Multi-model Deep Learning Framework for Thyroid Cancer Classification Using Ultrasound Imaging

Mariusz Rafało
Warsaw School of Economics
Warsaw, Poland *mrafalo@sgh.waw.pl*

Agnieszka Żyłka
Maria Skłodowska-Curie National Research Institute of Oncology
Warsaw, Poland

Abstract

This study presents the development and evaluation of a novel, multi-model AI system designed to train and deliver multiple deep learning models for the classification of focal lesions in thyroid, using ultrasound images. Leveraging a dataset of 484 images, we trained a diverse array of 1300 models encompassing advanced convolutional neural networks, including ResNet, DenseNet and VGG architectures. To minimize random errors, the training dataset was randomly sampled 20 times. The primary objective was to enhance the diagnostic accuracy in distinguishing benign from malignant thyroid nodules through automated analysis. The performance of our models was rigorously assessed, demonstrating promising results with an average area under the curve of 0.86 and sensitivity of 0.85. These findings highlight the significant potential of integrating deep learning techniques with ultrasound imaging to improve the classification of thyroid nodules.

Keywords: deep learning, thyroid cancer, image processing, US image classification.

1. Introduction

Thyroid nodules are prevalent, affecting over 50% of adults, with approximately 5% exhibiting malignancy, indicative of thyroid cancer [1]. An ultrasound (US) scan is the first-choice imaging technique in the diagnostic of the thyroid gland. US uses low-frequency sound waves, which makes it safe for the patient, unlike other imaging methods such as computed tomography (CT) or magnetic resonance imaging (MRI), which use radiation. US thyroid scans have a number of advantages: they are quick, non-invasive and widely available in many medical facilities. In addition, patients receive results immediately after the examination, which speeds up the diagnostic and treatment process. Moreover, based on the ultrasound image, the area for tissue collection for fine needle aspiration biopsy (FNAB) can be precisely selected, allowing for an accurate diagnosis of thyroid nodules.

However, ultrasound images are susceptible to speckle noise, which makes the diagnosis difficult. US image analysis depends largely on the parameters of the device on which the test is carried out and on the diagnostician who adjusts the device during the test. Moreover, the examination is carried out on a different scale, and the diagnostician's focus is directed at focal lesions, often without a reference point. These factors make US examinations difficult to compare with each other in terms of image standardization.

The use of artificial intelligence (AI) methods, including machine learning and deep neural networks can help in estimating the risk of malignancy of a focal lesion in the thyroid gland and in determining the optimal diagnostic and therapeutic procedure, including qualifying the lesion for fine-needle biopsy [16]. In order to use machine learning and AI for image diagnostics, these images are standardized and annotated to a form that can be used in training process. Deep

neural network is a mathematical technique that allows for identifying features in complex data structures, e.g. voice, images or videos. These networks can include convolutions, i.e. image operations that allow the detection of important image features (e.g. edges, shapes, colors, etc.). Convolutional neural networks are a type of deep neural networks mainly used to analyze data with high complexity and variable structure. They also implement activation functions that allow the detection of non-linear relationships. Furthermore, important feature in the training of a convolutional neural network is a pooling, which allows to change the size of the input image without losing important details [15].

Deep neural network's ability to automatically select features from image data has led to widespread utilization of image classification across various sectors, including education, industry, and notably, the medical field. It represents a novel approach to medical classification and have been widely embraced by researchers for diagnosing a diverse range of diseases. Convolutional neural networks are currently a common technique used in medical image analysis with a number of studies that confirm the usefulness of this technique. These algorithms are used in the analysis of x-ray images, computed tomography [21], ultrasound images [17], skin lesions[2], brain tumors [11] or liver lesions [18]. The use of deep learning methods in thyroid US image diagnosis is based on the use of advanced algorithms and data processing techniques [16] and the number of available algorithms is gradually increasing [13]. It is important to assess the usefulness of these methods in correlation with the results of thyroid ultrasound examination, using common malignancy risk classifiers like EU-TIRADS standard (Thyroid Imaging Reporting and Data System). TIRADS aims to minimize the number of unnecessary fine needle biopsies of the thyroid gland and aims to increase the sensitivity and specificity of the diagnosis. EU-TIRADS allows to classify focal lesions into five categories, depending on the risk of malignancy, the size of the focal lesion and its characteristics [4].

Beyond the use of convolutional neural networks, several studies have explored the application of other medical imaging techniques for classifying abnormal thyroid nodules. For example, some research use a transfer learning approach, which is a machine learning technique where a pre-trained model developed for one task is reused as the starting point for a model on a second task [22]. This approach allows you to shorten the model training time, because the base model is already adapted to identify selected features [7].

Moreover a visual transformer (ViT) neural network is an advanced deep neural network architecture that uses the concepts of analyzing image sequences. Image data is divided into small fragments, called segments, which are then transformed into vectors and processed by subsequent layers of the deep neural network. This allows for effective recognition of features in images. The transformer technique was developed in 2017 and has found wide application in natural language processing. In 2020, this algorithm was adapted for image analysis. Studies comparing convolutional networks with transformer networks show that in certain situations, the transformer is superior to the convolutional network. However, they require more computing power and (usually) more training data [23].

The goal of this paper is to to implement a multi-model deep learning system that enhances the diagnostic accuracy in distinguishing benign from malignant thyroid nodules through automated analysis. Our study fits into the research area of improving the effectiveness of models based on convolutional neural networks. In addition to the effectiveness of the models, we also focus on minimizing the variance of the results. Low variance allows the developed models to be used on various data sets. The idea is that the developed AI models could be effectively used in medical practice and not only in isolated analytical experiments.

2. Population

The study is a retrospective (therefore, there was no need to obtain any medical consent) analysis of ultrasound images of focal thyroid lesions in patients who underwent thyroid surgery at the

Department of Oncological Endocrinology and Nuclear Medicine of the Maria Skłodowska-Curie National Research Institute of Oncology in Warsaw, Poland. The initial dataset comprised 270 nodules from 270 patients who underwent diagnostic thyroid ultrasound examinations and US-guided fine needle aspiration of a focal thyroid nodule between 2019 and 2022. 155 patients were diagnosed from malignant and the rest (115) from benign lesions. The data set is therefore relatively balanced in relation to the category being searched for. The dataset was refined by excluding nodules with initial non-diagnostic or indeterminate cytologic results and without further histologic diagnosis (27 patients). Moreover, one patient had missing images in one orthogonal plane.

This refinement process results in a dataset of 242 nodules from 242 patients, which gives a total of 484 US images. Each focal lesion was presented in two images. The images present the same focal change, but were taken in a both orthogonal planes. Each image is then treated separately and annotated separately. We use two categories of nodules: malignant and benign. Each focal lesion was verified in histopathological examination after surgical treatment, which allows for confirmation of its nature (ground truth).

3. Method

In the study, we analyze the current state of common convolutional neural network models for thyroid cancer detection as well as developing our own models from scratch. We then compare 5 developed CNN models with several commonly used models, including three ResNet models (*ResNet50*, *ResNet101*, *ResNet152*), two VGG models (*VGG16* and *VGG19*), two DenseNet (*DenseNet121* and *DenseNet151*) and one InceptionNet model (*v3*).

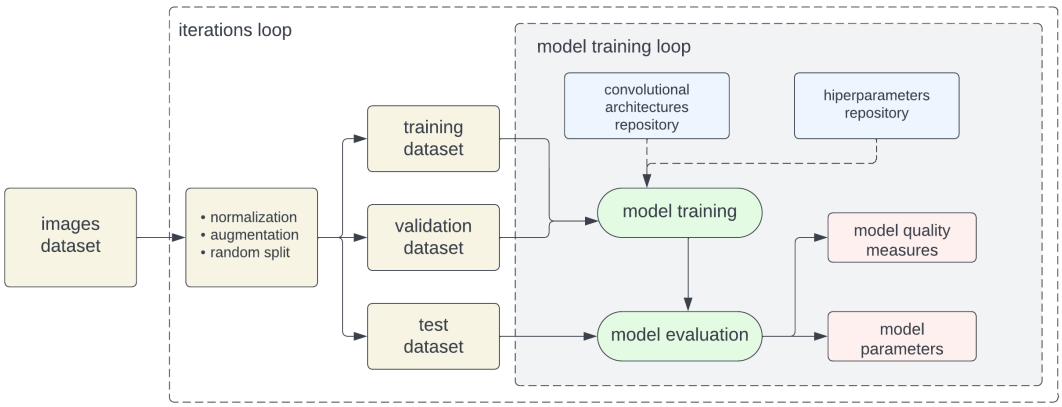

Fig. 1. Multi-model deep learning framework overview. Source: own study

We conduct the analysis according to the steps presented in Figure 1. These steps include exploratory analysis of image characteristics, input images preparation and standardization, development of multiple AI models based on convolutional neural networks and comparison of classification results.

During image pre-processing, the image data was normalized using the Z-score method (the resulting data has a mean of 0 and a standard deviation of 1). We also implemented models based on augmented data. For each image we generated $3 - 10$ slightly rotated or modified ones. However, augmentation did not contribute to a significant improvement in the quality of the models.

The developed system is used to conduct a series of simulations, including building, training and validating each model using Monte Carlo cross-validation technique (random selection) [14]. The results of each simulation iteration and each cross-validation iteration were recorded.

During the simulation, a total of 1377 convolutional neural networks were created, trained and validated. Each network was trained on 200 epochs. Each training process was performed 20 times (number of cross-validation iterations), which gives a total of over 27 000 trained models. In each iteration, the dataset was divided into training, test and validation datasets. The training dataset accounted for 76% (370 images) the test set was 15% (73 images) and validation dataset was 9% (41 images).

During the study, we developed five of our own convolutional neural network architectures. A summary pf these models (textitcnn1 - *cnn5*), along with an indication of the number of layers and number of parameters, is presented in Table 1. Individual networks differ in the number of layers, pooling scope and kernel sizes.

Table 1. Deep neural network parameter comparison. Source: own study.

Model name	Iterations	Parameters	Convolution layers	Pooling layers
cnn1	1368	132 775	7	5
cnn2	1368	314 230	6	5
cnn3	1368	30 111	5	4
cnn4	1368	1 322 775	7	5
cnn5	1368	134 599	7	5

For each model, we assess the metrics of accuracy, sensitivity, precision and area under ROC curve (AUC). The accuracy metric for a classification model is represented as the ratio of correctly predicted observations to the total observations. Sensitivity, also referred to as recall or the true positive rate, calculates the fraction of actual positive cases accurately detected by the model. This metric is vital in the context of medical diagnostics, as it is essential to capture all significant conditions without fail. Precision quantifies the accuracy of positive predictions made by the model. In the realm of medical imaging, achieving high precision is key to ensuring the dependability of diagnoses rendered by the model, thereby reducing the likelihood of false positives that might cause undue stress or lead to unwarranted medical interventions. Both sensitivity and precision are critical in the medical domain because they balance the trade-off between overlooking potential diagnoses and overdiagnosing patients. Optimizing these measures leads to more accurate and trustworthy models, ultimately improving patient care and outcomes in medical practice.

In deep learning, a loss function is a crucial component that measures the difference between the model's predicted output and the actual output. The primary goal of the loss function is to guide the training of the neural network by quantifying the error of the model, enabling the optimization algorithms to adjust the weights in a direction that minimizes this error. The choice of loss function directly impacts the model's performance and its ability to generalize from the training data. Optimizing the loss function using techniques such as gradient descent is fundamental to the learning process, enabling deep learning models to make accurate predictions. For this reason, and due to the aim of the work, which is to maximize the generalizability of the model's predictions, we decided to conduct a series of simulations for various loss functions. We focused on the following loss functions: categorical cross-entropy (binary cross-entropy in our case), focal loss, KL divergence loss and squared hinge loss.

The multi-model framework was implemented in Python programming language. The program is controlled by parameters, which allows for effective simulation in various conditions. The source codes are published in the author's public GitHub repository[1].

[1] https://github.com/mrafalo/thyroid

4. Results

The findings show that model *cnn1* has an AUC reaching 0.85, with a sensitivity of 0.91 and a precision of 0.75. The remaining, predefined models achieved a maximum AUC of 0.86. The standard deviation for the AUC measure for all iterations ranged from 0.06 to 0.09, depending on the models. The highest variance was achieved by the *VGG19* model, and the lowest by *ResNet152*. However, the differences are minimal. Therefore, the developed models achieved similar or better thyroid cancer classification quality, while maintaining much lower computational complexity (see Table 1 to compare number of parameters). Among the predefined models, the *ResNet101* model achieved high results, reporting AUC of 0.86. Most models achieved the highest performance for the categorical cross-entropy loss function. However, the focal loss function turned out to give good results for the *ResNet152* model, however with poor AUC score (0.66). Squared hinge loss function turned out to give good results for the *cnn4* model, presenting good sensitivity and precision (both around 0.79). KL divergence loss did not achieve acceptable results and was not included in the Table 2, however detailed results on how a given model depends on the loss function used are summarized in Table 3.

Table 2. Simulation summary for selected models. Source: own study.

Model	Optimizer	Loss function	AUC $\pm$ std. dev.	Sensitivity	Precision
ResNet101	SGD	categorical cross-entropy	0.86 ± 0.08	0.85	0.77
ResNet152	SGD	focal loss	0.66 ± 0.06	0.76	0.73
ResNet50	SGD	categorical cross-entropy	0.72 ± 0.07	0.59	0.80
VGG16	SGD	categorical cross-entropy	0.81 ± 0.08	0.81	0.81
VGG19	SGD	categorical cross-entropy	0.77 ± 0.09	0.71	0.79
cnn1	SGD	categorical cross-entropy	0.85 ± 0.09	0.91	0.75
cnn2	SGD	categorical cross-entropy	0.76 ± 0.08	0.71	0.79
cnn3	Adam	categorical cross-entropy	0.71 ± 0.07	0.83	0.78
cnn4	SGD	squared hinge	0.75 ± 0.06	0.79	0.79
cnn5	SGD	categorical cross-entropy	0.83 ± 0.08	0.62	0.81

Table 3. Simulation summary for selected models and loss functions. Source: own study.

Model	Categorical	Focal loss	KL div.	Squared hinge
ResNet101	0.86	0.62	0.72	0.62
ResNet152	0.65	0.66	0.59	0.60
ResNet50	0.72	0.68	0.65	0.57
VGG16	0.81	0.63	0.56	0.65
VGG19	0.77	0.68	0.62	0.67
cnn1	0.85	0.68	0.64	0.56
cnn2	0.76	0.64	0.50	0.58
cnn3	0.71	0.57	0.61	0.56
cnn4	0.73	0.66	0.60	0.75
cnn5	0.83	0.68	0.64	0.63

The study makes a significant advancement in aiding radiologists' in US image review. Existing studies typically have decent performance rates, achieving accuracy measure above 80% and AUC measure of up to 85% [8]. Moreover, there are number of predefined deep learning models commonly used for thyroid nodule classification tasks including VGG family models such as VGG16 and VGG19 [5], ResNet models [19, 20], InceptionNet models [12] and DenseNet models [13]. These architectures can achieve AUC results even above 90% [6]. These studies are mostly based on data from research centers (hospitals), mostly in China and Korea. Our research, compared to other reports, is summarized in Table 4. The proposed

multidimensional deep learning system enables relatively simple determination of important deep learing network parameters, while minimizing the complexity of the network (number of layers and number of parameters). This approach allows for significant optimization of the network training process. We verified 10 model architectures, through 27000 iterations. It turns out that an acceptable quality of the model can be achieved for a relatively simpler (and therefore easy to learn) architecture, especially in studies on a small number of images.

Table 4. Selected research on the use of CNN in the classification of thyroid cancer. Source: own study.

Research method	Input Image size	Patients	Images	Results
CNN [3]	$(160 \times 160 \times 1)$	1239	1278	AUC: 0.87
Inception [12]	$(224 \times 224 \times 3)$	298	450	Accuracy: 0.92
transfer learning [22]	$(299 \times 299 \times 1)$	1734	1759	AUC: 0.95
CAD [13]	$(224 \times 224 \times 1)$	8339	18049	AUC: 0.92
CNN-based CAD [9]		124	138	AUC: 0.87
DenseNet169 [21]	$(224 \times 224 \times 3)$	880	986	AUC: 0.9
ResNet18 [19]	$(224 \times 224 \times 1)$	508	508	AUC: 0.91
ResNet18 [20]		8079	10021	AUC: 0.88
Transfer learing [7]	$(299 \times 299 \times 1)$	5575	5575	AUC: 0.90
This study	$(180 \times 180 \times 1)$	242	484	AUC: 0.85

The selection of the appropriate neural network architecture depends on a number of factors, but most often individual frameworks are verified empirically. Our system may be an effective tool for such verification.

5. Conclusions

CNN models presented in the study have shown a potential in predicting the thyroid cancer. The incorporation of multi-model framework, allows to effectively find the optimal parameters of the neural network, also taking into account the complexity of the network and stability of prediction results.

Differences in AUC and sensitivity observed in various studies can often be traced back to disparities in methodologies and the demographics of the patient populations. Due to such a variance in research results and difficulties in adapting the developed models to other data sets, it is necessary to develop a tool that allows for the construction of relatively simple models that can be used in many medical facilities. Our system supports these assumptions, allowing the construction of many CNN models. This research emphasizes the value of incorporating AI into the clinical decision-making process for thyroid nodules. Although deep learning algorithms have demonstrated efficacy in some clinical results, applying AI directly in clinical settings without the supervision of medical professionals is not feasible, primarily due to the inconsistencies present in real-world data. However, still the efficacy of AI in comparison to the expertise of radiologists remains a topic of debate. Based on comprehensive and scientifically rigorous studies, AI's performance is generally on par with or slightly below that of experienced radiologists [3].

This paper contributes in several significant ways. Firstly, we conduct a comprehensive comparison of the developed models against other widely used AI models to validate their suitability. Secondly, the system was developed and tested on real-world diagnostic data from a single hospital in Poland. Finally, our study makes it possible to use the developed system on any other set of data by making our implementation available on GitHub repository.

In weighing the results of this study, several limitations should be considered. First, the dataset was relatively small ($n = 484$), especially compared to other studies. While our results

demonstrates good sensitivity and AUC compared to existing studies, limitations arise concerning precision scores. Moreover, maintaining the stability of results across different subsets of data is still challenging. This applies especially to relatively small datasets. However, the practice of medical research shows that the collections are usually unbalanced and often small in number. Therefore, future work involves testing our approach on more extensive datasets, containing images of different size and drawn from different population. Moreover, an approach based on regularization is possible; there are features in the focal lesions (e.g.: shape, uneven edges, etc.) that may indicate malignancy. Model regularization could therefore improve prediction, by taking these characteristics into account.

The practical advantage of our study is the possibility of adapting it to the EU-TIRADS system and, consequently, the possibility of using the classification results for the diagnostic process.

References

[1] Acharya, U.R., Swapna, G., Sree, S. V, Molinari F., et al.: A review on ultrasound-based thyroid cancer tissue characterization and automated classification. *Technol. Cancer Res. & Treat.*, vol. 13, no. 4, pp. 289–301, 2014.

[2] Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019.

[3] Buda, M., Wildman-Tobriner, B., Hoang, J.K., Thayer, D., Tessler, F.N., Middleton, W.D., Mazurowski, M.A.: Management of thyroid nodules seen on us images: Deep learning may match performance of radiologists. *Radiology*, vol. 292, no. 3, pp. 695–701, 2019.

[4] Dobruch-Sobczak, K., Adamczewski, Z., Dedecjus, M., Lewiński, A., Migda, B., Ruchała, M., Skowrońska-Szcześniak, A., Szczepanek-Parulska, E., Zajkowska, K., Żyłka, A.: Summary of meta-analyses of studies involving TIRADS classifications (EU-TIRADS, ACR-TIRADS, and K-TIRADS) in evaluating the malignant potential of focal lesions of the thyroid gland. *J. Ultrason.*, vol. 22, pp. e121–e129, 2022.

[5] Dov, D., Kovalsky, S.Z., Cohen, J., Range D. E., et al.: Thyroid cancer malignancy prediction from whole slide cytopathology images. in *Machine Learning for Healthcare Conference*, pp. 553–570, 2019.

[6] Gomes Ataide, E.J., Ponugoti, N., Illanes, A., Schenke S., et al.: Thyroid Nodule Classification for Physician Decision Support Using Machine Learning-Evaluated Geometric and Morphological Feature. *Sensors*, vol. 20, article 6110, 2020.

[7] Lee, J.H., Kim, Y.G., Ahn, Y., Park, S., Kong, H.J., Choi, J.Y., Kim, K., Nam, I.C., Lee, M.C., Masuoka, H., Miyauchi, A., Kim, S., Kim, Y.A., Choe, E.K., Chai, Y.J.: Investigation of optimal convolutional neural network conditions for thyroid ultrasound image analysis. Sci. Rep. 13 (1), 1–9 (2023)

[8] Li, X., Zhang, S., Zhang, Q., et al.: Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: A retrospective, multicohort, diagnostic study. *Lancet Oncol.*, vol. 20, no. 2, pp. 193–201, 2019.

[9] Liang, X., Huang, Y., Cai, Y., Liao Jianyi, et al.: A Computer-Aided Diagnosis System and Thyroid Imaging Reporting and Data System for Dual Validation of Ultrasound-Guided Fine-Needle Aspiration of Indeterminate Thyroid Nodules. *Front. Oncol.*, vol. 11, October, pp. 1–8, 2021.

[10] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal Loss for Dense Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327 (2020)

[11] Liu, M., Zhang, J., Adeli, E., Shen, D.: Deep multi-task multi-channel learning for joint classification and regression of brain status. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, pp. 3–11, 2017.

[12] Nguyen, D.T., Kang, J.K., Pham, T.D., Batchuluun, G., Park, K.R.: Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. *Sensors (Switzerland)*, vol. 20, no. 7, 2020.

[13] Peng, S., Liu, Y., Lv, W., Liu, L., Zhou, Q., Yang, H., et al.: Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *Lancet Digit. Heal.*, vol. 3, no. 4, pp. e250–e259, 2021.

[14] Rafało, M.: Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis. ICT Express. 8 (2), 183–188 (2022)

[15] Rguibi, Z., Hajami, A., Zitouni, D., Elqaraqoui, A., Bedraoui, A.: CXAI: Explaining Convolutional Neural Networks for Medical. *Electronics*, vol. 11, no. 11, pp. 1775–1794, 2022.

[16] Sorrenti, S., Dolcetti, V., Radzina, M., Bellini, M.I., Frezza, F., Munir, K., Grani, G., Durante, C., D'Andrea, V., David, E., Calò, P.G., Lori, E., Cantisani, V.: Artificial Intelligence for Thyroid Nodule Characterization: Where Are We Standing? *Cancers (Basel)*, vol. 14, no. 14, pp. 1–15, 2022.

[17] Tehrani, A.K.Z., Amiri, M., Rosado-mendez, I.M., Hall, T.J., Rivaz, H.: A Pilot Study on Scatterer Density Classification of Ultrasound Images Using Deep Neural Networks. in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2059–2062, Montreal, 2020.

[18] Todoroki, Y., Iwamoto, Y., Lin, L., Hu, H., Chen, Y.W.: Automatic Detection of Focal Liver Lesions in Multi-phase CT Images Using A Multi-channel & Multi-scale CNN. in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 872–875, 2019.

[19] Yang, J., Shi, X., Wang, B., Qiu, W., Tian, G., Wang, X., Wang, P., Yang, J.: Ultrasound Image Classification of Thyroid Nodules Based on Deep Learning. *Front. Oncol.*, vol. 12, July, pp. 1–9, 2022.

[20] Yao, S., Shen, P., Dai, T., Dai, F., Wang, Y., Zhang, W., Lu, H.: Human understandable thyroid ultrasound imaging AI report system — A bridge between AI and clinicians. iScience. 26 (4), 106530 (2023)

[21] Zhao, H.B., Liu, C., Ye, J., Chang, L.F., Xu, Q., Shi, B.W., Liu, L.L., Yin, Y.L., Shi, B. Bin: A comparison between deep learning convolutional neural networks and radiologists in the differentiation of benign and malignant thyroid nodules on CT images. *Endokrynol. Pol.*, vol. 72, no. 3, pp. 217–225, 2021.

[22] Zhou, H., Jin, Y., Dai, L., Zhang, M., Qiu, Y., Wang, K., Tian, J., Zheng, J.: Differential Diagnosis of Benign and Malignant Thyroid Nodules Using Deep Learning Radiomics of Thyroid Ultrasound Images. *Eur. J. Radiol.*, vol. 127, 2020.

[23] Zoph, B., Shlens, J.: Learning Transferable Architectures for Scalable Image Recognition. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.