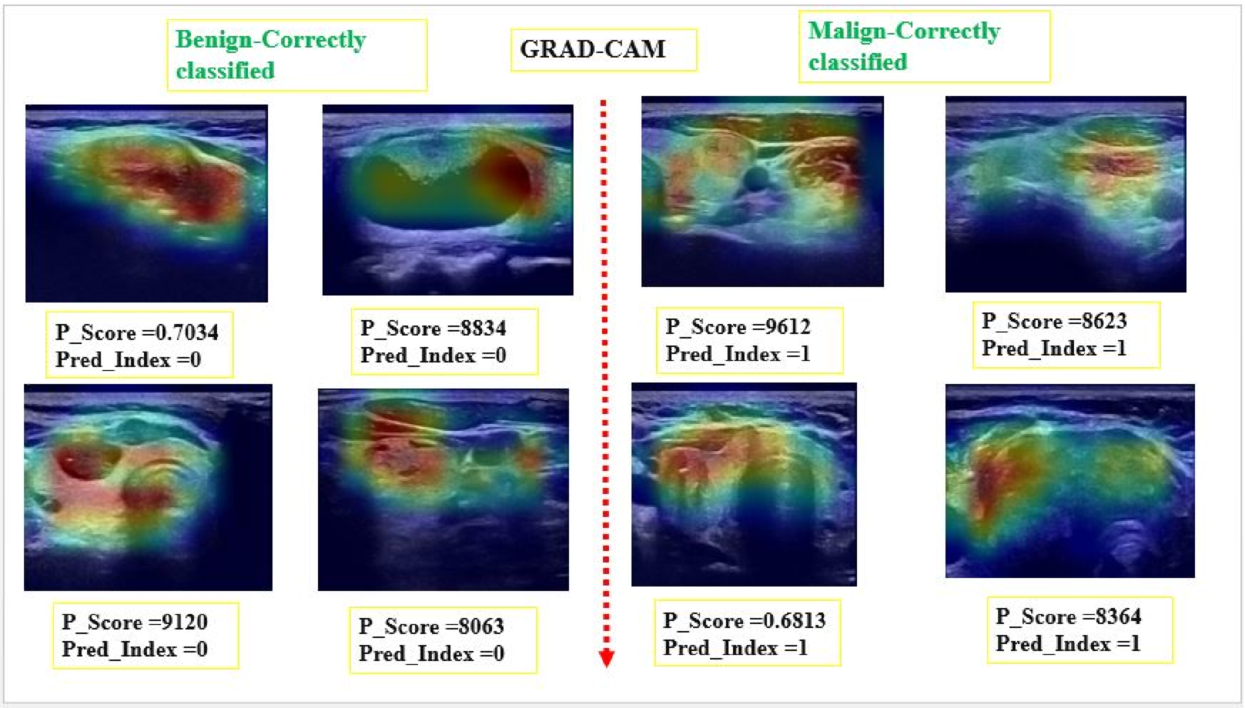


Figure 18: Illustration of correctly predicted images using Grad-CAM:**P-score:** probability score,**Pred-index:** prediction class

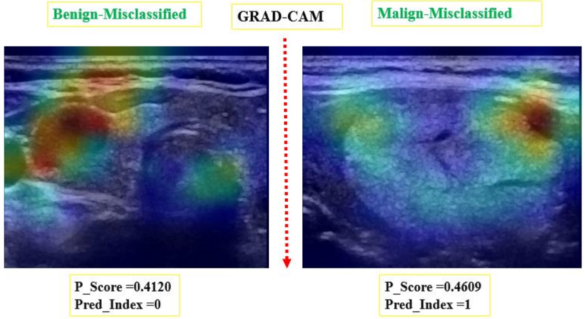


Figure 19: visualization of wrongly predicted images using Grad-CAM:**P-score:** probability score,**Pred-index:** prediction class

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Dong, Y., Su, H., Zhu, J., Zhang, B., 2017. Improving interpretability of deep neural networks with semantic information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4306–4314.

Frates, M.C., Benson, C.B., Charboneau, J.W., Cibas, E.S., Clark, O.H., Coleman, B.G., Cronan, J.J., Doubilet, P.M., Evans, D.B., Goellner, J.R., et al., 2005. Management of thyroid nodules detected at us: Society of radiologists in ultrasound consensus conference statement. Radiology 237, 794–800.

Goutte, C., Gaussier, E., 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, in: European conference on information retrieval, Springer. pp. 345–359.

Hambly, N.M., Gonen, M., Gerst, S.R., Li, D., Jia, X., Mironov, S., Sarasohn, D., Fleming, S.E., Hann, L.E., 2011. Implementation of evidence-based guidelines for thyroid nodule biopsy: a model for establishment of practice standards. American Journal of Roentgenology 196, 655–660.

Hang, Y., 2021. Thyroid nodule classification in ultrasound images by fusion of conventional features and res-gan deep features. Journal of Healthcare Engineering 2021.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.

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

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

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR. pp. 448–456.

Koh, J., Lee, E., Han, K., Kim, E.K., Son, E.J., Sohn, Y.M., Seo, M., Kwon, M.r., Yoon, J.H., Lee, J.H., et al., 2020. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. Scientific reports 10, 1–9.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation 1, 541–551.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Pang, B., Nijkamp, E., Wu, Y.N., 2020. Deep learning with tensorflow: A review. Journal of Educational and Behavioral Statistics 45, 227–248.

Patel, O., Maravi, Y.P., Sharma, S., 2013. A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement. arXiv preprint arXiv:1311.4033 .

Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E., 2015. An open access thyroid ultrasound image database, in: 10th International Symposium on Medical Information Processing and Analysis, International Society for Optics and Photonics. p. 92870W.

Peng, W., Liu, C., Xia, S., Shao, D., Chen, Y., Liu, R., Zhang, Z., 2017. Thyroid nodule recognition in computed tomography using first order statistics. Biomedical engineering online 16, 1–14.

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

Shin, J.H., Baek, J.H., Chung, J., Ha, E.J., Kim, J.h., Lee, Y.H., Lim, H.K., Moon, W.J., Na, D.G., Park, J.S., et al., 2016. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised korean society of thyroid radiology consensus statement and recommendations. Korean journal of radiology 17, 370–395.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Tessler, F.N., Middleton, W.D., Grant, E.G., 2018. Thyroid imaging reporting and data system (ti-rads): a user's guide. Radiology 287, 29–36.

Townsend, J.T., 1971. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics 9, 40–50.

Vakili, M., Ghamsari, M., Rezaei, M., 2020. Performance analysis and comparison of machine and deep learning algorithms for iot data classification. arXiv preprint arXiv:2001.09636 .

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Wu, H., Deng, Z., Zhang, B., Liu, Q., Chen, J., 2016. Classifier model based on machine learning algorithms: application to differential diagnosis of suspicious thyroid nodules via sonography. American Journal of Roentgenology 207, 859–864.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer. pp. 818–833.