algorithms

MDPI

Article

Explainable Artificial Intelligence Method (ParaNet+) Localises Abnormal Parathyroid Glands in Scintigraphic Scans of Patients with Primary Hyperparathyroidism

Dimitris J. Apostolopoulos [1], Ioannis D. Apostolopoulos [2,*], Nikolaos D. Papathanasiou [1], Trifon Spyridonidis [1] and George S. Panayiotakis [2]

[1] Department of Nuclear Medicine, University Hospital of Patras, University of Patras, 26504 Patras, Greece; dimap@med.upatras.gr (D.J.A.); nikopapath@upatras.gr (N.D.P.); tspyr@med.upatras.gr (T.S.)

[2] Department of Medical Physics, School of Medicine, University of Patras, 26504 Patras, Greece; panayiot@upatras.gr

* Correspondence: ece7216@upnet.gr

Abstract: The pre-operative localisation of abnormal parathyroid glands (PG) in parathyroid scintigraphy is essential for suggesting treatment and assisting surgery. Human experts examine the scintigraphic image outputs. An assisting diagnostic framework for localisation reduces the workload of physicians and can serve educational purposes. Former studies from the authors suggested a successful deep learning model, but it produced many false positives. Between 2010 and 2020, 648 participants were enrolled in the Department of Nuclear Medicine of the University Hospital of Patras, Greece. An innovative modification of the well-known VGG19 network (ParaNet+) is proposed to classify scintigraphic images into normal and abnormal classes. The Grad-CAM++ algorithm is applied to localise the abnormal PGs. An external dataset of 100 patients imaged at the same department who underwent parathyroidectomy in 2021 and 2022 was used for evaluation. ParaNet+ agreed with the human readers, showing 0.9861 on a patient-level and 0.8831 on a PG-level basis under a 10-fold cross-validation on the training set of 648 participants. Regarding the external dataset, the experts identified 93 of 100 abnormal patient cases and 99 of 118 surgically excised abnormal PGs. The human-reader false-positive rate (FPR) was 10% on a PG basis. ParaNet+ identified 99/100 abnormal cases and 103/118 PGs, with an 11.2% FPR. The model achieved higher sensitivity on both patient and PG bases than the human reader (99.0% vs. 93% and 87.3% vs. 83.9%, respectively), with comparable FPRs. Deep learning can assist in detecting and localising abnormal PGs in scintigraphic scans of patients with primary hyperparathyroidism and can be adapted to the everyday routine.

Keywords: deep learning; explainable artificial intelligence; parathyroid glands; hyperparathyroidism

check for updates

Citation: Apostolopoulos, D.J.; Apostolopoulos, I.D.; Papathanasiou, N.D.; Spyridonidis, T.; Panayiotakis, G.S. Explainable Artificial Intelligence Method (ParaNet+) Localises Abnormal Parathyroid Glands in Scintigraphic Scans of Patients with Primary Hyperparathyroidism. *Algorithms* **2023**, *16*, 435. https://doi.org/10.3390/a16090435

Academic Editors: Gilberto Ochoa-Ruiz, Raul Benitez and Christian Mata

Received: 24 July 2023
Revised: 5 September 2023
Accepted: 7 September 2023
Published: 11 September 2023

1. Introduction

Primary hyperparathyroidism (pHPT) is a common calcium metabolism disorder mainly affecting middle-aged females. It usually presents as a sporadic disease, while hereditary forms are much less common [1]. A hyper-functioning solitary parathyroid adenoma accounts for 80–85% of pHPT, while multiple abnormally functioning parathyroid glands (multiglandular disease) are responsible for the rest of the cases [2]. The diagnosis of pHPT is established biochemically by documenting increased serum calcium and parathyroid hormone levels and excluding causes of secondary HPT [3]. Despite using calcimimetic drugs to lower serum calcium and parathyroid hormone levels, the surgical excision of abnormal parathyroid glands (PGs) is currently the irreplaceable remedy for HPT [4]. Severe secondary HPT is caused primarily by end-stage renal failure. In this situation, all PGs are enlarged, each to a different degree.

Pre-operative localisation of abnormal PGs is highly desired to assist surgery [5,6]. Neck ultrasound, parathyroid scintigraphy, dynamic contrast-enhanced computerised tomography (CT), 4-D CT, and magnetic resonance imaging (MRI) are the imaging modalities for this task [7]. Employing neck ultrasound and scintigraphy as the first diagnostic approach is common. The other imaging methods are usually reserved for negative or ambiguous results [8].

Parathyroid scintigraphy is performed with the intravenous injection of the radioactive tracer ^{99m}Tc-Sestamibi (MIBI) [9]. Two scintigraphic techniques are usually employed, dual-phase and thyroid subtraction [10]. The dual-phase technique includes acquiring early (10 min post-MIBI administration) and late (2 h post-injection) images of the neck and the mediastinum. MIBI uptake by the thyroid gland challenges the identification of an underlying parathyroid adenoma in early images. However, prolonged tracer retention by most abnormal PGs facilitates their detection in late images because MIBI clears more rapidly from the normal thyroid [11].

Nevertheless, quick tracer washout from some PGs is a usual cause of false negative findings [11]. To address this issue, doctors use the thyroid subtraction technique, which involves the administration of a second radioactive tracer (^{123}I or ^{99m}Tc-pertechnetate) to depict the thyroid gland [12]. The thyroid image is digitally removed from early MIBI images. This helps avoid false-positive results caused by MIBI uptake in thyroid nodules. The techniques can be used separately or together. Single-photon emission computerised tomography (SPECT) or SPECT/CT imaging is also employed to enhance sensitivity and obtain precise location information in three-dimensional space [9].

Medical experts examine all the produced images. However, the findings are often not prominent or well visualised. As a result, composing the diagnostic report is a time-consuming procedure. Recent progress in artificial intelligence (AI), and more specifically in the subset entitled deep learning (DL) [13], has demonstrated pioneering methods for classifying and identifying findings of medical importance in medical images of various modalities.

However, due to the absence of human assistance and supervision, DL suffers from the issue of explainability [14–16]. The decisions of the algorithms are neither transparent nor interpretable. The latter catalysed the research for explainable methods that combine the capabilities of deep models with the transparency of more traditional approaches. Recently, post hoc explainability algorithms have been proposed to improve the explainability of deep models [14,17,18]. The most intuitive paradigm is that of the gradient-weighted class activation mapping (Grad-CAM) algorithm [17]. Moreover, the local interpretable model-agnostic explanations (LIME) [19] algorithm has recently gained attention as well.

The present study proposes a DL approach for identifying abnormal PGs in scintigraphic images. First, an innovative modification of the well-known Virtual Geometry Group (VGG19) convolutional neural network (CNN) network [20] is proposed to classify scintigraphic images into normal and abnormal classes. The Grad-CAM++ algorithm [21] is then applied to the trained model to highlight the crucial regions suggested by the model. As a result, the user can understand where the model bases its predictions.

This study is based on previous works by the author team ([22,23]). The latter showed that DL could localise abnormal PGs from parathyroid scintigraphy images by implementing the Grad-CAM algorithm. In [23], a patient-level accuracy of 94.8% in distinguishing between healthy and diseased subjects was observed. However, on a PG-level basis, the model yielded many false-positive findings, resulting in a PG-detection accuracy of 76.5%.

The current study proposes two methods for reducing false-positive findings. Firstly, an innovative modification of ParaNet [23], ParaNet+, allows more local feature extraction by introducing a dense fusion of extracted features by multiple layers. Secondly, the study employs an improvement of the vanilla Grad-CAM algorithm, which yields more precise results in localising abnormal PGs. Finally, the study performs extensive evaluations of ParaNet+ to assess its stability, parameters, and hyperparameters.

2. Materials and Methods

2.1. Research Methodology

The study's focal point is the detection of abnormal PGs using the MIBI-early, MIBI-late, and TcO4 images (thyroid scan) accompanying the scintigraphic outcome. Medical experts perform localisation in the everyday routine, compare the three images, and determine positive findings identified on the MIBI-late image. The study involves seven methodological steps, as presented in Table 1 and Figure 1.

Table 1. Research methodology.

Stage	Information
1	Data pre-processing—the creation of the dataset (years 2010–2020).
2	Develop ParaNet+. Train and validate the model using the human experts' labels as the reference. Inspect the agreement with the experts on a patient-level basis.
3	Employ Grad-CAM++ to localise abnormal PGs.
4	Experts perform the visual assessment and compute the agreement on a PG-level basis.
5	Evaluate the model on the external thyroidectomy dataset.
6	Conduct statistical significance tests and inspect the reproducibility of the experiments by performing multiple repetitions.
7	Compare ParaNet+ with state-of-the-art approaches.

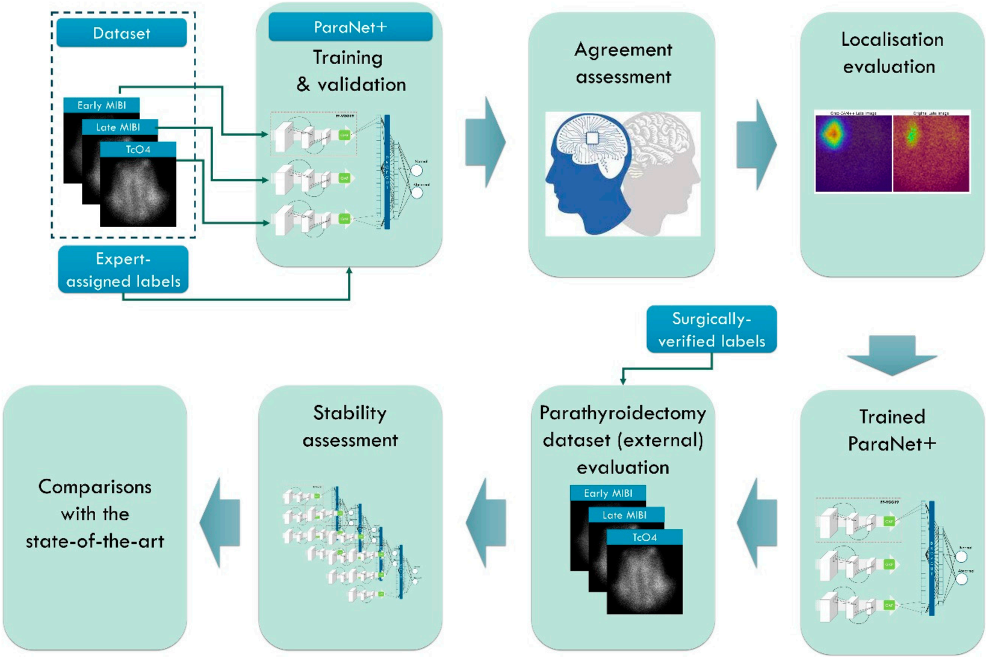

Figure 1. Research methodology. MIBI stands for the radioactive tracer ^{99m}Tc-Sestamibi and TcO4 stands for the ^{99m}Tc-pertechnetate for thyroid delineation.

The first stage involves the necessary data pre-processing to create the study's dataset. This stage includes region of interest (ROI) reduction and image normalisation. In the second stage, ParaNet+ is deployed to identify normal and abnormal scintigraphic images. Per-patient agreement with the experts is recorded and discussed. The Grad-CAM++ algorithm visualises the essential CNN-suggested features on the MIBI-late image. Medical experts perform a visual assessment during the fourth stage.

The fifth stage involves external testing using patient data from the year 2022. Patient data included 100 surgically verified cases. The assessment is on a patient-level and PG-level basis. The sixth and seventh stages include stability tests and comparisons with state-of-the-art approaches and previous works by the author group.

2.2. Dataset and Imaging Techniques

Patras is the third-biggest city in Greece, accounting for approximately 200,000 inhabitants. Medical services of our institution expand to a broader geographical area, including a population of almost 1,000,000. The retrospective study involves 648 confirmed participants with pHPT, 535 females and 113 males, aged 58.1 ± 12.5 years, who underwent parathyroid scintigraphy in the Department of Nuclear Medicine of the University Hospital of Patras, Greece. The period of the study ranges from January 2010 to December 2020. The planar dual-phase technique with ^{99m}Tc-Sestamibi was used in all participants.

Moreover, whenever judged necessary by medical experts, the thyroid subtraction technique was also used. The latter refers to 529 cases. ^{99m}Tc-pertechnetate (TcO4) for thyroid delineation was administered either after the conclusion of the dual-phase study or on another day.

The prospective external data set comprised 100 consecutive patients (87 females and 13 males), who also underwent scintigraphy in our department. This set differs from the retrospective group because the patients were subsequently subjected to parathyroidectomy in 2021–2022 in our or other surgical units. The results were available for the evaluation of the model on this external dataset. Table 2 summarises the dataset's characteristics.

Table 2. Characteristics of the study's datasets.

Information	Value
Training—validation dataset	
Hospital	University General Hospital of Patras, Greece
Department	Department of Nuclear Medicine
Date	2010–2020
Participants	648
Subjects labelled by human experts	648
Negative result	198
Positive result	450
Abnormal PGs identified by human experts	504
External dataset	
Hospital	University General Hospital of Patras, Greece
Department	Department of Nuclear Medicine
Date	2021–2022
Participants	100
Surgically verified subjects	100
Abnormal PGs	118

We used a pinhole collimator for planar imaging and placed it 10 cm over the neck. A SPECT/CT imaging session focused on the neck and the mediastinum using a high-sensitivity parallel-hole collimator took place approximately 30 min post-tracer injection. However, only planar imaging data have been included in the present study. Regarding the acquisition device technology, planar and SPECT/CT imaging were performed by the Hawkeye-4 system (General Electric Healthcare, Chicago, IL, USA).

The planar scintigraphic studies were evaluated retrospectively by three experienced Nuclear Medicine physicians (D.J.A, more than 20 years; N.D.P, more than 8 years; and T.I.S., more than 15 years). In a few ambiguous cases, the final decision was reached by consensus.

2.3. Data Pre-Processing and Augmentation

The data pre-processing and dataset creation is illustrated in Figure 2. The initial scintigraphic images contain artefacts embedded by the image acquisition technology. The reduction of focal regions is, therefore, mandatory. Moreover, the actual ROI is located at the centre of the image, surrounded by a black background, where no information of medical importance exists. Therefore, each image is cropped to 200 × 200 (width, height).

Inspection of the images revealed that this area is adequate to fit the complete thyroid without leaving out any area of clinical interest. Normalisation refers to transforming the pixel values, which range from 0 to 255, to the space [0, 1], which is preferable for machine learning (ML) and DL networks and speeds up the training process.

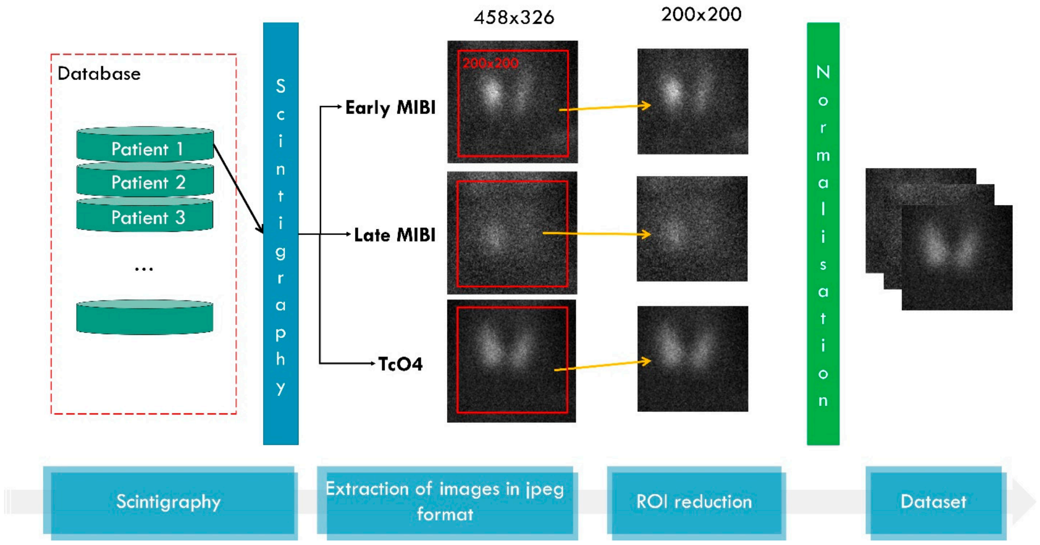

Figure 2. Data processing. MIBI stands for the radioactive tracer ^{99m}Tc-Sestamibi and TcO4 stands for the 99mTc-pertechnetate for thyroid delineation.

During network training, data augmentation is performed to increase the training images. Data augmentation is essential and benefits the network because it provides multiple versions of the same findings [24]. In the augmented images, slight geometrical transformations take place. These transformations do not distort the actual finding in the image but introduce variation and help the network focus on the region of interest. Therefore, the network learns to ignore redundant features, such as the orientation of a PG or other irrelevant spatial characteristics. The study proposes slight data augmentations, including rotations by ±5°, width and height shifts by ±10 pixels, and horizontal flips. The image's distortion and the generation of unrealistic PGs are avoided using these geometric transformations.

2.4. Deep Learning Model

The core of our idea is that a classification DL (ParaNet+) model can learn to identify abnormal PGs in the images and classify new images from the test set or an external set as normal or abnormal. To achieve this, the model should operate with three images as input and one output (the class). However, DL models do not provide explanations for their decisions. Hence, we used the Grad-CAM++ algorithm, which tracks back the decision of the model regarding the class of the image and tries to uncover the important areas of the image, where meaningful image features were found. In this way, we can understand if the model based the decision on the abnormal class based on an actual abnormal PG in the image or not.

2.4.1. ParaNet+

ParaNet+ is a multi-input framework consisting of three CNN components that process the inputs independently. The ParaNet+ version is an improved topology of ParaNet [23], which utilises the Feature-Fusion VGG19 (FF-VGG19) network [25] instead of the baseline VGG19.

Feature-Fusion VGG19 (FF-VGG19) is the modified version of the well-documented and successful VGG19 network. VGG19 has been employed several times for related tasks. The conception of FF-VGG19 is analytically presented in [20]. It is a uniform network of

19 layers, which include convolutional and max-pooling operations. The input image is incrementally filtered and hierarchically reduced in size. As a result, thousands of image features are extracted.

FF-VGG19 establishes a direct connection between each convolutional group and the classification layers. Eventually, the classification is based on more features than the baseline VGG19. The utilisation of global pooling layers enables this conception. Figure 3 illustrates the modified VGG19 component, and Figure 4 shows the entire ParaNet+ topology. The parameters and hyperparameters are presented in Table 3.

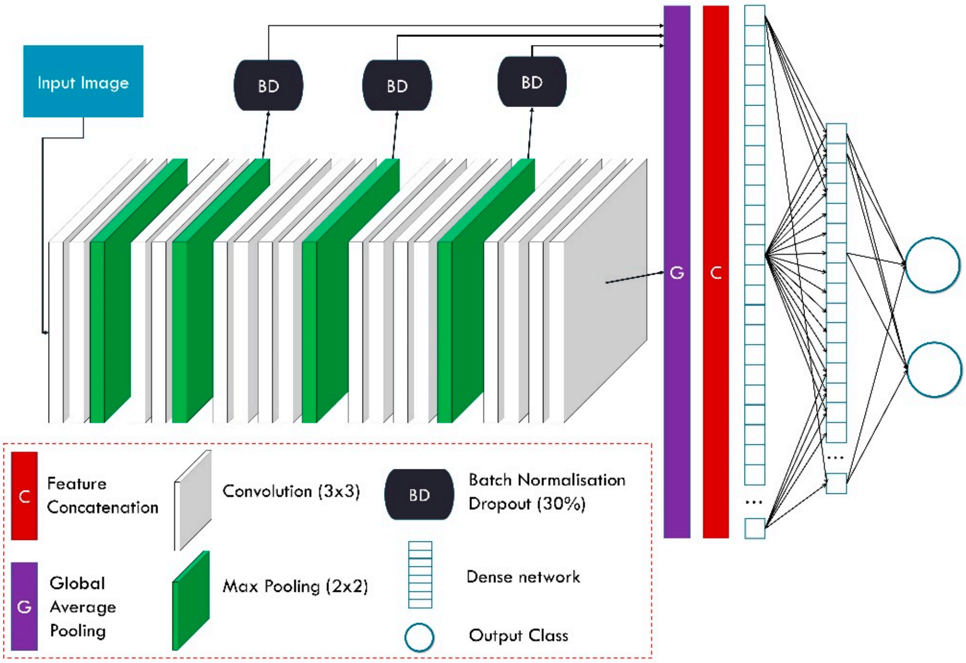

Figure 3. Feature-Fusion VGG19 component of ParaNet+.

Table 3. Parameters and hyperparameters of ParaNet+.

Parameter/Hyper-Parameter	Value
Activation function	Rectified linear unit
Final layer activation	Softmax
Loss function	Categorical cross-entropy
Batch normalisation	Yes
Dropout	40%
Epochs	500
Early stopping	At 98% validation accuracy
Batch size	32
Input image size	200 × 200 × 1 (height, width, channels)
Trainable parameters	3,079,628
Feature-fusion method	Concatenation
Optimisation	Adam [26]

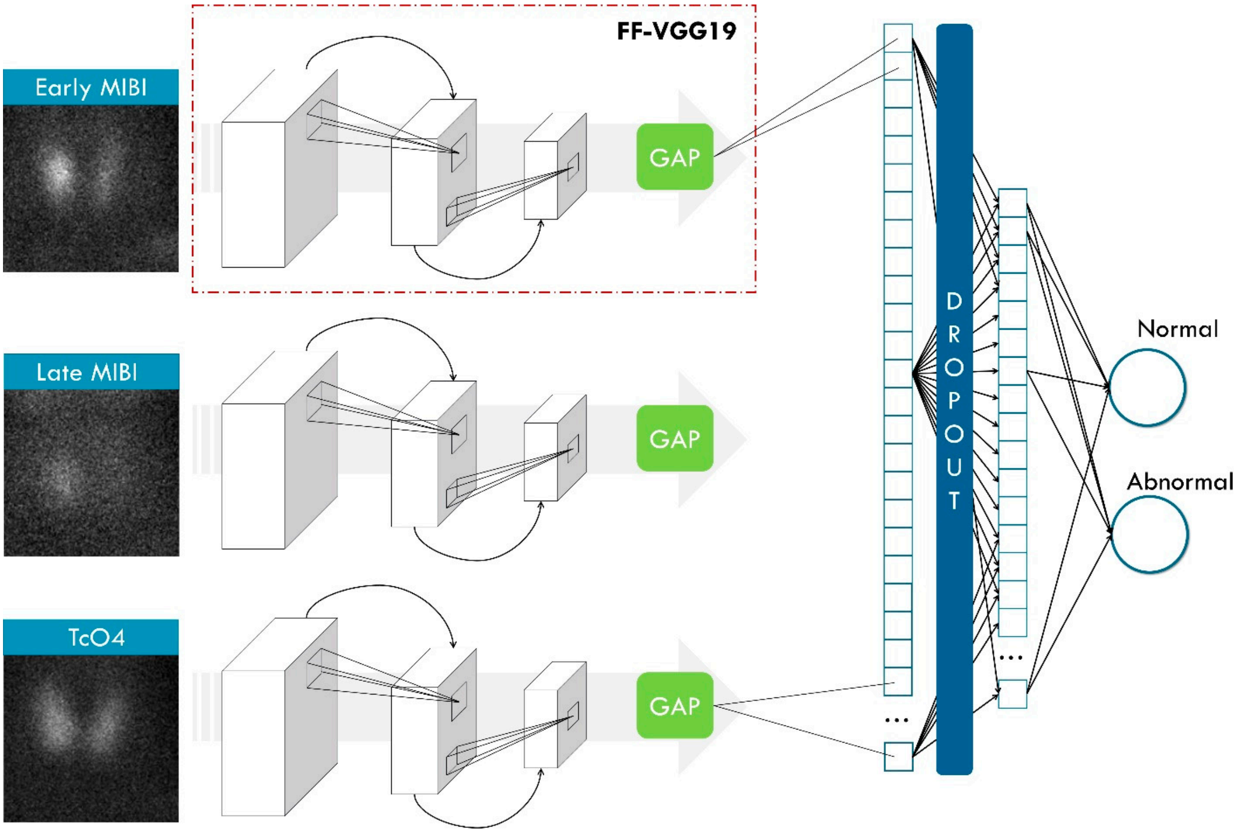

Figure 4. ParaNet+ topology. FF-VGG19 stands for the Feature-Fusion VGG19 network, which is an essential component of ParaNet+. GAP stands for global average pooling. MIBI stands for the radioactive tracer ^{99m}Tc-Sestamibi.

2.4.2. Gradient-Weighted Class Activation Mapping (Grad-CAM++)

Gradient-based explanation or interpretation constitutes one of the most efficient and lightweight methods for explaining deep networks' decisions. These methods use gradients to understand how a slight change in the input would affect the output. For example, in CNNs for image classification, local areas of the image are inspected to determine their significance based on the latter conception.

The Grad-CAM algorithm is a technique used for visualising the regions of an image that contribute most to the output of a neural network. The Grad-CAM algorithm generates a class activation map, highlighting the regions of an image that are most relevant for predicting a particular class. First, the network's output gradients for the final convolutional layer are calculated to create this map. These gradients are then global-average-pooled to obtain the importance weights for each feature map in the last convolutional layer. Finally, the feature maps are weighted by their corresponding importance weights and summed to obtain the class activation map.

The Grad-CAM++ algorithm [21] is an improved version of Grad-CAM [17] that improves the localisation of the object of interest in the image. It achieves this by incorporating the second-order gradients of the output for the final convolutional layer, which is done using a Taylor series approximation to estimate the second-order gradients and is then used to refine the importance weights obtained in the first step of the Grad-CAM algorithm.

2.5. Experiment Setup and Evaluation Metrics

The experiments are performed on an NVidia RTX 380 GPU computer with 64 GB of RAM and an Intel Core i9 processor. In addition, the Tensorflow-gpu library has been employed in a Python 3.8 environment.

We considered the 10-fold cross-validation methodology for evaluating the agreement between the experts and the model, using images labelled by the human experts. The performance is assessed based on the total number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These scores are used to calculate the accuracy, sensitivity, specificity, and positive and negative predictive values (PPV and NPV, respectively). Moreover, the F1 and AUC scores are reported. The F1 score is a commonly used metric in machine learning to evaluate the performance of a classification model. It measures the model's accuracy, considering both precision and recall. Precision is the ratio of true positive predictions to the total number of positive predictions. At the same time, recall is the ratio of true positive predictions to the total number of actual positive instances. The F1 score is the harmonic mean of precision and recall.

3. Results

3.1. Agreement with the Experts on a Patient-Level and PG-Level Basis in the 2010–2020 Group

According to experts' diagnosis, the 2010–2020 dataset consisted of 198 subjects with no abnormal findings and 450 with at least one abnormal finding. The experts identified 504 abnormal PGs in the 450 abnormal scintigraphic scans. These findings were the ground truth for estimating the model's performance. Table 4 presents the results.

Table 4. Agreement with the experts on a patient-level and PG-level basis using the three-image input in the retrospective group of 648 patients.

Agreement	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**	**F1**
Patient-level	0.9861	0.9889	0.9798	0.9911	0.9749	0.9900
PG level	0.8831	0.9167	0.8125	0.9112	0.8228	0.9139

ACC: accuracy; SENS: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value; F1: F1 score.

On a patient-level basis, the model attains an accuracy of 0.9861, a sensitivity of 0. 9889, and a specificity of 0.9798. The PPV and NPV values are 0.9911 and 0.9749, respectively. The high F1 score (0.99) indicates that the model is not biased towards the majority class despite the class imbalance issue. The number of false positives was four, and the number of false negatives was five on a patient-level basis. The reader should note that the reported metrics are computed based on human-reader labelling. Therefore, the current experiment measures the agreement between the model and the experts' verdict.

There is a decline in accuracy when inspecting the PG-level performance of the model. The latter inspection was performed with the aid of the Grad-CAM++ algorithm. Examples of the outputs are shown in Figures 5 and 6. Specifically, the model identified 462 of the 504 PGs. Eventually, 195 PGs were correctly identified as normal (true negatives). Subsequently, the model achieved an accuracy of 0.8831 (Table 4), 0.9167 sensitivity, and 0.8125 specificity. The positive predictive value was 0.9112, and the negative predictive value was 0.8228. The false-positive findings were 45, and the false-negative findings (abnormal PGs that the model ignored) were 42.

3.2. Grad-CAM++

Figure 5 shows cases with true positive examples containing prominent findings. Two images accompany each case. The right is the original late MIBI image, and the left is the fused feature map. The red colour is added by Grad-CAM++ and highlights the features that hold an essential role in the predicted class. Green areas correspond to medium importance, and the blue regions to minor importance. The model correctly localises the abnormal PGs of these cases. However, considering red and green areas as important, the suggested area union is larger than the actual finding. This may confuse a human reader when trying to understand what the model considers a finding. It is also observed that the factual finding is not always located where the red areas appear, but very close by, which may be an inherent drawback of Grad-CAM++ and would require further investigation. Heat maps containing no red areas imply ambiguous decisions and can be considered suspicious for false-positive findings. Figure 6 illustrates more challenging true-positive cases.

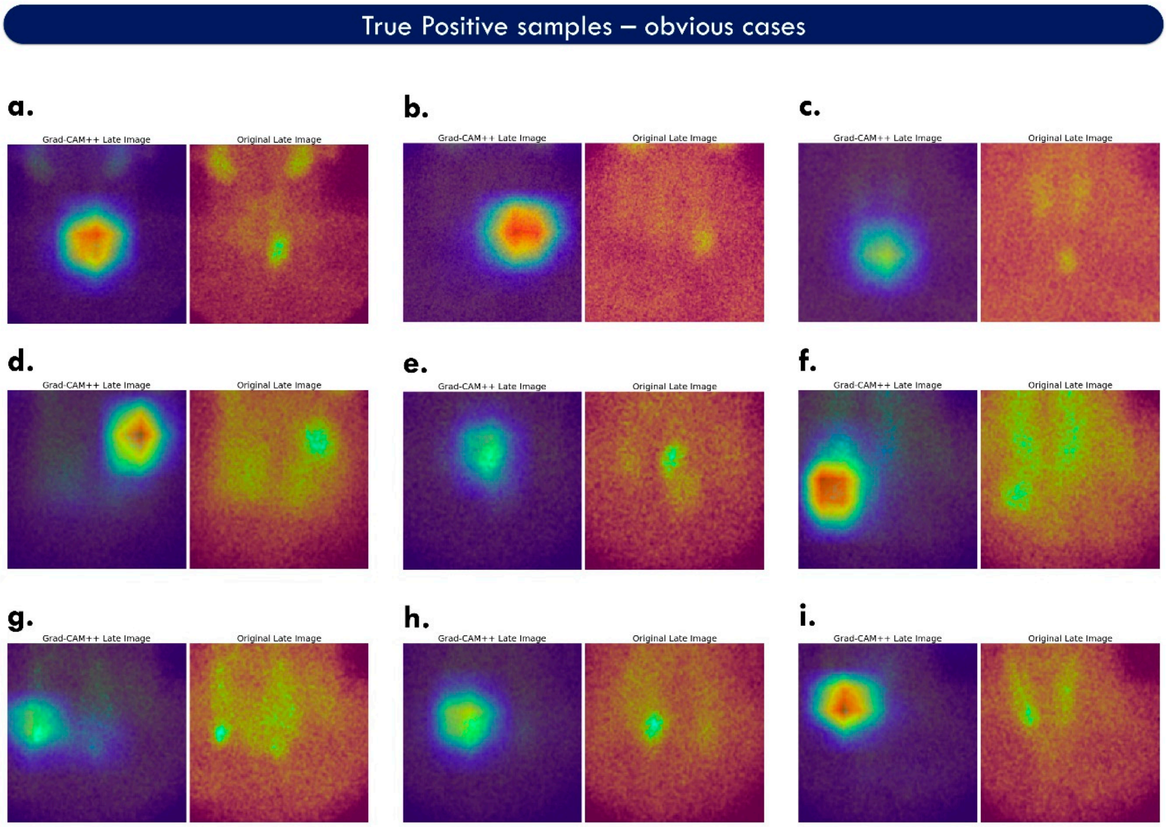

Figure 5. Grad-CAM++ areas of interest corresponding to true-positive samples. Each case (**a**–**i**) is accompanied by the original image (on the right) and the associated heat map (on the left). This figure illustrates obvious cases, wherein the abnormal PG is visible in the original image.

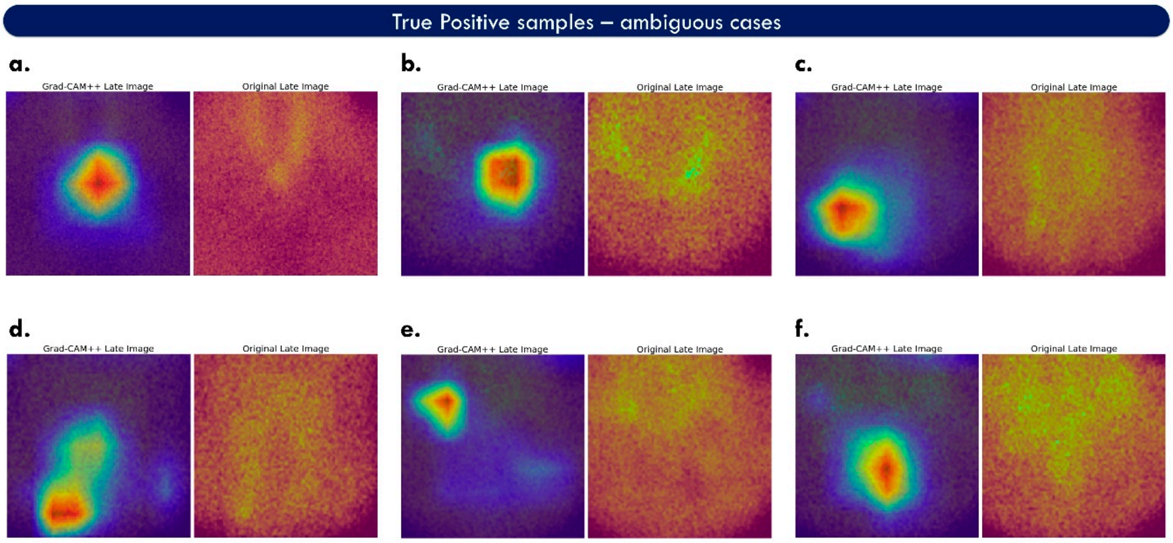

Figure 6. Grad-CAM++ areas of interest corresponding to true-positive samples. Each case (**a**–**f**) is accompanied by the original image (on the right) and the associated heat map (on the left). This figure illustrates ambiguous cases, wherein the abnormal PG is not easily spotted in the original image.

ParaNet+ performs well on inconclusive images, such as case f of Figure 6. However, the model is also confused in cases such as case d of the same figure, where the suggested areas are ill-defined. The upper part of the heat map of case d is considered a false positive. In contrast, the lower area is a true-positive example.

3.3. Ablation Study

External data were used to inspect how ParaNet+ performs on surgically verified samples. The external dataset consists of 100 participants. All participants had at least one abnormal finding on surgery. Therefore, there are no subjects populating the normal class. Consequently, only sensitivity and false-positive rates are assessed in this group. On surgery, 118 abnormal PGs were identified and excised, 86 corresponding to solitary adenomas (in 86 patients), and 32 to multiglandular disease (in 14 patients).

The experts identified 93/100 abnormal cases on the patient level and 99/118 PGs. Eleven findings of the human reader were false positives on the PG level.

ParaNet++ identified 99/100 cases on a patient basis and 103/118 on a PG basis. There were 13 false-positive PG findings. On both patient and PG bases, the model yielded somewhat higher sensitivity than the human reader (99.0% vs. 93.0% and 87.3% vs. 83.9%, respectively) with comparable FPRs (11.2% vs. 10.0%, respectively). The results are also summarised in Table 5.

Table 5. Experts' and ParaNet++'s diagnostic results for 100 patients subjected to surgical parathyroidectomy. PTs: patients, PGs: parathyroid glands; TOT: total; S: solitary adenoma; MGD: multiglandular disease.

	Experts				Model					
	With reference to surgical findings of the external dataset				With reference to experts' diagnosis on the external dataset		With reference to surgical findings of the external dataset			
			PGs						PGs	
	PTs	Total	SA	MGD	PTs	TOT PGs	PTs	Total	SA	MGD
Correct findings	93/100	99/118	78/86	21/32	93/94	98/106	99/100	103/118	84/86	19/32
Percentage	93	83.9	90.7	65.6	98.9	92.5	99	87.3	97.7	59.4

PTs: patients; SA: solitary adenoma; MGD: multiglandular disease.

The model successfully detected solitary adenomas (97.7% sensitivity). However, it performed sub-optimally in cases with multiglandular disease (59.4% sensitivity). It is worth noticing that the human readers performed better than the model for the latter case, yielding a sensitivity of 65.6%. On the other hand, some false positives from the model, judged with reference to experts' diagnoses, proved to be true positives based on the surgical findings.

3.4. Parametrisation

3.4.1. Freedom of Learning

The FF-VGG19 components of ParaNet+ can be employed under three scenarios, as follows: (i) pre-trained (with their weights determined by their initial training in the ImageNet challenge database [27]); (ii) trained from scratch, wherein all the learning layers are trainable; (iii) fine-tuned, wherein some of the learning layers retain the parameters of their initial training, and some other are made trainable. The exact number of trainable and untrainable layers can be defined by extensive experiments.

Freedom of learning plays a significant role in transfer learning [28]. Training a model entirely from scratch is ideal when the available datasets are adequate to learn from and would require large-scale sets because deep CNNs involve millions of trainable parameters. On the other hand, training a deep CNN from scratch may result in underfitting, as observed in Table 6, wherein ParaNet+ obtains 0.6790 accuracy.

Table 6. Classification results of ParaNet+ under multiple learning scenarios.

Type	ACC	SEN	SPE	PPV	NPV	F1
Transfer learning (zero Freedom)	0.6620	0.6711	0.6414	0.8097	0.4618	0.7339
1-layer Freedom	0.9306	0.9711	0.8384	0.9318	0.9274	0.9510
2-layer Freedom	0.9861	0.9889	0.9798	0.9911	0.9749	0.9900
4-layer Freedom	0.8843	0.8556	0.9495	0.9747	0.7431	0.9112
6-layer Freedom	0.7623	0.7578	0.7727	0.8834	0.5840	0.8158
8-layer Freedom	0.7423	0.7822	0.6515	0.8361	0.5683	0.8083
Training from scratch (19-layer Freedom)	0.6790	0.6844	0.6667	0.8235	0.4818	0.7476

ACC: accuracy; SENS: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value; F1: F1 score.

Borrowing the weights obtained by training on the ImageNet dataset was ineffective (0.629 accuracy). The latter performance was expected since the initial training had been performed on irrelevant images. Progressively allowing some training improved the performance of the model. It is observed that two-layer freedom yields the best accuracy of 0.9861.

3.4.2. Optimisation

This experiment determines the suitable optimiser of ParaNet+. The right optimiser can affect the learning capabilities of the model because it is responsible for how the weights are updated to reduce the losses. Additionally, different optimisers may increase or decrease the training times and the training stability. Table 7 presents the performance metrics of ParaNet+ under other optimisation methods.

Table 7. Classification results of ParaNet+ under different optimisation settings.

Type	ACC	SEN	SPE	PPV	NPV	F1
Adam	0.9861	0.9889	0.9798	0.9911	0.9749	0.9900
Adagrad	0.6960	1.0000	0.0051	0.6955	1.0000	0.8204
AdaDelta	0.6975	0.9933	0.0253	0.6984	0.6250	0.8202
Nadam	0.8148	0.9800	0.4394	0.7989	0.9063	0.8802
RMSProp	0.9074	0.9356	0.8434	0.9314	0.8520	0.9335

ACC: accuracy; SENS: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value; F1: F1 score.

Adam and RMSprop stand out, obtaining 0.9861 and 0.9074 accuracy, respectively. In addition, Adam speeds up training compared to the rest of the algorithms, achieving complete dataset training in 869 s.

3.4.3. Data Augmentation

Data augmentation holds a notable role in CNN training for classification tasks. It provides CNNs with augmented data to increase their performance. CNNs benefit from data augmentation because they can learn to ignore geometrical and positional features that introduce variations among objects of the same class. In addition, CNNs learn to ignore noisy representations and focus on discovering significant features.

Strong augmenting in medical imaging is avoided because it generates unrealistic representations, which may cause underfitting. In the experiment, width shift, height shift, rotation, shear, and horizontal flip are applied incrementally. Table 8 summarises the performance of ParaNet+ under different data augmentation settings.

Table 8. Classification results of ParaNet+ under different data augmentation methods.

Type	ACC	SEN	SPE	PPV	NPV	F1
WS	0.8519	0.8600	0.8333	0.9214	0.7237	0.8897
HS	0.8827	0.8911	0.8636	0.9369	0.7773	0.9134
WS + HS	0.9059	0.9156	0.8838	0.9471	0.8216	0.9311
WS + HS + R	0.9136	0.9222	0.8939	0.9518	0.8349	0.9368
WS + HS + R + SH	0.9707	0.9822	0.9444	0.9757	0.9590	0.9790
WS + HS + R + HF	0.9861	0.9889	0.9798	0.9911	0.9749	0.9900

ACC: accuracy; SENS: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value; F1: F1 score; WS: width shift; HS: height shift; R: rotation; SH: sheer; HF: horizontal flip.

It is observed that the best accuracy is obtained when combining width and height shifts, rotations, and horizontal flips.

3.4.4. Batch Size, Early Stopping

This experiment evaluates different batch sizes and early training-stopping settings. Small batch sizes generally go with small learning rates [29]. On large-scale, the suggested batch sizes range from 8 to 64 or more [29].

Early stopping is a method to avoid overfitting and unnecessary training. During each training epoch, training will stop if a specific criterion is met, and the model will be returned to its most effective condition until this event is triggered. The criterion in the particular setup is the validation accuracy computed after each epoch. Table 9 presents the performance of ParaNet+ under various batch sizes and early stopping settings.

Table 9. Classification results of ParaNet+ under different batch sizes and early stopping settings. The abbreviation for Ab. is "Aborted due to computational resources limit".

Type	ACC	SEN	SPE	PPV	NPV	F1
Batch Sizes						
4	0.9660	0.9711	0.9545	0.9798	0.9356	0.9754
8	0.9506	0.9511	0.9495	0.9772	0.8952	0.9640
16	0.9861	0.9889	0.9798	0.9911	0.9749	0.9900
32	Ab.	Ab.	Ab.	Ab.	Ab.	Ab.
64	Ab.	Ab.	Ab.	Ab.	Ab.	Ab.
Early Stopping at validation accuracy						
0.99	0.9799	0.9867	0.9646	0.9845	0.9695	0.9856
0.98	0.9861	0.9889	0.9798	0.9911	0.9749	0.9900
0.97	0.9645	0.9711	0.9495	0.9776	0.9353	0.9744
0.96	0.9830	0.9867	0.9747	0.9889	0.9698	0.9878
0.95	0.9645	0.9644	0.9646	0.9841	0.9227	0.9742

ACC: accuracy; SENS: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value; F1: F1 score.

The optimal batch size is observed to be 16. However, further increasing the batch size was not allowed due to computational resource limitations. The model exhibited its best accuracy when early stopping was triggered at a 0.98 validation accuracy.

3.5. Reproducibility

The experiment was conducted 40 times to validate the stability of the model in reproducing the results. A T-test was performed to investigate the statistical significance of the discrepancy (Table 10). The experiment showed that, at the 0.05 level, the population mean (0.9859) is not significantly different from the test mean (0.9861).

Table 10. Statistical significance test for the evaluation of the model's stability.

Type	Mean	Standard Deviation
Accuracy	0.9859	0.0043
t-value	−1.25969	
Outcome	At the 0.05 level, the population mean (0.9859) is not significantly different from the test mean (0.9861)	

3.6. Train, Test, and Visualisation Times

The training, evaluation, and visualisation time is an essential aspect of each DL model intended for everyday practice in real environments. Ideally, the time taken for a trained model to predict the class and the PGs of a scintigraphic image has to be negligible, enabling such models to be deployed in real time. Table 11 showcases the time it took the model to complete a full training, a 10-fold cross-validation procedure, a single prediction, and the generation of the accompanying visualisation.

Table 11. Operation times.

Operation	Time (s)
Complete training (no cross-validation)	869
Complete 10-fold cross-validation	8759
Prediction of class (single image)	0.4
Visualisation (Grad-CAM++) (single image)	0.5

Though training times depend heavily on the computational infrastructure, it is observed that an ordinary personal computer can complete an entire dataset training in less than 900 s. In addition, the trained model took less than a second to process a new image and produce the prediction class and the Grad-CAM++ output. The latter times highlight that the model is suited to hospital workstations or personal computers.

4. Discussion

Previous studies by the author group established DL methods for identifying and localising abnormal PG in scintigraphic images [22,23]. The conception of these methods was based on recent literature reviews [30]. However, despite the remarkable performance metrics in distinguishing between normal and abnormal patient cases, previous approaches yielded many false-positive PGs, which were revealed by investigating the sensitivity maps produced by Grad-CAM.

The present study improves the algorithm for false-positive reduction and more precise localisation. Based on the conception that an abnormal PG can be identified using low-level features, usually extracted by the first convolutional layers of a CNN, the study proposes modifications in the baseline ParaNet to allow for more local feature extraction. Local, low-level, and high-level features are now extracted in a non-sequential manner and are fused. The study employed a post hoc explainability method (Grad-CAM++) to localise the DL-suggested abnormal PGs. Grad-CAM++ produced the sensitivity maps, highlighting the areas of the image where the model grounded its predictions. As a result, ParaNet++ showed better classification accuracy, obtaining 0.9868 on a patient basis and when using the expert's diagnostic yield as the reference. Scrutiny on a PG-level basis revealed an 88.31% agreement between the model and the experts. Hence, the latter classification accuracy measured the agreement between DL and the physicians. In addition, there was a significant reduction in false positives compared to our previous work [22].

The proposed network performs well on external testing involving surgically verified samples. Though the network's training and initial validation had been performed using human interpretation as the reference, its generalisation capability is adequate to classify 99% of the parathyroidectomy-labelled images correctly, which is considered among the main strengths of the framework. On a PG-level basis, the model showed 87.3% sensitivity in detecting abnormal PGs versus 83.9% by the experts.

Direct comparisons with related works may be inconclusive due to variations in the test data, the image acquisition devices, and the data sizes. Nevertheless, the available comparisons populate Table 12.

Table 12. Comparisons with the recent literature. ACC: accuracy, SEN: sensitivity.

Study	Reference	Evaluation Data Size	Patient-Level Performance	PG-Level Performance
Apostolopoulos et al.	[22]	632	96.56% accuracy	-
Apostolopoulos et al.	[22]	50 (operated)	92% accuracy	-
Apostolopoulos et al.	[23]	632	94.8% accuracy	76.5% accuracy
Apostolopoulos et al.	[23]	472 (operated)	84.51% accuracy	59.43% accuracy
Yoshida et al.	[31]	44	-	83% sensitivity
Imbus et al.	[32]	2010	94.1% accuracy in distinguishing between single-gland and multi-gland adenoma	
Avci et al.	[33]	47	-	95.7% precision, 90.5% recall
Avci et al.	[34]	90	-	89% precision and recall
This study		648	98.61% accuracy	88.31%
This study		100 (external-operated)	99% sensitivity	87.3% sensitivity

The superiority of ParaNet+ compared to the baseline ParaNet is verified. ParaNet achieved an accuracy of 0.948 on a patient-level basis, while ParaNet+ reached 0.98681. The research's results are also consistent with the literature [22,32–34].

Limitations

We must underline that the model showed suboptimal performance in patients with multiple abnormal PGs in both the retrospective and the prospective data set. This shortcoming should be addressed with better tuning of the Grad-CAM++ algorithm and more data containing multiple findings. The false-positive findings of the model (FPR = 11.2%) were comparable to those of the human reader (10.0%) and were mainly attributed to thyroid nodules. This level of FPR is instead an inherent drawback of MIBI scintigraphy, particularly in geographic areas with a high prevalence of thyroid nodularity, such as Greece, than a flaw of the model.

Future studies must involve further external testing using more surgical and histological verification cases. However, to achieve better results in parathyroidectomy cases, the training datasets must be populated with surgically verified samples. The absence of such cases in the training set of the current study is a limitation.

In addition, the network's robustness to image acquisition device variation is currently questionable. Clinical and demographic information integration may improve precision and offer a more holistic approach. However, the actual influence of clinical and demographic factors is ambiguous. These factors are not expected to improve the localisation precision. However, they can improve the per-patient classification, distinguishing between normal and abnormal incidents.

Finally, the scarcity of related works available for comparison is an unavoidable limitation of the current study. The limited number of studies focusing on the same research area reduces the opportunity for direct comparison of findings and validation of results. Furthermore, the existing studies often employ different datasets, lacking a global standard dataset, which makes it challenging to establish a fair and consistent benchmark. Additionally, variations in image acquisition techniques and devices utilised in these studies introduce further complexity in drawing accurate comparisons. Consequently, the absence of a standardised dataset and consistent imaging devices hinders the ability to perform a comprehensive and unbiased evaluation across different methodologies. Therefore, caution should be exercised when interpreting the results of this study in the broader context of the field, and further research with a more extensive and standardised dataset is warranted to facilitate a more fair and reliable comparison among different approaches.

5. Conclusions

The study proposed a DL approach for identifying healthy and diseased subjects in parathyroid scintigraphy.

The results draw some conclusions, as follows: (i) the model showed excellent agreement with the experts (98.61%) when trained to predict the human-assigned labels per patient; (ii) the model adequately identified the abnormal PGs (88.31%), with significantly reduced false positives compared to previous studies; (iii) the model retained its performance per patient when asked to predict the labels of parathyroidectomy-verified samples of an external dataset (99%); (iv) the model shows slight performance decrease in identifying the abnormal PGs in the parathyroidectomy images (87.3%); (v) the model exhibited slightly better performance compared to the human readers concerning the external parathyroidectomy-verified dataset (99% model—93% human on a patient-level basis and 87.3% model—83.9% human on the PG-level); (vi) the model yielded suboptimal results in detecting multiglandular disease (multiple abnormal PGs—59.4% accuracy in the external test set).

Author Contributions: Conceptualisation, D.J.A. and I.D.A.; Data curation, N.D.P.; Formal analysis, D.J.A. and G.S.P.; Investigation, D.J.A. and T.S.; Methodology, I.D.A. and T.S.; Resources, D.J.A. and N.D.P.; Software, I.D.A.; Supervision, G.S.P.; Validation, D.J.A., N.D.P., and T.S.; Visualisation, N.D.P.; Writing—original draft, I.D.A.; Writing—review and editing, D.J.A., I.D.A., T.S. and G.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Ethical reasons prohibit the public availability of the dataset. However, it can be confidentially communicated to the reviewers and the Journal's Editor if necessary.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Walker, M.D.; Silverberg, S.J. Primary Hyperparathyroidism. *Nat. Rev. Endocrinol.* **2018**, *14*, 115–125. [CrossRef]
2. Bilezikian, J.P.; Cusano, N.E.; Khan, A.A.; Liu, J.-M.; Marcocci, C.; Bandeira, F. Primary Hyperparathyroidism. *Nat. Rev. Dis. Primer* **2016**, *2*, 16033. [CrossRef] [PubMed]
3. Fisher, S.B.; Perrier, N.D. Primary Hyperparathyroidism and Hypertension. *Gland Surg.* **2020**, *9*, 142–149. [CrossRef]
4. Ali, D.S.; Dandurand, K.; Khan, A.A. Primary Hyperparathyroidism in Pregnancy: Literature Review of the Diagnosis and Management. *J. Clin. Med.* **2021**, *10*, 2956. [CrossRef] [PubMed]
5. Argirò, R.; Diacinti, D.; Sacconi, B.; Iannarelli, A.; Diacinti, D.; Cipriani, C.; Pisani, D.; Romagnoli, E.; Biffoni, M.; Di Gioia, C.; et al. Diagnostic Accuracy of 3T Magnetic Resonance Imaging in the Preoperative Localisation of Parathyroid Adenomas: Comparison with Ultrasound and 99mTc-Sestamibi Scans. *Eur. Radiol.* **2018**, *28*, 4900–4908. [CrossRef] [PubMed]
6. De Pasquale, L.; Lori, E.; Bulfamante, A.M.; Felisati, G.; Castellani, L.; Saibene, A.M. Evaluation of Wisconsin and CaPTHUS Indices Usefulness for Predicting Monoglandular and Multiglandular Disease in Patients with Primary Hyperparathyroidism through the Analysis of a Single-Center Experience. *Int. J. Endocrinol.* **2021**, *2021*, 1–8. [CrossRef]
7. Khafif, A.; Masalha, M.; Landsberg, R.; Domachevsky, L.; Bernstine, H.; Groshar, D.; Azoulay, O.; Lockman, Y. The Role of F18-Fluorocholine Positron Emission Tomography/Magnetic Resonance Imaging in Localizing Parathyroid Adenomas. *Eur. Arch. Otorhinolaryngol.* **2019**, *276*, 1509–1516. [CrossRef] [PubMed]
8. Yeh, R.; Tay, Y.-K.D.; Tabacco, G.; Dercle, L.; Kuo, J.H.; Bandeira, L.; McManus, C.; Leung, D.K.; Lee, J.A.; Bilezikian, J.P. Diagnostic Performance of 4D CT and Sestamibi SPECT/CT in Localizing Parathyroid Adenomas in Primary Hyperparathyroidism. *Radiology* **2019**, *291*, 469–476. [CrossRef]
9. Petranović Ovčariček, P.; Giovanella, L.; Carrió Gasset, I.; Hindié, E.; Huellner, M.W.; Luster, M.; Piccardo, A.; Weber, T.; Talbot, J.-N.; Verburg, F.A. The EANM Practice Guidelines for Parathyroid Imaging. *Eur. J. Nucl. Med. Mol. Imaging* **2021**, *48*, 2801–2822. [CrossRef]
10. Iwen, K.A.; Kußmann, J.; Fendrich, V.; Lindner, K.; Zahn, A. Accuracy of Parathyroid Adenoma Localization by Preoperative Ultrasound and Sestamibi in 1089 Patients with Primary Hyperparathyroidism. *World J. Surg.* **2022**, *46*, 2197–2205. [CrossRef]

11. Assante, R.; Zampella, E.; Nicolai, E.; Acampa, W.; Vergara, E.; Nappi, C.; Gaudieri, V.; Fiumara, G.; Klain, M.; Petretta, M.; et al. Incremental Value of Sestamibi SPECT/CT Over Dual-Phase Planar Scintigraphy in Patients With Primary Hyperparathyroidism and Inconclusive Ultrasound. *Front. Med.* **2019**, *6*, 164. [CrossRef] [PubMed]
12. Hassler, S.; Ben-Sellem, D.; Hubele, F.; Constantinesco, A.; Goetz, C. Dual-Isotope 99mTc-MIBI/123I Parathyroid Scintigraphy in Primary Hyperparathyroidism: Comparison of Subtraction SPECT/CT and Pinhole Planar Scan. *Clin. Nucl. Med.* **2014**, *39*, 32–36. [CrossRef] [PubMed]
13. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
14. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
15. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Yates, A.N.; Greene, K.; Broniatowski, D.A.; Przybocki, M.A. *Four Principles of Explainable Artificial Intelligence*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2021. [CrossRef]
16. van der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.A.; Viergever, M.A. Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef] [PubMed]
17. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
18. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: Removing Noise by Adding Noise. *arXiv* **2017**, arXiv:1706.03825.
19. Palatnik de Sousa, I.; Maria Bernardes Rebuzzi Vellasco, M.; Costa da Silva, E. Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors* **2019**, *19*, 2969. [CrossRef]
20. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
21. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
22. Apostolopoulos, I.D.; Papathanasiou, N.D.; Apostolopoulos, D.J. A Deep Learning Methodology for the Detection of Abnormal Parathyroid Glands via Scintigraphy with 99mTc-Sestamibi. *Diseases* **2022**, *10*, 56. [CrossRef]
23. Apostolopoulos, D.J.; Apostolopoulos, I.D.; Papathanasiou, N.D.; Spyridonidis, T.; Panayiotakis, G.S. Detection and Localisation of Abnormal Parathyroid Glands: An Explainable Deep Learning Approach. *Algorithms* **2022**, *15*, 455. [CrossRef]
24. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
25. Apostolopoulos, I.D.; Tzani, M.A. Industrial Object and Defect Recognition Utilizing Multilevel Feature Extraction from Industrial Scenes with Deep Learning Approach. *J. Ambient Intell. Humaniz. Comput.* **2022**, *14*, 10263–10276. [CrossRef]
26. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:14126980.
27. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
28. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
29. Kandel, I.; Castelli, M. The Effect of Batch Size on the Generalizability of the Convolutional Neural Networks on a Histopathology Dataset. *ICT Express* **2020**, *6*, 312–315. [CrossRef]
30. Apostolopoulos, I.D.; Papandrianos, N.I.; Papageorgiou, E.I.; Apostolopoulos, D.J. Artificial Intelligence Methods for Identifying and Localizing Abnormal Parathyroid Glands: A Review Study. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 814–826. [CrossRef]
31. Yoshida, A.; Ueda, D.; Higashiyama, S.; Katayama, Y.; Matsumoto, T.; Yamanaga, T.; Miki, Y.; Kawabe, J. Deep Learning-Based Detection of Parathyroid Adenoma by 99mTc-MIBI Scintigraphy in Patients with Primary Hyperparathyroidism. *Ann. Nucl. Med.* **2022**, *36*, 468–478. [CrossRef]
32. Imbus, J.R.; Randle, R.W.; Pitt, S.C.; Sippel, R.S.; Schneider, D.F. Machine Learning to Identify Multigland Disease in Primary Hyperparathyroidism. *J. Surg. Res.* **2017**, *219*, 173–179. [CrossRef]
33. Avci, S.N.; Isiktas, G.; Berber, E. A Visual Deep Learning Model to Localize Parathyroid-Specific Autofluorescence on Near-Infrared Imaging: Localization of Parathyroid Autofluorescence with Deep Learning. *Ann. Surg. Oncol.* **2022**, *29*, 4248–4252. [CrossRef]
34. Avci, S.N.; Isiktas, G.; Ergun, O.; Berber, E. A Visual Deep Learning Model to Predict Abnormal versus Normal Parathyroid Glands Using Intraoperative Autofluorescence Signals. *J. Surg. Oncol.* **2022**, *126*, 263–267. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.