ThyExp: an explainable AI-assisted decision making toolkit for thyroid nodule diagnosis based on ultra-sound images

ThyExp: An explainable AI-assisted Decision Making Toolkit for Thyroid Nodule Diagnosis based on Ultra-sound Images

Jamie Morris
The University of Reading
United Kingdom, Reading
jkm013@hotmail.com

Zehao Liu
Newcastle University
United Kingdom, Newcastle
z.liu83@newcastle.ac.uk

Huizhi Liang
Newcastle University
United Kingdom, Newcastle
huizhi.liang@newcastle.ac.uk

Sidhartha Nagala
Royal Berkshire Hospital
United Kingdom, Reading
siddini@me.com

Xia Hong
The University of Reading
United Kingdom, Reading
x.hong@reading.ac.uk

ABSTRACT

Radiologists have an important task of diagnosing thyroid nodules present in ultra sound images. Although reporting systems exist to aid in the diagnosis process, these systems do not provide explanations about the diagnosis results. We present **ThyExp** – a web based toolkit for it use by medical professionals, allowing for accurate diagnosis with explanations of thyroid nodules present in ultrasound images utilising artificial intelligence models. The proposed web-based toolkit can be easily incorporated into current medical workflows, and allows medical professionals to have the confidence of a highly accurate machine learning model with explanations to provide supplementary diagnosis data. The solution provides classification results with their probability accuracy, as well as the explanations in the form of presenting the key features or characteristics that contribute to the classification results. The experiments conducted on a real-world UK NHS hospital patient dataset demonstrate the effectiveness of the proposed approach. This toolkit can improve the trust of medical professional to understand the confidence of the model in its predictions. This toolkit can improve the trust of medical professionals in understanding the models reasoning behind its predictions.

CCS CONCEPTS

• **Information systems** → **Expert systems**; • **Applied computing** → **Health care information systems**; • **Computing methodologies** → **Feature selection**; **Image representations**.

KEYWORDS

Artificial Intelligence Assisted disease diagnosis, Thyroid Nodule

ACM Reference Format:
Jamie Morris, Zehao Liu, Huizhi Liang, Sidhartha Nagala, and Xia Hong. 2023. ThyExp: An explainable AI-assisted Decision Making Toolkit for Thyroid Nodule Diagnosis based on Ultra-sound Images. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615131

CIKM'23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615131

1 INTRODUCTION

Thyroid cancer is the most common malignant endocrine tumour, with an annual incidence in the UK of 5.8 per 100,000 per year. Thyroid nodules may have benign or malignant pathology. Current diagnostic work involves ultrasound and needle biopsy. However, despite repeated needle biopsies, up to 30% of lumps only yield indeterminate test results. The risk of malignancy within these indeterminate lumps is 20%–30%. Patients are often recommended for diagnostic surgery to rule out cancer. Better preoperative diagnosis would reduce unnecessary operations and improve the management of patients with thyroid lumps. British Thyroid Association (BTA) recommends research to develop more accurate techniques in diagnosing thyroid nodules. Ultrasound classification of thyroid nodules is routinely performed in clinical practice. However, inter-observer reliability to classify thyroid nodules on ultrasound exists. The development of a robust AI software system would assist clinicians with their decision-making in classifying thyroid nodules.

To prevent excessive treatment and over-diagnosis, risk stratification is important. The the American College of Radiology (ACR) Thyroid Imaging Reporting and Data System (ACR TI-RADS) [10] shown in Fig. 1 provides radiologists with a standardised TI-RADS risk stratification system. Radiologists utilise their learnt knowledge of features and characteristics to classify a thyroid nodule as either malignant or benign. The key characteristics and features of an ultrasound image consist of the *Composition*, *Echogenicity*, *Shape*, *Margin* and *Echogenic foci*. Each feature has a corresponding TR score. The combination of these features and the added total TR scores of these features help to determine which TR-level an image is and whether it is benign or malignant as well as whether a biopsy is recommended for the nodule or not. For example, a radiologist could determine a thyroid nodule to have the *Shape* of *"wider than tall"* or *"taller than wide"*. If the shape of the thyroid nodule is *"wider than tall"*, then 0 points are added to the total score. Alternatively, if the nodule is *"taller than wide"*, 3 points are added to the total score, resulting in the TR-level label classification being at least a TR3, mildly suspicious.

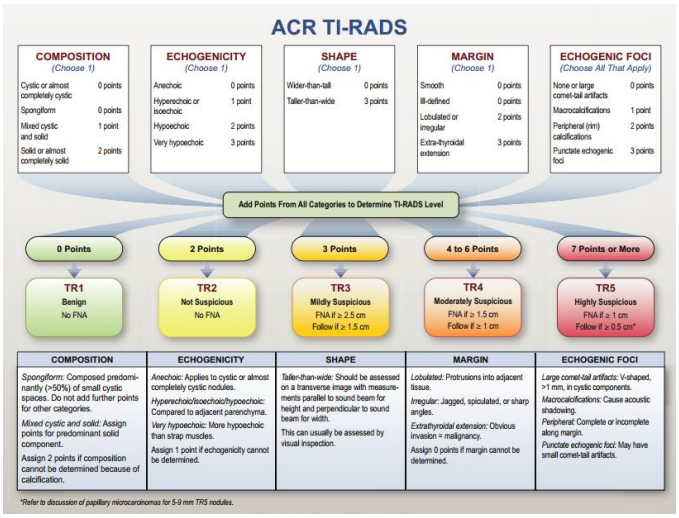

Figure 1: ACR TI-RADS flow chart [10]

Artificial Intelligence (AI) models and tools have been developed to aid radiologists in their diagnosis of thyroid nodules [1, 3]. Many existing well-performed deep learning models [12] only conduct TR-level classification or binary classification (malignant or benign). Deep learning models are usually black box models and lacking explanations, which brings hurdles for the practical and trustworthiness of these systems. This paper demonstrates **ThyExp**, a web-based explainable AI-assisted decision making toolkit for thyroid nodule diagnosis. The toolkit can achieve high accuracy results on a real-world patient dataset collected from a UK NHS hospital. Together with the results, it can provide explanations about the diagnosis results to users.

2 RELATED WORK

Existing systems available to the public are limited in this area of work, they mostly consists of theoretical implementations of Artificial Intelligence models in the diagnosis of thyroid nodules. ThyNet [9] allowed for the classification of thyroid nodules as either malignant or benign. ThyNet uses a combination of 3 already existing AI architectures, ResNet, ResNeXt and DenseNet. Another study [11] attempted to develop an end-to-end network based on ResNet and YOLOv2. These models did not provide explanations to users about their results.

Currently, there are only three other FDA-approved AI software (Koios [6], Samsung S-Detect 2 [4], AmCAD [2]). All the current commercially available software has its own limitations. In summary, the existing works do not provide explanatory diagnostic results, and instead focus on achieving a high binary diagnosis accuracy. The authors are among the few in the UK who have developed working AI software in this field. Our software bridges the gap of providing highly accurate diagnosis predictions with explanations to users.

3 THE PROPOSED TOOLKIT

The proposed toolkit **ThyExp** is a web-based application that consists of both frontend web pages and a backend server API. The toolkit mainly includes the annotation tool and the diagnosis tool. The annotation tool provides the ability to annotate, update, and delete the annotations of ultrasound images. The diagnosis tool allows for the model training and classification of ultrasound images. It is implemented as an API server. The API receives the necessary parameters and completes the model training and returns their results. We discuss the two tools in the following sub sections.

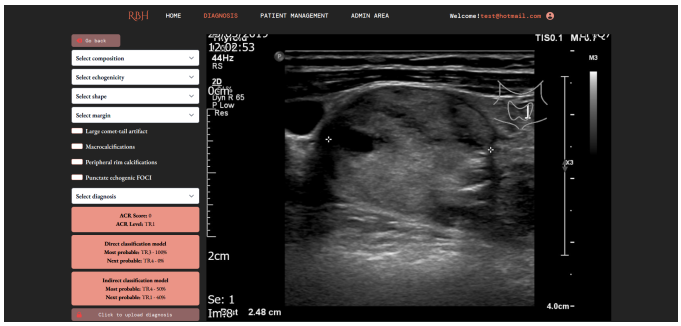

Figure 2: The screenshot of Diagnosis tool

3.1 Annotation Tool

Registered medical professional users can access this tool. They can see the uploaded images, the decision label options, and the diagnostic TR-level label options. The decision label options consist of the *Composition* (4 classes), *Echogenicity* (4 classes), *Shape* (2 classes), *Margin* (4 classes), and *Echogenic FOCI* (4 classes). This tool provide selection options of each class label name and its corresponding TR score. Following the ACR TI-RADS diagnostic rules, it will calculate the scores and suggests the ACR TI-RADS TR-level label based on the selected class label options of each decision label. These labels or annotations are then stored in the database. After quality checking by domain experts, the annotated images can be used as ground truth dataset and becomes a part of the training data for machine learning (ML) models.

From the implementation point of view, initially the tool's backend server converts the annotation ultrasound image to its Base64 image equivalent, these data are then sent in a POST request to the REST API server. The server API processes the annotation data uploaded. The ultrasound image base64 image data is converted to a JPEG, a universally unique identifier (UUID) is generated. The decision labels of the thyroid nodule, alongside the generated UUID are stored in the database. The converted JPEG is stored on the servers disk space, using its UUID as the filename for identification. Upon accessing the annotation tool, the stored annotations are loaded from the server by requesting the UUIDS from the database, and obtaining the decision labels and JPEG images by reversing the storage process. Thus converting the JPEGs to Base64 equivalents and returning these values to the front-end website.

3.2 Diagnosis Tool

The diagnosis tool allows for a medical professional to upload an ultrasound image and be presented with diagnosis TR-level label classification of a thyroid nodule. Fig. 2 shows the screenshot of the diagnostic tool. Different with other models or tools, this tool can

present the decision labels as the explanation of why this image has been classified as the predicted diagnosis TR-level label.

This tool can incorporate different types of ML based prediction models. The users can select a ML model, and begin training of the selected model for a user selected epoch count. Early stopping is implemented to prevent over-fitting of the model during this training process. The users can select a ML model and set up training percentage and number of training epochs. To achieve high accuracy performances, in this paper, we adopt the state-of-the-art model LTQ (Local texture quantization) [5] model, as the ultrasound image dataset are noisy and limited in quantity. The LTQ model transforms images into index grids, using its quantized local texture and allows for the classification of an image based on its index grids [5]. The LTQ model directly predicts the TI-RADS TR-level without utilising the decision labels, which provides no explanation of their results to users.

In this study, we propose a method of TR-level prediction with explanation called **LTQ-E** to improve both explanability and prediction accuracy of ultrasound image diagnosis. **LTQ-E** model, incorporates methodologies extended from the main LTQ [5] model, involving two steps:

- Step 1: Apply LTQ models to predict each of the four decision labels;
- Step 2: Predict the final diagnosis TR-level label based on the predicted decision labels of Step 1.

In Step 1, predictions are made on each decision label of a given image using the trained LTQ model. Since the TR-level scores can be calculated from decision labels, the presenting of decision labels helps to explain the TR-level prediction. The same model structure is used for the TR-level prediction, but only modified the output layer to fit the classification class of each decision label.

Step 2 is to get the TR-level label based on the predicted decision labels of Step 1. The proposed **LTQ-E** model extracts features (embedding) from each trained decision label classification models. The useful information contained within the decision label model following the training process was pre-trained by taking the weights (i.e., embedding) of the last second layer of each decision label model and aggregating them into a single feature in a unified format. This feature was subsequently used as an input for classifiers such as linear classification models and a neural network model to generate TR-level label predictions.

4 EXPERIMENTS AND IMPLEMENTATION

This section discusses experiments and implementation details.

4.1 Experimental Setup

RBH Dataset. A data set comprised a total of 831 ultrasound images is obtained from 307 patients from Royal Berkshire Hospital. All identifiable information was removed from the images to ensure anonymity. Each patient had between 1 and 6 images that were used to assess their TR-level diagnosis. Most patients had 3 or 2 images. The ACR TI-RADS rules were applied to convert the TR score range into TR-level labels, where $TR1$ corresponds to a score range of 0-2, $TR2$ corresponds to a score range of 2-3, $TR3$ corresponds to a score range of 3-4, $TR4$ corresponds to a score range of 4-7, and $TR5$ corresponds to a score range of 7-14. Fig. 4 displays the distribution

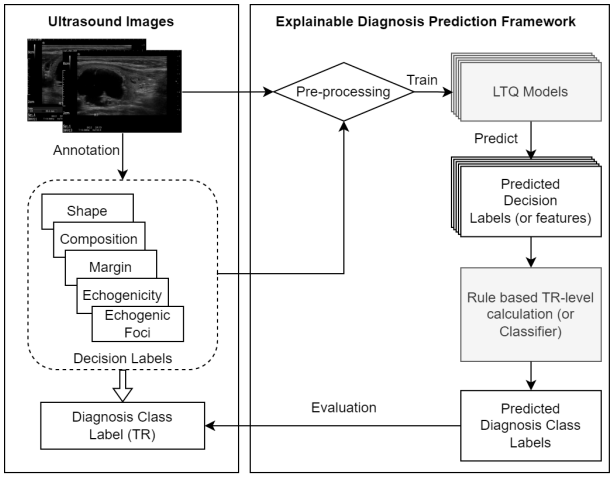

Figure 3: The framework of the proposed LTQ-E model

of TI-RADS (TR) labels at the image level. It can be seen that the majority of the ultrasound images had TI-RADS (TR) levels of TR1, TR2, TR3 and TR4, while only a small number of images were TR5.

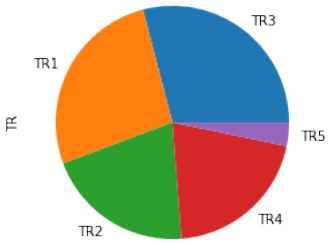

Figure 4: The distribution of Image Level TR-level labels

Pre-processing. In preparation for model training, all ultrasound images were resized to 315×500 pixels. The dataset was then split into a training set, comprised of 80% of the data and a test set comprised of the remaining 20%. To address class imbalance in the dataset, an upsampling data augmentation technique, by applying random horizontal flip to generate new images for each minor TR-level classes, was employed. This approach matched the maximum number of images for TR-level label. It ensured equal number of training samples among the classes. We only chose horizontal flipping for data augmentation, as vertical flipping could impact the 'shape' features. After upsampling, the maximum number of images per TR-level label was set to 220.

4.2 Experimental Results

We discuss the results of decision label prediction and TR-level diagnosis label prediction in this sub section.

Results of Decision Label prediction. Since the TR-level scores can be calculated from decision labels, the prediction of decision

Table 1: Results of decision labels prediction

Decision label Type	Number of Classess	Accuracy
Composition	4	0.8583
Echogenicity	4	0.823
Echogenicfoci	4	0.8398
Margin	4	0.8263
Shape	2	0.7844

labels helps to explain the TR-level prediction. To explain the diagnosis decisions, predictions are made using LTQ model on each decision label . Table 1 shows the prediction accuracy of each decision label. It can be seen that the LTQ model achieved high classification accuracy above 80% for each decision label.

Results of Diagnosis class label (TR-level) prediction. To evaluate the performances of the proposed model for the task of TR-level label prediction, The following approaches are compared on the RBH Dataset:

- LTQ model: This state-of-the-art approach predicts the TR-level label directly. This model can not provide explanations.
- LTQ-E model: The proposed model with explanations in the form of presenting decision labels. It uses the embedding of each decision label model and adopts classifiers to predict TR-level labels. We developed 4 variations using different classifiers. $LTQ\text{-}E^1$ adopts Logistic Regression as the classifier. $LTQ\text{-}E^2$ adopts SVM as the classifier, $LTQ\text{-}E^3$ adopts Random Forest as the classifier. $LTQ\text{-}E^4$ adopts a fully connected neural network as the classifier.
- LTQ-R model: The proposed model that uses TI-RADS rules to calculate the total TR score of predicted decision labels to get the TR-level label. The same with LTQ-E model, this approach explains the TR-level label in the form of presenting decision labels of each image.

The prediction accuracy results are shown in Table 2. It can be seen that LTQ model achieved an accuracy of 78.25% for the task of predicting the TR-level labels. The proposed LTQ-R and LTQ-E model had higher accuracy than the LTQ model. Among all the 4 variations, $LTQ\text{-}E^1$ achieved the best results. LTQ-E model also has better explanation than LTQ model via providing decision labels to explain the key features or characteristics that lead to the prediction of TR-level labels. LTQ-E model complies with the human decision making diagnosis process.

Table 2: Results of TR-level Prediction

Method	Model Description	Accuracy
LTQ	Predict TR-level label directly	0.7825
LTQ-R	Decision Label + TR Rules	0.82
$LTQ\text{-}E^1$	Decision label Embedding + LR	**0.87**
$LTQ\text{-}E^2$	Decision label Embedding + SVM	0.86
$LTQ\text{-}E^3$	Decision label Embedding + RF	0.85
$LTQ\text{-}E^4$	Decision label Embedding + NN	0.78

In the demo system, the best performing model is selected, that is $LTQ\text{-}E^1$ with accuracy of 87% to make TR-level diagnosis predictions.

4.3 Implementation

The ML models are implemented in Python and Pytorch. The website framework primarily utilises NodeJS and NextJS, a React framework allowing for the development of single page full-stack applications. It utilises TailwindCSS for the website styling. Prisma ORM is used for its database client and easily accessible query functionality required for the storage, and retrieval of user, account, annotation and diagnostic data. The REST API handles the classification, training of ML models, storage and retrieval of annotation data. The API was developed in Python due to its compatibility with PyTorch. PyTorch was the main library utilised for the ML models. Alongside the REST API, a Redis-server is utilised for the temporary storage of model training progress. The website uses Auth0 to handle the authorization and authentication of users for the platform. Vercel is used for hosting of the website framework. An AWS EC2 t2.large instance with 2 vCPUs and 8G of RAM were used for the hosting of the REST API. The PostgreSQL database holding all data required for the user authentication, diagnosis and annotation storage was hosted using AWS RDS db.t3.micro. The prototype system can be accessed with this link [7]. The demo video can be visited with this link [8].

5 CONCLUSIONS

This paper presents a web-based toolkit **ThyExp**. The key components and functions of the proposed toolkit consist of the ability to submit ultrasound images, annotate the images, and receive the predicted TR-level label classifications with explanations in the form of presenting decision labels. The proposed solution attempts to provide more explanations through the classification of key features that contribute to the final TR-level diagnosis decisions such as *Composition*, *Echogenicity*, *Shape*, *Margin* and *Echogenic foci*, as well as the final diagnosis decisions such as binary classification (benign and malignant) and 5-class classification (TI-RADS labels). The experimental results conducted on a real-world UK NHS hospital patient dataset show that the proposed toolkit can achieve high accuracy diagnosis of thyroid nodules.

Although the proposed prediction models are able to produce a high accuracy diagnosis, the results should be tested with bigger datasets. As well as developing our own software, we will look to further improve the usability, accuracy, and explainability of the software for the UK population.

ACKNOWLEDGEMENTS

This project is funded by the Collaborative Innovation Fund of Royal Berkshire NHS Foundation Trust and University of Reading.

REFERENCES

[1] Ziyu Bai, Luchen Chang, Ruiguo Yu, Xuewei Li, Xi Wei, Mei Yu, Zhiqiang Liu, Jie Gao, Jialin Zhu, Yulin Zhang, Shuaijie Wang, and Zhuo Zhang. 2020. Thyroid nodules risk stratification through deep learning based on ultrasound images. *Medical Physics* 47, 12. https://doi.org/10.1002/mp.14543

[2] AmCad BioMed. 2019. *AmCAD-UT.* Retrieved June 16, 2023 from https://www.amcadbiomed.com/product/ut

[3] M. Han, E.J. Ha, and J.H. Park. 2021. Computer-Aided Diagnostic System for Thyroid Nodules on Ultrasonography: Diagnostic Performance Based on the Thyroid Imaging Reporting and Data System Classification and Dichotomous Outcomes. *American Journal of Neuroradiology* 42, 3 (2021), 559–565. https://doi.org/10.3174/ajnr.A6922 arXiv:https://www.ajnr.org/content/42/3/559.full.pdf

[4] Samsung Healthcare. 2021. *S-Detector*. Retrieved June 16, 2023 from https://www.samsunghealthcare.com/en/products/digital_radiography/S-Detector

[5] Xiao Li, Huizhi Liang, Sidhartha Nagala, and Jane Chen. 2022. Improving Ultrasound Image Classification with Local Texture Quantisation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1211–1215. https://doi.org/10.1109/ICASSP43922.2022.9747883

[6] Koios Medical. 2023. *Koios DS thyroid*. Retrieved June 16, 2023 from https://koiosmedical.com/products/koios-ds-thyroid/

[7] Jamie Morris, Zehao Liu, Huizhi Liang, Sidhartha Nagala, and Xia Hong. 2023. *AI-Assisted disease diagnosis based on images*. Retrieved June 16, 2023 from https://rbh-web-xlbt.vercel.app/

[8] Jamie Morris, Zehao Liu, Huizhi Liang, Sidhartha Nagala, and Xia Hong. 2023. *Project demonstration*. Retrieved June 16, 2023 from https://drive.google.com/drive/folders/12Sj78gU-Ns14E1vCW0YD2dIY_AZ-N9QS?usp=sharing

[9] Sui Peng, Yihao Liu, Weiming Lv, Longzhong Liu, Qian Zhou, Hong Yang, Jie Ren, Guang-Jian Liu, Xiaodong Wang, Xuehua Zhang, Qiang Du, Nie Fangxing, Gao Huang, Yuchen Guo, Jie Li, Jin-Yu Liang, Hang-Tong Hu, Han Xiao, Ze-Long Liu, Fenghua Lai, Qiuyi Zheng, Haibo Wang, Yanbing Li, Erik K. Alexander, Wei Wang, and Haipeng Xiao. 2021. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. *The Lancet Digital Health* 3. https://doi.org/10.1016/S2589-7500(21)00041-8

[10] Franklin N. Tessler, William D. Middleton, Edward R. Grant, Jenny K. Hoang, Lincoln L. Berland, Sharlene A. Teefey, John E. Cronan, Michael D. Beland, Terry S. Desser, Mary C. Frates, Lynwood Hammers, Ulrike M. Hamper, Jill E. Langer, Carl C. Reading, Leslie M. Scoutt, and A. Thomas Stavros. 2017. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *Journal of the American College of Radiology* 14, 5. https://doi.org/10.1016/j.jacr.2017.01.046

[11] Lei Wang, Shujian Yang, Shan Yang, Cheng Zhao, Guangye Tian, Yuxiu Gao, Yongjian Chen, and Yun Lu. 2019. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. *World Journal of Surgical Oncology* 17, Article 12. https://doi.org/10.1186/s12957-019-1558-z

[12] Heng Ye, Jing Hang, Xiaowei Chen, Di Xu, Jie Chen, Xinhua Ye, and Dong H. Zhang. 2020. An intelligent platform for ultrasound diagnosis of thyroid nodules. *Scientific Reports* 10, 1, Article 13223. https://doi.org/10.1038/s41598-020-70159-y