J of Nuclear Medicine Technology, first published online December 7, 2021 as doi:10.2967/jnmt.121.263081

Re-Modelling 99m-Technetium Pertechnetate Thyroid Uptake; Statistical, Machine Learning and Deep Learning Approaches

Geoff Currie[1,2], Basit Iqbal[3]

[1]Charles Sturt University, Wagga Wagga, Australia
[2]Baylor College of Medicine, Houston, USA
[3]Gujranwala Institute of Nuclear Medicine & Radiotherapy, Gujranwala, Pakistan

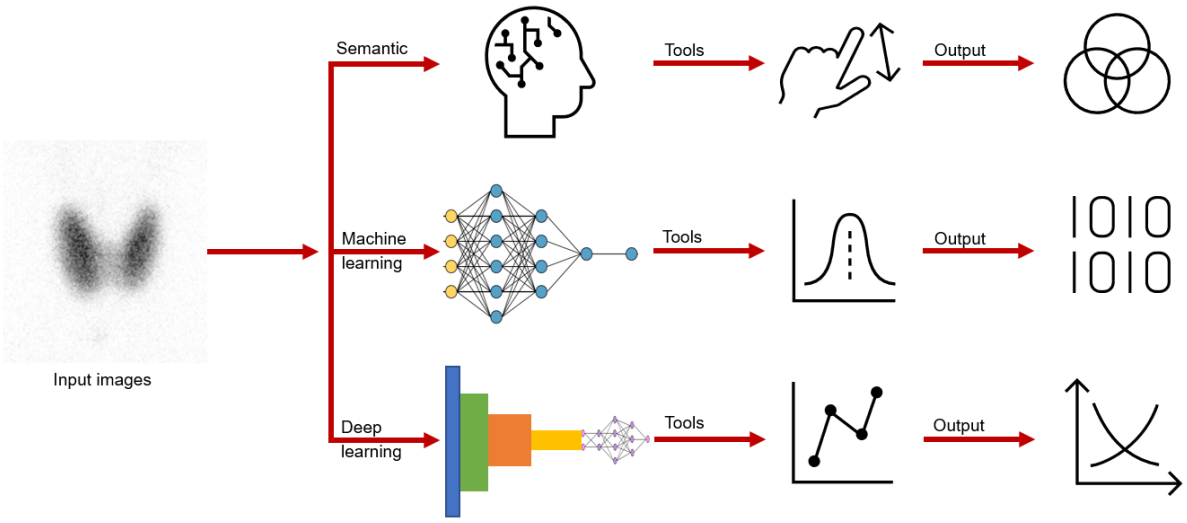
Key words: thyroid uptake, hyperthyroidism, machine learning, deep learning, artificial intelligence

Declarations

There are no funding or conflicts of interest to declare.

The retrospective analysis was approved by institutional ethics committee and a waiver for consent for use of de-identified data was authorised.

Graphical Abstract



Abstract

Background: While normal ranges for 99mTc thyroid percentage uptake vary, the seemingly intuitive evaluation of thyroid function does not reflect the complexity of thyroid pathology and biochemical status. The emergence of artificial intelligence (AI) in nuclear medicine has driven problem solving associated with logic and reasoning that warrant re-examination of established benchmarks in thyroid functional assessment.

Method: There were 123 patients retrospectively analysed in the study sample comparing scintigraphic findings to grounded truth established through biochemistry status. Conventional statistical approaches were used in conjunction with an artificial neural network (ANN) to determine predictors of thyroid function from data features. A convolutional neural network (CNN) was also used to extract features from the input tensor (images).

Results: Analysis was confounded by sub-clinical hyperthyroidism, primary hypothyroidism, sub-clinical hypothyroidism and T3 toxicosis. Binary accuracy for identifying hyperthyroidism was highest for thyroid uptake classification using a threshold of 4.5% (82.6%), followed by pooled physician 6interpretation with the aid of uptake values (82.3%). Visual evaluation without quantitative values reduced accuracy to 61.0% for pooled physician determinations and 61.4% classifying on the basis of thyroid gland intensity relative to salivary glands. The machine learning (ML) algorithm produced 84.6% accuracy, however, this included biochemistry features not available to the semantic analysis. The deep learning (DL) algorithm had an accuracy of 80.5% based on image inputs alone.

Conclusion: Thyroid scintigraphy is useful in identifying hyperthyroid patients suitable for radioiodine therapy when using an appropriately validated cut-off for the patient population (4.5% in this population). ML ANN algorithms can be developed to improve accuracy as second readers systems when biochemistry results are available. DL CNN algorithms can be developed to improve accuracy in the absence of biochemistry results. ML and DL do not displace the role of the physician in thyroid scintigraphy but could be used as second reader systems to minimize errors and increase confidence.

.

Introduction

In 1967, Atkins and Richards (*1*) evaluated the potential role of 99m-technetium (99mTc) pertechnetate in evaluating thyroid function as an alternative to sodium iodide with 131-iodine (131I) on the basis that 99mTc uptake in the thyroid reflects the gland's trapping function. This landmark work utilized a probe detector rather than gamma camera imaging approach for the uptake calculation. A small number of hypothyroid patients were included and all had percentage uptakes below 0.5%. Only 2 of 15 hyperthyroid patients fell below 4% while 4 of 133 euthyroid patients had uptake above 4%. Thus, a cut-off for normality was set at **0.4% to 4.0%** to provide 87% accuracy in hyperthyroidism, 97% accuracy in euthyroid, and 100% accuracy in hypothyroidism.

Later work in 1973 by Maisey et al (*2*) utilized a gamma camera, pinhole collimation and interfaced computer to generate regions of interest for calculation of 99mTc pertechnetate uptake in the thyroid. Euthyroid patients ranged from 0.2% to 3.6% or 0.3% to 6.2% in the presence of a goitre, 2.8% to 8.8% for hyperthyroidism, and 0.1% to 0.3% for hypothyroidism to establish a normal range of **0.3% to 3.4%**. More recently, 99mTc pertechnetate uptake in euthyroid was characterized in the range **0.4% to 1.7%** in 47 clinically normal patients (*3*). It is widely acknowledged that normal values change with geography and time, particularly in relation to iodine deficiency (*4*). While it is common for widespread use of international standards (0.5% to 4.5% for example), these values may not reflect either the technique used (probe versus gamma camera) or population characteristics (eg. iodine deficiency). In Namibia, investigators found the normal range to be **0.15% to 2.14%** (*4*) although the study only included 76 patients and all were euthyroid. A UK study (*5*) used 60 euthyroid patients to estimate the local normal range as **0.2% to 2.0%**.

While normal ranges for percentage uptake vary, the method for calculation of thyroid function on 99mTc scintigraphy also varies (*6*). The seemingly intuitive evaluation of thyroid function has also referenced as a visual evaluation of thyroid activity relative to salivary gland activity (figure 1). This does not reflect the complexity of thyroid pathology and biochemical status. This simplification is intuitive when the bulk of patients are

euthyroid or hyperthyroid but fails to accommodate sub-clinical hyperthyroidism which can produce low thyroid accumulation of 99mTc, T3 toxicosis which can have high or low 99mTc uptake, sub-clinical hypothyroidism which can have elevated or normal 99mTc accumulation and primary hypothyroidism which can have normal or elevated 99mTc accumulation. Thus, the accuracy of 99mTc uptake may be more dependent on the pathological cross section of patients than the technique itself.

The emergence of artificial intelligence (AI) in nuclear medicine has driven problem solving associated with logic and reasoning (*7,8*). Developments in machine learning (ML) and deep learning (DL) provide valuable research tools, particularly for image segmentation and interpretation (*9*). The artificial neural network (ANN) provides the backbone for both ML and DL algorithms. The ANN relies on input of specific data (features) and is generally referred to a ML. More complex ANNs can produce deep architectures (high number of layers and nodes) and refers to DL. Deep ANNs are generally associated, in medical imaging, with convolutional neural networks (CNN) that use convolution and pooling layers so features can be extracted from input tensors (images) (*9,10*). While there have been historical uses of neural networks for classification of thyroid based ophthalmologic conditions and for evaluation of invitro laboratory tests, it has only been recently that DL approaches have been applied to thyroid scintigraphy. Using single photon emission computed tomography (SPECT) thyroid scintigraphy, three DL models based on AlexNet, VGGNet and ResNet architectures trained on 1430 clinical studies were modeled and compared to residents in nuclear medicine (*11*). While the investigators concluded that DL approaches perform well in thyroid scintigraphy, the role might be limited to assisting the physician in training more so than any specific clinical utility. The algorithms marginally out-performed first year residents but did not perform as well as second year residents, let alone experienced physicians. Concurrent use of the DL approaches improved the performance of residents in the order of 5% and reduced reporting time. Nonetheless, there is a need to explore potential clinical and research applications and the less complex nature of planar thyroid scintigraphy may be better suited to DL approaches. The performance of these algorithms was enhanced by a sanitized data set with the case population comprising normal (175), Grave's disease

(834) or subacute thyroiditis (421). The three DL architectures reported a high degree of recall for subacute thyroiditis, poor accuracy for normality and moderate accuracy for Grave's disease (*11*).

The aim of this investigation was to correlate each of the following with biochemical status and compare performance:

1. the percentage uptake of 99mTc,
2. visual correlation of thyroid activity in the thyroid,
3. machine learning (ML) algorithms using an artificial neural network (ANN) and
4. deep learning (DL) approaches using a convolutional neural network (CNN).

Method

There were 123 patients retrospectively analysed in the study sample (90.2% female) with a mean age of 35 years (range of 10-70 years). The mean intravenous dose of 99mTc was 153.4 MBq. The 99mTc based thyroid uptake was determined using background corrected thyroid regions of interest and a measured standard. All calculations were decay corrected and accounted for residual dose in the syringe post injection. Image features extracted included both background corrected and non-corrected total thyroid, left side and right side area (cm^2), counts and counts per pixel. The ratio of right to left lobe area (cm^2), counts and counts per pixel was also determined with and without background correction. Additionally, thyroid to background for total thyroid, right lobe and left lobe were determined (trapping index). The dose relative to the total counts was also calculated and visual classification of thyroid activity relative to the salivary glands was recorded. Biochemical features included the free T4 (pmol/L), free T3 (pmol/L) and TSH (μIU/mL). The biochemical status of the patient was determined (table 1) which was further stratified as ternary (hypothyroid, euthyroid or hyperthyroid) and binary (hyperthyroid or not hyperthyroid) (*1-6,12,13*). Other imaging features were also recorded (eg. hot or cold nodule, multi-nodular goitre etc). Only 96 patients had both imaging features and biochemical status. The investigation was approved by institutional ethics committee.

Conventional statistical analysis was undertaken using JMP 15.2.1 (SAS Institute) software. The statistical significance was calculated using Chi-Square analysis for nominal data and Student's *t* test for continuous data. The Pearson Chi-Square (X^2) test was employed for categorical data with normal distribution and the Likelihood Ratio Chi-Square (G^2) test for categorical data without normal distribution. The *F* test analysis of variances was used to determine statistically significant differences within grouped data. A *P* value less than 0.05 was considered significant. Inter-observer correlation was evaluated with Chi-Square analysis and inter-observer reliability measured using Cohen's Kappa coefficient.

The data was also evaluated using an ANN (Neural Analyser version 2.9.5). There were 42 input variables in 123 patients (instances) using a binary classification of hyperthyroid or euthyroid. A 50:25:25 split of 96 valid instances (excluded missing biochemistry data) was used for training, selection and testing. The initial network architecture included 16 scaling layer inputs, 3 hidden layers of 6, 4 and 3 nodes respectively, using a logistic activation function (defines the output of each node based on its input) for a single probabilistic layer (binary). The weighted squared error method was used to determine the loss index and the neural parameters norm was used for the regularisation method. A Quasi-Newton training method was employed using gradient information to estimate the inverse Hessian for each iteration of the algorithm (no second derivatives). The loss function associated with the training phase estimates the error associated with the data the neural network observes.

The single anterior neck image for the 96 patients was evaluated by three independent expert physicians blinded to other image and biochemical features. Each scan was recorded based on visual appearances as euthyroid, hypothyroid or hyperthyroid. On completion of the stratification, each physician re-evaluated the ternary status with the visual inspection supplemented by the calculated thyroid uptake (%). Physician rating was determined by majority group consensus.

Individual, non-annotated, anterior neck images representative of each patient was evaluated using a CNN classifier (MatLab R2020b Deep Learning Toolkit Deep Network Designer App). Given the lack of discriminatory power of either visual evaluation or thyroid uptake quantitation using various cut-off values to identify hypothyroidism, the CNN classifier was designed to identify hyperthyroidism or not hyperthyroidism (euthyroid and hypothyroid). Given the lack of complexity in the image data, the architecture used for the CNN was initially modelled on a binary version of AlexNet with 25 layers but optimised using a model that resembled the VGG-19 CNN architecture with a binary output and 30 layers (table 2 and figure 2). All patient files were trained and validated thrice (70:30 data random split) for each of three image types; white on black greyscale, black on white greyscale, and the magnitude spectrum of the Fourier transformation of each image (figure 3). Specific parameters included an ADAM (adaptive movement estimation) stochastic gradient descent optimiser algorithm, and initial learn rate of 0.001, a maximum of 50 Epochs (1 Epoch = 1 iteration) and randomisation with each Epoch.

Situation analysis was undertaken using the confusion matrix for classifier prediction including true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs). A number of performance indicators can be gleaned from the confusion matrix including:

- Precision = TPs / (TPs + FPs)
- Recall = TPs / (TPs + FNs)
- Accuracy = (TPs + TNs) / (TPs + TNs + FPs + TNs)
- F1-score = 2 × TPs / (2 × TPs + FPs + FNs)

Results

Statistical Analysis

For the 123 patients the mean thyroid uptake was 4.4% (95% CI 3.3-5.5%) with a median of 2.2% (table 3). Among the visual findings, 9 patients had increased uptake associated with primary hypothyroidism, 22 increased uptake for Grave's disease, 9 multinodular goitres and 2 nodular thyroids, 28 normal morphology, 3 goitres, 11 had reduced or absent uptake, 7 had autonomous glands with contralateral suppression (6 on the right),

there were 24 cold nodules (16 on the right), 8 hot nodules (4 on the right). Table 4 summarises other key demographic data.

The mean age of hypothyroid patients (48.0 years) was statistically higher than for biochemically euthyroid patients (33.7 years) (P=0.041) but not against hyperthyroid patients (36.7 years). There was also a weak positive correlation between age and thyroid size (P<0.001; R^2=0.117). No other statistically significant relationships were noted for patient age. Men demonstrated a statistically higher mean thyroid area (48.5 cm^2) than women (32.2 cm^2) (P=0.003). There was also a statistically significant difference in the biochemical status (P=0.019) with a disproportionately high representation of hyperthyroidism for males and lower euthyroid rate. Given then lower representation of males in the thyroid scan population, this observation may reflect lower presentation rates for males in the absence of markedly abnormal thyroid function driving more pressing symptoms. No other statistically significant relationships were noted for patient gender or patient dose (MBq).

There was no statistically significant correlation with right lobe to left lobe ratio (P=0.672), thyroid area (P=0.166) or background CPP (P=0.416). The increase in thyroid uptake associated with increasing total counts (P<0.001; R^2=0.458) and total CPP (P<0.001; R^2=0.356) were expected. There was also statistically significant relationships between increasing thyroid uptake and increasing thyroid to background ratios (P<0.001; R^2=0.376). The mean thyroid uptake was statistically higher (P<0.001) when the scan appearance showed, relative to thyroid appearance, no salivary activity (9.1%) than for faint thyroid activity (2.5%), less than thyroid activity (1.7%), equal to thyroid activity (1.1%) and greater than thyroid activity (0.4%). A positive correlation between thyroid uptake and both free T4 (P<0.001; R^2=0.351) and free T3 (P<0.001; R^2=0.365) was noted, however, no correlation was noted between thyroid uptake and TSH (P=0.695; R^2=0.002).

Biochemical status demonstrated a statistically significant difference (P<0.001) for the mean thyroid uptake stratified as hyperthyroid (9.5% with 95% CI 7.1-12.0%), hypothyroid

(4.0% with 95% CI 1.3-6.7%) and euthyroid (2.5% with 95% CI 0.9-4.2%). Hypothyroid studies had a higher mean thyroid uptake than euthyroid because of the primary hypothyroidism cases. Excluding primary hypothyroidism, there is no statistically significant difference in thyroid uptake between hypothyroid and euthyroid, or with sub clinical hyperthyroid and suppressed hyperthyroid. While 4.5% is a cut-off that is 100% sensitivity for standard hyperthyroidism, clinically hyperthyroid with suppression and sub clinical hyperthyroidism (both biochemically) are not identified by this normal range.

The optimised cut-offs for thyroid uptake against biochemical status was 0.45% to 4.5% although the lower cut-off is a poor discriminator for hypothyroidism against euthyroid. For biochemical hyperthyroidism, 70.8% of cases had an uptake greater than 4.5% while 29.3% fell below 4.5%. 100% of those below 4.5% were biochemically subclinical hyperthyroidism or T3 toxicosis. 100% of “true” hyperthyroidism cases biochemically had uptake above 4.5%. Conversely, 27.8% of hypothyroidism cases had uptakes above 4.5%. There were no (zero) hypothyroidism cases that had uptake values below the 0.45% cut-off (all values below this were hyperthyroid or euthyroid biochemically). In the biochemically euthyroid range, only 6% had an uptake above 4.5% and 2% below 0.45%.

Using the ternary classification, a thyroid uptake above 4.5% had a sensitivity of 70.8% for detecting hyperthyroidism and a specificity of 88.2%. A thyroid uptake below 0.45% had a sensitivity for hypothyroidism of 0% and specificity of 95.9% (figure 4, left). A broader biochemical classification of hyperthyroidism saw the sensitivity of the 4.5% cut-off reach 100% with specificity of 88.2% (figure 4, right).

Based on the ternary biochemical status, there was a statistically higher thyroid area for hyperthyroidism (40.7 cm^2) than hypothyroidism (29.5 cm^2) and euthyroid (33.0 cm^2) (P=0.049). With reference to figure 1, scintigraphic appearances of thyroid activity relative to salivary gland activity correctly identified 70.3% of hyperthyroid studies, 0% of hypothyroid studies and 62.7% of euthyroid studies (table 5). Excluding sub-clinical hyperthyroidism and T3 toxicosis, 94.1% of hyperthyroidism studies were identified using

the visual criteria. Table 5 also provides an outline of true positive rate (recall) for each set of cut-off values against the biochemical status.

Machine Learning

There were 42 input variables in 96 patients (instances) using a binary classification of hyperthyroid or euthyroid. The heat map / correlation matrix identified a number of redundant variables and the highest correlation scores associated with TSH (0.888), appearance of salivary glands on scans (0.627), free T4 (0.575), percentage uptake (0.501) and free T3 (0.491); consistent with the conventional statistical analysis. The network architecture included 16 scaling layer inputs, 3 hidden layers of 6, 4 and 3 nodes respectively. The initial value of the training loss was 1.5473, and the final value after 105 iterations is 0.0172. The initial value of the selection loss was 1.5570, and the final value after 105 iterations is 1.1895.

A growing inputs method was used to calculate the correlation for every input against each output in the data set. Beginning with the most highly correlated inputs, progressively decreasing correlated inputs were added to the network until the selection loss increased. The final architecture of the neural network reflects the optimised subset of inputs with the lowest selection loss. In this case, the selection loss and the training loss identified the optimal number of inputs to be 4 with a training loss optimised at 0.0298 and the selection loss of less than 0.0001. The final architecture was 4 scaling layer inputs, 3 hidden layers of 6, 4 and 1 nodes respectively, unscaling layer and a single binary probabilistic layer (figure 5).

A number of metrics were employed to test the final architecture using a subset of the original patient data. Receiver operator characteristics (ROC) analysis demonstrated an area under the curve (AUC) of 0.933. This correlates with a sensitivity of 100%, a specificity of 80% and a classification accuracy of 0.846. This was consistent with scores of 0.60 for precision, 0.75 for F1 score (harmonic mean of sensitivity and precision), 0.693 for Matthew's correlation (correlation between targets and outputs), and 0.8 for Youden's index (probability of a correct decision as opposed to guessing). The cumulative gain

analysis demonstrates the benefit of using the developed model over a random guess. The positive cumulative gain shows the percentage of positive instances found (Y-axis) against the percentage of population (X-axis). Similarly, the negative cumulative gain shows the percentage of the negative instances found against the percentage of population. The straight line represents a random classifier. The broader the separation, the better the predictive model (figure 6). Since the instances ratio provides maximum separation (maximised percentage of positive and negative instances), the instances ratio 0.40 has a maximum gain score of 0.8. Specifically, but individually, hyperthyroidism is predicted by a 99mTc uptake value over 5.7%, free T4 below 20 or above 34 pmol/L, free T3 above 9.8 pmol/L and TSH less than 5.5 μIU/mL. In combination, these scaled and weighted input features of the neural network can be expressed mathematically enhancing the collective predictive capability.

Deep Learning

Preliminary network development demonstrated over-fitting beyond 30 iterations (Epochs) and, therefore, the maximum Epoch number was re-set to 30. The results of the triplicated training and validation passes are summarised in table 6. The variations in validation accuracy reflect the smaller dataset and the random assignment of cases to training and validation. No statistically significant differences (grouped F test) were noted between training or validation accuracy for different types of input tensors (P=0.161 for training accuracy and P=0.531 for validation accuracy) despite the higher accuracy for white on black and the lower accuracy for the magnitude spectrum. A direct comparison of white on black against the magnitude spectrum showed P=0.068 for training accuracy and P=0.280 for validation accuracy).

Discussion

While thyroid scintigraphy is a well-established technique for the assessment of thyroid function, there is variable opinion on the role in identifying low thyroid uptake compared to high thyroid uptake to guide radionuclide therapy. Thyroid scintigraphy is useful in the evaluation of hyperthyroidism to differentiate causes and guide therapy (*14*). While the specific scintigraphic patterns associated with thyroid pathology do not easily differentiate

biochemical status of the patient (figure 7), scintigraphic imaging does provide useful information to identify patients suitable for radioiodine therapy (*14*). Despite being in widespread use for the purpose internationally, 99mTc-pertechnetate based thyroid uptake is not considered suitable in some circles for guiding therapeutic dosage of radioiodine (*14*). Consistent with the observations of this study, scintigraphy has a limited role in hypothyroidism (*15*).

The challenges and limitations of thyroid scintigraphy are highlighted by poor agreement of physician interpretation. It should be noted that the physician interpretation is not under normal conditions with the exclusion of patient history and biochemistry results. For the purpose of this study, however, the constrained interpretation provides a useful benchmark. Using the thyroid uptake cut-off of 0.45-4.5%, there was only 63.5% agreement with physician interpretation and utilizing the salivary gland appearance had an agreement of just 53.1% with the physician interpretation. Agreement between physicians was not strong with a range of 59.4% to 86.5% and the agreement with biochemistry grounded truth ranged from 42.7% to 68.8%. This, combined with the poor prediction utility of the salivary gland appearance contradicts the simplicity of thyroid imaging depicted in figure 1.

Using the ternary classification of euthyroid, hyperthyroid and hypothyroid, a thyroid uptake above 4.5% had a sensitivity of 70.8% for detecting hyperthyroidism and a specificity of 88.2%. A thyroid uptake below 0.45% had a sensitivity for hypothyroidism of 0% and specificity of 95.9%. Specific biochemical classification of hyperthyroidism that excluded T3 toxicosis and sub-clinical hyperthyroidism improved sensitivity of the 4.5% cut-off to 100% with specificity of 88.2%. This highlights the value of thyroid uptake with a cut-off of 4.5% in identifying patients suitable for radioiodine therapy. Given this is the primary goal and the limited role of scintigraphy in hypothyroidism of the adult population, a binary (hyperthyroidism or not hyperthyroidism) provides a more suitable evaluation. The value of an appropriate thyroid uptake cut-off is highlighted in table 5 where, for this population, binary accuracy was highest for 4.5% (82.6%) and physician interpretation augmented by the uptake value (82.3%), and poor for salivary gland appearance alone

(59.4%) and blind physician interpretation (61.0%). Indeed, the value and accuracy of 4.5% as the cut-off is reinforced by the similarity in physician interpretation with and without the uptake augmented information.

While ML was able to demonstrate improved accuracy to 100%, the algorithm relied on biochemistry not available for physician interpretation. Indeed, the grounded truth was reliant on the additional value of biochemistry insights with physician insights. In the absence of availability of biochemistry results, the ML algorithm is reliant on uptake alone. Conversely, the physician interpretation would improve substantially with the additional insights from biochemistry. In this study, regardless of the apparent performance results, ML augmentation only outperforms physician interpretation because the physician is blinded to the biochemistry results available for the ML algorithm. Nonetheless, the role of ML is not and should not be to displace physician reporting but rather to improve accuracy by eliminating error. In this instance, the ML algorithm has been shown to be an accurate second reader system that could be automated with minimal cost and resources to identify hyperthyroid patients suitable for radioiodine therapy.

In contrast to the success of ML algorithm development, the DL CNN performed poorer than both the 4.5% cut-off discriminator and the uptake augmented physician interpretation. The best results were achieved using the white on black images (80.5%). While this represents only a marginal decrease in performance compared to uptake alone (82.6%) and physician interpretation (82.3%), it should be kept in mind that the CNN was only trained on a single anterior neck image. The CNN did not have inputs for either the thyroid uptake percentage or the biochemistry results. As a result, the comparative performance should be considered the physician rating without uptake values. In this regard, the 80.5% binary accuracy of the CNN was superior to the physician interpretation (61.0%) and the visual classification against salivary gland appearance (61.5%). While this does not suggest displacement of physician interpretation, it does indicate that accuracy of physician reporting could be improved using the CNN algorithm in circumstances where biochemistry results are not available.

Conclusion

Thyroid scintigraphy is useful in identifying hyperthyroid patients suitable for radioiodine therapy. Physician interpretation relies on an accurate thyroid function assessment (uptake) and an appropriately validated cut-off for the patient population (4.5% in this population). An inappropriate cut-off significantly undermines accuracy. ML ANN algorithms can be developed to improve accuracy as second readers systems when biochemistry results are available. DL CNN algorithms can be developed to improve accuracy in the absence of biochemistry results. ML and DL do not displace the role of the physician in thyroid scintigraphy but could be used as second reader systems to minimize errors and increase confidence.

Acknowledgement

The authors would like to thank the 3 physicians who performed the visual analysis of the images. The authors would also like to thank Hugo Currie from the Riverina Anglican College in Wagga Wagga, Australia, for producing the Fourier magnitude spectrum images for analysis.

Key Points

Question: Can ML and DL approaches improve semantic evaluation of thyroid scintigraphy and uptake in hyperthyroidism?

Pertinent findings: ML algorithms can be developed to improve accuracy as second readers systems when biochemistry results are available. DL CNN algorithms can be developed to improve accuracy in the absence of biochemistry results.

Implications for patient care: ML and DL do not displace the role of the physician in thyroid scintigraphy but could be used as second reader systems to minimize errors and increase confidence.

References

1. Atkins H, Richards P. Assessment of thyroid function and anatomy with technetium-99m as pertechnetate. *J Nucl Med*. 1967;9:7-15.

2. Maisey MN, Natarajan TK, Hurley PJ, Wagner HN Jr. Validation of a rapid computerized method of measuring 99mTc pertechnetate uptake for routine assessment of thyroid structure and function. *J Clin Endocrinol Metab*. 1973;36:317-322.

3. Ramos CD, Wittmann DEZ, de Camargo Etchebehere ECS, Tambascia MA, Silva CAM, Camargo EE. Thyroid uptake and scintigraphy using 99mTc pertechnetate: standardization in normal individuals. *Sao Paulo Med J*. 2002;120:45-48.

4. Hamunyela RH, Kotze T, Philotheou GM. Normal reference values for thyroid uptake of technetium-99m pertechnetate for the Namibian population. *J Endocrin, Metab & Diabetes of South Africa*. 2013;18:142-147.

5. Macauley M, Shawgi M, Ali T, et al.. Assessment of normal reference values for thyroid uptake of technetium-99m pertechnetate in a single centre UK population, *Nucl Med Communications*. 2018;39:834-838.

6. Currie G, Dixon C, Vu T. Validation of a normal range for trapping index in thyroid scintigraphy. *ANZ Nucl Med J*. 2004;35:11-16.

7. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. *J Med Imag Rad Sci*. 2019;50:477-487.

8. Currie G. Intelligent Imaging: artificial intelligence augmented nuclear medicine. *J Nucl Med Technol*. 2019;47:217-222.

9. Currie G. Intelligent Imaging: anatomy of machine learning and deep learning. *J Nucl Med Technol*. 2019;47:273-281.

10. Currie G, Rohren E. Intelligent imaging in nuclear medicine: the principles of artificial intelligence, machine learning and deep learning. *Sem Nucl Med.* 2021;51:102-111.

11. Qiao T, Liu S, Cui Z, et al. Deep learning for intelligent diagnosis in thyroid scintigraphy. *J Int Med Res*. 2021 Jan;49(1):300060520982842. doi: 10.1177/0300060520982842. PMID: 33445994; PMCID: PMC7812409.

12. Alswat K, Assiri SA, Althaqafi RMM, et al. Scintigraphy evaluation of hyperthyroidism and its correlation with clinical and biochemical profiles. *BMC Research Notes*. 2020;13:324. https://doi.org/10.1186/s13104-020-05164-5

13. Wagieh S, Salman K, Bakhsh A, et al. Retrospective study of Tc-99m thyroid scan in patients with Graves' disease: is there significant difference in lobar activity? *Indian J Nucl Med*. 2020;35:122–129.

14. Mariani G, Tonacchera M, Grosso M, Orsolini F, Vitti P, Strauss HW. The role of nuclear medicine in the clinical management of benign thyroid disorders, part 1: hyperthyroidism. *J Nucl Med*. 2021;62:304-312.

15. Mariani G, Tonacchera M, Grosso M, et al. The role of nuclear medicine in the clinical management of benign thyroid disorders, part 2: nodular goiter, Hypothyroidism, and Subacute Thyroiditis. *J Nucl Med*. 2021 Jul 1;62:886-895.

Tables

Table 1: Biochemical stratification of patient studies and findings (*1-6,12,13*).

Free T3 (2-7 pmol/L)	**Free T4 (12-30 pmol/L)**	**TSH (0.45-4.5 μIU/mL)**	**Biochemical Status**	**99mTc uptake (%)**	**Comment on uptake normal range**
High	High	Low	Hyperthyroidism	> 4.5	0% false negative rate
Normal	Normal	Low	Subclinical hyperthyroidism	< 4.5 including < 0.45 or absent	0% true positive, comprised false negative or false positive hypothyroidism
High	Normal	Low	T3 toxicosis	> 4.5 or < 0.45	False positive hypothyroidism
Normal	High	Low	Thyroiditis		No cases
Low	Low	Low	Secondary hypothyroidism		No cases
Normal	Normal	High	Subclinical hypothyroidism	> 0.45 but < 4.5	100% false negative
Low or normal	Low	High	Primary hypothyroidism	> 0.45 and in over 50% of cases > 4.5	100% false negative
Normal	Normal	Normal	Euthyroid	< 4.5%	9% false positive rate (6% hyperthyroid, 3% hypothyroid)

Table 2: CNN architecture, activations and parameters.

Layer	Name	Activations	Parameters
1	Tensor Input Layer	[725,725,3]	
2	2D Convolution Layer	[239,239,64]	Weights [11,11,3,64], Bias [1,1,64]
3	Batch Normalization	[239,239,64]	Offset and scale [1,1,64]
4	ReLU Layer	[239,239,64]	
5	Max Pooling Layer	[119,119,64]	Size [3,3], Stride [2,2], Padding [0,0,0,0]
6	2D Convolution Layer	[40,40,128]	Weights 5,5,64,128], Bias [1,1,128]
7	Batch Normalization	[40,40,128]	Offset and scale [1,1,128]
8	ReLU Layer	[40,40,128]	
9	Max Pooling Layer	[19,19,128]	Size [3,3], Stride [2,2], Padding [0,0,0,0]
10	2D Convolution Layer	[19,19,256]	Weights [3,3,128,256], Bias [1,1,256]
11	Batch Normalization	[19,19,256]	Offset and scale [1,1,256]
12	ReLU Layer	[19,19,256]	
13	Max Pooling Layer	[9,9,256]	Size [3,3], Stride [2,2], Padding [0,0,0,0]
14	2D Convolution Layer	[9,9,192]	Weights [3,3,256,192], Bias [1,1,192]
15	Batch Normalization	[9,9,192]	Offset and scale [1,1,192]
16	ReLU Layer	[9,9,192]	
17	Max Pooling Layer	[4,4,192]	Size [3,3], Stride [2,2], Padding [0,0,0,0]
18	2D Convolution Layer	[4,4,192]	Weights [3,3,256,192], Bias [1,1,192]
19	Batch Normalization	[4,4,192]	Offset and scale [1,1,192]
20	ReLU Layer	[4,4,192]	
21	Max Pooling Layer	[1,1,192]	Size [3,3], Stride [2,2], Padding [0,0,0,0]
22	Fully Connected Layer	[1,1,192]	Weights [192,192], Bias [192,1]
23	ReLU Layer	[1,1,192]	
24	Dropout Layer	[1,1,192]	0.5
25	Fully Connected Layer	[1,1,86]	Weights [86,192], Bias [86,1]
26	ReLU Layer	[1,1,86]	
27	Dropout Layer	[1,1,86]	0.5
28	Fully Connected Layer	[1,1,2]	Weights [2,86], Bias [2,1]
29	Softmax Layer	[1,1,2]	
30	Classification Layer		Cross entropy loss function

Table 3: Summary of ternary classification of thyroid function based on various published normal ranges.

Normal range	Euthyroid	Hyperthyroid	Hypothyroid	Reference
0.45-4.5%	67.5%	26.8%	7.7%	*6*
0.4-1.7%	35.0%	61.0%	4.0%	*3*
0.4-4.0%	65.0%	31.0%	4.0%	*4*
0.3-3.4%	57.7%	38.2%	4.1%	*2*
0.2-2.0%	43.1%	52.8%	4.1%	*5*
Biochemical status	53.1%	27.1%	19.8%*	*11*
Salivary classification	44.8%	50.0%	5.2%	-
Physician visual rating	51.0%	43.8%	5.2%	-
Physician rating with uptake value	64.6%	29.2%	6.3%	-

*15.6% were hypothyroid without suppression of uptake (2.1% autonomous, 2.1% secondary hypothyroidism, 11.5% primary hypothyroidism, 4.2% subclinical hypothyroidism).

Table 4: Summary of key variables.

	Mean	**95% CI**
Right lobe activity to left lobe total count ratio	1.5	1.03-2.02
Right lobe activity to left lobe CPP ratio	1.29	0.98-1.60
Area	33.8 cm^2	31.1-36.5
Size right	3092 pixels	2848-3340
Size left	2937 pixels	2662-3212
Thyroid : background ratio	4.06	3.43-4.69
Right	4.01 CPP	3.49-4.52
Left	4.08 CPP	3.28-4.89
Dose to total counts ratio	4.85	3.44-6.26
FT4	21.1 pmol/L	18.1-24.2
FT3	7.1 pmol/L	5.1-9.1
TSH	4.2 pmol/L	2.3-6.1

Table 5: Summary of ternary classification of thyroid function based on recall against biochemical status. Accuracy is also provided for binary classification.

Normal range	Euthyroid	Hyperthyroid (*)	Hypothyroid	Accuracy**
0.45-4.5%	71.4%	66.6% (100%)	0%	82.6%
0.4-1.7%	49.0%	74.1% (94.1%)	0%	51.0%
0.4-4.0%	86.3%	63.0% (94.1%)	0%	77.1%
0.3-3.4%	74.5%	63.0% (94.1%)	0%	68.8%
0.2-2.0%	58.8%	74.1% (94.1%)	0%	59.4%
Salivary classification	62.7%	70.3% (94.1%)	0%	61.4%
Physician rating	72.5%	63.0% (89.5%)	0%	61.0%
Physician rating with uptake	88.2%	70.3% (100%)	0%	82.3%

*excluding sub-clinical hyperthyroidism and T3 toxicosis.
**binary accuracy for reference to table 6

Table 6: Summary of triplicate training and validation binary results (hyperthyroid or not hyperthyroid) for the 30-layer CNN architecture. The corresponding binary accuracies of the best performing thyroid uptake cut-offs, visual classification against salivary activity relative to thyroid activity, and physician rating are included for comparison.

Input tensor	**Training accuracy**	**Training loss**	**Validation accuracy**	**Validation loss**	**Mean validation accuracy**
White on black	82.1%	0.420	75.9%	0.536	80.5%
White on black	94.0%	0.225	79.3%	0.602	
White on black	91.0%	0.218	86.2%	0.414	
Black on white	83.6%	0.383	82.8%	0.405	78.2%
Black on white	80.6%	0.452	72.4%	0.544	
Black on white	91.0%	0.232	79.3%	0.690	
Magnitude spectrum	76.1%	0.459	75.9%	0.530	75.9%
Magnitude spectrum	74.6%	0.508	72.4%	0.542	
Magnitude spectrum	85.1%	0.306	79.3%	0.380	
Mean	84.2%	0.356	78.2%	0.516	
Initial 25-layer CNN					69.0%
Conventional metrics					**Binary accuracy**
Normal cut-off 4.5%					82.6%
Normal cut-off 4.0%					77.1%
Salivary classification					61.5%
Physician rating					61.0%
Physician rating with uptake					82.3%

List of figures

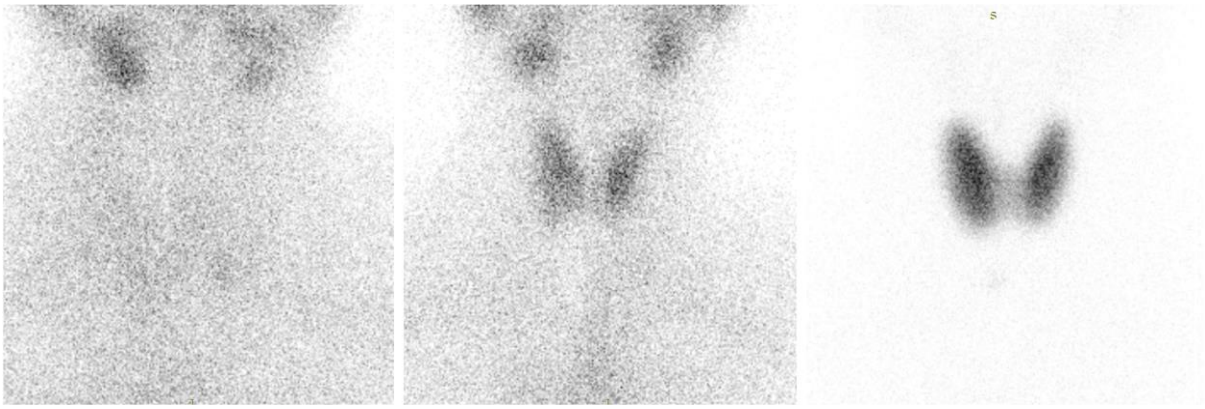


Figure 1: Intuitive, but sometimes inaccurate, visual evaluation of thyroid status relative to salivary gland activity. Left with salivary gland activity exceeding thyroid gland activity suggests hypothyroidism. Middle with salivary gland activity and thyroid gland activity being similar (within the same scale) suggests euthyroid. Right with salivary gland activity not apparent relative to thyroid activity suggests hyperthyroid. All images are 99mTc pertechnetate using high resolution, parallel hole imaging.

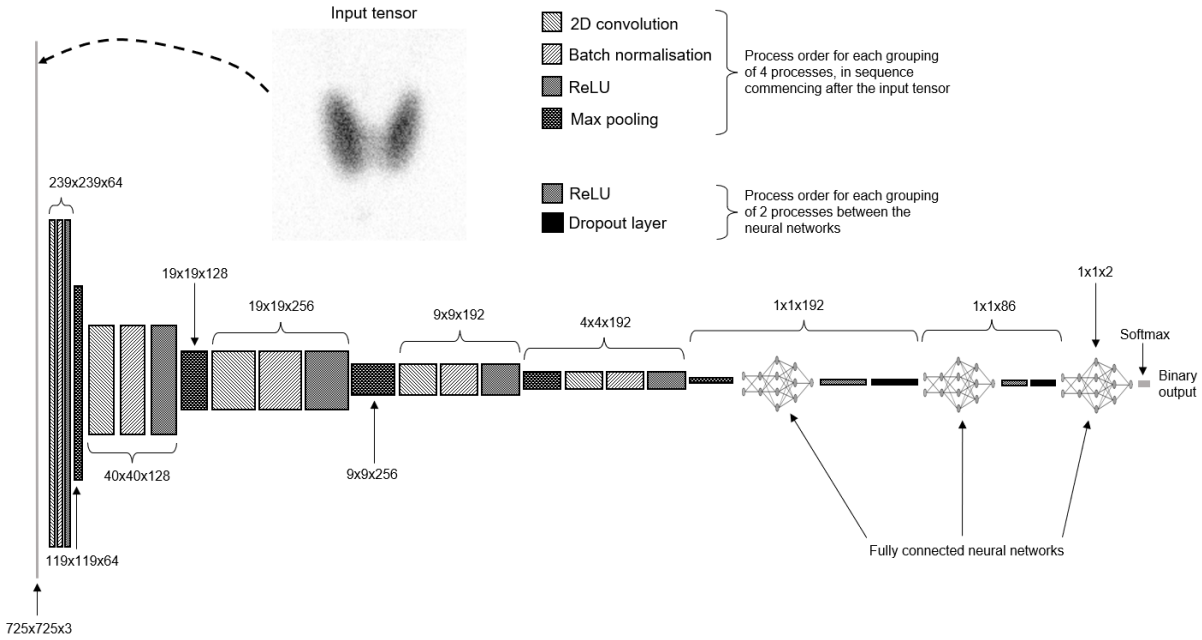


Figure 2: CNN architecture.

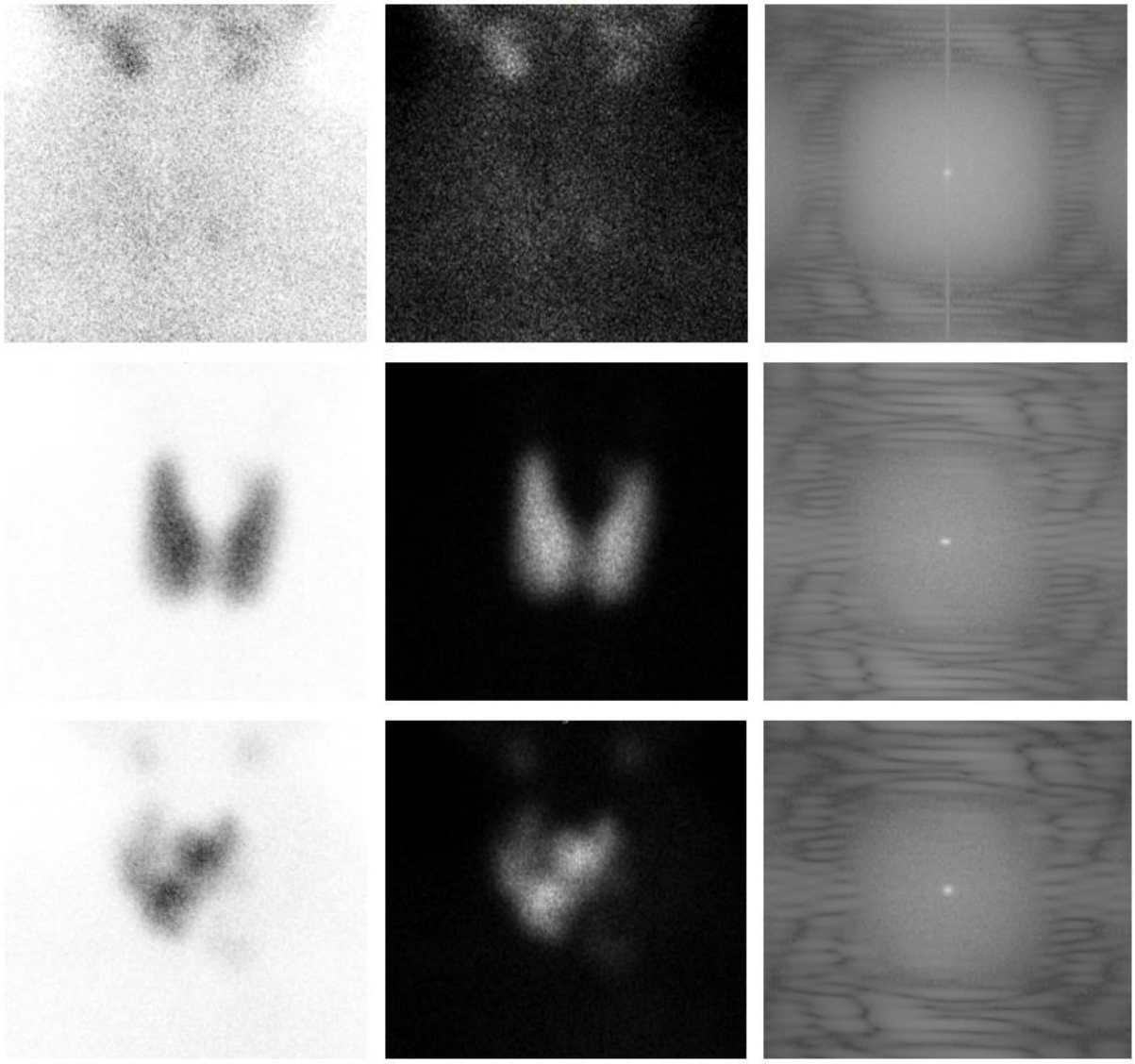


Figure 3: Three example patients (top, middle and bottom) with each of black on white (left), white on black (centre), and magnitude spectrum from Fourier transformation (right) used as inputs for the CNN.

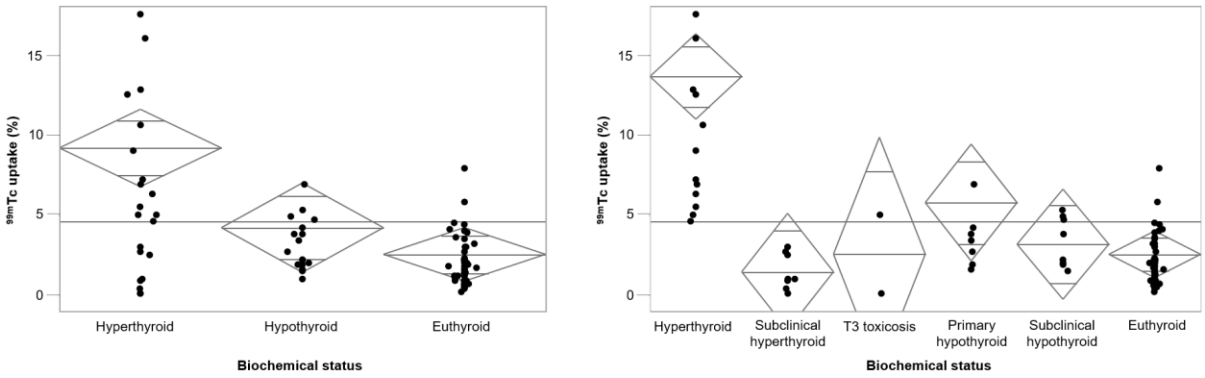


Figure 4: Left; ternary biochemical status classification against thyroid uptake. Right; broader biochemical status classification against thyroid uptake. The horizontal line represents overall mean while the diamonds represent the class mean and 95% confidence intervals.

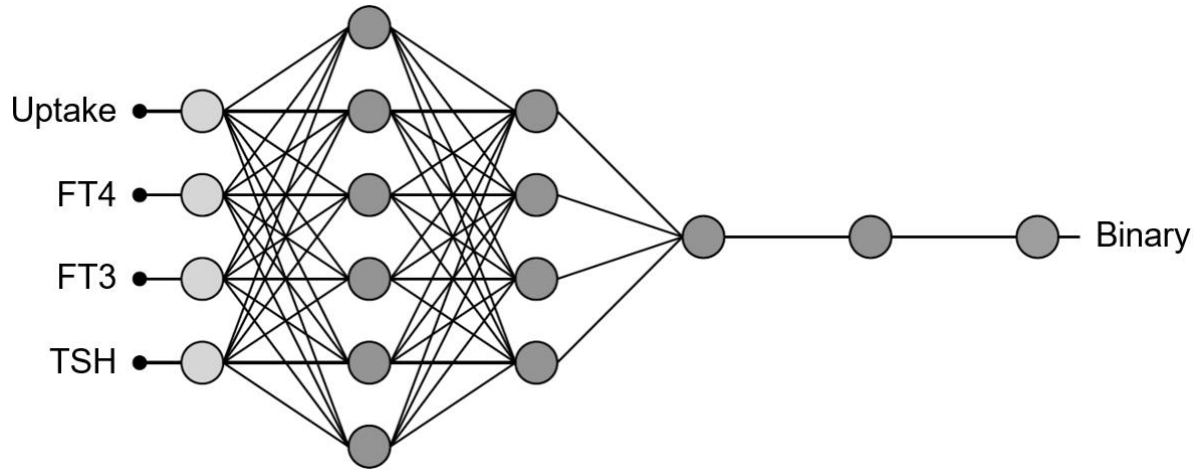


Figure 5: Final architecture of the trained and validated neural network.

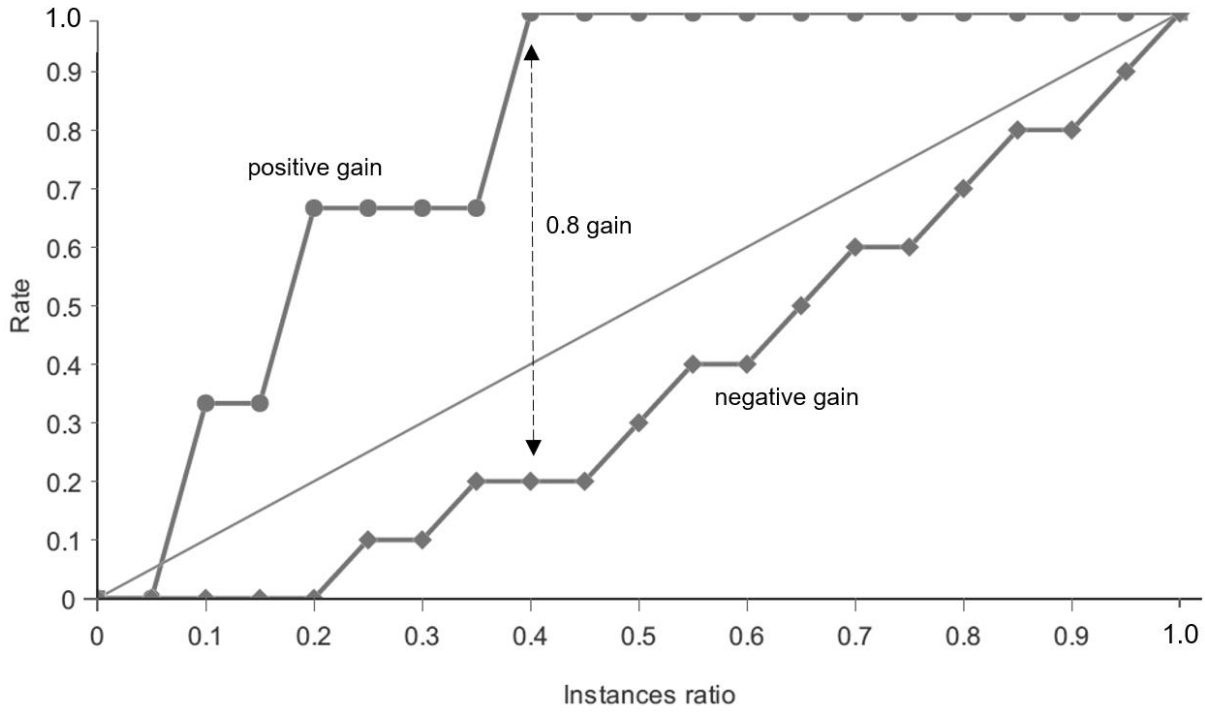


Figure 6: The cumulative gain chart demonstrating maximum separation of positive and negative curves to provide a cumulative gain score of 0.8 and instances ratio of 0.4 (black dashed arrow).

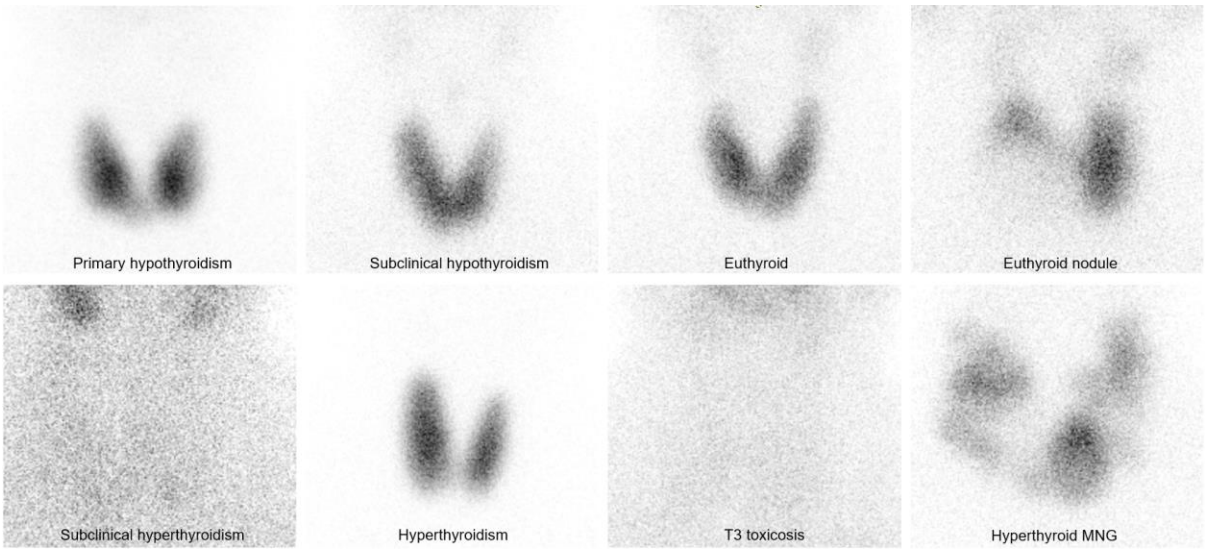


Figure 7: Various scintigraphic appearances of thyroid pathology using parallel hole (high resolution) collimation and 99mTc pertechnetate.