EPiC Series in Engineering

Volume 4, 2023, Pages 72–79

Proceedings of Emerging Tech Conference: Edge Intelligence 2022

Explainable Deep Learning for localising abnormal Parathyroid Glands in parathyroid scintigraphy

Ioannis D. Apostolopoulos[a*], Dimitris J. Apostolopoulos[b], Trifon Spyridonidis[b], George S. Panayiotakis[a]

[a]Department of Medical Physics, School of Medicine, University of Patras, Greece, GR 265-00
[b]Department of Nuclear Medicine, University General Hospital of Patras, School of Medicine, University of Patras, Greece, GR 265-00

ece7216@upnet.gr

Abstract

Parathyroid scintigraphy with 99mTc-sestamibi (MIBI) is an established technique for localising abnormal Parathyroid Glands (PGs). However, the identification and localisation of PGs require much attention from medical experts and are time-consuming. Artificial Intelligence methods can offer an assisting solution and can be embedded at the edge of Medical Decision Support Systems and hospital computer stations. This retrospective study enrolled 632 patients, who underwent parathyroid scintigraphy with double-phase and thyroid subtraction techniques. The study proposes a three-path approach, employing the state-of-the-art Convolutional Neural Network called VGG19. Image input to the model involved a set of three scintigraphic images in each case: MIBI early phase, MIBI late phase, and 99mTcO4 thyroid scan. A medical expert's diagnosis provided the ground truth for positive/negative results. Moreover, the visualised suggested areas of interest produced by the Grad-CAM algorithm are examined to evaluate the PG-level agreement between the model and the experts. Medical experts identified 545 abnormal glands in 452 patients. On a patient basis, the Deep Learning (DL) model attained an accuracy of 94.8% (sensitivity 93.8%; specificity 97.2%) in distinguishing normal from abnormal scintigraphic images. On a PG basis and in achieving identical positioning of the findings with the experts, the model correctly identified and localised 453/545 glands (83.1%) and yielded 101 false focal results (false positive rate 18.23%). Concerning surgical findings, the expert's sensitivity was 89.68% on patients and 77.6% on a PG basis, while that of the model reached 84.5% and 67.6%, respectively. Deep Learning in parathyroid scintigraphy can potentially assist medical experts in identifying abnormal findings. Despite the time-consuming training procedure and the high computational demands, these limitations apply only to the training procedure. Once the model is trained, it can deliver its

D. Mourtzis (ed.), ETCEI2022 (EPiC Series in Engineering, vol. 4), pp. 72–79

predictions swiftly, due to its lightweight nature. This fact enables the utilisation of such models in a plethora of systems, even wearable devices.

1 Introduction

Parathyroid adenoma is part of a spectrum of parathyroid proliferative disorders, including parathyroid hyperplasia, parathyroid adenoma, and parathyroid carcinoma [1]. Approximately 85 per cent of primary hyperparathyroidism (HPPT) is caused by a parathyroid adenoma, followed by parathyroid hyperplasia with a percentage of 15. Parathyroid carcinoma is rare [2]. However, recent evidence does no longer support the entity of hyperplasia in primary HPPT with multiple abnormal glands and, in this setting, suggests the presence of two or more parathyroid adenomas. Therefore, according to the WHO 2022 classification, the term "hyperplasia" should be confined to secondary HPPT, while primary HPPT should be replaced by "primary HPPT-related multiglandular parathyroid disease" [3]. Severe secondary HPPT is caused primarily by end-stage renal failure. In this situation, all Parathyroid Glands (PGs) are enlarged, each to a different degree. Tertiary HPTT denotes the persistence of HPPT after successful renal transplantation. Despite the current use of calcimimetic drugs, which succeed in lowering serum calcium and parathyroid hormone levels, the definitive cure of HPPT is the surgical excision of abnormal PGs. The surgical approach relies highly on imaging modalities' preoperative localisation of enlarged glands. Preoperative localising methods include neck ultrasound (U/S), parathyroid scintigraphy, dynamic contrast-enhanced computerised tomography, 4-D C.T., and magnetic resonance imaging (MRI). Depending on local experience and expertise, U/S and scintigraphy are used first, while 4-D C.T. and MRI are usually reserved for negative or ambiguous cases.

Parathyroid scintigraphy is performed with the intravenous injection of the radioactive tracer 99mTc-Sestamibi (MIBI). The dual-phase technique includes acquiring early (10 minutes post- MIBI administration) and late (2 hours post-injection) images of the neck and the mediastinum. In early images, MIBI uptake by the thyroid gland may obscure the detection of an underlying parathyroid adenoma. However, most abnormal PGs exhibit prolonged tracer retention and appear prominent in late images, while MIBI clears more rapidly from the thyroid gland. Fast clearance of MIBI from some parathyroid adenomas and many hyperplastic glands is a common cause of false negative scans. The thyroid subtraction technique requires the administration of a second radioactive tracer (123I or 99mTc-pertechnetate) to delineate the thyroid gland. Then, the thyroid image is subtracted digitally from the early MIBI image. Early MIBI uptake by some thyroid nodules is a common cause of false positive findings. The two techniques can be applied alone or can be combined. In addition to planar images, SPECT or SPECT/CT can also be implemented to increase the method's sensitivity and offer a more precise localisation of findings in the 3-D space [4].

Computer-Aided Diagnostic (CAD) assistance in parathyroid adenoma identification could alleviate human tiredness and routine in everyday clinical practice, allowing medical experts to put their efforts into nontrivial tasks. Still, human expertise is indispensable to evaluating computer suggestions, which is a lot simpler task.

Recent advances in Deep Learning (DL) algorithms demonstrate substantial performance in detecting and classifying medical images [5–7]. DL introduced a feature extraction revolution from image data, enabling encapsulating millions of potentially significant image features. DL algorithms can learn to detect and distinguish important features that characterise an image according to a pre-defined label. For example, such methods have attained remarkable success in various cancer-detection studies on various imaging modalities [8–10].

Recent clinical studies report novel optical technologies that enhance PGs' localisation or viability assessment. These technologies could complement the surgeon's eyes and improve surgical outcomes in thyroidectomy and parathyroidectomy [11]. Most of the studies focus on developing surgeon-assisting tools for accurately detecting PGs. Those studies' contribution to the field is beyond question. However, little has been investigated regarding the non-invasive detection of parathyroid using medical image acquisition devices and before the surgery.

The study proposes a multi-path DL pipeline to simultaneously process the MIBI early phase, MIBI late phase, and the 99mTcO4 thyroid scan (Figure 1). To this end, the study employs the state-of-the-art Convolutional Neural Network (CNN) called Virtual Geometry Group (VGG) to furnish a multi-path pipeline, which performs a per-patient classification between normal and abnormal scans. Furthermore, the Grad CAM algorithm is employed to visualise the important local areas of each image according to the model.

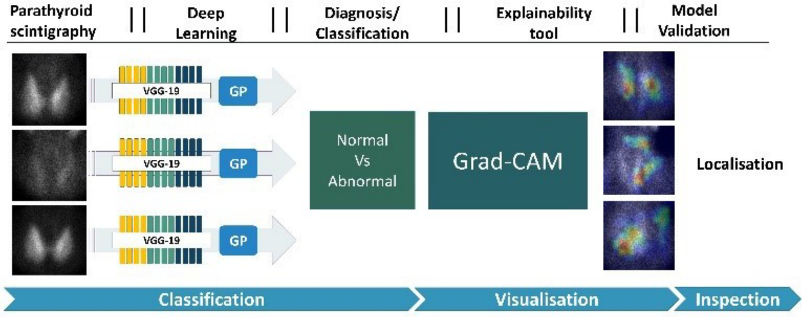

Figure 1: Overview of the study

2 Methods

2.1 Dataset and preprocessing

From January 2010 to December 2021, 632 patients with HPPT were referred to our department for parathyroid scintigraphy. Six hundred seven had biochemical evidence of primary HPPT, and 25 had refractory secondary or tertiary HPPT due to end-stage renal failure. We used the planar dual-phase technique in all patients and, whenever the medical experts judged it necessary, the thyroid subtraction technique in 514 cases (81.3%). In addition, 99mTc-pertechnetate (TcO4) for thyroid delineation was administered either after the conclusion of the dual-phase study or on another day. We used a pinhole collimator placed 10 cm over the neck for planar imaging. A SPECT/CT imaging session focused on the neck, and the mediastinum using a high-sensitivity parallel-hole collimator took place approximately 30 minutes post tracer injection. However, only planar imaging data have been included in the present study. Planar and SPECT/CT imaging were performed by the Hawkey-4 system (G.E. Healthcare). Two senior medical experts retrospectively evaluated the planar scintigraphic studies. In a few ambiguous cases, the final decision was achieved by consensus.

The original images are 1400x1050 pixels and contain five sub-figures. The informative details gather in the early MIBI, the late MIBI, and the thyroid TcO4 image Figure 2. Next, the annotations and the irrelevant artifacts are removed by reducing the area of attention. The final images are 350x350 pixels in jpeg format. Data pre-processing has been performed using the OpenCV library, written for the Python programming language

2.2 Parathyroid Network

CNNs are capable of portraying high-level abstract representations from non-linear information. CNNs belong to the broader area of deep Neural Networks [12]. CNNs utilise convolution layers to process and filter the input data distributions. Convolution layers transform the input data distributions and extract many image-related features [13]. Auxiliary layers, such as pooling layers, aid in dimensionality reduction, overfitting prevention, regularisation, and more [14]. In classification problems, the extracted feature maps are commonly processed by densely connected layers that filter out the irrelevant features based on a pre-defined desired outcome. The problem of PG identification is addressed by cross- examining three images, as presented earlier. To this end, a three-path CNN is suggested. The MIBI early phase image, MIBI late phase image, and the 99mTcO4 thyroid image are processed independently by the three paths of the network, and the extracted features from each path are fused at the later processing stages. Each path is responsible for extracting meaningful information from a single input image. Therefore, the overall approach contains three independent CNN components.

For each CNN component, the VGG architecture with 19 convolutional layers (VGG19) is suggested (Figure 2). VGG19 is a very consistent and successful CNN for relevant medical imaging tasks [5–7,15]. Initially, this network is designed to perform multi-class classification on non-medical images. However, its uniform architecture and feature extraction capabilities have also made it suitable for medical imaging tasks. The Triple-VGG19, called ParaNet, contains 3,079,628 trainable parameters and 52,999,836 non-trainable. At the top (the last convolutional layer) of each VGG19 component, a Global Average Pooling layer has been applied. The input image size (350x350x1) is incrementally reduced to (21x21x512), where 512 represents the number of filters of the last convolutional layer. The output of the Average Max Pooling layer is connected to a dense layer of 750 nodes, which is followed by a Dropout layer that randomly disconnects half of the nodes. Next, a dense layer of 256 nodes is connected to the previous layer, and a final densely connected layer of two nodes (as many as the output classes) follows.

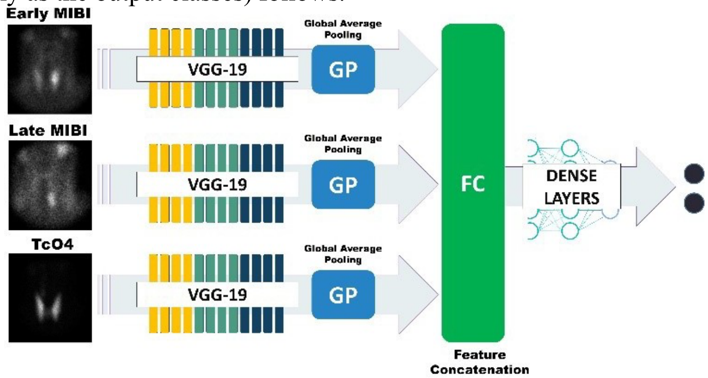

Figure 2: The architecture of Parathyroid Network. Three VGG-19 components process the input images. The extracted features are fused and supplied to the densely connected network, which performs the classification between the normal and abnormal class

10-fold stratified cross-validation is used to train and evaluate ParaNet. Stratified cross-validation splits the complete dataset into ten non-overlapping folds. During each iteration, a unique fold is used as a test set, whereas the remaining folds constitute the training sets. The stratified validation ensures class-aware creation of those folds to avoid biased sets. At the end of the iterations, each fold of the original dataset has been used as the test set once. The final metrics are calculated based on the average metrics of the ten iterations. The final confusion matrix is populated from each iteration.

The proposed network is trained for a maximum of 400 epochs. Early stopping is applied to avoid redundant training epochs. More specifically, the model instantly aborts the remaining training epochs when the validation accuracy has reached 94%. Each epoch of training is performed in mini- batches of 50 examples.

The agreement rating reflects the overall accuracy (ACC) score and Cohen's Kappa measure. However, considering medical expertise as the ground truth, the total number of True Positive (T.P.), True Negative (T.N.), False Positive (F.P.), and False Negative (F.N.) samples are reported. The model is also evaluated using the corresponding Sensitivity, Specificity, Positive Predicting Value (PPV), Negative Predicting Value (NPV), F-1 score, and Area Under Curve Score (AUC).

During each fold, the discussed metrics are recorded. Those metrics are based on the number of True Positive, False Positive, True Negative, and False Negative subjects of the fold. After the 10th fold, the average values of the above metrics are recorded, and the final confusion matrix is extracted.

3 Results

According to medical experts' decision, 180 cases were classified as negative (28.48%). In the remaining 452 patients, 545 abnormal PGs were identified in various positions.

The model has been evaluated on a patient-level basis following 10-fold stratified cross-validation. In this section, the model's metrics are presented. It is highlighted that the present evaluation takes place by opposing the predicted labels to the actual labels. It does not refer to cross-examination using the resulting suggested areas, as illustrated by Grad-CAM. The results demonstrate significant agreement between the model and the human experts. More specifically, the DL model obtains 94.8% accuracy and an F1 score of 0.96. The model achieves high sensitivity and specificity rates (93.8% and 97.2%, respectively). PPV and NPV values are 98.8% and 86.2% respectively. Cohen's Kappa statistic score is found to be 0.91.

During the stratified 10-fold cross-validation, the ten test sets are used to evaluate the model's accuracy. The Grad-CAM algorithm integration ensures that the model identifies each test image group's suggested areas of interest. At the end of the ten iterations, each fold has participated in the evaluation set only once. Selected samples from the Grad-CAM results are visually provided and discussed. It is highlighted that the initial agreement of 94.8%, as presented in the earlier sections, has decreased to 76.5% on a PG level basis. In addition, the visualisations revealed cases wherein the model yielded correct predictions, the suggested area of interest is irrelevant. Re- assessment of the agreement rating between the experts and the model's suggestions has been performed following this observation. The reader shall recall that the PG level assessment involves a case-to-case examination of the 545 PGs in the images. Table 1 summarises the results.

Agreement	ACC	SEN	SPE	PPV	NPV	F1
	(%)	(%)	(%)	(%)	(%)	(%)
Patient-level (632 subjects)	94.8	93.8	97.2	98.8	86.2	96.3
PG level (545 PGs)	76.5	83.1	63.5	81.8	65.7	82.4

Table 1: Results on a patient-level basis and a PG-level basis.

In Figure 3, confirmed abnormal PGs are presented. The arrows point at positive scintigraphic findings. As observed from case 1b, the visualisation reveals irrelevant areas suggested by the model, even in cases where the model's predicted class is correct ("Normal" class). The model correctly identifies multiple positive findings in cases 1a, 1b, and 1c. In case 2a, which corresponds to positive cases, the model predicts them as normal. Still, the model identifies some PGs (e.g., case 2b). However, those findings of the model are not correctly characterised as abnormal. Therefore, the overall effectiveness of the model has to be re-assessed following the inspection of the Grad-CAM results. Nevertheless, the model demonstrates some promising visualisation results, as observed from 1a and 1c.

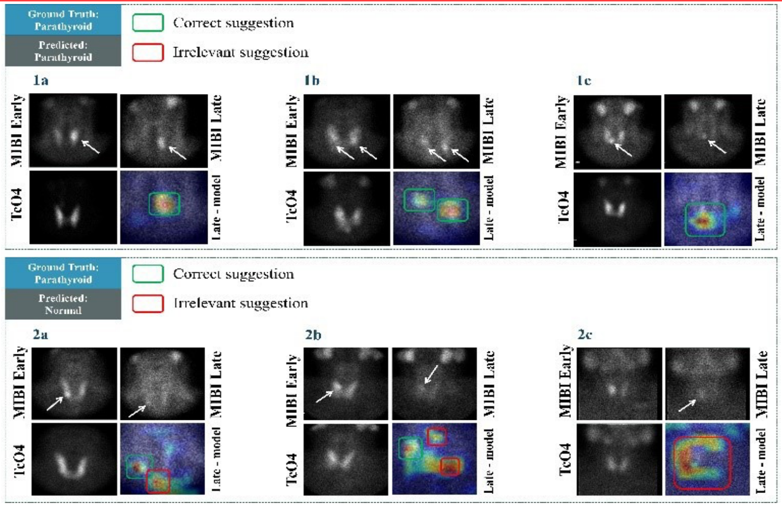

Figure 3: Examples of the Grad-CAM algorithm outcomes.

4 Discussion and Conclusions

The contributions of this study are two-fold. Firstly, an effective DL framework has been proposed to detect abnormal parathyroid scintigraphy images of patients with suspected HTTP. The efficiency of the presented ParaNet topology reached approximately 95% in detecting images with abnormal PGs. Secondly, the Grad-CAM algorithm is successfully employed to assist human experts in explaining the model's decisions. However, an extensive evaluation on a PG level basis revealed that the model could not correctly identify the factual findings' actual location while producing several False Positive findings. Furthermore, an extensive evaluation on a PG level basis revealed that the model exhibited lower sensitivity than the experts in the whole study cohort (83.1%) and surgically confirmed cases (67.6% vs 77.6% of the experts) while producing several false positive findings (18.2% and 16.9% vs 4.5% of the experts).

The findings of this study revealed that deep networks might yield remarkable accuracy and minimum losses in terms of metrics, but their proper understanding may be limited. False positive reduction is necessary to improve the diagnostic efficiency of the model and is a matter of future research. The sub-optimal specificity caused by the overwhelming number of F.P. findings can be explained by two decisive factors constraining the model's learning capacity. Firstly, there is a strong data imbalance issue. Normal scans are under-represented (28.48%) in the dataset, thereby introducing susceptibility to biased training and results. Data augmentation has reduced the effect of this issue in model training. However, the imbalance issue remains and may not be circumvented completely. Secondly, the efficiency of the Grad-CAM algorithm is questionable in a variety of cases, as reported in the literature [16]. More specifically, Grad-CAM may fail to recognise multiple findings of the same class in the same image. In addition, grad-CAM may poorly visualise the exact location of the important features on some occasions. Therefore, future research involves employing more sophisticated approaches, such as the Grad-CAM++ algorithm [16].

Nevertheless, the actual agreement with the human expertise reached an acceptable rate (76.5% agreement on a PG level and 95% on a patient level). The study suffers some limitations. Firstly, the study employed state-of-the-art models solely. Though such models are of undeniable robustness, designing task-specific DL topologies and training them from scratch would potentially exhibit better results and reveal more significant regions of interest. For example, integrating an attention mechanism may enhance the model's ability to seek important features in vital areas of the image. Moreover, designing a three-component Siamese Network [17], which aggregates the distances of the three input images and computes the gradients based on a carefully designed loss function, may improve the results further. Secondly, the study used the experts' diagnostic yield as the ground truth. This limitation constrains the horizons of the experiments because we can only measure the agreement with the experts and not the prediction's precision compared to surgical and histopathologic results. On the other hand, surgical results can provide a minimal number of negative cases, which poses serious limitations for the training purposes of every DL model. Finally, more data could aid in the re-assessment of the proposed method.

These limitations cannot degrade the importance of the findings. With the absence of related works that use the same image source, this study is the first attempt to introduce Deep Learning approaches for the localisation of PGs in Parathyroid scintigraphy with 99mTc-sestamibi (MIBI) studies. It is demonstrated that DL can at least compete with human expertise in the specific task, which is very desirable when developing Medical Decision Support Systems. Despite the time-consuming training procedure and the high computational demands, these limitations apply only to the training procedure. Once the model is trained, it can deliver its predictions swiftly, due to its lightweight nature. This fact enables the utilisation of such models in a plethora of systems, even wearable devices.

References

[1] Wieneke JA, Smith A. Parathyroid Adenoma. Head and Neck Pathol 2008;2:305–8. https://doi.org/10.1007/s12105-008-0088-8.

[2] Thakker RV. Genetics of parathyroid tumours. J Intern Med 2016;280:574–83. https://doi.org/10.1111/joim.12523.

[3] Erickson LA, Mete O, Juhlin CC, Perren A, Gill AJ. Overview of the 2022 WHO Classification of Parathyroid Tumors. Endocr Pathol 2022;33:64–89. https://doi.org/10.1007/s12022-022-09709-1.

[4] Petranović Ovčariček P, Giovanella L, Carrió Gasset I, Hindié E, Huellner MW, Luster M, et al. The EANM practice guidelines for parathyroid imaging. Eur J Nucl Med Mol Imaging 2021;48:2801–22. https://doi.org/10.1007/s00259-021-05334-y.

[5] Apostolopoulos ID, Apostolopoulos DI, Spyridonidis TI, Papathanasiou ND, Panayiotakis GS. Multi-input deep learning approach for Cardiovascular Disease diagnosis using Myocardial Perfusion Imaging and clinical data. Physica Medica 2021;84:168–77. https://doi.org/10.1016/j.ejmp.2021.04.011.

[6] Apostolopoulos ID, Pintelas EG, Livieris IE, Apostolopoulos DJ, Papathanasiou ND, Pintelas PE, et al. Automatic classification of solitary pulmonary nodules in PET/CT imaging employing transfer learning techniques. Med Biol Eng Comput 2021;59:1299–310. https://doi.org/10.1007/s11517-021-02378-y.

[7] Apostolopoulos ID, Papathanasiou ND. Classification of lung nodule malignancy in computed tomography imaging utilising generative adversarial networks and semi-supervised transfer learning. Biocybernetics and Biomedical Engineering 2021;41:1243–57. https://doi.org/10.1016/j.bbe.2021.08.006.

[8] Astaraki M, Zakko Y, Toma Dasu I, Smedby Ö, Wang C. Benign- malignant pulmonary nodule classification in low-dose CT with convolutional features. Physica Medica 2021;83:146–53. https://doi.org/10.1016/j.ejmp.2021.03.013.

[9] Haggenmüller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill RL, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. European Journal of Cancer 2021;156:202–16. https://doi.org/10.1016/j.ejca.2021.06.049.

[10] Lee S-Y, Kang H, Jeong J-H, Kang D. Performance evaluation in Florbetaben brain PET images classification using 3D Convolutional Neural Network. PLoS ONE 2021;16:e0258214. https://doi.org/10.1371/journal.pone.0258214.

[11] Abbaci M, De Leeuw F, Breuskin I, Casiraghi O, Lakhdar AB, Ghanem W, et al.

Parathyroid gland management using optical technologies during thyroidectomy or parathyroidectomy: A systematic review. Oral Oncology 2018;87:186–96. https://doi.org/10.1016/j.oraloncology.2018.11.011.

[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[13] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

[14] LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, France: IEEE; 2010, p. 253–6. https://doi.org/10.1109/ISCAS.2010.5537907.

[15] Apostolopoulos ID, Apostolopoulos DJ, Papathanasiou ND. Deep Learning Methods to Reveal Important X-ray Features in COVID-19 Detection: Investigation of Explainability and Feature Reproducibility. Reports 2022;5. https://doi.org/10.3390/reports5020020.

[16] Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, p. 839–47. https://doi.org/10.1109/WACV.2018.00097.

[17] Li B, Yan J, Wu W, Zhu Z, Hu X. High performance visual tracking with siamese region proposal network. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, p. 8971–80.