



MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES



**AN IMAGING FEATURES APPROACH
TO TIRADS SCORE ESTIMATION
BY TRANSFER LEARNING**

MURAT CAN TÜRTÜK

MASTER THESIS

Department of Electrical and Electronics Engineering

Thesis Supervisor

Assoc.Prof. Gökhan Bora ESMER

ISTANBUL, 2021

Table of Contents

1.INTRODUCTION	1
1.1 Motivation and Overview	1
1.2 Scope of Thesis	3
2. MATERIALS AND METHODS	4
2.1 Thyroid Nodules and TIRADS	4
2.2 Dataset Preprocessing	6
2.2.1 ROI Extraction by Connected Component Labeling	7
2.2.2 Template Matching and Removal	10
2.2.3 Reconstruction by Image Inpainting	12
2.2.4 Reduction of Computational Complexity	13
2.2.5 Derivation of New Datasets for Thyroid Ultrasound Image Features	14
2.3 Neural Networks	14
2.3.1 ResNet-50	15
2.3.2 Loss Function.....	16
2.3.3 Hyperparameter Fine-Tuning.....	18
2.3.4 Partial Augmentation	20
3.RESULTS AND DISCUSSION	21
3.1 ResNet-50 Training and Results	21
3.1.1 Effects of Focal Loss	23
3.1.2 Effects of Partial Augmentation.....	23
4.CONCLUSION.....	29
REFERENCES	31

ÖZET

TRANSFER ÖĞRENMESİ İLE TİRAD PUAN TAHMİNİNE YÖNELİK GÖRÜNTÜLEME ÖZELLİKLERİ YAKLAŞIMI

Yapılan çalışmada, Tiroid Görüntüleme Raporlama ve Veri Sistemi (Thyroid Imaging Reporting and Data System (TIRADS)) puanının tahminine yönelik yeni bir görüntüleme özellikleri yaklaşımı önerilmiş ve bu yöntemin gerçekleştirilebilirliği en önemli iki görüntüleme özelliği analiz edilerek incelenmiştir. Elde edilen ilk sonuçlar irdelenmiş, yardımcı metodlar ile iyileştirilmiş, ve yardımcı metodların etkileri belirtilmiştir. Yöntemin mevcut güncel yöntemlerden farkları, üstünlükleri ve eksiklikleri belirlenmiştir.

İlk aşamada, tiroid ultrason görüntüleri işlenerek, görüntüler ultrason cihazı tarafından eklenen yapay işaretler ve diğer gereksiz kısımlardan temizlenmiştir. Böylece, tiroid görüntüsünün hesaplamalar için gerekli olan kısmı bulunmuştur. Önerilen yöntem, geleneksel ikili sistemde yapılan sınıflandırma problemini çoklu sınıflandırma problemine çevirmektedir. Bu yöntemin uygulanabilmesi için, ikili sistemde sınıflandırma yapmaya yönelik hazırlanmış görüntü setinden, farklı farklı görüntüleme özellikleri için yeni birer görüntü seti oluşturulması gerekmektedir. Bu nedenle, orijinal görüntü seti tekrar işlenerek, TIRADS puanı belirlenmesinde kullanılan her bir özellik için (kireçlenmeler (kalsifikasyon), ekojenite vb.) ayrı bir görüntü seti türetilmiştir. Bu görüntüler, önceden eğitilmiş sinir ağlarına çoklu sınıflandırma yapmak üzere girdi olarak verilip, çıktılar analiz edilmiştir. Sınırlı sayıda görüntü içeren ve sınıflar arasında dengeli bir dağılıma sahip olmayan bir görüntü seti kullanılmasına rağmen, kalsifikasyonlar yaklaşık %85 doğrulukla sınıflandırılabilmiştir. Daha az sayıda örneğe, ve yine sınıflar arasında dengesiz dağılıma sahip olan ekojenite özelliğinde ise doğruluk %80 seviyesinde kalmıştır.

Önerilen sistemin her bir görüntüleme özelliği için yeterli sayıda örneğe ve sınıflar arasında dengeli dağılıma sahip bir görüntü seti kullanılarak eğitilmesiyle, yöntemin yüksek bir sınıflandırma doğruluğu ve yüksek çıktı çözünürlüğü ile uygulanabilir olacağı öngörülmektedir.

ABSTRACT

AN IMAGING FEATURES APPROACH TO TIRADS SCORE ESTIMATION BY TRANSFER LEARNING

A novel method for Thyroid Imaging Reporting and Data System (TIRADS) score estimation using the imaging features is proposed. Applicability of the method is investigated by analyzing the two most important imaging features. Initial results were evaluated and improved by using auxiliary methods, and the effects of auxiliary methods were stated. Advantages and disadvantages of the proposed method when compared to state-of-the-art methods were determined.

In the first step, images were processed to remove artificial markers added by the ultrasound device and other redundant data. As a result, the region of interest (ROI) is found in the thyroid image. The proposed method turns the traditional binary classification problem into multiclass classification. In order to be able to apply the proposed method, the dataset was processed further and multiple different datasets were derived, one for each of the image features (calcifications, echogenicity etc.) used when determining the TIRADS scores of thyroid nodules. Then, these derived datasets were fed into pretrained neural networks for multiclass classification, and the results were evaluated. Even with a dataset of limited size and biased samples, the calcification property classification accuracy turns out to be 85%, whereas for echogenicity property, again with a limited number of biased samples, stays around 80%.

By using a dataset which has more number of samples for each imaging feature and a better balanced distribution among classes, it is anticipated that the proposed method may become applicable, with a high overall classification accuracy, and a high output resolution.

SYMBOLS

y_i : Expected Output

\hat{y}_i : Predicted Output

n : Number of data points in error calculation

p_t : Prediction accuracy

α_t, γ : Focal loss parameters

ABBREVIATIONS

US : Ultrasound

TIRADS : Thyroid Imaging Reporting and Data System

FNA : Fine Needle Aspiration

FNAB : Fine Needle Aspiration Biopsy

NN : Neural Network

ANN : Artificial Neural Network

CNN : Convolutional Neural Network

ROI : Region of Interest

MSE : Mean Square Error

CE : Cross-Entropy Error

FL : Focal Loss

LIST OF FIGURES

Figure 1.1 Traditional binary classification framework.....	2
Figure 1.2 Overview of the proposed method.....	3
Figure 2.1 Unprocessed thyroid ultrasound image	7
Figure 2.2 Overview of the ROI extraction process	8
Figure 2.3 Thyroid ultrasound image after ROI extraction	9
Figure 2.4 Artificially generated US imaging device brand marker.....	11
Figure 2.5 Example image with US device brand marker	11
Figure 2.6 Example thyroid US image after template matching and watermark removal.....	12
Figure 2.7 A thyroid US image restored with image inpainting method.....	13
Figure 2.8 Variation of focal loss with respect to its parameters.....	17
Figure 2.9 Partial augmentation scheme	20
Figure 3.1 Training, validation accuracies and losses for echogenity	22
Figure 3.2 Training and validation process for calcifications with cross-entropy loss.....	25
Figure 3.3 Training and validation process for calcifications with focal loss	26
Figure 3.4 Calcification training and validation without partial augmentation	27
Figure 3.5 Calcification training and validation with partial augmentation, applied on the macrocalcification class	28

LIST OF TABLES

Table 2.1 Features and associated scores in TIRADS	5
Table 2.2 TIRADS categories vs suspicion	6
Table 2.3 Number of samples for calcification feature	14
Table 2.4 Depths and number of parameters for different neural networks	16
Table 3.1 Number of correct classifications for calcifications without focal loss and partial augmentation.....	24
Table 3.2 Number of correct classifications for calcifications with focal loss and partial augmentation.....	24

1.INTRODUCTION

1.1 Motivation and Overview

Thyroid nodules are commonly observed on humans throughout their life span. Even though the rate of malignant nodules is relatively low, some nodules, such as thyroid cancer, have a risk of malignancy. Thyroid Imaging Reporting and Data System (TIRADS) is widely used for classifying thyroid nodules according to their key characteristic features. These features can be defined as calcifications, echogenicity, composition, shape and margin.

Intensive research is still carried on with different approaches, using image processing methods, neural networks, or a mixture of both. Successful results were obtained with deep learning methods, which are presented in [1, 2]. Liu et. al achieved feasible overall accuracy by using hybrid features [4]. However, taking samples from the nodule is the most important step in diagnosis for cancer [16]. In order to find out if a nodule is malignant or benign, a fine needle aspiration biopsy (FNAB, or FNA biopsy) under the guidance of ultrasound (US) imaging must be performed [16]. On the other hand, lots of unnecessary biopsies are performed on low risk patients every year because of having misleading conclusions from thyroid US images.

In this thesis, a multiclass classification approach to the thyroid nodule classification problem using thyroid US imaging features and transfer learning is proposed, and it's applicability is investigated. Even if there is a vast amount of research on the binary classification problem of thyroid nodules, we develop a multiclass classification approach to determine the TIRADS scores of thyroid nodules. Instead of having a binary output such as malignant or benign, the output is the TIRADS score of the thyroid nodule, such as TIRADS 1, 2, 3, 4 or 5. This multiclass scheme allows an output with increased resolution compared to the traditional methods, which generally assume during the neural network training that the nodules with scores 2 or 3 are classified as benign, whereas nodules with scores 4 and above are classified as malignant. Schematic overviews of traditional binary classification, and the proposed TIRADS score multiclass classification methods are given in Figures 1.1 and 1.2, respectively.

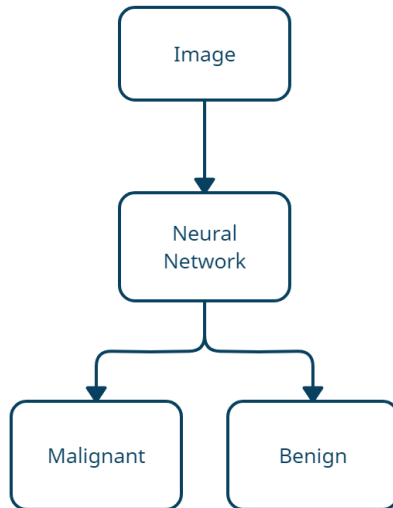


Figure 1.1 Traditional binary classification framework

Binary classification of thyroid nodules such as malignant or benign may provide simpler implementation and better results, but from a medical perspective, diagnosing a malignant nodule as benign can cause unacceptable results, so even if modern binary classification systems can reach very high overall accuracies, they are still susceptible to error. Determining the risk category of a thyroid nodule by estimating the TIRADS score can also not be done with perfect accuracy. However, the proposed method can provide predictions in close proximity of the actual TIRADS score even if individual predictions over the characteristic features are wrong. Therefore, this method can be safely used in medical diagnosis procedure as a support system since it can provide an insight on the risk category of the nodule instead of directly classifying it as malignant or benign.

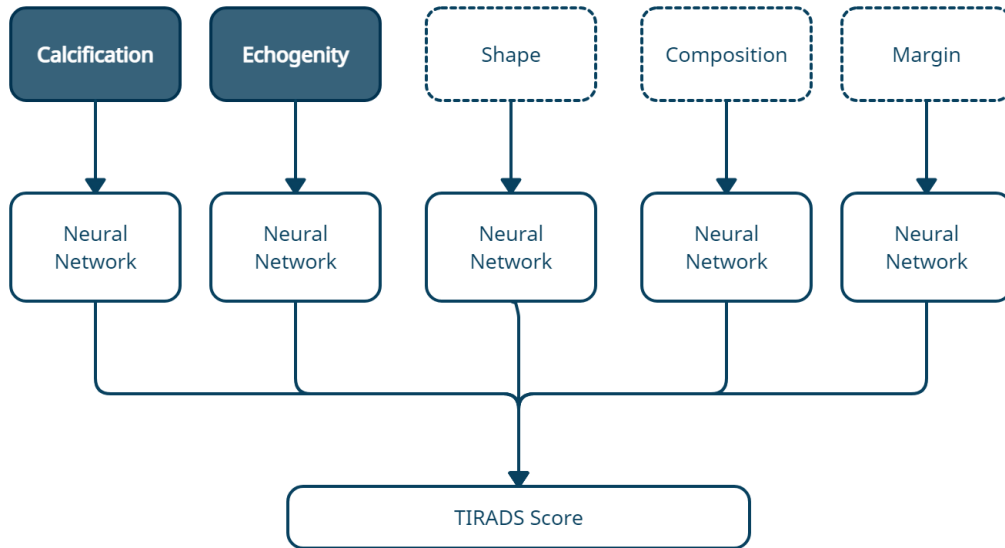


Figure 1.2 Overview of the proposed method

Furthermore, estimating the risk category of patients paves the way to reduce the number of unnecessary biopsies performed on very low risk patients. As a result of this, patients can be saved from the pain and the costs of biopsy procedures. Also, it will be useful to save the precious time of radiologists and other medical staff.

1.2 Scope of Thesis

Applicability of the proposed method on the calcification and echogenity features was investigated as presented in Figure 1.2 by using the Digital Database of Thyroid Images (DDTI) [34], which is an open access dataset, available on the Internet. Images in the dataset were first preprocessed to extract the region of interest (ROI), which is the US image of the thyroid gland. Then, in order to transform the binary classification problem into a multiclass classification problem, new datasets, which are suitable for multiclass classification are derived from the original one.

These new datasets are then used to fine-tune the parameters of pretrained ResNet-50 CNNs, one for each of the features. Effects of different loss functions as well as the effect of using a new partial data augmentation scheme are also discussed. Also, advantages and disadvantages of the proposed method compared to state-of-the-art methods are presented. Finally, possible future improvements to the method are stated as future work.

2. MATERIALS AND METHODS

Thyroid US image dataset used in this thesis has open access and it is publicly available from [34]. The dataset has a major drawback due to the fact that it was prepared for the binary classification of thyroid US images. In order to convert the binary classification problem to a multiclass classification problem, new datasets were derived from the original one, using the feature properties of thyroid US images provided with the DDTI dataset. Images were sorted and labeled according to the calcification and echogenity properties provided with the dataset.

OpenCV, Matlab and Octave were used for different kinds of operations such as dataset preprocessing, image sorting, image preprocessing, template matching, regional reconstruction, neural networks training.

Detailed information on the preprocessing steps, as well as thyroid nodules and TIRADS are given in the following sections.

2.1 Thyroid Nodules and TIRADS

Thyroid nodules are commonly observed on males and females throughout the life span, regardless of the age. Most of these nodules are benign, but some have malignant characteristics and pose a risk of thyroid cancer.

Thyroid nodules may have various different characteristics. They can be tiny, or large, shallow or deeply embedded within the thyroid gland. These nodules might also have a round, circular shape, but some can be taller than wide (TTW). The region around the thyroid nodules might have microcalcifications, macrocalcifications or none at all. TIRADS is the gold standard used in the diagnosis of thyroid nodules and it utilizes these characteristics of thyroid nodules while classifying them according to their risk category.

TIRADS score of a thyroid nodule is determined by the characteristic features of the nodules. Each category has multiple different properties, one of which describes the characteristics of the thyroid nodule and each property has an intermediate score associated with it. The scores of each

feature(category) are summed up to determine the TIRADS score of the nodule. These features and associated scores are given in Table 2.1.

Feature	Property	Score
Calcifications	Macrocalcifications	1
	Peripheral(rim) calcifications	2
	Microcalcifications	3
	None	0
Echogenicity	Hyperechoic	1
	Isoechoic	1
	Hypoechoic	3
	Anechoic	0
Composition	Cystic	0
	Spongiform	0
	Mixed cystic and solid	1
	Solid	2
Shape	Wider than tall	0
	Taller than wide	3
Margin	Smooth or ill defined	0
	Lobulated or irregular	2
	Extra-thyroidal extension	3

Table 2.1 Features and associated scores in TIRADS

Overall, many different characteristic features can be seen in thyroid nodules, but some of them are more important than the others while determining the risk category of the nodule. Calcification and echogenicity features have major effect on determining the TIRADS score. The reason behind this is that, properties associated with 3 points occur much more frequently for calcification and echogenicity features compared to the other features.

In order to calculate the TIRADS score, intermediate scores of each category should be summed, and the result is classified according to Table 2.2. TIRADS categories TR1 and TR2 indicate the nodules which have almost no suspicion of being malignant. Therefore, no FNA biopsy procedure is applied to patients with these types of nodules. Other risk categories TR3, TR4 and TR5 mostly require a

biopsy, depending on the nodule diameter. For example, FNA biopsy is requested from a patient with TR4 risk category if the nodule diameter is greater than or equal to 15 mm, whereas for a patient with TR5 risk category, FNA biopsy is requested if nodule diameter is greater than or equal to 10 mm.

Points	Category	Suspicion
0	TR1	Benign
2	TR2	Not suspicious
3	TR3	Mildly suspicious
4 - 6	TR4	Moderately suspicious
7 or more	TR5	Highly suspicious

Table 2.2 TIRADS categories vs suspicion

Even though TIRADS score of a thyroid nodule provides very important information about its characteristics, biopsy procedure is the only way to certainly determine if a nodule is malignant or benign. Even though many biopsies are performed on many patients annually, some biopsies may turn out to be inconclusive. Thus, doctors may request another biopsy to analyze the sample. Main reason of having inconclusive biopsies is failing to get a sample from the actual nodule. Sometimes the nodule is very tiny, or it is deeply embedded within the thyroid gland. Hence, the radiologists may not hit the nodule under the US guidance. As a result of this, many unnecessary FNA biopsy procedures can be applied on patients, even if they are in the very low risk categories such as TR1 or TR2 [5].

2.2 Dataset Preprocessing

The DDTI dataset consists of thyroid ultrasound images and corresponding XML files that contain the necessary label information about the thyroid imaging features in Table 2.1. Some of these files did not contain any information about one or more of the thyroid properties such as calcifications, echogenicity, etc. As a result, the samples with missing properties were excluded from the work.

Multiple different operations were applied on the images and the dataset in order to convert everything into a suitable format for multiclass classification. The preprocessing steps can be

summarized as: ROI extraction and cropping, reduction of computational complexity by discarding redundant channels, template matching and removal, and regional reconstruction with image inpainting by fast marching method. Details of these steps are presented in the following subsections.

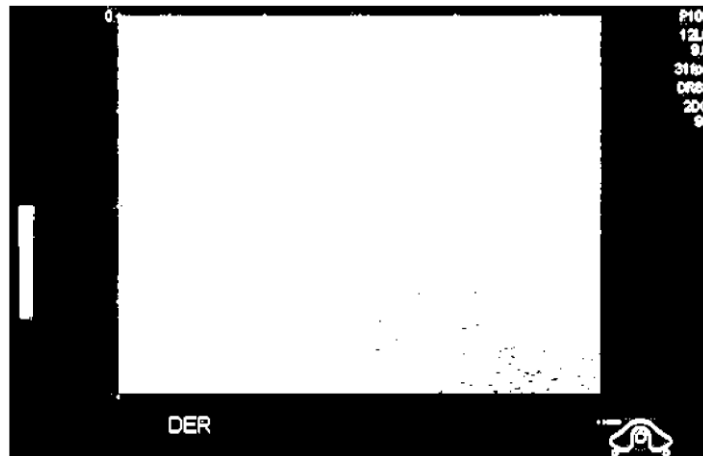
2.2.1 ROI Extraction by Connected Component Labeling

Images included in the DDTI dataset contain redundant information in addition to the ROI. This redundant information must be removed before any further processing.

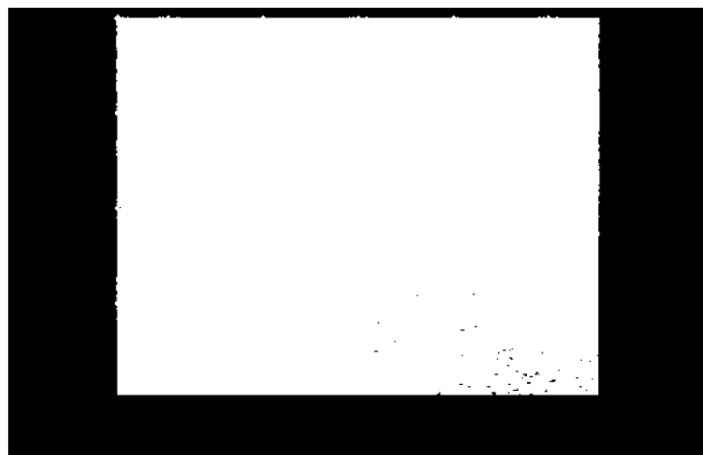


Figure 2.1 Unprocessed thyroid ultrasound image

In Figure 2.1, the ROI is located in the middle of the unprocessed image. This rectangular ROI is surrounded by artificially generated black pixels with zero intensity value, and some other artificially generated white pixels. Converting the image in Figure 2.1 to a binary image with a very low constant threshold such as 5% (of the maximum intensity 255) produces the image provided in Figure 2.2a. The ROI can be identified as the huge white blob in the middle of the binary image provided in Figure 2.2a. First, connected component detection is applied to this binary image to detect and label the 8-connected regions in Matlab. Then, the areas are calculated by counting the number of pixels belonging to each label separately. The pixels belonging to the label with the largest area are kept as



(a)



(b)



(c)

Figure 2.2 Overview of the ROI extraction process

(a) Conversion to binary with a fixed threshold

(b) Labeling the ROI using its area

(c) ROI after closing and opening morphological operations

one, and the rest of the pixel values are set to zero. Result of this operation is provided in Figure 2.2b. In the obtained binary image, only the pixels which belong to a coarse ROI are set to one.

This coarse ROI does not have a perfect rectangular shape due to the non-idealities around its perimeter. A series of morphological operations such as opening and closing are performed on this binary image to remove these non-idealities around the perimeter to obtain a perfect rectangular shape. First, closing operation is applied with a 5x5 structuring element to remove the small gaps(zero pixels) within the ROI, by replacing their values with one. Then, two opening operations are applied with 3x200 and 200x3 structuring elements. These opening operations remove the non-idealities around the perimeter of the ROI. As a result, a perfect rectangular ROI is obtained and it is provided in Figure 2.2c. Sizes of all structuring elements were determined experimentally during implementation and all structuring elements consist of only ones and no zeros.

Corners of the rectangular ROI presented in Figure 2.2c are found by a simple pixel coordinate comparison algorithm. Nonzero pixel positions are compared and the x_{min} , x_{max} , y_{min} , y_{max} values are found. These values are used to determine (x_i, y_i) pairs which are the coordinates of the corners of the ROI. The original image is cropped from these coordinates to extract the ROI provided in Figure 2.3. As a result, most of the redundant information is removed and the ROI is extracted from the original unprocessed image.

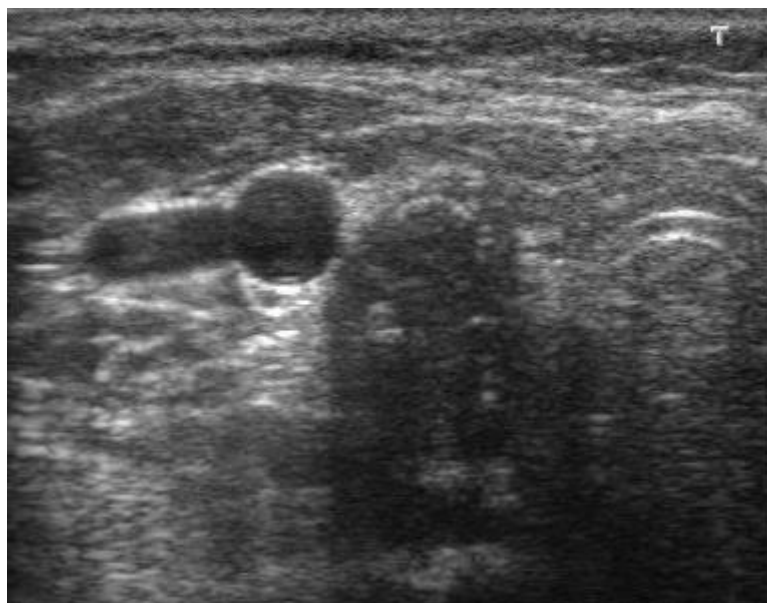


Figure 2.3 Thyroid ultrasound image after ROI extraction

2.2.2 Template Matching and Removal

Template matching is a digital image processing technique used to find regions of a larger image that match a smaller template image. Some application areas of template matching are manufacturing quality control [20], robot motion control [21] and edge detection [22]. In this thesis, template matching is used to detect the location of the T-shaped marker given in Figure 2.4, in the upper-right region of Figure 2.5. This marker indicates the brand of the US device. It is artificially created by the US imaging device and inserted into the image as a visible watermark. The locations of these markers in all images were found by use of template matching algorithm, and then marker regions were removed from the images, by setting the pixel intensities to zero.

Different matching metrics can be used to determine how similar the inspected region of the larger image is to the template. The most straightforward approach is to check the sum of absolute differences between the source image regions and the template image. The region in the source image which has the minimum absolute difference to the template is the best match.

A small template image (an 11x11 mask) which contains the artificially generated marker is moved throughout all pixels in the ROI, similar to a 2D convolution operation. At each pixel position, sum of absolute differences are calculated for the overlapping pixels in the mask and the larger image. As the algorithm runs, these calculated values are temporarily stored in memory along with their pixel position information. Lower sum of absolute differences is equivalent to more similarity between the template and the source image. Therefore, the template location is determined as the location with the minimum sum of absolute differences.

This kind of matching metric is susceptible to error if multiple matches occur for a given source image. In order to prevent this, the following precautions were taken. The region in which the T-shaped marker occurs is around the upper-right corner of the original image. Therefore, template matching is applied only in a 40x40 pixel region in the upper-right part of the US image. This approach greatly reduces the computational complexity as well as the probability of getting multiple matches. Also, an if-condition is added to the end of the matching algorithm which is triggered when there are multiple matches with the same sum of absolute differences. This if-condition introduces weights to some of the pixels and performs template matching again, only for the multiple best match points. Only the pixel intensities which form the T-shape in the template are multiplied by 1.1

(weight), and the sum of absolute differences are calculated for these best matching points again. The weight is gradually increased and computation is performed iteratively until only one best matching position remains. Even though this precaution was taken, none of the images had multiple best matches during implementation, due to restricting the template matching area into a 40x40 pixel region in the upper-right part of the thyroid US images.



Figure 2.4 Artificially generated US imaging device brand marker

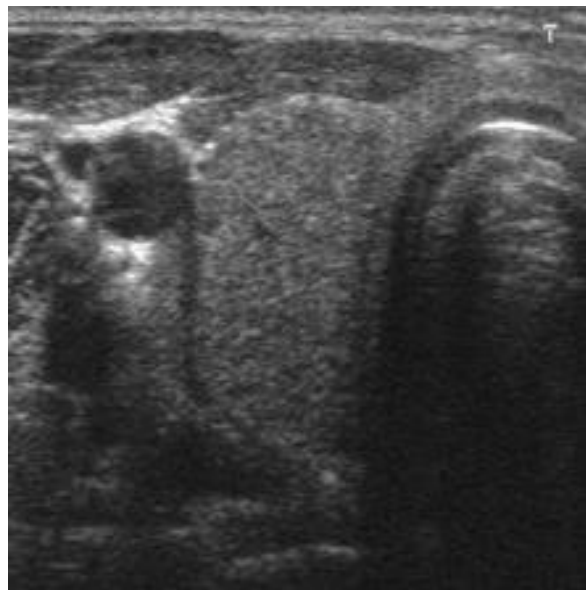


Figure 2.5 Example image with US device brand marker

In Figure 2.5, watermark of the US imaging device can be seen in the upper-right region. In order to remove this redundant information, template matching algorithm was used with minimum absolute difference matching metric. As a result of this, markers in all images were detected and removed by setting the pixel intensities to zero. An example thyroid US image after template matching and removal operation of the watermark can be seen in Figure 2.6.

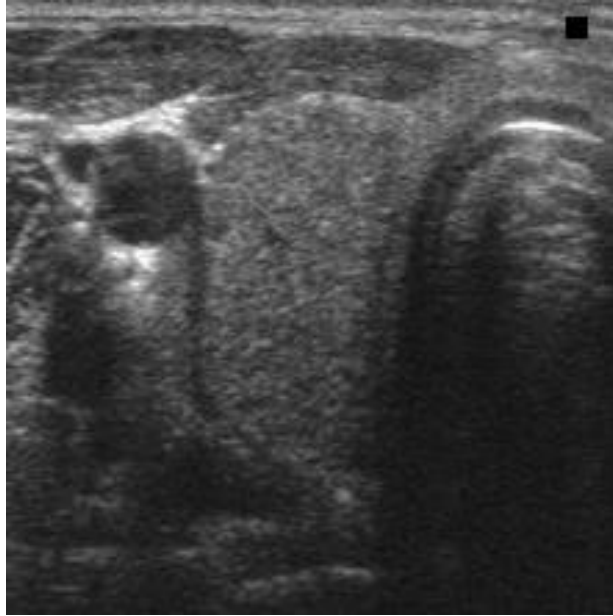


Figure 2.6 Example thyroid US image after template matching and watermark removal

2.2.3 Reconstruction by Image Inpainting

Image inpainting is used in image restoration, for reconstructing degraded areas in images by using the similarities with the pixels around the degraded neighborhood. First, the region to be inpainted is found and represented as a closed contour. This closed contour is represented by an 11×11 square which has the center position found in the previous template matching step. Then, the algorithm starts from the boundary of this region and iterates through all pixels. Each pixel is replaced by the normalized weighted average of the other known pixels in the predefined neighborhood. Once a pixel is painted, the Fast Marching Method (FMM) [3, 25] algorithm moves to the nearest pixel and eventually iterates through all pixels. By using this approach, it is ensured that the unknown pixels which are nearest to the known pixels are painted before any other pixel in the reconstructed region.

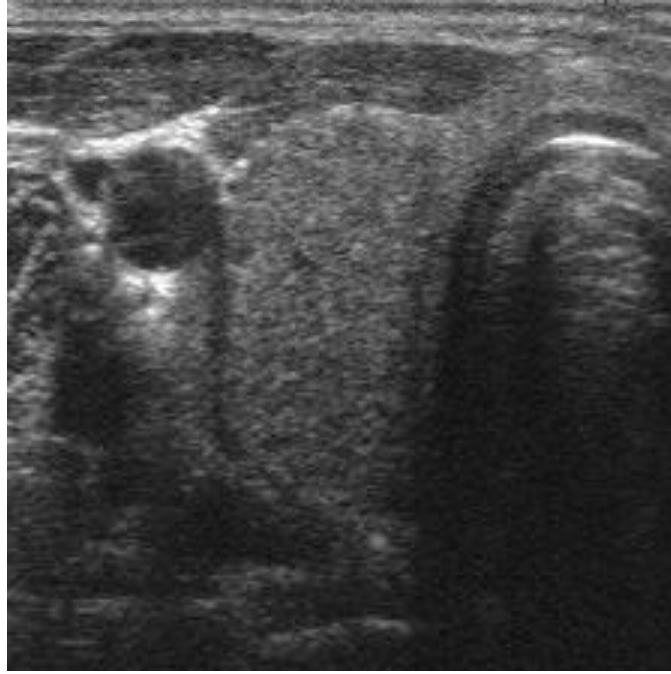


Figure 2.7 A thyroid US image restored with image inpainting method.

In the previous step, center of the artificial marker position was found by template matching algorithm and size of the removed region is also known. Implementation was done in C++ by OpenCV due to the fact that it provides the image inpainting by FMM algorithm. An example thyroid US image after template matching and watermark removal is given in Figure 2.6, and the same image after the application of image inpainting is presented in Figure 2.7.

2.2.4 Reduction of Computational Complexity

All images included in the dataset were in three-channel 24-bit RGB format, prepared with lossless JPEG compression. However, the ROI in the US image is actually in grayscale, thus, having three channels provides redundant information. The cropped ROIs are converted from 24-bit RGB to 8-bit grayscale inherently lossless PNG format. All three channels had the same data, so the R channel is picked as the reference.

This operation greatly reduces the neural network model complexity and saves a significant amount of training and validation time, since the input matrix size to the ResNet-50 model can be reduced by approximately 67% after dropping two out of three redundant channels.

2.2.5 Derivation of New Datasets for Thyroid Ultrasound Image Features

Before training the CNNs, new separate datasets for each of the image features (calcifications, echogenicity, etc.) should be derived from the original dataset. The XML files containing the thyroid US image features are parsed, and the images in the dataset are sorted according to these image features. For example, for the calcification feature, a new dataset is created by sorting the images by different labels according to their calcification property. Thyroid images with macrocalcifications are collected into one folder, whereas images with microcalcifications are collected into another, respectively. This kind of sorting is repeated for the other features as well.

It should be noted that some image features occur more frequently than the others. Therefore, the derived datasets had unbalanced distribution among different classes. An example of this type of result is presented in Table 2.3, which gives an overview of the number of samples for the calcification feature.

Feature	Classes	Number of Samples
Calcifications	Macrocalcifications	43
	Microcalcifications	216
	None	137

Table 2.3 Number of samples for calcification feature

2.3 Neural Networks

Neural networks (NN) use labelled events from a set of examples, and establish a relationship between the given set of inputs to predict the outcomes of the other events. Artificial neural networks (ANN) consist of different layers with specific functionalities and weighted connections between them. The learning process is basically storing information as these weights between connections and optimizing the weights according to the data provided to the network.

In general, there are three different types of learning strategies such as:

- Supervised Learning
- Reinforcement Learning
- Unsupervised Learning

Supervised learning is the most commonly used learning method with ANNs. Samples are provided to the network along with the expected outputs, and the network makes predictions. The output of the network (predicted output) is compared to the expected output. Then, the weights (hyperparameters) are tuned according to a loss function and according to the difference between the predicted and the expected output (error). The described procedure is performed iteratively while trying to minimize the error. In the presented work, mentioned iterative learning method is used to train the ResNet-50 CNNs.

In reinforcement learning, the sample outputs are not directly provided to the system. However, the correctness of the prediction is provided after having the prediction. The learning algorithm is designed such that the system takes the correctness of its previous prediction into account. Then it optimizes the parameters which will be used in the next prediction calculations.

Unsupervised learning is the type of learning that does not supply the output values to the system, and expects the system to learn the relationship between the parameters by itself.

2.3.1 ResNet-50

In this thesis, supervised learning strategy is used in a transfer learning scheme. Initial testing and performance comparison was performed between GoogleNet and ResNet-50 CNN architectures. ResNet-50 was chosen due to its slightly better overall performance compared to GoogleNet. The main reason behind this result is that, GoogleNet CNN architecture consists of 22 layers, while ResNet-50 consists of 50 layers. Therefore, ResNet-50 has 25.6 million hyperparameters while GoogleNet has 7 million. This hyperparameter comparison is given in Table 2.4. As a result of this, ResNet-50 performs better while detecting deep features in the thyroid US images. Relative prediction times and accuracies of different types of NN architectures trained with the ImageNet database were compared [27], and it can be seen that ResNet-50 has a balance between low prediction time (also training time) and relatively high accuracy. This was also tested with the DDTI thyroid image dataset and confirmed, during the initial testing.

Using a more complicated NN architecture with better performance than ResNet-50 was not possible due to the memory limitation of the GTX 970 GPU used during training. Using a more complex

NN model would have required multiple GPUs with much more memory than the GTX 970. Conclusively, the research and implementation proceeded with ResNet-50.

Network	Depth	Number of Parameters(Millions)
GoogleNet	22	7.0
ResNet-18	18	11.7
ResNet-50	50	25.6
Inception-ResNet-v2	164	55.9

Table 2.4 Depths and number of parameters for different neural networks

2.3.2 Loss Function

Making predictions, comparing them to the real expected outputs and trying to minimize the error is the key concept behind neural networks. Selection of the optimum loss function is a crucial factor in minimizing the error. Generally, Mean Square Error (MSE) or Cross-Entropy Error (CE) perform well in minimizing the error between the expected output and the predicted one. MSE and CE are calculated as:

$$\text{MSE} = \frac{1}{n} \sum_{k=1}^n (y_k - \widehat{y}_k)^2 \quad (1)$$

$$\text{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise} \end{cases} \quad (2)$$

In Eq. (1), y_k is the expected output and \widehat{y}_k refers to the predicted output. In Eq. (2), p refers to the predicted probability of a class and y defines the label (class) for binary classification. Thus, it is either equal to one or zero. We can modify the CE equation and say that p_t represents the prediction accuracy, such that, if p_t is equal to 1, prediction accuracy is 100% and CE is zero. The Cross-Entropy Error after this modification is given in Eq. (3).

$$CE(p, y) = CE(p_t) = -\log(p_t) \quad (3)$$

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases}$$

And the Focal Loss is defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (4)$$

There is a significant amount of similarity between the Focal Loss expression provided in Eq. (4) and the CE. As prediction accuracy p_t converges to 1, the $(1 - p_t)$ term approaches zero, which makes the focal loss expression also converge to zero. As prediction accuracy p_t decreases, the $(1 - p_t)$ term increases and converges to 1. Then, the α_t and γ parameters significantly affect the FL expression.

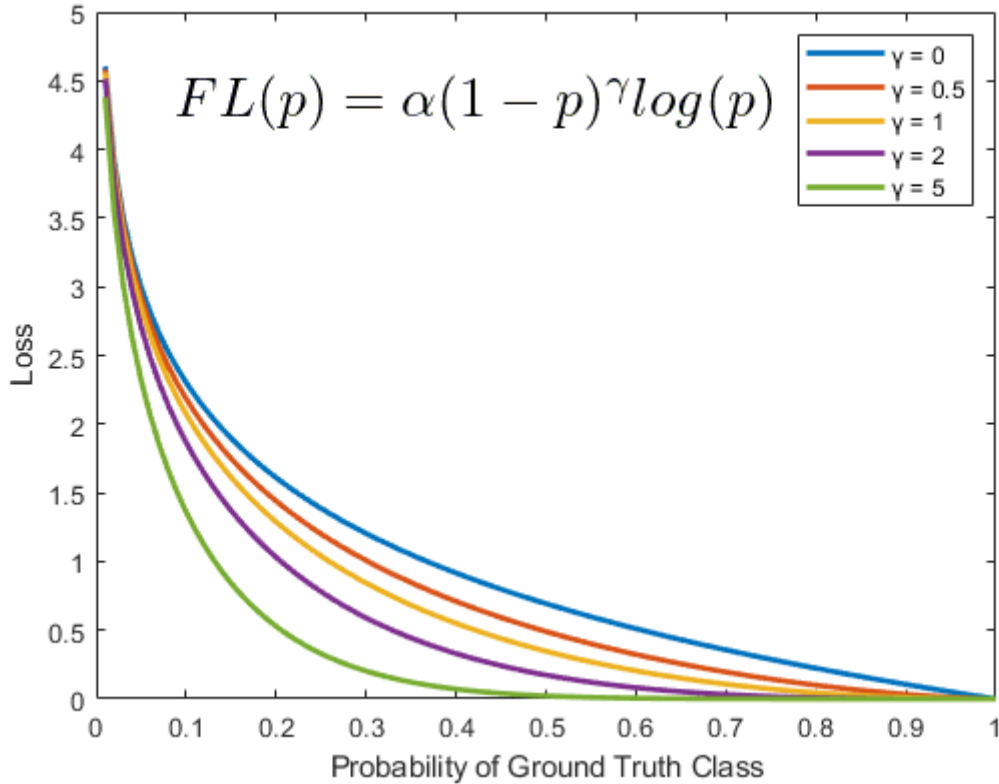


Figure 2.8 Variation of focal loss with respect to its parameters

To summarize, Focal Loss decreases as prediction accuracy increases, and increases as prediction accuracy decreases. That behavior is taken advantage of in the thesis, due to the following reasons. First reason is related to having small number of samples in the dataset. Second reason is about the distribution of samples among classes. Using Focal Loss makes the wrong predictions more taxing, which is essentially useful in the training phase while learning to make predictions about classes with relatively less number of samples when compared to the others.

Generally, Focal Loss parameters α_t and γ in Eq. (4) are determined by experimental trials. In Matlab, default values for α_t and γ are defined as 2.0 and 0.25 respectively. After some experimentation with these parameters, it was decided that $\alpha_t = 5.0$ and $\gamma = 0.25$ provide better results than the default values. Effect of these parameters on Focal Loss can be seen in Figure 2.8.

2.3.3 Hyperparameter Fine-Tuning

Neural networks typically require thousands of sample images from each class so that they are able to make correct predictions, with high accuracy. This requirement could be easily satisfied when classifying everyday objects such as mug, pen and pencil, notebook, etc. Animals can also be easily classified due to the fact that there are millions of sample images publicly available on the Internet. This is not the case for thyroid US images. Most of the images created by US imaging devices also have classified patient information, and these health records can not be made public. So, it is practically almost impossible to design a brand-new neural network architecture and train it with thousands of thyroid US images from scratch, with high accuracy at the end. As a solution to this problem, transfer learning becomes extremely beneficial.

Transfer learning is based on utilizing the prior information obtained previously. Furthermore, it uses that information in solving a problem. The ResNet-50 model used in the thesis was pretrained with the ImageNet dataset, with approximately 14 million images for the classification of 1000 different classes [35]. This training process is basically the tuning of the weights in the CNN model, and the model learns to extract useful features which define each of the classes in the ImageNet dataset. That information is stored as hyperparameters and they are utilized in learning process for classifying different thyroid US images. Training a neural network from scratch for the multiclass classification of thyroid ultrasound images would require a huge amount of samples as well as an extremely long

training time if transfer learning method is not employed.

In the thesis, all the advantages of transfer learning scheme are taken by “freezing” the weights of the first few layers of the ResNet-50 model during training process. The hyperparameters in the first 10 layers of the ResNet-50 model were kept constant (frozen) during the training phase. As a result, training time is marginally reduced, with a negligible loss in prediction accuracy.

ResNet-50 input image size is 224x224, so the preprocessed thyroid US images were rescaled to that resolution with bicubic interpolation before the training phase. Moreover, since the utilized ResNet-50 is trained for the classification of 1000 different classes, its output layer needed modifications for the model to become suitable for the multiclass classification of thyroid US images. The last two layers of the ResNet-50 model are a fully connected and a softmax layer. The fully connected layer connects each neuron in this layer to all the neurons in the next layer. The softmax layer selects the highest probability of occurrence among all classes and sets it to 1. The rest of them are set to zero. Basically the softmax layer picks the class with highest probability.

Fully connected and the softmax layers had to be modified for the dataset that was used in transfer learning. These two layers are replaced by the ones with suitable dimensions depending on the dataset utilized for the fine-tuning process. If the NN is going to be used for classifying the calcification feature of thyroid US images, the fully connected layer size should match the number of different classes for the calcification feature. This procedure is repeated for the other thyroid US image features, according to the number of different cases(classes) present for each feature.

By following that transfer learning procedure for each of the imaging features of thyroid US images, CNN parameters are fine-tuned to classify different thyroid image features. The outputs of these models can be used to predict the final TIRADS score of the thyroid nodule, which designates its risk category.

2.3.4 Partial Augmentation

The number of samples between different classes are not uniformly distributed, thus, there is a bias problem in the dataset. As a result of this, the predictions made by the NNs tend to be biased towards the class which dominates the dataset. Even though by using Focal Loss and traditional dataset augmentation with x and y translation, rotation, scaling, results can be improved a bit, but the bias towards the dominating class can not be fully eliminated since the traditional image augmentation is applied equally on all classes.

The bias effect can be minimized by a different kind of image augmentation. Chi et. al divided the thyroid images into different segments around the location where the thyroid nodule is located. Then the sub-images are used as a novel augmentation scheme [2]. Similarly, in the presented work, the nodule perimeter is found from the Cartesian coordinates provided in the dataset, and its center of mass location is calculated. The images are divided into 4 sub-images around the center of the nodule, and these additional images are used to augment the full scale ones, but only for the classes with very low number of samples, in order to have a better balanced distribution among all classes. A side-by-side example of the original image and the sub-images are given in Figure 2.9.

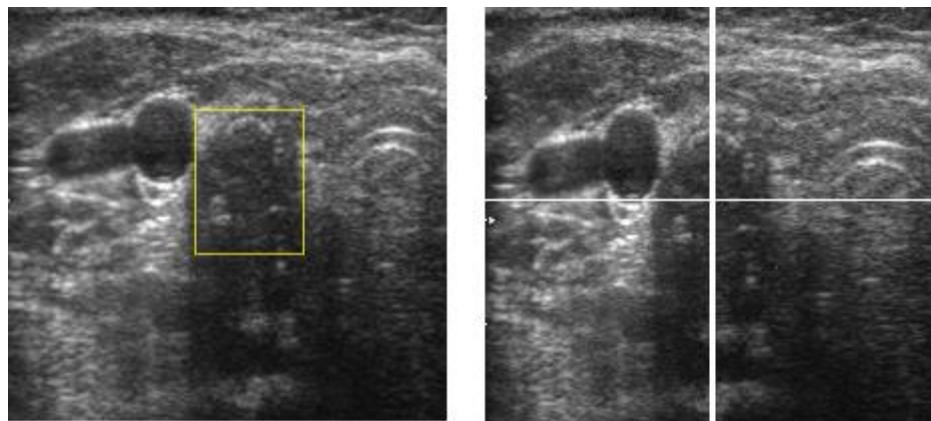


Figure 2.9 Partial augmentation scheme

$\begin{bmatrix} a & b & c \\ d & & e \end{bmatrix}$ (a) Original image, (b) Top-left sub-image, (c) Top-right sub-image, (d) Bottom-left sub-image, (e) Bottom-right sub-image

3.RESULTS AND DISCUSSION

The proposed method for the multiclass classification of thyroid nodules to estimate the TIRADS scores was partially implemented and its applicability was investigated by analyzing the results obtained for the calcification and echogenicity features. ResNet-50 CNNs were trained, results were compared and improvements obtained by the application of Focal Loss and the partial augmentation scheme were presented. Advantages, disadvantages and possible future improvements and modifications to the proposed method were also discussed.

3.1 ResNet-50 Training and Results

The full implementation of the proposed method was not possible due to the fact that features other than calcification and echogenicity had much more bias in between classes and the validation accuracies of the neural networks were unable to go above 65%. Results of training and validation for echogenicity and calcification properties are provided in Figure 3.1 and Figure 3.5 respectively.

When dividing the datasets into validation and training sets, 30% vs. 70% ratio was used. All neural networks were trained with Adam Optimizer, and a learning rate of 0.0001. Implementation was performed by using Matlab and the models were trained on an Nvidia GTX970 general purpose GPU with 3500MB memory. Experimentally determined Focal Loss parameters $\alpha = 5$ and $\gamma = 0.25$ were chosen.

Different ResNet-50 models were trained with different datasets, for different number of epochs, and different validation frequencies, with and without partial augmentation. Maximum training time is observed as 40 minutes, when all classes in the calcification dataset were augmented with 4 sub-images each, and validated after 2 epochs, with a total of 100 epoch training time.

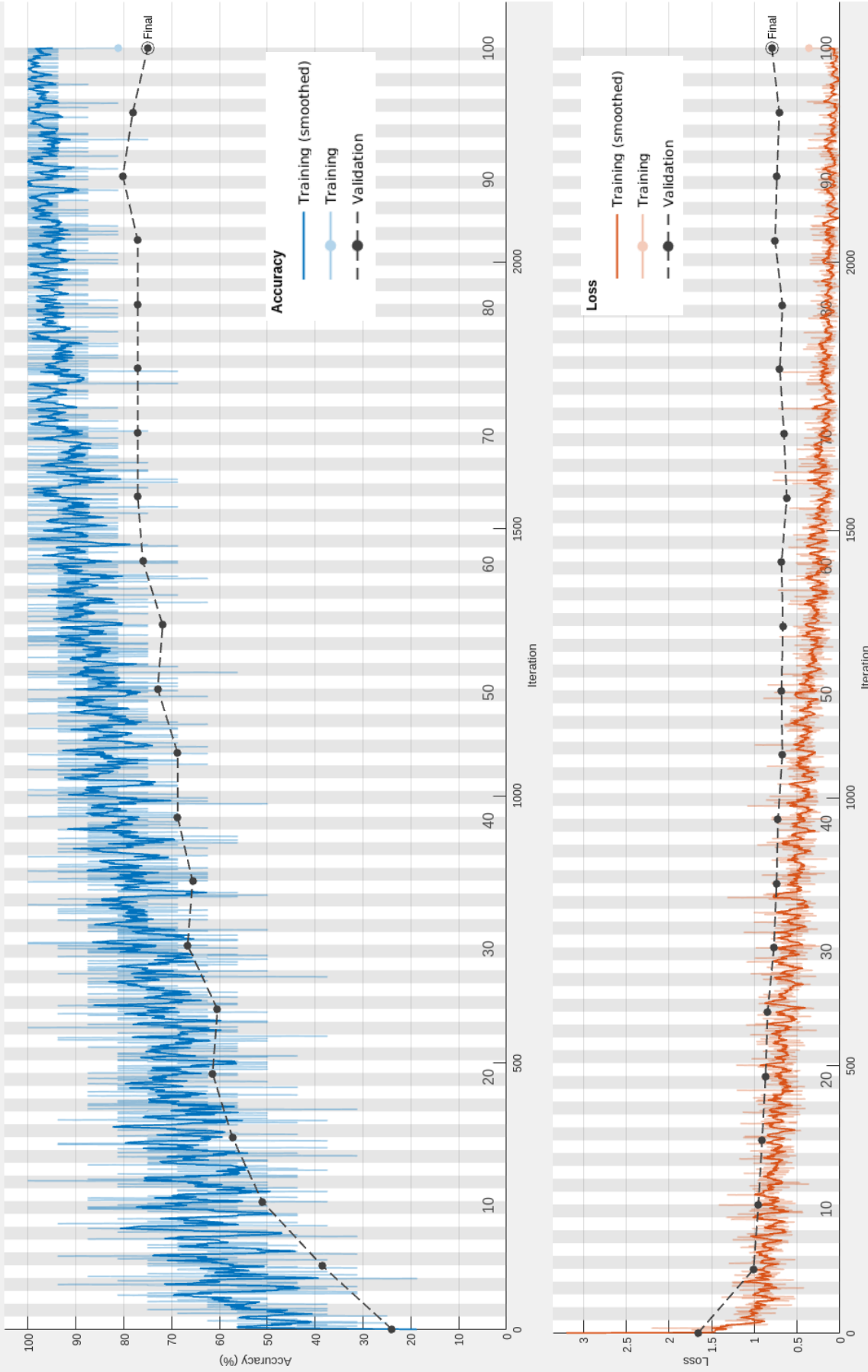


Figure 3.1 Training, validation accuracies and losses for echogenity

3.1.1 Effects of Focal Loss

Focal Loss Function with $\alpha = 5$ and $\gamma = 0.25$ was utilized, and its effects on calcification scores can be seen in Figure 3.2 and Figure 3.3. Figure 3.2 demonstrates the training process for calcification property with Cross-Entropy, whereas Figure 3.3 shows the same training process with Focal Loss.

Main advantage of using Focal Loss is that, training accuracy starts to diverge from validation accuracy much later compared to Cross-Entropy Loss. As a result of this, the NN model starts overfitting the training data much later during the training process when Focal Loss is used. This result can be seen when the training accuracies in Figure 3.2 and Figure 3.3 are compared. Another advantage of using Focal Loss can be observed by comparing the loss graphs provided in Figure 3.2 and Figure 3.3. In both loss graphs, training losses (plotted in orange color) approach zero as training accuracies increase, but the validation loss when Focal Loss is used is significantly lower than the validation loss with Cross Entropy. Interpretation of this result is that, the misclassified images are in a much closer proximity of the ground truth class when Focal Loss is used instead of Cross-Entropy Loss.

3.1.2 Effects of Partial Augmentation

As explained in Section 2.3.4, a new partial augmentation scheme was used to minimize the effect of imbalanced number of samples between different classes. This can be seen from the difference between the number of samples belonging to the macrocalcification and microcalcification classes. Number of sample thyroid US images with macrocalcifications are 43, whereas there are 216 sample images with microcalcifications as presented in Table 2.3.

Augmenting the images belonging to both macrocalcification and microcalcification classes just increases the number of samples for each class, but it does nothing to minimize the bias towards the microcalcification class with dominant number of samples. Instead of augmenting all of them, only the images belonging to the macrocalcification class are partially augmented with 4 sub-images. Then, the number of samples belonging to that class becomes 215.

As a result of partial augmentation, prediction accuracies have significantly increased. Figure 3.4 has a peak prediction accuracy around 78% without partial augmentation for the calcifications whereas Figure 3.5 has its peak at 88% with partial augmentation applied on macrocalcification class. Partial augmentation provided a significant 12.8% improvement from 78% prediction accuracy to 88%. This improvement can also be observed from Table 3.1 and Table 3.2 by comparing the number of correct classifications for each class. Partial augmentation does not only improve the classification performance of the NN on the class with small number of samples. It also improves the overall performance by providing a better balanced dataset.

Feature	Samples	Correct Classifications
Macrocalcifications	43	30
Microcalcifications	216	181
None	137	103

Table 3.1 Number of correct classifications for calcifications without focal loss and partial augmentation

Feature	Samples	Correct Classifications
Macrocalcifications	215	189
Microcalcifications	216	197
None	137	117

Table 3.2 Number of correct classifications for calcifications with focal loss and partial augmentation

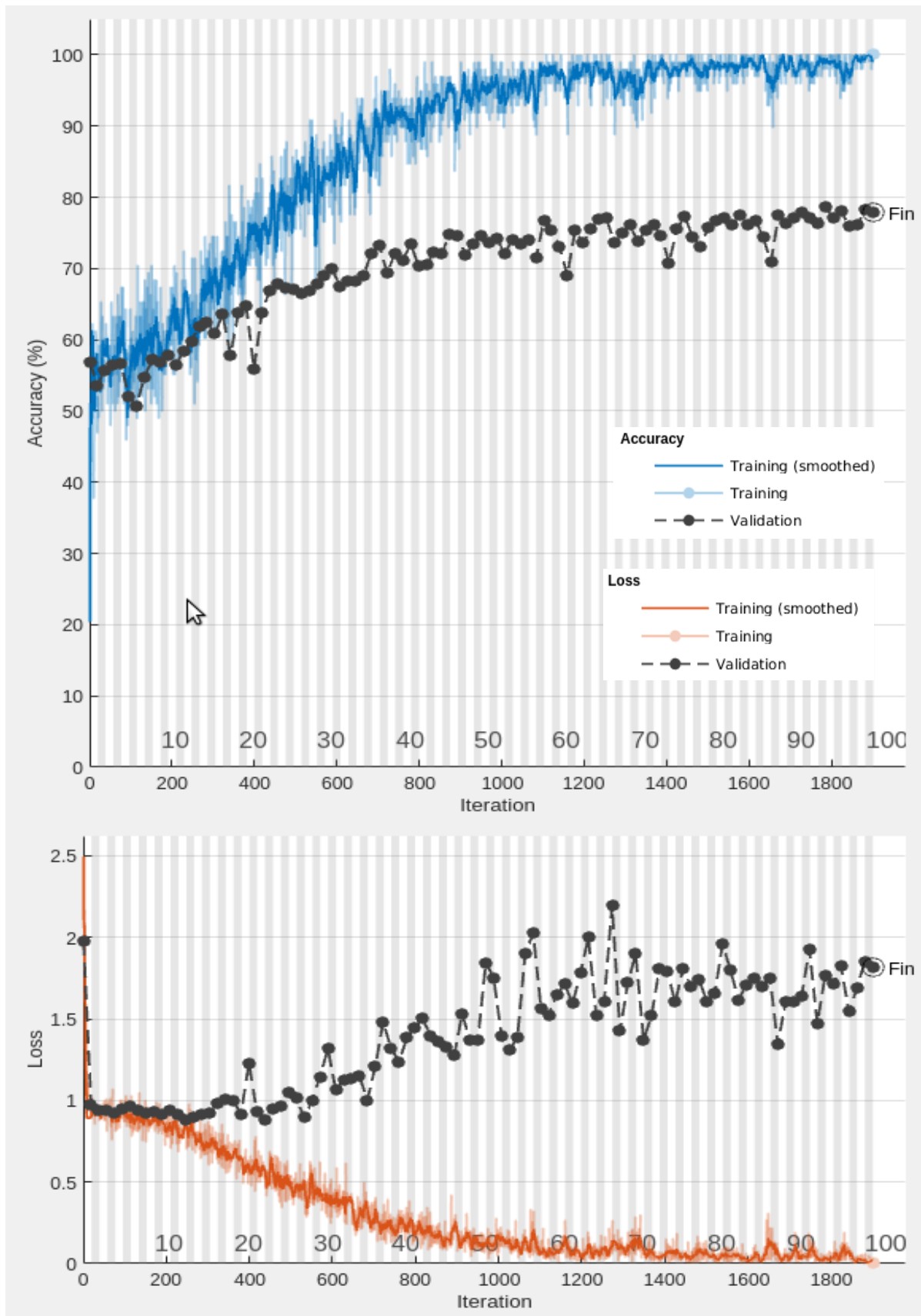


Figure 3.2 Training and validation process for calcifications with cross-entropy loss

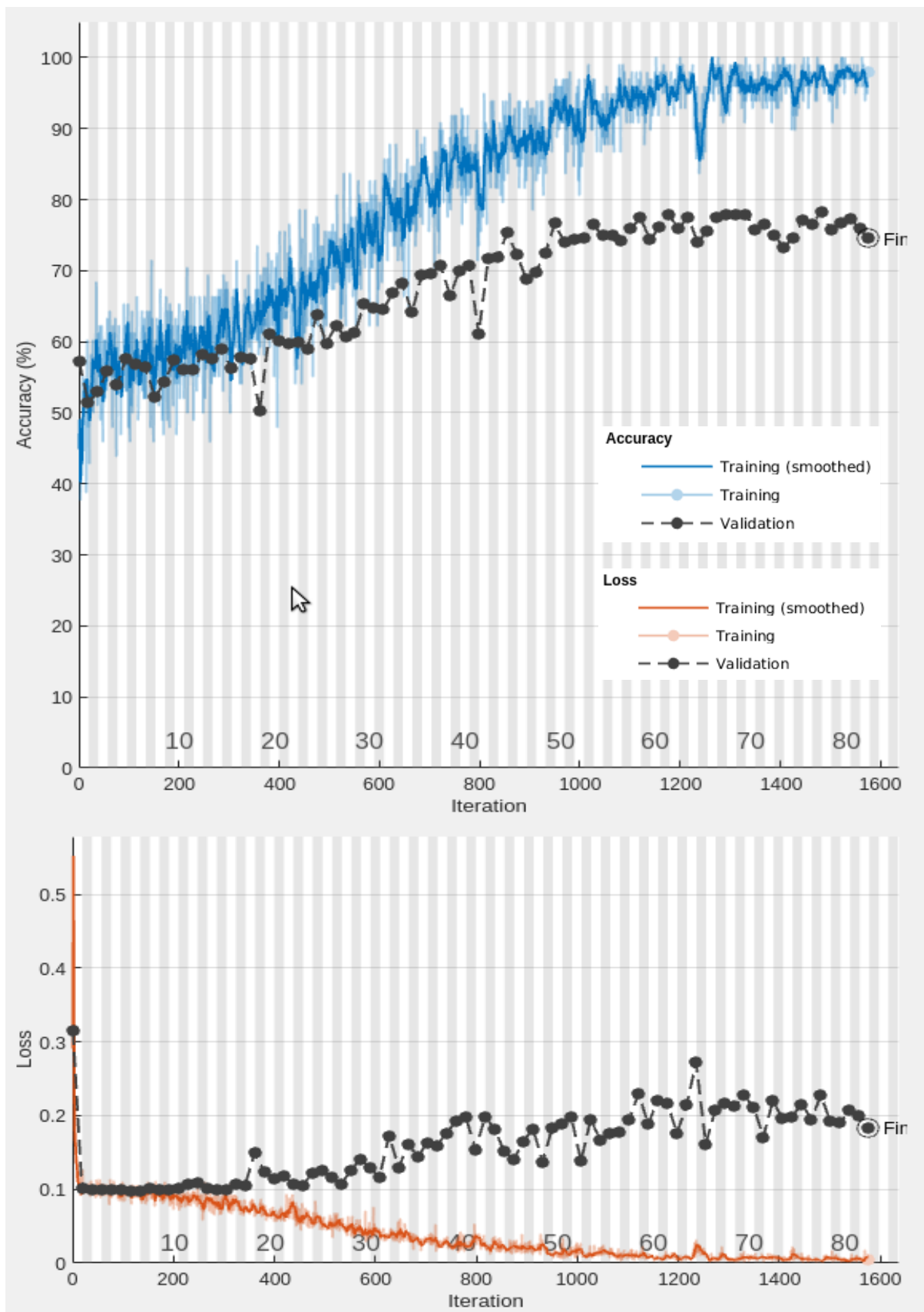


Figure 3.3 Training and validation process for calcifications with focal loss

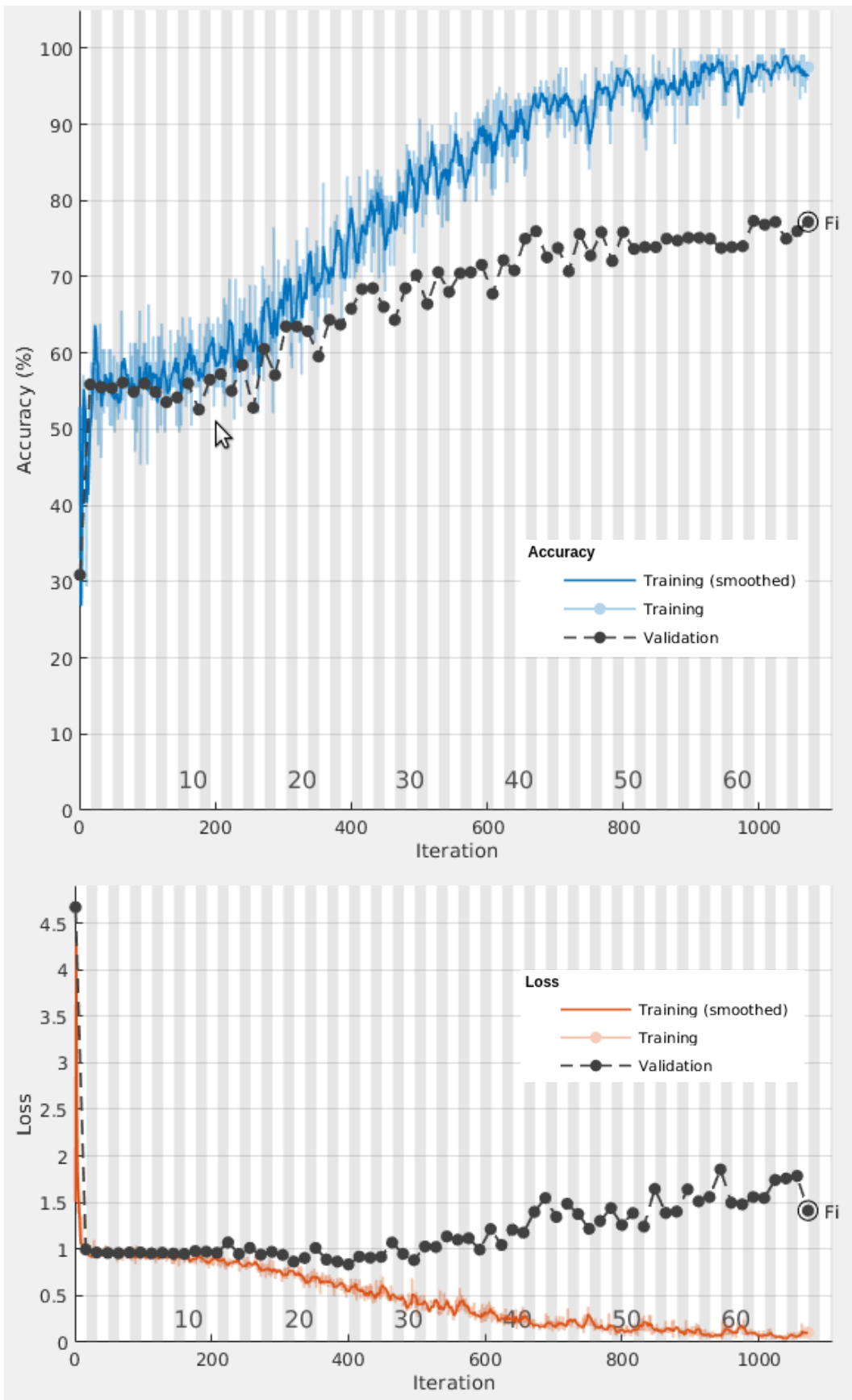


Figure 3.4 Calcification training and validation without partial augmentation

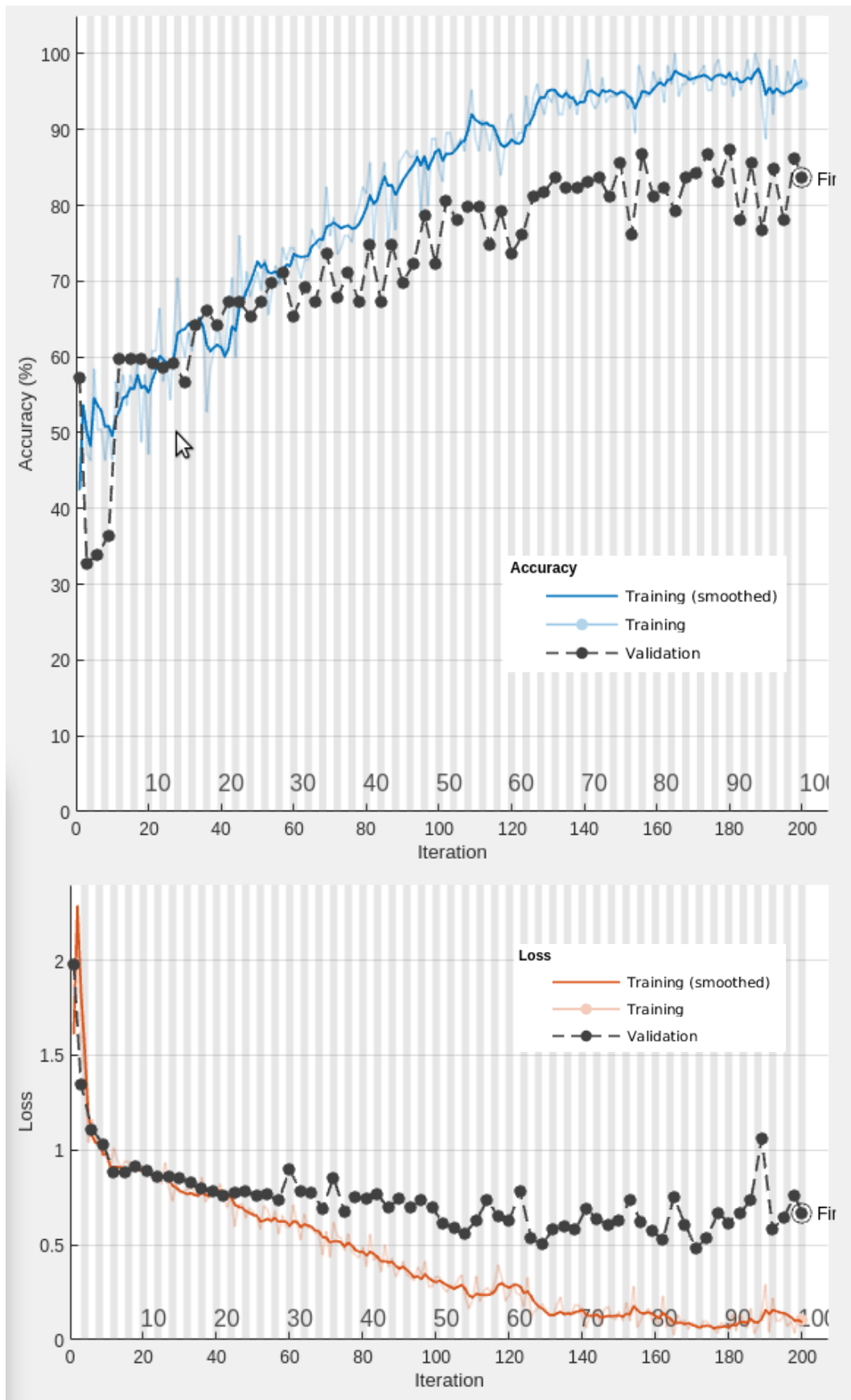


Figure 3.5 Calcification training and validation with partial augmentation, applied on the macrocalcification class

4.CONCLUSION

In the presented work, a novel method for estimating the TIRADS scores from thyroid US image features was proposed, and its applicability was investigated by classifying the two most important features and analyzing the results. The proposed method converts the binary classification problem of thyroid US images into a multiclass classification one. Conversion of the classification was achieved by the presented image preprocessing steps such as connected component labeling, template matching, image inpainting and derivation of new datasets suitable for multiclass classification and a partial image augmentation method. In the final step, CNNs were trained and obtained results were discussed. Moreover, the effects of choosing Focal Loss over Cross-Entropy Loss were stated. The effect of partial image augmentation on classification accuracy for calcifications was demonstrated.

One disadvantage of this proposed method is, it requires a much larger dataset in order to reach almost perfect classification accuracy, due to the fact that it can provide multiclass classification. However, it must be noted that, the proposed method requires much less samples than application of multiclass classification with just one neural network. Another disadvantage of the proposed method is related to memory requirement. It would need more memory in order to run on a computer or an embedded device. However, as the amount of memory even on embedded devices is increasing tremendously, that would not be a problem in the near-future. The main advantage of the proposed method is being able to perform the classification of TIRADS scores of thyroid nodules with a high resolution.

Transfer learning and fine-tuning for the derived datasets were performed and the obtained results were presented. Even though the number of samples are limited and the available data is biased, the results seem promising and successful, with validation accuracy of 85% for calcification feature and 80% validation accuracy for the echogenity feature. Using Focal Loss and partial augmentation provide a significant advantage in reaching these classification accuracies with a limited number of biased thyroid US images. Therefore, the implementation of this method with a larger and unbiased dataset can achieve a very high overall accuracy in classifying thyroid US images into different TIRADS score categories.

Training a system with the proposed method, with a good amount of diverse samples, the system would save the precious time of radiologists and other medical staff, since it has the potential to reduce the amount of unnecessary FNAB procedures applied on very low risk patients. Moreover, it would also save the patients from the pain and costs of conducting a biopsy on their thyroid. It also has the potential to be used as a training system for medical students during the education process.

In conclusion, the presented method can be used to predict the risk category of thyroid nodules, with small number of samples and high output resolution when trained with a more suitable dataset. The presented method could also be turned into a hybrid method. For example, composition parameter can be predicted by using deterministic methods and image processing techniques with much higher accuracy than a CNN. In addition, it might be possible to determine the echogenicity parameter with higher accuracy by using wavelet-based techniques. Finally, results obtained for each different feature with different techniques such as image processing and neural networks may be combined to achieve a high overall prediction accuracy. Investigating different approaches and their applicability to different thyroid US image features, and merging the best possible results to achieve a high overall multiclass classification accuracy are left as the focus of further research studies.

REFERENCES

- [1] L. Wang, S. Yang, S. Yang, C. Zhao, G. Tian, Y. Gao, Y. Chen, and Y. Lu, "Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network," *World Journal of Surgical Oncology*, vol. 17, no. 1, Jan. 2019, doi: 10.1186/s12957-019-1558-z.
- [2] J. Chi, E. Walia, P. Babyn, J. Wang, G. Groot, and M. Eramian, "Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network," *Journal of Digital Imaging*, vol.30, no.4, pp.477–486, Jul.2017, doi: 10.1007/s10278-017-9997-y.
- [3] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of Graphics Tools*, vol. 9, no. 1, pp. 23–34, Jan. 2004, doi: 10.1080/10867651.2004.10487596.
- [4] T. Liu, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, doi: 10.1109/icassp.2017.7952290.
- [5] J. T. Wang, P. Babyn, G. Groot, and R. Otani, "Electronic synoptic reporting of thyroid nodules: Potential for reduction in number of patients undergoing thyroid nodule biopsies," *Open Journal of Radiology*, vol. 06, no. 03, pp. 233–242, 2016, doi:10.4236/ojrad.2016.63031.
- [6] D. E. Maroulis, M. A. Savelonas, S. A. Karkanis, D. K. Iakovidis and N. Dimitropoulos, "Computer-aided thyroid nodule detection in ultrasound images," *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, 2005, pp. 271-276, doi: 10.1109/CBMS.2005.44.
- [7] Y. Zhu, Z. Fu and J. Fei, "An image augmentation method using convolutional network for thyroid nodule classification by transfer learning", *3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 1819 -1823, doi: 10.1109/Comp.Comm.-2017.8322853.

- [8] D. Koundal, S. Gupta, and S. Singh, "Computer aided thyroid nodule detection system using medical ultrasound images," *Biomedical Signal Processing and Control*, vol. 40, pp. 117–130, Feb. 2018, doi: 10.1016/j.bspc.2017.08.025.
- [9] B. Wildman-Tobriner, M. Buda, J. K. Hoang, W. D. Middleton, D. Thayer, R. G. Short, F. N. Tessler, and M. A. Mazurowski, "Using artificial intelligence to revise acr ti-rads risk stratification of thyroid nodules: Diagnostic accuracy and utility," *Radiology*, vol. 292, no. 1, pp. 112–119, 2019, PMID: 31112088, doi:10.1148/radiol.2019182128.
- [10] D.Koundal, "Computer-aided diagnosis of thyroid nodule: A review," *International Journal of Computer Science & Engineering Survey*, vol. 3, no. 4, pp. 67–83, Aug. 2012, doi:10.5121/ijcses.2012.3406.
- [11] Y. J. Choi, J. H. Baek, H. S. Park, W. H. Shim, T. Y. Kim, Y. K. Shong, and J. H. Lee, "A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment," *Thyroid*, vol. 27, no. 4, pp. 546–552, Apr. 2017, doi: 10.1089/thy.2016.0372.
- [12] S. Gitto, G. Grassi, C. D. Angelis, C. G. Monaco, S. Sdao, F. Sardanelli, L. M. Sconfienza, and G. Mauri, "A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound," *La radiologia medica*, vol. 124, no. 2, pp. 118–125, Sep. 2018, doi: 10.1007/s11547-018-0942-z.
- [13] J. Wang, S. Li, W. Song, H. Qin, B. Zhang and A. Hao, "Learning from Weakly-Labeled Clinical Data for Automatic Thyroid Nodule Classification in Ultrasound Images," *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3114-3118, doi: 10.1109/ICIP.2018.8451085.
- [14] D. Bibicu, L. Moraru, and A. Biswas, "Thyroid nodule recognition based on feature selection and pixel classification methods," *Journal of Digital Imaging*, vol. 26, no. 1, pp. 119–128, May 2012, doi: 10.1007/s10278-012-9475-5.

- [15] D. S. Dean and H. Gharib, “Epidemiology of thyroid nodules,” *Best Practice & Research Clinical Endocrinology & Metabolism*, vol. 22, no. 6, pp. 901–911, Dec. 2008, doi: 10.1016/j.beem.2008.09.019.
- [16] G. Popoveniuc and J. Jonklaas, “Thyroid nodules,” *Medical Clinics of North America*, vol. 96, no. 2, pp. 329–349, Mar. 2012, doi: 10.1016/j.mcna.2012.02.002.
- [17] D. Koundal, S. Gupta, and S. Singh, “Automated delineation of thyroid nodules in ultrasound images using spatial neutrosophic clustering and level set,” *Applied Soft Computing*, vol. 40, pp. 86–97, Mar. 2016, doi: 10.1016/j.asoc.2015.11.035.
- [18] M. S. Yildirim, H. Atasoy, C. Ceylan, and A. Akan, “Computerized tomography based novel features in thyroid cancer,” in *2017 Medical Technologies National Congress (TIPTEKNO)*, 2017, pp. 1–4.
- [19] Teng, D., Fu, P., Li, W. et al. Learnability and reproducibility of ACR Thyroid Imaging, Reporting and Data System (TI-RADS) in postgraduate freshmen. *Endocrine* 67, 643–650 (2020), doi: 10.1007/s12020-019-02161-y.
- [20] Aksoy, M.S., Torkul, O. & Cedimoglu, I.H. An industrial visual inspection system that uses inductive learning. *Journal of Intelligent Manufacturing* 15, 569–574 (2004). doi: 10.1023/B:JIMS.0000034120.86709.8c.
- [21] “Template Matching” [Online]. Available: https://en.wikipedia.org/wiki/Template_matching. [Accessed 28 December 2020].
- [22] “OpenCV: Template Matching” [Online]. Available: https://docs.opencv.org/master/d4/dc6/tutorial_py_template_matching.html. [Accessed 28 December 2020].
- [23] “Inpainting” [Online]. Available: <https://en.wikipedia.org/wiki/Inpainting>. [Accessed 12 January 2021].

[24] Elharrouss, O., Almaadeed, N., Al-Maadeed, S. et al. Image Inpainting: A Review .*Neural Processing Letters* 51, 2007–2028 (2020). doi: 10.1007/s11063-019-10163-0.

[25] “OpenCV: Image Inpainting” [Online]. Available: https://docs.opencv.org/master/df/d3d/tutorial_py_inpainting.html. [Accessed 12 January 2021].

[26] A. Geron, *Hands-On Machine Learning with Scikit-Learn and Tensorflow*, O’Reilly, 2017.

[27] “Matlab Documentation – Deep Learning Toolbox: Pretrained Deep Neural Networks” [Online]. Available: <https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html>. [Accessed 13 January 2021].

[28] “Demystifying Focal Loss I: A More Focused Cross Entropy Loss” [Online]. Available: <https://medium.com/ai-salon/demystifying-focal-loss-i-a-more-focused-version-of-cross-entropy-loss-f49e4b044213>. [Accessed 16 January 2021].

[29] “Demystifying Focal Loss II: A Distance-aware Cross Entropy Loss” [Online]. Available: <https://medium.com/ai-salon/demystifying-focal-loss-ii-a-distance-aware-cross-entropy-loss-7cecb6bf7cf7>. [Accessed 16 January 2021].

[30] Fürnkranz J. (2011) Decision Tree. In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA, doi: 10.1007/978-0-387-30164-8_204

[31] “Train Deep Learning Network to Classify New Images” [Online]. Available: <https://www.mathworks.com/help/deeplearning/ug/train-deep-learning-network-to-classify-new-images.html>. [Accessed 16 January 2021].

[32] “Transfer Learning Using Pretrained Network” [Online]. Available: <https://www.mathworks.com/help/deeplearning/ug/transfer-learning-using-pretrained-network.html>. [Accessed 17 January 2021].

[33] C. Szegedy et al., "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[34] Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., & Romero, E. (2015). An open access thyroid ultrasound image database. In E. Romero & N. Lepore (Eds.), *10th International Symposium on Medical Information Processing and Analysis*. SPIE, doi: 10.1117/12.2073532.

[35] "ImageNet" [Online]. Available: <https://en.wikipedia.org/wiki/ImageNet>. [Accessed 16 January 2021].