Intelligent Oncology 1 (2025) 299–307

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/

Full Length Article

Deep learning-based localization of preoperative parathyroid glands in secondary hyperparathyroidism patients using dual-phase unenhanced and contrast-enhanced computed tomography data

Fuqiang Li[a,#], Yao Zhang[b,#], Yijing Wen[c], Xiaomei Wang[b,*], Hao Yang[d,*]

[a] Department of Thyroid Surgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou Zhejiang 310003, China
[b] Hangzhou PathoAI Technology Co., Ltd., Hangzhou Zhejiang 311100, China
[c] University of Strathclyde, Glasgow G1 1QN, United Kingdom
[d] Hangzhou Meiaiying Biomedical Technology Co., Ltd., Hangzhou Zhejiang 311100, China

ARTICLE INFO

Keywords:
Preoperative localization
Deep learning
Dual-phase CT
Parathyroid glands
Secondary hyperparathyroidism

ABSTRACT

Introduction: Accurate preoperative localization of parathyroid glands (PGs) is crucial for patients with secondary hyperparathyroidism scheduled for parathyroidectomy. However, despite its importance, localization remains challenging since current preoperative imaging modalities for PG localization vary in sensitivity and accessibility.

Materials and methods: In this study, we developed a deep learning model for PG identification using a dual-phase computed tomography (CT) dataset, including unenhanced CT and contrast-enhanced (CE) CT data, and validated the model's sensitivity in clinical application. A retrospective study was conducted using 94 CT images obtained from 47 patients (one plain CT image and one CE CT image per patient). The data were randomly assigned to a training set (38 cases, 76 CT images) and a test set (9 cases, 18 CT images) based on per-patient splits. A three-dimensional U-Net model was trained using the training set and then validated using the test set. An analysis was conducted to compare the model's and clinicians' sensitivity in detecting PGs based on various imaging modalities. An error analysis and an intermodal imaging complementarity analysis were performed to provide references for subsequent model enhancement and application.

Results: The dual-phase CT model identified PGs with a diagnostic sensitivity of 94.44%. This was significantly higher than the sensitivity achieved by clinicians using ultrasonography (61.11%, $P = 0.0013$) and CT (72.22%, $P = 0.0238$). Additionally, the sensitivity achieved using the dual-phase CT model was comparable to that achieved using technetium-99m-methoxyisobutylisonitrile single-photon emission CT/CT (86.11%, $P = 0.429$). We also found that combining predictions from this model with other imaging modalities further improved the PG detection rates.

Conclusions: The study findings suggest that using a deep learning model with plain and CE CT data could improve PG identification prior to thyroidectomy or parathyroidectomy.

1. Introduction

Parathyroid carcinoma is a rare, low-grade but locally aggressive malignancy whose prognosis hinges on complete en-bloc resection at first surgery. When entirely removed, 5-year disease-specific survival reaches 78%–86%; if residual disease remains, approximately 30% experience local recurrence within 3 years and 25% develop distant metastases (lung, bone, etc), dropping 10-year survival to 49%–77%[1]. Tumor size and invasion into adjacent structures (thyroid, recurrent laryngeal nerve, cervical muscles) are the strongest predictors of recurrence and survival, outweighing age, sex, or nodal status. Thus, parathyroidectomy (PTx)—especially in patients with secondary hyperparathyroidism (sHPT)—demands exquisitely accurate preoperative localization of the parathyroid glands (PGs).This condition affects approximately 90% of individuals with chronic kidney disease and significantly increases mortality risk.[2–5] However, preoperative

* Corresponding authors.
E-mail addresses: christinewang@pathoai.com (X. Wang), hnmoon025@163.com (H. Yang).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.intonc.2025.09.004
Received 5 June 2025; Received in revised form 19 September 2025; Accepted 23 September 2025
Available online 26 September 2025
2950-2616/© 2025 The Author(s). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

localization remains challenging: reliable imaging traditionally requires extensive physician expertise and time-intensive manual reviews, increasing surgical uncertainty and the risk of complications.[4] Preoperative localization failures increase the risk of inadvertent PG resection during PTx, which is the primary cause of postoperative hypoparathyroidism (a common complication in thyroid cancer surgery). Transient and permanent postoperative hypoparathyroidism occur in 6.9%–49% and 0.4%–33% of cases, respectively, with permanent postoperative hypoparathyroidism severely impairing quality of life.[6] Currently, this condition is mainly managed with medication; however, long-term supplementation with calcium and vitamin D analogs may lead to progressive nephrocalcinosis and, subsequently, renal insufficiency.[7–9] Thus, reliable preoperative PG localization is crucial for guiding intraoperative protection, reducing accidental resection, and minimizing iatrogenic injury during thyroid cancer surgery.[10–13]

Conventional modalities for PG localization include ultrasonography (US), methoxyisobutylisonitrile (MIBI) scintigraphy, magnetic resonance imaging, and computed tomography (CT).[14–17] technetium-99m-MIBI (^{99m}Tc-MIBI) scintigraphy is widely used for accurate localization at low radiation doses but exhibits low standalone detection rates, limiting its sensitivity for multiple PG lesions.[18] Fusing single-photon emission CT/CT (SPECT/CT) with MIBI to integrate functional and anatomical imaging can improve accuracy,[18] but ^{99m}Tc-MIBI SPECT/CT equipment is costly, restricting accessibility and imposing financial burdens on patients. US, although valuable, is operator-dependent and has low resolution. CT is less operator reliant, not constrained by the insonation angle, and more cost-effective and accessible than MIBI nuclear scanning.

Recent advances in artificial intelligence (AI) have shown promise for enhancing PG localization. Previous studies exploring AI models based on different algorithms and imaging techniques have demonstrated their potential to promote surgical outcomes.[19–21] For example, AI-based multimodal strategies improved the localization of bilateral upper and lower PGs before PTx.[18] Specifically, Greene et al.[22] compared the accuracy of a bidirectional long short-term memory network and a support vector machine in identifying parathyroid adenomas using four-dimensional CT scans of patients with hyperparathyroidism, finding that the bidirectional long short-term memory network outperformed the support vector machine when using both morphology and enhancement information. Ezenekwe et al.[23] identified quantifiable dimensional and texture features in four-dimensional CT images to distinguish adenomas from surrounding tissues. However, these machine learning methods lose partial three-dimensional (3D) voxel information, limiting their application to spatial data. Griffith et al.[24] further noted that preoperative localization of PGs in primary hyperparathyroidism patients using dual-phase CT was challenging for physicians, with sensitivity and specificity for multigland disease being 22.9% and 79.5%, respectively.

Notably, there are additional difficulties in establishing multimodal frameworks: cross-modality registration issues, data collection barriers (strict governance and patient privacy concerns), reliance on expert surgeon annotations for accurate labeling, and high equipment costs. Unlike prior studies focusing on the intraoperative identification and preservation of healthy PGs, the aim of this study was to achieve precise preoperative identification of all PGs in sHPT patients using AI-integrated dual-phase CT (with non-contrast [plain] and contrast-enhanced [CE] data). The primary objective was to develop a simplified system to enhance PG localization sensitivity while reducing equipment and patient costs. The secondary objective was to analyze the complementarity between the model's and clinicians' PG identification sensitivity across imaging modalities.

Integrating preoperative diagnosis with AI-based dual-phase CT modeling showed promise for improving localization sensitivity, making three key contributions:

(1) **An end-to-end 3D U-Net segmentation model based on dual-phase CT**: The model directly modeled 3D voxel blocks of CT and output segmentation predictions for PGs, fully preserving CT spatial information, implicitly extracting "basic anatomical contour features" from plain CT scans and "vascular-glandular enhancement difference features" from CE CT scans.
(2) **Complementarity analysis across multiple imaging modalities**: Quantitative analysis clarified the necessity of multimodal integration by defining their complementary roles.
(3) **A practical preoperative localization tool for PTx**: Segmentation results showed that automatically generated 3D visualization of PGs could assist physicians with preoperative planning and reduce intraoperative exploration time.

Despite these advances, challenges remain, including distinguishing PGs from adjacent tissues and shortening surgeons' learning curves. Developing an intelligent algorithm capable of detecting all PGs via CT would significantly advance surgical practice, streamline training, and guide precise dissection.

Beyond technical performance, the present study addresses three unmet clinical needs. First, persistent hypoparathyroidism remains the most frequent severe complication after total thyroidectomy or PTx, with permanent rates of 0.4%–33% and a claims-database prevalence as high as 20% in malignant disease.[25–28] Every additional gland that is inadvertently removed or devascularized increases this risk; thus an ultra-sensitive, CT-based AI map generated the day before surgery can be immediately fused with neuronavigation or mixed-reality headsets to provide real-time "do-not-cross" boundaries, reducing iatrogenic injury without prolonging operative time. Second, global inequity in MIBI-SPECT/CT availability leaves many sHPT patients—especially those on long-term hemodialysis—under-localized and subjected to repeated neck explorations. A deep-learning model that runs on standard dual-phase CT (available in most district hospitals) offers "imaging-level accuracy at CT-level cost", making high-quality localization accessible to low-resource regions and sparing patients the extra radiation, travel and expense of nuclear scans.Third, from a surgical training perspective, the 3D heat-map output can be integrated into preoperative briefings or AR simulators, shortening the learning curve for junior surgeons who traditionally need >150 cases to reach expert recognition rates. In high-volume endocrine centers the same tool may serve as a quality-assurance metric, benchmarking individual surgeons against an objective AI reference and thereby standardizing care. Collectively, translating this algorithm into routine workflow could shift the current paradigm from "experience-dependent anatomical search" to data-driven, patient-specific gland preservation, ultimately lowering the incidence of permanent hypoparathyroidism, reducing re-operations and improving long-term quality of life.

2. Materials and methods

2.1. Data collection and annotation

This study was based on retrospectively collected data (January 2020–April 2022) provided by the Department of Endocrine Surgery at the First Affiliated Hospital of Zhejiang University School of Medicine in China. The Institutional Review Board approved the study (Approval No.: IIT20230210B-R1). Patients diagnosed with sPHT who underwent both plain and CE CT examinations for preoperative PG localization were included in the study. Individuals lacking complete diagnostic data were excluded, resulting in 47 patients and 94 CT scans. The dataset was randomly partitioned into training and test sets at an 8:2 patient ratio. Specifically, the training set contained 76 scans from 38 patients, while the test set included 18 scans from 9 patients. For each patient, both plain and CE CT images were acquired using a multidetector CT scanner (Philips Brilliance iCT, Amsterdam, Netherlands) with the following scanner parameters: 3.0 mm collimation, 3.0 mm

Table 1
Demographic and clinical details of the study patients.

Characteristics	Training set (n = 38)	Test set (n = 9)
Gender, n (%)		
Male	21 (55.3%)	5 (55.6%)
Female	17 (44.7%)	4 (44.4%)
Age (mean ± SD) (years)	48.3 ± 10.61	49.0 ± 10.22
Duration of onset (mean ± SD) (years)	8.8 ± 4.52	6.89 ± 3.1
Number of PGs	151	36

Abbreviations: 1. PGs: parathyroid glands 2. SD: standard deviation.

reconstruction interval, 200 mA tube current, and 120 kVp voltage. All CE CT scans were obtained using an intravenous contrast agent (Iohexol; GE Healthcare, Milwaukee, WI). The injection protocol specified an injection rate of 3 mL/s with a total injection volume calculated as 1.0 mL/kg of the patient's weight. All images were reconstructed with a matrix size of 512 × 512 pixels. The median number of slices per scan (using both plain and CE CT) was 83 in the training set and 67 in the test set. Table 1 summarizes the demographic and clinical characteristics of the patients in both cohorts.

The PG ground truth was pathologically confirmed by the surgeon during thyroidectomy or PTx. The original data in DICOM format were converted to NIfTI format using SimpleITK software to facilitate annotation and subsequent modeling. Subsequently, the surgeon who performed the PTx delineated the PG locations slice by slice on both plain and CE CT images using ITK-SNAP software (Figure 1A).[29] The annotated PG regions were assigned a value of 1 to represent the foreground class, whereas the remaining background regions were assigned a value of 0. These binarized masks then served as learning targets for the subsequent deep learning-based model.

2.2. Model details

The modeling pipeline involved four phases: data preprocessing, network architecture selection, model training, and inference. For the training and inference setups, CT images were preprocessed. Data augmentation was performed to enhance model generalizability using rotation, mirroring, scaling, Gaussian noise, Gaussian blur, brightness and contrast adjustment, low-resolution simulation, and gamma correction. U-Net architecture (Figure 1B)—widely used for computer visioning—was employed to capture image features.[30] Through an encoder–decoder network, U-Net extracted features from training data via mechanisms such as instance normalization, the leaky rectified linear unit function, skip connections (to avoid loss of information), and deep supervision (to enhance regularization for accuracy and improve feature learning from small-sized training data). Specifically, a 3D full-resolution U-Net served as the backbone to minimize the loss of image resolution. The Dice and cross-entropy loss functions were used to optimize the model's parameters during the training phase, with loss weights progressively reduced. The final training loss (L) is defined as the sum of losses at five resolution levels, with weights (w) decreasing by half at each level:

$L = w_1 \times L_1 + w_2 \times L_2 + \ldots + w_5 \times L_5$. Weights were normalized to sum to 1, and w_5 was set at 0 due to its lowest information content at the final resolution. By satisfying $w_2 = ½ \times w_1$, $w_3 = ¼ \times w_1$, …, and $w_1 + w_2 + \ldots + w_5 = 1$, the respective weights were derived as 0.53, 0.27, 0.13, 0.07, and 0 (rounded to two decimal places), facilitating the differential exploitation of information across layers.

Based on the training set (Table 1), the model was optimized using the stochastic gradient descent optimization algorithm with Nesterov momentum ($\mu = 0.99$), a polynomial learning rate schedule (initial learning rate = 0.01), and a batch size of 2. Additional pipeline details included 1,000 training epochs (250 mini-batches per epoch), foreground oversampling, an inference procedure involving sliding windows with half-patch size overlap, and Gaussian patch center weighting.

Furthermore, to optimize the training process by assessing model overfitting and adjusting training strategies, and to evaluate model stability during the training phase, five-fold cross-validation was performed on the training set (Table 1). Due to the limited amount of data, cross-validation is used not only to optimize the training hyperparameters, but also to assess the algorithm's robustness against data selection bias. During the five-fold cross-validation, the models achieving the highest Dice scores for their respective validation sets across multiple training iterations within each fold were then evaluated using the test set. After determining the training hyperparameters based on the training set in Table 1, the final model was obtained using the training

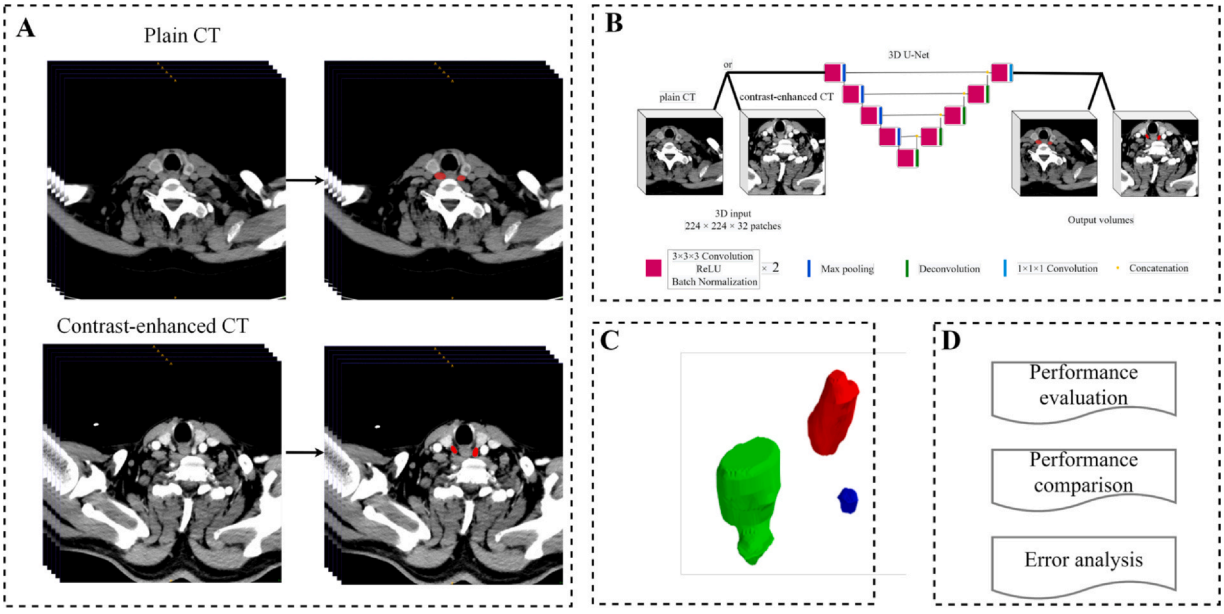


Figure 1. Overall pipeline schematic. A: Data labeling illustration chart. B: End-to-end training of 3D U-Net, with "2 ×" denoting two stacked "3 × 3 × 3 convolution → ReLU → Batch Normalization" blocks. C: Model prediction visualization. D: Performance evaluation and comparison error analysis. Abbreviations: 1. CT: computed tomography 2. 3D: three-dimensional 3. ReLU: rectified linear unit.

Table 2
Performance evaluation of the model using the test set at the pixel and PG levels.

Level	Dice score (%)	Precision (%)	Recall (%)	F1 score (%)	FDR (%)	FNR (%)
Pixel	58.74	72.12	51.72	60.24	27.88	48.28
PG	/	87.14	74.39	80.26	12.86	25.61

Abbreviations: 1. FDR: false discovery rate 2. FNR: false negative rate 3. PG: parathyroid gland.

configuration described above. This model was utilized for performance evaluation and comparison.

When the trained model inferred CT voxel data, it output a confidence score for each voxel, representing the probability that the voxel is a PG. Using a 0.5 confidence threshold, voxels with scores ≥ 0.5 were labeled as PGs (1), and those with scores < 0.5 as background (0), yielding a binary mask.

2.3. Performance evaluation

To assess the model's predictive performance, its efficacy was evaluated using a test set at both the pixel and PG levels (Figure 1D). At the pixel level, a pixel that the model predicted to be a PG was considered a true positive if it was located within the manually annotated region; otherwise, it was deemed a false positive. For the PG-level evaluation, a Python package (connected-components-3d, version 3.10.3)[31] was used to generate 3D multilabel connected-component-labeled images based on the binary mask output by the model. The predicted PGs were then extracted from these images based on distinct mask values (Figure 1C). 50% volumetric overlap is typically required to define a true positive. To align with standard organ-level detection metrics, a predicted PG was considered a true positive only if its voxel-level intersection with the manually labeled region exceeded 50% of the total voxels within that label; otherwise, the prediction was classified as a false positive.

Metrics, including the Dice score, precision (equivalent to a positive predictive value), recall, F1 score, false discovery rate (FDR), and false negative rate (FNR), were used to evaluate performance at both the pixel and PG levels. Since the Dice score is primarily used as a pixel-level metric for semantic segmentation, it was not calculated for the PG level; instead, the F1 score was adopted as a comprehensive metric. Notably, when combining detection results across different modalities to analyze PG identification sensitivity and intermodality complementarity, a PG was considered successfully identified if it was detected via any modality; otherwise, it was classified as missed.

2.4. Statistical analysis

Statistical analysis and data visualization were performed using Python and GraphPad Prism version 9 (GraphPad Software, San Diego, CA, United States) software. Fisher's exact test was used to compare the categorical variables between the groups. Potential significant differences in PG identification sensitivity between models and individuals across different modalities were evaluated. Furthermore, whether combining two modalities (rather than using a single modality) significantly improves PG identification sensitivity was assessed. A two-sided P value < 0.05 was considered statistically significant.

2.5. Implementation details

The implementation of the modeling process described in the previous section can be summarized as follows. Original DICOM files were converted to NIfTI format using SimpleITK 2.2.0.[32] The model was implemented in Python 3.9.13 based on the PyTorch 1.9.0 framework.[33] Model training and testing were conducted on NVIDIA Tesla V100 graphics processing units (GPUs) with 32 GB of GPU memory, using graphics card driver version 510.73.08 and NVIDIA® CUDA® Toolkit version 11.6. Data augmentation was performed using the Python package batch generator 0.24. The additional Python libraries used included NumPy 1.24.2, SciPy 1.10.1, and pandas 1.5.3.[34,35]

3. Results

3.1. The AI model's parathyroid identification efficiency

The model performance was evaluated at two levels (pixel and PG levels) using the test set. As shown in Table 2, regarding the model's pixel-level performance, the Dice score, precision, recall, F1 score, FDR, and FNR were 58.74%, 72.12%, 51.72%, 60.24%, 27.88%, and 48.28%, respectively. At the PG level, the model achieved a precision of 87.14%, a recall of 74.39%, an F1 score of 80.26%, an FDR of 12.86%, and an FNR of 25.61%. In terms of inference speed, the model required approximately 30 s to generate a prediction for a single CT scan. For dual-phase CT data (one plain CT and one CE CT), the model's prediction time was approximately 1 min. To demonstrate the algorithm's robustness against potential performance bias caused by random data splitting, the model that achieved the best Dice score in each fold of cross-validation was evaluated using the test set (Table 3). The average F1 scores at the pixel level (60.29%, $P = 0.9505$) and the PG level (79.47%, $P = 0.7154$) exhibited no significant differences from the corresponding values of the current model (60.24% and 80.26%, respectively). The results of one-sample t-tests indicated that the model's

Table 3
Performance of the five models generated by cross-validation using the test set at the pixel and PG levels.

Level	Folds	Dice score (%)	Precision (%)	Recall (%)	F1 score (%)	FDR (%)	FNR (%)
Pixel	1	60.78	71.54	55.20	62.32	28.46	44.80
	2	56.07	56.09	60.39	58.16	43.91	39.61
	3	58.24	65.72	55.15	59.97	34.28	44.85
	4	57.45	69.18	51.45	59.01	30.82	48.55
	5	58.54	68.42	56.69	62.01	31.58	43.31
	mean	58.22	66.19	55.78	60.29	33.81	44.22
PG	1	/	87.14	74.39	80.26	12.86	25.61
	2	/	68.75	80.49	74.16	31.25	19.51
	3	/	89.55	73.17	80.54	10.45	26.83
	4	/	85.07	69.51	76.51	14.93	30.49
	5	/	95.52	78.05	85.91	4.48	21.95
	mean	/	85.21	75.12	79.47	14.79	24.88

Abbreviations: 1. FDR: false discovery rate 2. FNR: false negative rate 3. PG: parathyroid gland.

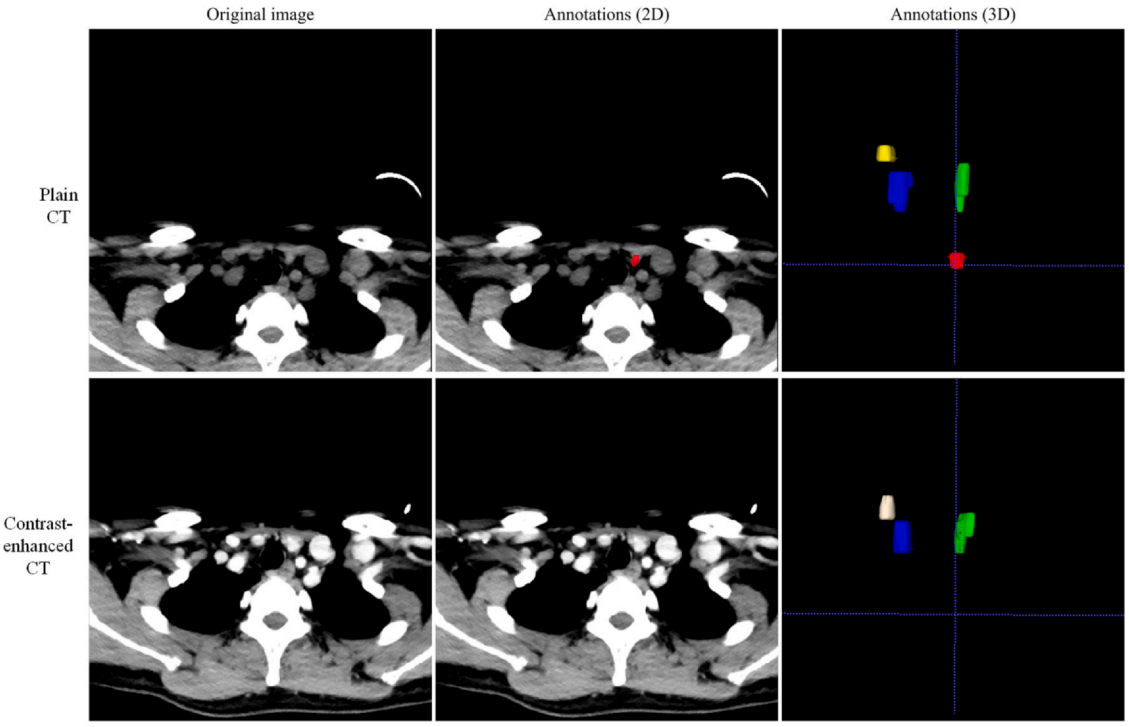


Figure 2. Example of complementarity between plain and contrast-enhanced CT images. Abbreviations: 1. CT: computed tomography 2. 2D: two-dimensional 3. 3D: three-dimensional.

performance remained consistent across multiple randomized trials, suggesting robustness against data selection bias. However, due to limited data, employing multiple models for cross-validation to assess robustness against data selection bias represented a compromise. Further validation using a larger prospective multicenter dataset would have been a more optimal approach.

In addition, we measured the centroid distance to evaluate the consistency of PG identification between the AI model and the manually labeled ground truth by surgeons. The results showed that the pixel-level overlapping area of PGs identified by the two approaches was small, and the PG-level centroid distance between the two approaches was approximately 1.28 ± 1.34 cm. Potential reasons for this inconsistency included differences in PG imaging characteristics between plain and CE CT images, which resulted in varying degrees of blurring at the PG boundaries. The identification difficulty also varied between the two imaging types. Moreover, the subjective maturity of human labeling introduced additional uncertainty. Even for clinicians who performed PTx and had access to surgical records, there could be discrepancies in labeling the morphology and size of PGs between the two modalities for the same patient (Figure 2). Thus, we focused on identifying the primary location of the PGs rather than their exact size, although the estimated size of the labeled PGs remained a valuable reference.

3.2. Comparison of identification rates among CT, US, and ^{99m}Tc-MIBI SPECT/CT

Table 4 and Figure 3 present experienced clinicians' rates of PG detection using various imaging modalities. Tc-MIBI SPECT/CT achieved the highest sensitivity at 86.11%, significantly outperforming that of US (61.11%, $P = 0.0309$). However, its sensitivity did not differ statistically from that of CT (72.22%, $P = 0.2454$).

For the AI model, the sensitivity scores were 77.78% and 69.44% for plain CT and CE CT data alone, respectively. Notably, when predictions based on both plain and CE CT data were combined, the sensitivity increased to 94.44%, which was significantly higher than the sensitivity of CT ($P = 0.0238$), US ($P = 0.0013$), or CE CT alone ($P = 0.012$). No significant differences in sensitivity were observed between the other modalities.

Plain CT exhibited a higher sensitivity than CE CT, although this difference was not statistically significant; the two modalities remained complementary. Combining them increased the number of detected PGs by 6 (relative to plain CT alone) and 9 (relative to CE CT alone), with corresponding sensitivity improvements of 16.66% and 25% compared to their individual sensitivities. Although the sensitivity of plain CT did not improve significantly ($P = 0.0847$), that of CE CT improved significantly ($P = 0.012$).

Table 4
Comparison of PG detection sensitivity between models and clinicians for different modalities.

Modalities	No. of identified PGs	No. of missed PGs	Sensitivity (%)	FNR (%)
CT	26	10	72.22	27.78
US	22	14	61.11	38.89
^{99m}Tc-MIBI SPECT/CT	31	5	86.11	13.89
AI (pCT)	28	8	77.78	22.22
AI (CE CT)	25	11	69.44	30.56
AI (p&CE CT)	34	2	94.44	5.56

Abbreviations: 1. CT: computed tomography 2. US: ultrasonography 3. ^{99m}Tc-MIBI SPECT/CT: technetium-99m-methoxyisobutylisonitrile single-photon emission CT/CT 4. pCT: plain CT 5. CE CT: contrast-enhanced CT 6. p&CE CT: plain CT and contrast-enhanced CT integrated prediction 7. PGs: parathyroid glands 8. FNR: false negative rate 9. AI: artificial intelligence.

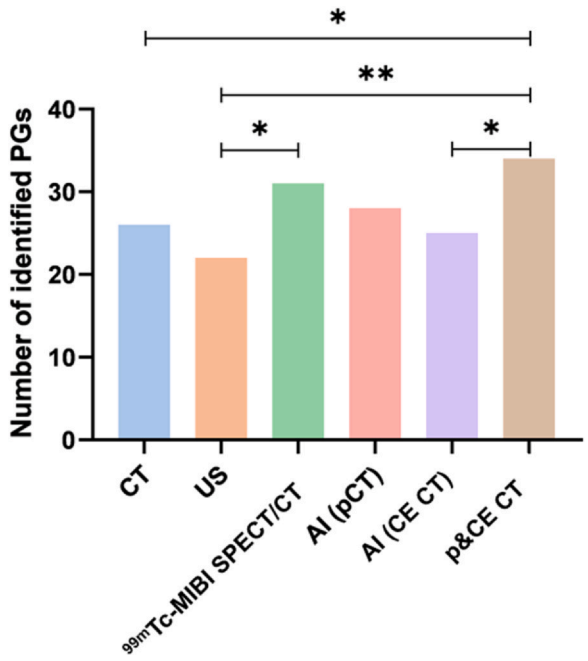


Figure 3. Significance levels for sensitivity differences between different modalities. *, ** denote $P < 0.05$ and $P < 0.01$, respectively. Abbreviations: 1. PGs: parathyroid glands 2. CT: computed tomography 3. US: ultrasonography 4. ^{99m}Tc-MIBI SPECT/CT: technetium-99m-methoxyisobutylisonitrile single-photon emission CT/CT 5. AI: artificial intelligence 6. pCT: plain CT 7. CE CT: contrast-enhanced CT, 8. p&CE CT: plain CT and contrast-enhanced CT integrated prediction.

3.3. Error analysis

As shown in Figure 4, we compared the PG locations predicted by the model with the corresponding ground truth. For the five PGs included in the error analysis, the model detected five PGs, whereas the ground truth detection confirmed only three PGs (Figure 4A and 4B). Specifically, the model accurately identified two of the three ground truth PGs, yielding three false positives and one false negative (Figure 4C).

UpSet plots (Figure 5) clearly visualized disparities between the PGs detected by the model and those detected by experienced clinicians across modalities. The model missed two PGs, one of which was undetected by all modalities (CT, US, and ^{99m}Tc-MIBI SPECT/CT). In contrast, experienced clinicians had higher rates of missed detection when interpreting images across modalities, raising cosiderable concerns about diagnostic accuracy. The results suggest that although different modalities were, to some extent, complementary, the model offered greater complementarity, providing clinicians with more precise information about PG localization and ultimately improving PG detection sensitivity.

3.4. Analysis of intermodal complementarity

Overall, four PGs were visible on the plain CT images, whereas only three were visible on the CE CT images (Figure 2). One PG was indistinct from the surrounding tissues, rendering it difficult to identify. When both CT modalities were combined, all PGs were successfully detected, indicating the complementarity of the modalities and a significant improvement in the PG detection rate.

To further explore intermodality complementarity, we combined two modalities with comparable sensitivities. Sensitivities before and after the combination were compared to assess whether a significant improvement was achieved. Based on the available data, the results showed that AI support significantly enhanced the sensitivity of US imaging (Table 5).

4. Discussion

4.1. Performance of 3D U-Net

It is crucial for PGs in patients with sHPT to be accurately localized before PTx to minimize surgical risks and perioperative trauma, thereby facilitating postoperative recovery. Preoperative PG localization commonly relies on medical imaging modalities, including US, CT, and ^{99m}Tc-MIBI SPECT/CT. However, the efficacy of a single modality is limited: US is cost-effective but suffers from lower accuracy due to patient-specific variations, CT is widely accessible but has limited sensitivity, and ^{99m}Tc-MIBI SPECT/CT is effective but less accessible and involves radiation exposure from radioactive isotopes.

In this article, we proposed a dual-phase model to leverage plain and CE CT data based on a 3D U-Net-based semantic segmentation algorithm. The model leverages 3D voxel blocks as inputs and preserves the spatial topological information of CT data (e.g., morphological continuity across slices and spatial relationships with the thyroid gland and blood vessels) via 3D convolutional operations. It harnesses the complementary effects of plain CT (which provides basic anatomical information) and CE CT (which provides dynamic enhancement information). The model directly outputs a 3D segmentation mask, facilitating end-to-end mapping from raw CT data to the spatial locations of PGs. Quantitative analysis of intermodality complementarity clearly demonstrated the need for multimodal integration.

The model exhibited good sensitivity for PG identification, making it a useful tool for surgical planning. By combining plain and CE CT, the model achieved a sensitivity of 94.44%, outperforming clinician sensitivity based on CT or US alone or their combination with ^{99m}Tc-MIBI SPECT/CT. However, two PGs were not detected by the model: one was a small PG in a 60-year-old female patient, which adhered tightly to the thyroid gland and was thus indistinct on CT images; the other was an ectopic PG in a 48-year-old male patient, located deep in the left upper anterior mediastinum—an anatomical location underrepresented in the training data. Notably, all four PGs in the latter case were successfully identified by clinicians. Analyzing false predictions, particularly when incorporating spatial distribution features to reduce false positives

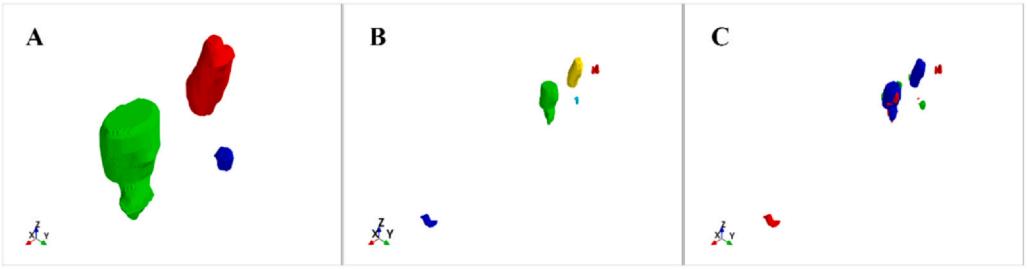


Figure 4. Three-dimensional visualization of model predictions and gold standards. A: Gold standards. B: The model's predictions. C: Prediction results of the model concerning the gold standards in the same spatial space. The colors blue, red, and green indicate a hit, a false positive, and a false negative, respectively.

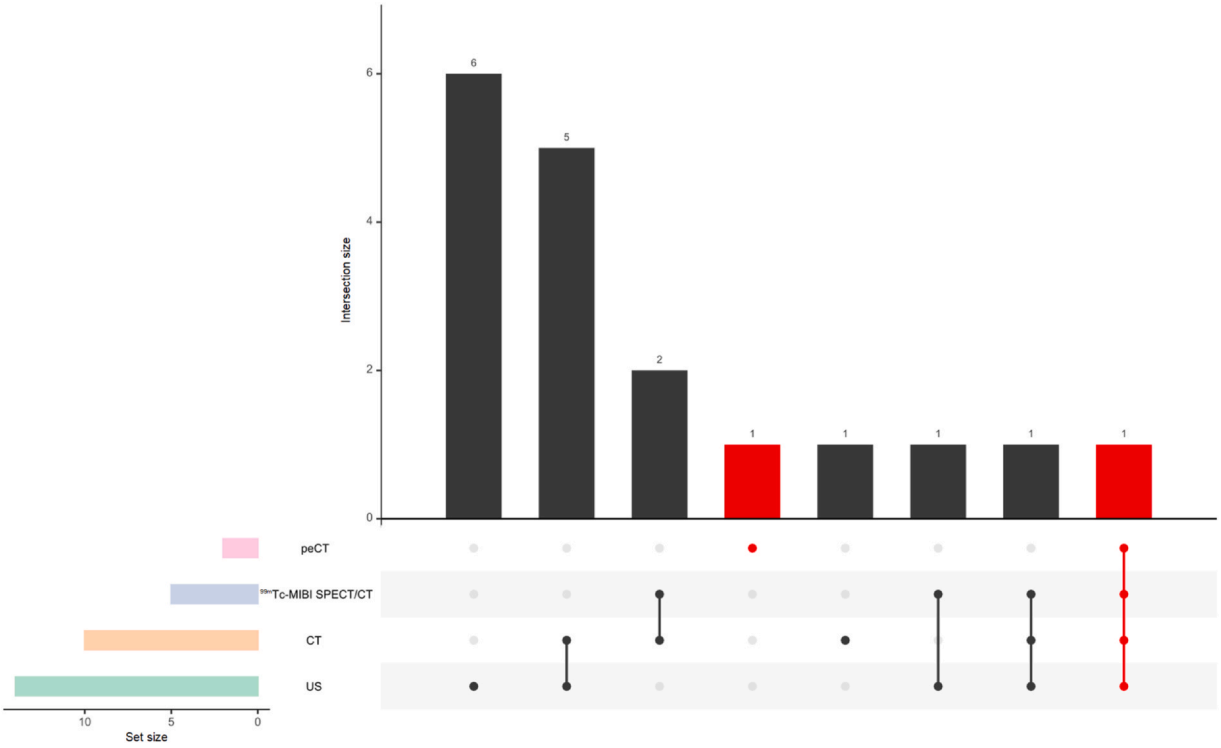


Figure 5. UpSet plot of the distribution of PGs missed by clinicians and the model using different modalities. The distribution of red and black solid dots represents the distribution of PGs missed by clinicians and the model across different modalities, respectively. p&CE CT denotes diagnoses missed by the model based on plain CT and contrast-enhanced CT, whereas other modalities indicate diagnoses missed by clinicians. Abbreviations: 1. CT: computed tomography 2. p&CE CT: plain CT and contrast-enhanced CT integrated prediction 3. ^{99m}Tc-MIBI SPECT/CT: technetium-99m-methoxyisobutylisonitrile single-photon emission CT/CT 4. US: ultrasonography.

Table 5
Analysis of the gains in performance due to modal complementarity.

Modality_1	Modality_2	Modality_1&2	Sensitivity_1&2	*P* value_1	*P* value_2	Sig_level_1	Sig_level_2
CT	US	CT + US	80.56	0.5798	0.1187	-	-
CT	^{99m}Tc-MIBI SPECT/CT	CT + ^{99m}Tc-MIBI SPECT/CT	88.89	0.1348	1.0000	-	-
CT	pCT	CT + pCT	88.89	0.1348	0.3434	-	-
CT	CE CT	CT + CE CT	86.11	0.2454	0.1552	-	-
US	pCT	US + pCT	86.11	**0.0309**	0.5414	*	-
US	CE CT	US + CE CT	88.89	**0.0130**	0.0795	*	-
^{99m}Tc-MIBI SPECT/CT	pCT	^{99m}Tc-MIBI SPECT/CT + pCT	91.67	0.7101	0.1887	-	-
^{99m}Tc-MIBI SPECT/CT	CE CT	^{99m}Tc-MIBI SPECT/CT + CE CT	91.67	0.7101	**0.0346**	-	*
^{99m}Tc-MIBI SPECT/CT	p&CE CT	^{99m}Tc-MIBI SPECT/CT + p&CE CT	97.22	0.1987	1.0000	-	-
pCT	CE CT	pCT + CE CT	94.44	0.0847	**0.0120**	-	*
pCT	p&CE CT	pCT + p&CE CT	94.44	0.0847	1.0000	-	-

Table headers: modality_1 (first modality), modality_2 (second modality), modality_1&2 (combination), sensitivity_1&2 (combined sensitivity), *P* value_1 and sig_level_1 (the *P* value and significance level of the sensitivity difference: modality_1&2 vs. modality_1), *P* value_2 and sig_level_2 (the *P* value and significance level of the sensitivity difference: modality_1&2 vs. modality_2). Symbols: - (no statistical significance), * ($P < 0.05$). Abbreviations: 1. CT: computed tomography 2. pCT: plain CT 3. CE CT: contrast-enhanced CT 4. p&CE CT: plain CT and contrast-enhanced CT integrated prediction 5. ^{99m}Tc-MIBI SPECT/CT: technetium-99m-methoxyisobutylisonitrile single-photon emission CT/CT 6. US: ultrasonography.

associated with ectopic PGs, helped improve the model. Overall, the dual-phase CT-based model showed potential as a valuable tool for preoperative PG localization to aid PTx planning.

Nevertheless, this study has several limitations. First, the sample size was small, and the data sets were derived from a single-center retrospective cohort; thus, the generalizability of the model and the robustness of our conclusions require further validation using larger prospective multicenter datasets. Additionally, the integration of dual-phase CT data into the current model relied solely on data-level fusion, which enabled the model to implicitly learn discriminative features from both phases. This implementation represented a relatively low-level, straightforward approach that did not allow image registration between plain and CE CT scans, placed no explicit constraints on interphase feature interactions, and failed to fully capture the inherent correlations and complementarity between the two CT phases.

4.2. Limitations and future directions

A principal limitation of the present study is the small test cohort ($n = 9$ patients, 18 CT scans), which reflects the real-world difficulty of

curating paired plain and CE CT examinations together with surgeon-verified voxel-level annotations. To address this bottleneck, we have initiated a prospective, multicenter expansion that will enroll over 200 patients with sHPT from at least three tertiary hospitals in different geographic regions of China. Each center will follow a unified imaging protocol (identical CT acquisition parameters, contrast injection scheme and reconstruction kernels) and an independent, blinded annotation workflow reviewed by two endocrine surgeons and one radiologist. This larger dataset will be split into an external-validation cohort (60%) for performance confirmation and a clinical-utility cohort (40%) for on-line deployment testing.

The primary end-points of the multicenter trial are:

(1) To maintain $\geq 90\%$ patient-level sensitivity for detecting at least one PG while keeping false-positive rate ≤ 0.5 glands per scan;

(2) To demonstrate non-inferiority to ^{99m}Tc-MIBI SPECT/CT (margin 5%) in a head-to-head reader study;

(3) To measure the impact of AI-assisted localization on surgical decision-making (rate of minimally invasive vs. bilateral exploration, operative time, and 30-day permanent hypoparathyroidism).

Parallel technical efforts will include semi-supervised learning with pseudo-labeling to exploit unlabeled CT pairs, reducing annotation burden by about 40%; federated learning to preserve patient privacy across centers without centralizing raw data; integration with commercial neuronavigation and mixed-reality headsets for real-time overlay during thyroidectomy or PTx. The algorithm will be launched as a cloud service that allows surgeons to refine oncologic resection while safeguarding parathyroid function.

5. Conclusions

This study established a deep learning model to leverage dual-phase CT (plain and CE CT) for preoperative PG localization in sHPT patients. Our model achieved a sensitivity of 94.44% for PG identification, significantly outperforming clinician-interpreted US (61.11%, $P = 0.0013$) and CT (72.22%, $P = 0.0238$) sensitivity. Notably, the model demonstrated accuracy comparable to that of ^{99m}Tc-MIBI SPECT/CT (86.11%, $P = 0.429$). The results confirmed that fusion of plain and CE CT features via deep learning unlocked critical complementary information for PG detection, addressing a key clinical challenge in PTx planning. Moreover, synergy with other modalities revealed the potential for further sensitivity improvements based on hybrid diagnostic workflows. This model offers a broadly accessible, cost-effective alternative to specialized nuclear imaging, particularly for use in resource-constrained settings. Future research will validate these findings using larger prospective multicenter cohorts and continuous model refinement will further elevate performance.

CRediT authorship contribution statement

Hao Yang: Writing – review & editing, Project administration, Formal analysis, Conceptualization. **Fuqiang Li:** Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Data curation, Conceptualization. **Yao Zhang:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Yijing Wen:** Writing – review & editing, Data curation. **Xiaomei Wang:** Writing – review & editing, Resources, Project administration.

Ethical statement

The study design was approved by the Institutional Review Board of the First Affiliated Hospital, Zhejiang University School of Medicine (Approval No.: IIT20230210B-R1).

Funding

This research received no financial support.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships, which may be considered as potential competing interests: Xiaomei Wang and Yao Zhang are currently employed by Hangzhou PathoAI Technology Co., Ltd.; Hao Yang is currently employed by Hangzhou Meiaiying Biomedical Technology Co., Ltd.

Acknowledgments

We thank Jin Han, Shiqiao Zhang, and Bojun Cai of Hangzhou PathoAI Technology Co., Ltd. for their technical support and discussion.

References

1. Zhang K, Su A, Wang X, et al. Non-linear correlation between tumor size and survival outcomes for parathyroid carcinoma: A SEER population-based cohort study. *Front Endocrinol (Lausanne)*. 2022;13:882579.
2. Eckardt KU, Coresh J, Devuyst O, et al. Evolving importance of kidney disease: from subspecialty to global health burden. *Lancet*. 2013;382(9887):158–169.
3. Hill NR, Fatoba ST, Oke JL, et al. Global prevalence of chronic kidney disease - A systematic review and meta-analysis. *PLoS One*. 2016;11(7):e0158765.
4. Strambu V, Bratucu M, Garofil D, et al. The value of imaging of the parathyroid glands in secondary hyperparathyroidism. *Chirurgia (Bucur)*. 2019;114(5):541–549.
5. Trainor D, Borthwick E, Ferguson A. Perioperative management of the hemodialysis patient. *Semin Dial*. 2011;24(3):314–326.
6. Ozogul B, Akcay MN, Akcay G, Bulut OH. Factors affecting hypocalcaemia following total thyroidectomy: a prospective study. *Eurasian J Med*. 2014;46(1):15–21.
7. Ru Z, Mingliang W, Maofei W, Qiaofeng C, Jianming Y. Analysis of risk factors for hypoparathyroidism after total thyroidectomy. *Front Surg*. 2021;8:668498.
8. Dughiero S, Torresan F, Censi S, Mian C, Carrillo Lizarazo JL, Iacobone M. Risk and protective factors of postoperative and persistent hypoparathyroidism after total thyroidectomy in a series of 1965 patients. *Cancers (Basel)*. 2024;16(16):2867.
9. Takahashi T, Yamazaki K, Shodo R, Ueki Y, Horii A. Actual prevalence of hypoparathyroidism after total thyroidectomy: a health insurance claims-database study. *Endocrine*. 2022;78(1):151–158.
10. Cannizzaro MA, Lo Bianco S, Picardo MC, Provenzano D, Buffone A. How to avoid and to manage post-operative complications in thyroid surgery. *Updates Surg*. 2017;69(2):211–215.
11. Reinisch A, Malkomes P, Liese J, et al. Education in thyroid surgery: a matched-pair analysis comparing residents and board-certified surgeons. *Langenbecks Arch Surg*. 2016;401(2):239–247.
12. Sanabria A, Kowalski LP, Tartaglia F. Inferior thyroid artery ligation increases hypocalcemia after thyroidectomy: A meta-analysis. *Laryngoscope*. 2018;128(2):534–541.
13. Gerardi I, Verro B, Amodei R, Richiusa P, Saraniti C. Thyroidectomy and Its Complications: A Comprehensive Analysis. *Biomedicines*. 2025;13(2):433.
14. Kettle AG, O'Doherty MJ. Parathyroid imaging: how good is it and how should it be done? *Semin Nucl Med*. 2006;36(3):206–211.
15. Russell MD, Orloff LA. Ultrasonography of the thyroid, parathyroids, and beyond. Sonographie der Schilddrüse, der Nebenschilddrüsen und darüber hinaus. *HNO*. 2022;70(5):333–344.
16. Taieb D, Hindie E, Grassetto G, Colletti PM, Rubello D. Parathyroid scintigraphy: when, how, and why? A concise systematic review. *Clin Nucl Med*. 2012;37(6):568–574.
17. Alessa M, Algouhi A, Alsowailmi G, Arafat A. Preoperative localization for primary hyperparathyroidism surgery: comparison of imaging techniques at a tertiary center. *Indian J Surg*. 2024;86(1):151–160.
18. Jiang SQ, Yang T, Zou Q, et al. The role of ^{99m}Tc-MIBI SPECT/CT in patients with secondary hyperparathyroidism: comparison with ^{99m}Tc-MIBI planar scintigraphy and ultrasonography. *BMC Med Imaging*. 2020;20(1):115.
19. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24–29.
20. Wang B, Zheng J, Yu JF, et al. Development of Artificial Intelligence for Parathyroid Recognition During Endoscopic Thyroid Surgery. *Laryngoscope*. 2022;132(12):2516–2523.
21. Avci SN, 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*. Published online March 28, 2022. https://doi.org/10.1245/s10434-022-11632-y
22. Greene C, Fujima N, Sakai O, Andreu-Arasa VC. Comparing accuracy of machine learning approaches to identifying parathyroid adenomas: Lessons and new directions. *Am J Otolaryngol*. 2024;45(2):104155.
23. Ezenekwe CH, Zhang M, Hussain IT, Ginat D. Utilizing radiomic features of arterial phase computed tomography for delineating parathyroid adenomas from surrounding anatomical structures. *J Clin Transl Res*. 2024;10(2):165–171.
24. Griffith B, Chaudhary H, Mahmood G, et al. Accuracy of 2-phase parathyroid CT for the preoperative localization of parathyroid adenomas in primary hyperparathyroidism. *AJNR Am J Neuroradiol*. 2015;36(12):2373–2379.

25. Bunch PM, Aribindi S, Gorris MA, Randle RW. Opportunistic CT assessment of parathyroid glands: utility of radiologist-recommended biochemical evaluation for diagnosing primary hyperparathyroidism. *AJR Am J Roentgenol.* 2023;221(2):218–227.
26. Lincoln CMM. Editorial comment: value-added assessment for parathyroid adenomas on routine CT examinations. *AJR Am J Roentgenol.* 2023;221(2):226.
27. Morris MA, Saboury B, Ahlman M, et al. Parathyroid imaging: past, present, and future. *Front Endocrinol (Lausanne).* 2022;12:760419.
28. Petranović Ovčariček P, Giovanella L, Carrió Gasset I, et al. The EANM practice guidelines for parathyroid imaging. *Eur J Nucl Med Mol Imaging.* 2021;48(9):2801–2822.
29. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage.* 2006;31(3):1116–1128.
30. Isensee F, Jaeger PF, Kohl SAA, Petersen J. Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2021;18(2):203–211.
31. Sutheebanjard P. Decision tree for 3-D connected components labeling. *2012 International Symposium on Information Technologies in Medicine and Education.* Vol 2. IEEE; 2012:709–713.
32. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of SimpleITK. *Front Neuroinform.* 2013;7:45.
33. Paszke A, Gross S, Massa F, Pytorch, et al. An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst.* 2019:32.
34. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature.* 2020;585(7825):357–362.
35. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261–272.