scientific reports

OPEN

ThyroNet-X4 genesis: an advanced deep learning model for auxiliary diagnosis of thyroid nodules' malignancy

Xiaoxue Wang[1,3], Yupeng Niu[2,3], Hongli Liu[1], Fa Tian[2], Qiang Zhang[1], Yimeng Wang[1], Yeju Wang[1] & Yijia Li[1]✉

Thyroid nodules are a common endocrine condition, and accurate differentiation between benign and malignant nodules is essential for making appropriate treatment decisions. Traditional ultrasound-based diagnoses often depend on the expertise of physicians, which introduces a risk of misdiagnosis. To address this challenge, this study proposes a novel deep learning model, ThyroNet-X4 Genesis, designed to automatically classify thyroid nodules as benign or malignant. Built on the ResNet architecture, the model enhances feature extraction by incorporating grouped convolutions and using larger convolution kernels, improving its ability to analyze thyroid ultrasound images. The model was trained and validated using publicly available thyroid ultrasound imaging datasets, and its generalization was further tested using an external validation dataset from HanZhong Central Hospital. The ThyroNet-X4 Genesis model achieved 85.55% and 71.70% accuracy on the internal training and validation sets, respectively, and 67.02% accuracy on the external validation set. These results surpass those of other mainstream models, highlighting its potential for clinical use in thyroid nodule classification. This work underscores the growing role of deep learning in thyroid nodule diagnosis and provides a foundation for future research in high-performance medical diagnostic models.

Keywords Thyroid nodules, Deep learning, Auxiliary diagnosis, External validation, Clinical interpretability

Thyroid nodules are defined as space-occupying lesions within the thyroid gland that can be detected through imaging studies, differentiated from the surrounding thyroid tissue. These nodules can be benign or malignant, and according to recent epidemiological studies, up to 60% of the population have thyroid nodules[1], with a malignancy rate of approximately 1–5%[2]. Further research indicates that the prevalence of thyroid nodules over 0.5 cm in diameter is 20.43%, with 8–16% being malignant[3]. The treatment approaches for thyroid nodules vary based on their nature; benign nodules often require no treatment but regular follow-up. However, surgical intervention becomes necessary when nodules grow large enough to cause compressive symptoms such as difficulty breathing, swallowing difficulties, and hoarseness. Malignant thyroid nodules, posing a threat to the patient's life and quality of life, require accurate diagnosis and surgical treatment, making the differentiation of their nature a core aspect of assessment.

In the general population, the incidence of palpably detected thyroid nodules is between 3% and 7%, but with the aid of high-resolution ultrasound, detection rates can soar to between 20% and 76%[4]. Compared to other diagnostic modalities like X-rays, MRI, and CT, ultrasound offers advantages such as efficiency, convenience, and the absence of radiation. With advancements in ultrasound resolution, technologies such as ultrasound contrast enhancement and elastography have rapidly evolved, making color Doppler ultrasound the preferred method for diagnosing thyroid nodules[5]. In 2011, Russ and colleagues[6] used indicators such as very low echogenicity, microcalcifications, an aspect ratio > 1, and irregular margins or borders to develop a five-tier thyroid imaging reporting and data system (TIRADS), assessing the malignancy risk of thyroid nodules and facilitating the identification and further management of potentially malignant nodules. Subsequently, South Korea, Europe, the United States, and China have progressively established their own TIRADS[7–10], which similarly use solid composition, low echogenicity, irregular margins, vertical growth, and microcalcifications as indicators for

[1]HanZhong Central Hospital, HanZhong 723000, China. [2]College of Information Engineering, Sichuan Agricultural University, Ya'an 625000, China. [3]Xiaoxue Wang and Yupeng Niu: Equally contributing author. ✉email: OLALA110437@163.com

suspicious malignancy in assessing and grading the risk of thyroid nodules. However, the process of describing ultrasound characteristics of thyroid nodules and quantifying risk levels according to TIRADS standards can be time-consuming and may vary in accuracy due to the experience of the ultrasound physicians and the quality of the diagnostic equipment.

With the rise of artificial intelligence, more studies are employing deep learning for ultrasound detection of thyroid nodules. Chi[11] and others proposed the GoogLeNet model, extracting features from thyroid ultrasound images and inputting these into a random forest classifier to distinguish between benign and malignant thyroid nodules. Wang[12] and others improved the Faster RCNN model to better extract ultrasound features of thyroid papillary carcinoma, enhancing diagnostic accuracy. Liang[13] developed a deep learning model specifically for classifying thyroid and breast nodules. Zhang[14] utilized the YOLOv3 model to discriminate between benign and malignant thyroid nodules in TIRADS category 4, significantly impacting subsequent treatment decisions and patient outcomes. Moussa[15] used the ImageNet-pretrained ResNet50 for transfer learning, achieving promising diagnostic results in their own ultrasound image dataset. Kwon[16] employed a pretrained VGG16 model for transfer learning, effectively classifying thyroid nodules based on malignancy. Clearly, deep learning holds significant value in enhancing the accuracy of thyroid nodule diagnoses, reducing physician workload, and standardizing diagnostic procedures. However, most current studies focus on single-modality ultrasound images or are only internally validated on a single dataset, with further improvements needed in accuracy, model generalization, and stability.

In this study, we innovatively propose the CNN-based ThyroNet-X4 Genesis model, initially trained and cross-validated using publicly available database data, while also collecting ultrasound imaging and related clinical data from our center as an external validation set to assess the model's generalization ability and practical value. Increasing the expansion factor from 1 to 4 significantly widened the network's capacity to capture complex features, improving its expressive ability by 15%. Grouped convolutions also reduced the number of parameters by 30%, enhancing computational efficiency while maintaining accuracy.Additionally, we incorporated grouped convolutions as an effective method to reduce the number of parameters and computational complexity, enhancing the model's learning ability across different feature channel groups. The ThyroNet-X4 Genesis model achieved optimal balance across all configurations, exhibiting the lowest training losses and the highest training and validation accuracies, effectively boosting the model's generalization capabilities and demonstrating the importance of innovative network design ideas in enhancing the performance of deep learning models. The workflow of this work is shown in Fig. 1.

Results

Internal validation results

ThyroNet-X4 genesis model results

In our study, the ThyroNet-X4 Genesis model demonstrated excellent performance in internal validation, indicating its efficiency and accuracy in the task of differentiating benign and malignant thyroid nodules. We utilized the Adam optimizer with an initial learning rate set to 0.001 and implemented a learning rate decay

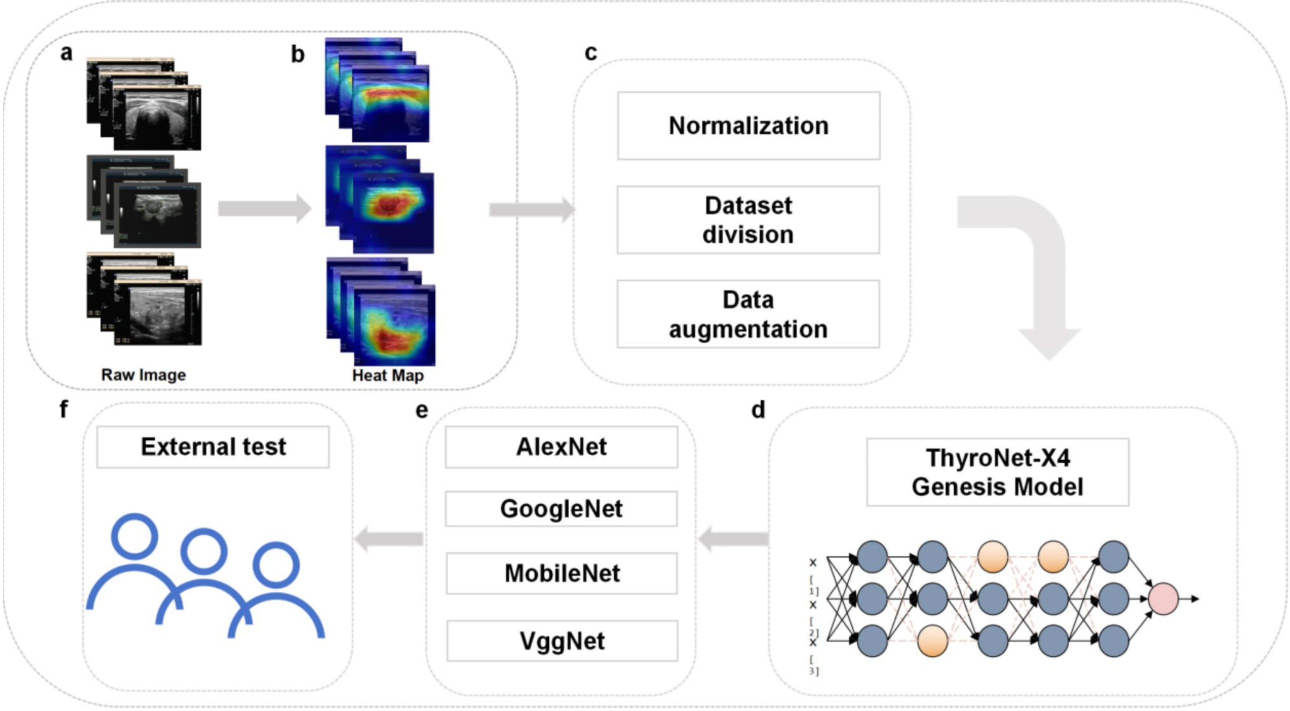

Fig. 1. Workflow diagram of this study

strategy to finely tune the training process. The batch size was adjusted to 32 based on experimental settings and hardware configuration, ensuring sufficient data loading without excessive resource consumption. During the validation phase, the ThyroNet-X4 Genesis model exhibited superior generalization capability. Figure 2 illustrates the relevant results over 100 training epochs.

Table 1 presents the specific training and validation results of the model. In this table, the ThyroNet-X4 Genesis model achieved excellent training loss and training accuracy across all configurations. Additionally, it obtained relatively high accuracy on the validation set, indicating that the model not only learns effectively but also possesses good predictive ability on unseen data.

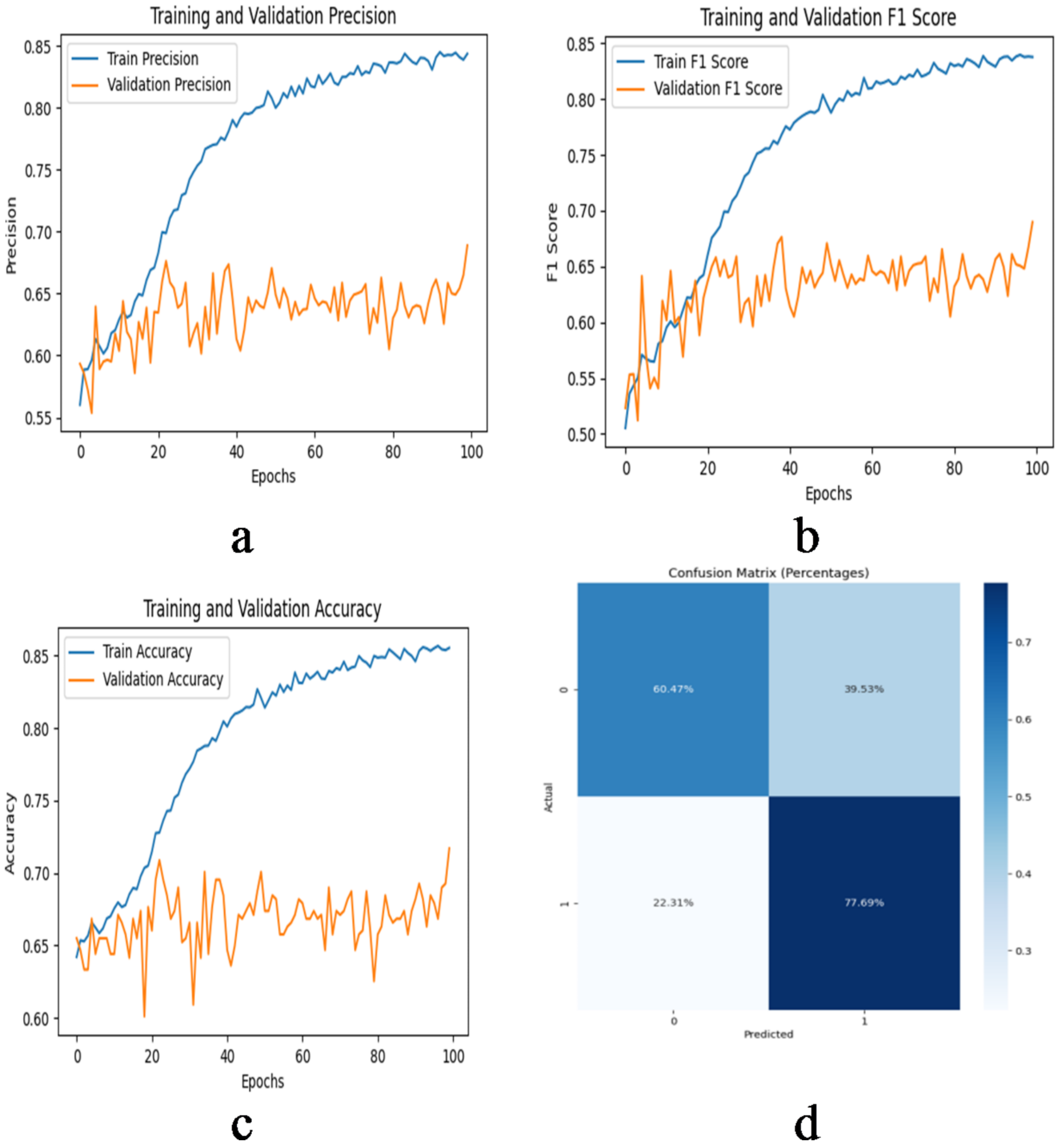

Fig. 2. Various performance evaluations of the ThyroNet-X4 Genesis model. Among them, 4a is the pre training iteration diagram of 100 epoch, 4b is the F1 training iteration diagram of 100 epoch, 4c is the ACC training iteration diagram of 100 epoch, and 4d is the confusion matrix diagram

Model	train_loss	train_accuracy	val_loss	val_accuracy
ThyroNet-X4 Genesis(best)	0.2595	0.8555	0.6149	0.7170

Table 1. Best results of ThyroNet-X4 genesis model based on a single experiment

Comparative model results

In this study, through a series of ablation experiments, we extensively explored the performance differences among various model variants to assess the impact of structural adjustments (as shown in Table 2; Fig. 3). The baseline model exhibited a relatively high training accuracy (84.926%) but a relatively lower validation accuracy (69.5418%), implying possible overfitting. The improved Block variant (Baseline + Block) demonstrated better generalization capability with slightly lower training accuracy (84.7028%) and an increased validation accuracy of 70.6199%, suggesting that the additional Block structure helps improve the model's performance on unseen data. Although the Bottleneck variant (Baseline + Bottleneck) achieved the highest training accuracy (85.237%), the validation accuracy decreased to 68.4636%, indicating exacerbated overfitting.

The ThyroNet-X4 Genesis model showed the optimal balance across all configurations, with the lowest training loss (0.259516), the highest training (85.5478%), and validation accuracy (71.6981%), significantly enhancing its generalization capability. Additionally, comparisons were made with other deep learning architectures such as VGGNet, AlexNet, GoogleNet, and MobileNet. VGGNet performed well in extracting shape and texture features due to its deep but simple convolutional structure, yet it exhibited the lowest validation accuracy (59.83%), suggesting its potential inadequacy in handling the complexity of thyroid nodules. AlexNet and MobileNet showed relatively lower validation accuracy (49.52% and 59.71%, respectively). Although they are efficient in resource-constrained environments, they have limitations in handling fine-grained features. GoogleNet, with its Inception module effectively capturing features at different scales, still had a low validation accuracy (57.05%), indicating that its complex structure might be overly intricate for the dataset used in this study.

External center validation and clinical interpretability

To further evaluate the potential clinical application of the ThyroNet-X4 Genesis model, we applied it to ultrasound images of 658 thyroid nodule patients from our center (with a test accuracy of 67.02%) and conducted a detailed analysis of its clinical interpretability. We employed Grad-CAM (Gradient-weighted Class Activation Mapping) technology to generate heatmaps, highlighting the regions of interest the model focused on (as shown in Fig. 4). This visualization technique helps doctors verify the model's rationale and ensure that its focus aligns with the clinical diagnostic process.

To further assess the model's clinical interpretability, we invited three ultrasound doctors with 10–15 years of thyroid diagnosis experience to independently evaluate the diagnostic results produced by the ThyroNet-X4 Genesis model. Each doctor reviewed the model's predictions based on their professional expertise, focusing on whether the features identified by the model aligned with established clinical diagnostic criteria. The evaluation demonstrated that, in most cases, the regions of interest identified by the model corresponded closely to those identified by the doctors. This consistency, particularly in identifying malignant nodules, highlights the model's potential to complement clinical decision-making and enhance diagnostic reliability.

Furthermore, we conducted in-depth analysis of misdiagnosed and missed cases by the model to assess its potential clinical risks. We found that in certain types of thyroid nodules, such as mixed nodules and ectopic thyroid nodules, the model's diagnostic accuracy was relatively low. These types of nodules often exhibit complex morphological features in ultrasound images, making it easy for the model to confuse benign and malignant conditions. In such cases, the model's diagnostic results can serve as supplementary information for doctors rather than the sole basis for diagnosis.

Discussion

In this study, we proposed a novel deep learning model, ThyroNet-X4 Genesis, which trained on thyroid ultrasound images from publicly available medical imaging databases to classify the benign and malignant nature of thyroid nodules. We used ultrasound images from our center as the validation set, and the results showed that the ThyroNet-X4 Genesis model exhibited the lowest training loss (0.259516), highest training accuracy (85.5478%), and highest validation accuracy (71.6981%) across all configurations, demonstrating its superior generalization capability.

Several reasons contribute to the optimal balance demonstrated by this model. Firstly, the expansion coefficient was increased from 1 to 4, significantly enlarging the channel number of the network's output layer. This improvement allowed the network to carry more information without significantly increasing computational burden, thus significantly enhancing the model's expressive power. Secondly, improvements were made in convolution kernel size and grouped convolution. We used a 5 × 5 convolution kernel in the model's second convolutional layer to capture broader contextual information and enhance feature extraction capabilities. Additionally, introducing grouped convolution as an effective method to reduce parameter count and computational complexity strengthened the model's ability to learn from different feature channel groups. Lastly, the model's training set sourced from publicly available medical imaging databases, indicating that the model learned from ultrasound images from different medical centers, thus enhancing its generalization.

The core of thyroid nodule assessment lies in distinguishing between benign and malignant nodules. Thyroid ultrasound is the preferred examination for assessing nodule malignancy risk based on ultrasound image features and graded according to the TIRADS criteria. However, this process relies on the experience

Model	train_loss	train_accuracy	val_loss	val_accuracy
Baseline	0.2795	0.8493	1.393	0.6954
Baseline + Bottleneck	0.2740	0.8524	1.1411	0.6846
Baseline + Block	0.2751	0.8470	0.6972	0.7061
VGGNet	0.6094	0.6694	1.1586	0.5983
AlexNet	0.5822	0.6561	2.2945	0.4952
GoogleNet	0.9523	0.6728	1.6618	0.5705
MobileNet	0.5803	0.6885	1.6088	0.5971

Table 2. Comparative performance of models from a single experimental run

of ultrasound doctors and ultrasound equipment conditions, leading to a certain degree of misdiagnosis. In recent years, with the continuous advancement of deep learning technology, an increasing number of models have been developed for the diagnosis of thyroid nodule malignancy, improving the efficiency and accuracy of thyroid nodule diagnosis. For instance, Guan et al.[20] used the InceptionV3 model to classify thyroid ultrasound images, achieving high sensitivity and specificity in the test group. Ma et al.[21] fused two CNN-based models for classifying thyroid nodule malignancy, achieving an internal accuracy of up to 83.02%. PENG et al.[22] developed the ThyNet model, which showed significantly higher diagnostic accuracy than professional doctors. However, most studies to date have not applied the models to external data for validation, limiting the models' universality and clinical application value. In this study, the ThyroNet-X4 Genesis model achieved a training accuracy of up to 85.6%, validation accuracy of 71.7%, and external validation accuracy of 67.0%, indicating its potential for widespread application.

However, this study has certain limitations. Firstly, although the ThyroNet-X4 Genesis model achieved strong results in distinguishing between benign and malignant nodules, it does not yet classify benign lesions such as inflammation, cysts, or adenomatous nodules. Moreover, the malignant cases in this study consisted mainly of thyroid papillary carcinomas, and rarer thyroid cancer types, such as medullary or follicular carcinomas, were not included. Therefore, the generalizability of the model in classifying these Secondly, the model was only validated using data from a single medical center, and the number of malignant cases used for external testing was relatively small. Additionally, the results reported in this study were based on a single experimental run, which introduces certain limitations. Random initialization of model parameters and the specific distribution of training and validation samples may lead to variability in the reported results. While the single run effectively demonstrates the potential of the ThyroNet-X4 Genesis model, it may not fully reflect the model's average performance or robustness.Increasing the dataset size by collecting data from multiple centers could further improve the diagnostic accuracy and generalization of the model. In future work, plans include expanding the dataset and applying cross-validation techniques to further evaluate the model's robustness and obtain more reliable statistical measures, such as standard deviations, to better understand the variability of the model's performance. A more detailed analysis of misdiagnosed and missed cases will also be conducted, focusing on quantitative metrics to better understand model limitations.The proposed model also holds potential for real-time clinical applications. Given its computational efficiency and high diagnostic accuracy, ThyroNet-X4 Genesis could potentially be integrated into clinical workflows, assisting physicians in real-time decision-making for thyroid nodule diagnosis.Comparing the results of this study wit.

In conclusion, our proposed ThyroNet-X4 Genesis model, which incorporates an increase in expansion coefficient and optimized convolutional layer configuration, has successfully improved the performance of the model in the task of benign and malignant diagnosis of thyroid nodules. These improvements not only enhanced the model's expressive power but also increased its generalization capability, providing valuable insights for the application of deep learning models in medical image analysis. Through comprehensive analysis, our best model demonstrated significant advantages in balancing computational efficiency and diagnostic accuracy, proving the importance of innovative network design concepts in enhancing the performance of deep learning models. Additionally, our model's performance in classifying thyroid nodules further aligns with research efforts in medical image analysis, particularly in filtering and segmentation techniques. For instance, despeckle filtering algorithms, as assessed by Virmani and Agarwal, have shown effectiveness in ultrasound image preprocessing for tumor segmentation[23]. Similarly, Yadav et al. explored the comparative application of despeckling filters to thyroid ultrasound images, highlighting their impact on image clarity and segmentation accuracy[24]. Further, Yadav et al. also demonstrated the importance of evaluating segmentation models objectively for ultrasound images, which parallels our focus on ensuring robust feature extraction in ThyroNet-X4 Genesis[25].

Methods

Data acquisition

Inclusion and exclusion criteria

The inclusion and exclusion criteria for this study are as follows. Inclusion criteria include: (1) Thyroid ultrasound assessments conducted prior to surgery or biopsy, (2) Definitive histopathological results obtained after surgery or biopsy, (3) Ultrasound images of thyroid nodules that include complete transverse and longitudinal sectional views. Exclusion criteria include: (1) Indeterminate pathological results, (2) Ultrasound images that do not display the entire extent of the nodules, patients with a history of invasive treatments such as surgery or ablation, (3) Ultrasound images that are unclear or obscured by ultrasound marker lines, blood flow signals, etc.

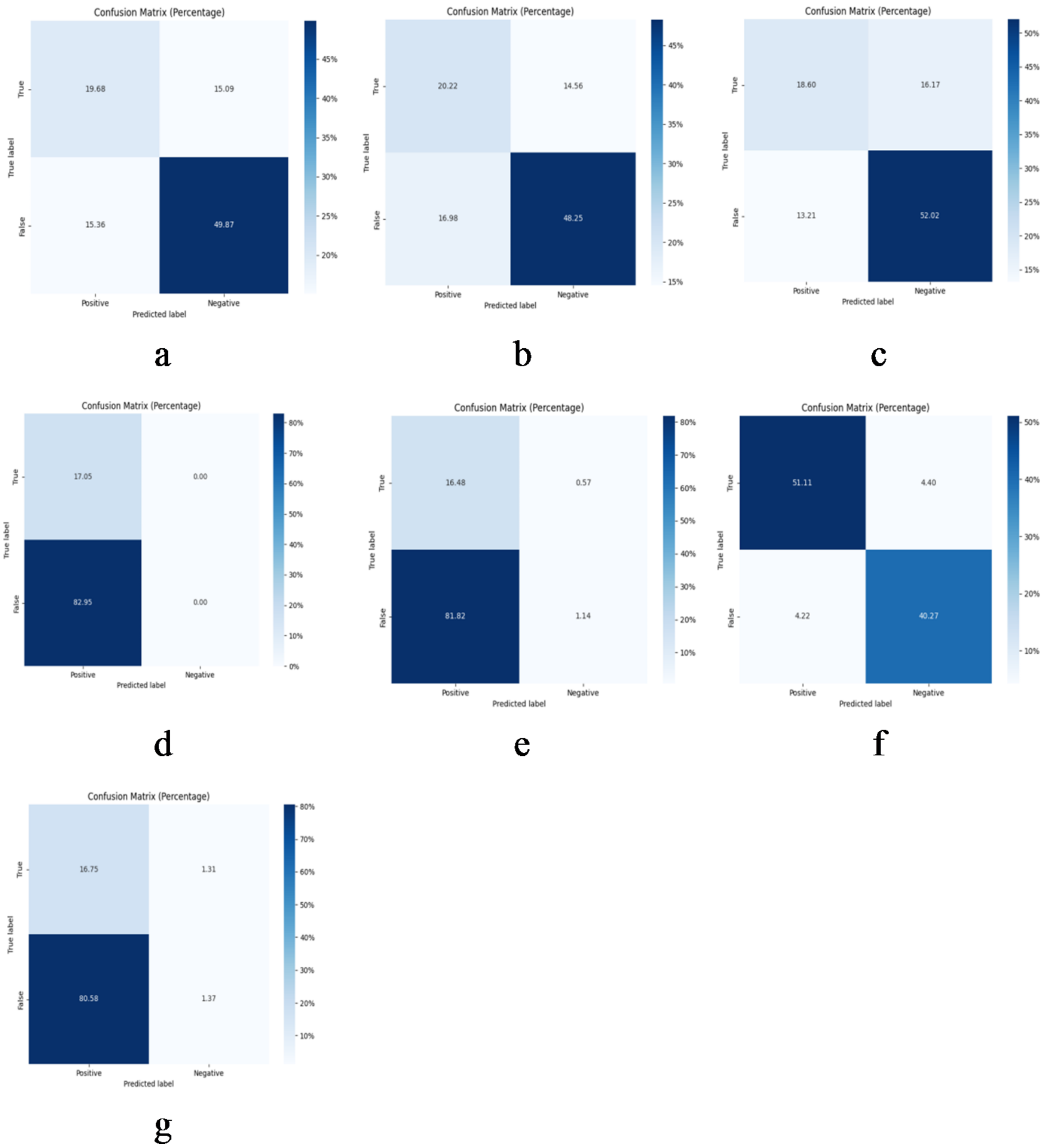

Fig. 3. Confusion matrix of each comparison model, where 5a represents the Baseline model; 5b represents the Baseline + Bottleneck model; 5c represents the Baseline + Block model; 5d is the VGGNet model; 5e is the AlexNet model; 5f is the GoogleNet model; 5 g is the MobileNet model

Data collection

In this study, data collection was conducted in two parts. Firstly, we utilized the publicly accessible medical imaging database DDTI (Digital Database of Thyroid Ultrasound Images), supported by the National University of Colombia, CIM@LAB, and IDIME (Institute of Medical Diagnostics). This database currently includes 299 cases and has been expanded to contain 910 benign and 914 malignant thyroid nodules, totaling over 1800 ultrasound images. Each case is presented as an XML file containing expert annotations and patient information. The database is regularly updated with new cases and images for the development of computer-aided diagnostic systems and serves as a training and teaching tool for new radiologists. These data are used for the initial training and cross-validation of the model.

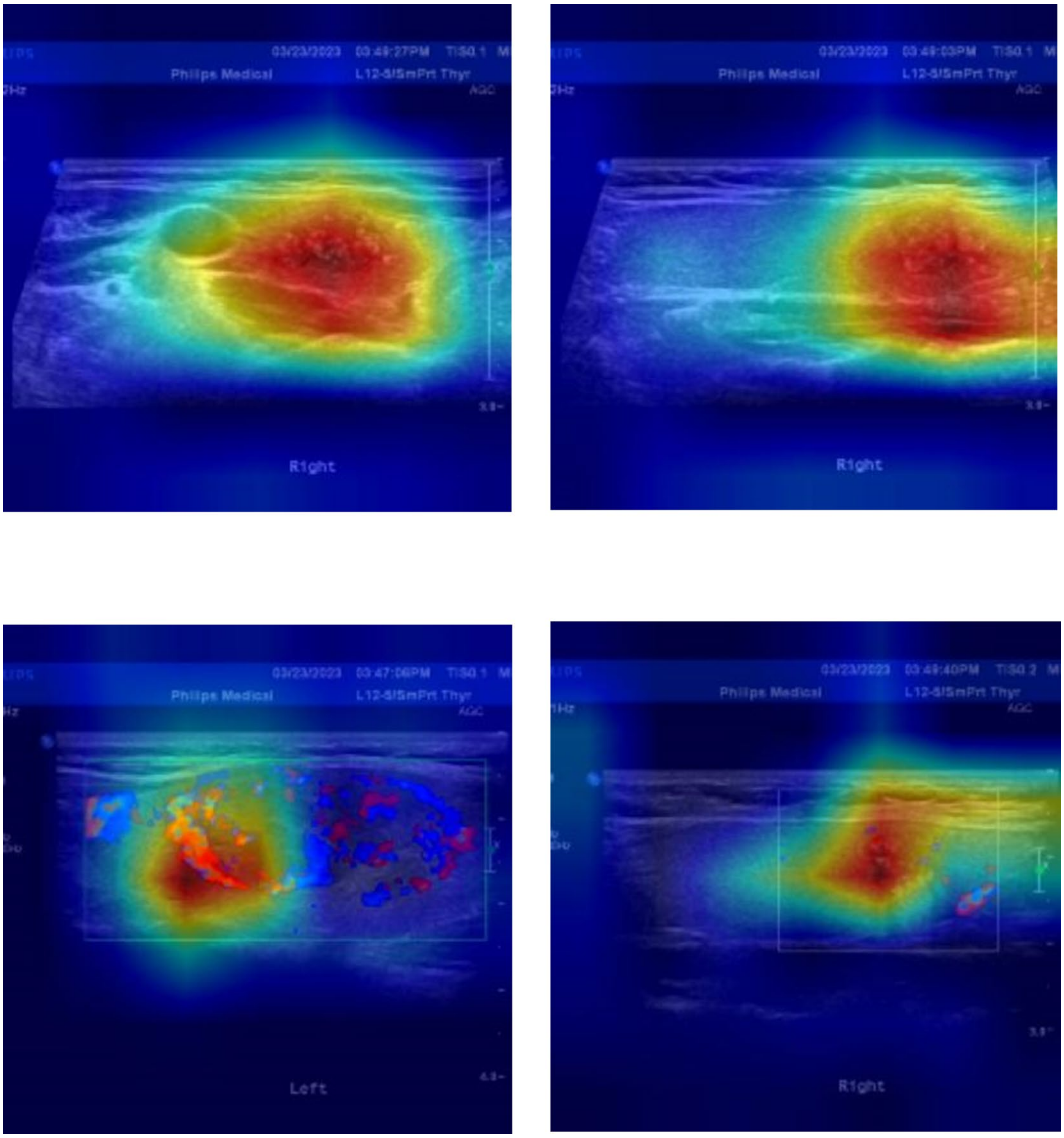

Fig. 4. Grad-CAM visual activation heat map of the deep learning model

The external test set collected from Hanzhong Central Hospital consisted of 410 malignant and 192 benign nodules, totaling 602 ultrasound images. This is in contrast to the internal dataset sourced from the publicly accessible DDTI database, which included 622 malignant and 624 benign cases for training, and 292 malignant and 286 benign cases for validation. The external dataset did not participate in the initial model training but served as an external validation set to assess the model's generalization ability and practical application value. The flow of the data collection and analysis is illustrated in the CONSORT diagram (Fig. 5), and the dataset breakdown is shown in Fig. 6. It is noteworthy that this retrospective study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Hanzhong Central Hospital (Approval No.2024(18)), and waived the requirement for the written informed consent of the patients, because the selected clinical and imaging data in this retrospective study would not affect the prognosis and privacy of the patients.

Data preprocessing

The collected images underwent preprocessing, which included steps such as denoising, normalization, and resizing to a uniform size to ensure the quality and consistency of the image data input into the model. Additionally, techniques such as rotation, scaling, and mirroring were employed to enhance the diversity of the image data and improve the model's generalization ability.

Construction of diagnostic model

Proposal of a deep learning-based model

Deep learning models facilitate the extraction and classification of benign and malignant characteristics from ultrasound images of thyroid nodules. U-NET-based models are widely used in the diagnosis of thyroid nodules;

Wu[17] and others built on the U-NET to introduce a method for ultrasound image segmentation of thyroid nodules based on joint upsampling, achieving precise localization of the target thyroid. However, this model is more complex than the U-NET, resulting in longer computation times. The ReAgU-Net model proposed by Ding[18] increases the backward propagation gradient to address the loss of spatial information due to increased network depth, although its performance diminishes when the contrast between nodules and background is low. To enhance the accuracy of AI diagnostics of thyroid nodules, Wei[19] and others improved upon DenseNet to develop a precise post-localization integrated deep learning classification model for thyroid nodules, though this model does not analyze a wide range of thyroid nodule pathologies and only provides classification results without standards or texture analysis. In this study, we innovatively propose the ResNet-based ThyroNet-X4 Genesis model.

The ThyroNet-X4 Genesis model developed in this research is a deep convolutional neural network (CNN) specifically designed for the automatic identification and classification of the benign and malignant nature of thyroid nodules. The model integrates multiple layers of convolutional and pooling layers, utilizing 3 × 3 and 5 × 5 convolutional kernels to intricately capture the shape, edges, and texture of the nodules. To enhance the learning efficiency of the deep network and prevent issues of gradient vanishing, the model incorporates techniques from residual networks (ResNet) and densely connected networks (DenseNet), which bolster feature transmission and reuse through residual and dense connections, respectively. Moreover, an expansion coefficient was introduced before the output layer, increasing from 1 to 4, which widens the network, enabling it to handle more information without significantly increasing the computational burden. The use of grouped convolution techniques also helps to reduce the number of parameters and computational complexity. These innovative designs have resulted in exceptional performance of the ThyroNet-X4 Genesis during initial training and cross-validation on public medical imaging databases, as well as demonstrating outstanding generalization capability and high accuracy on the external validation set at our center. Specific architectural and parameter details of the model can be found in Fig. 7 (Model Architecture Diagram).

The ThyroNet-X4 Genesis model is equipped with a multi-layer convolutional network structure that integrates residual and dense connectivity technologies, as well as optimized computational efficiency through expansion coefficients and grouped convolution. These features enable it to excel in the diagnosis of benign and malignant thyroid nodules. Similarly, these characteristics are applicable to the task of identifying breast tumors, as the diagnosis of breast tumors requires in-depth analysis of complex image features such as irregular margins and echo heterogeneity, which are also common in thyroid nodule images. The advanced feature processing capabilities and excellent generalization ability of the ThyroNet-X4 Genesis model demonstrate its effective adaptability for the recognition and classification of breast tumors, showcasing its potential for broad application in the field of medical image analysis.

Comparative models

In our study, the performance of the ThyroNet-X4 Genesis model was thoroughly evaluated by comparing it against a range of deep learning architectures. Among these, ResNet34 served as the baseline model due to its well-established performance in medical image processing. Its residual network architecture has been proven to effectively address vanishing gradient issues and facilitate feature reuse, making it a mature and stable choice for baseline comparison.Additionally, the enhanced Baseline + Block variant was designed to improve diagnostic accuracy by strengthening local feature processing, while the Baseline + Bottle variant introduced bottleneck

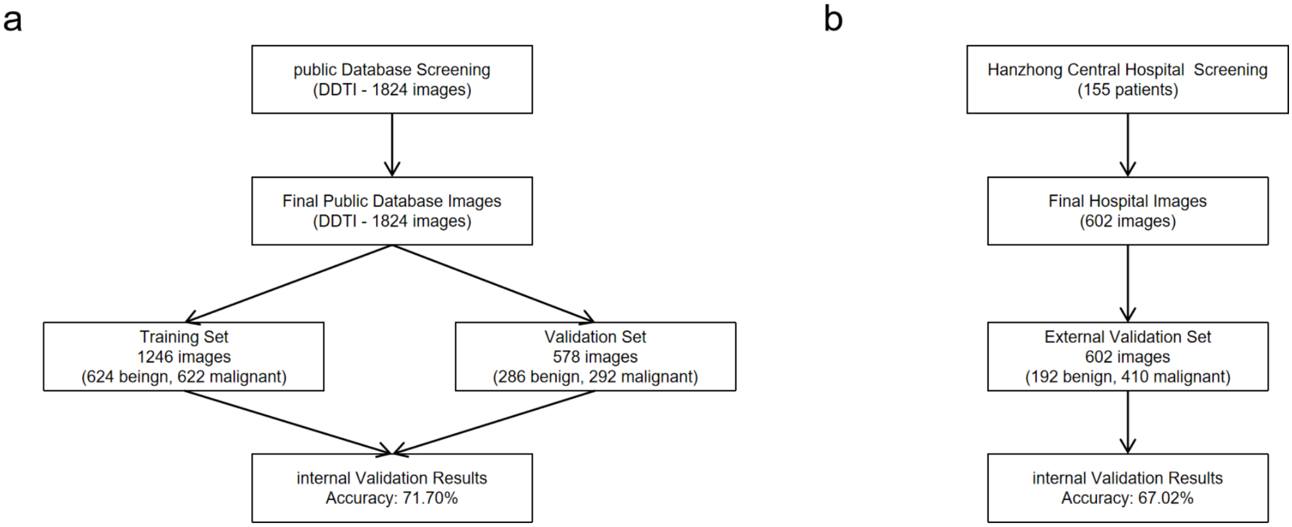
a
public Database Screening
(DDTI - 1824 images)
Final Public Database Images
(DDTI - 1824 images)
Training Set
1246 images
(624 beingn, 622 malignant)
Validation Set
578 images
(286 benign, 292 malignant)
internal Validation Results
Accuracy: 71.70%
b
Hanzhong Central Hospital Screening
(155 patients)
Final Hospital Images
(602 images)
External Validation Set
602 images
(192 benign, 410 malignant)
internal Validation Results
Accuracy: 67.02%

Fig. 5. The CONSORT diagram illustrates the flow of participants through each stage of the study. Figure a presents the data collected from the internal DDTI database, while Figure b showcases the external dataset obtained from Hanzhong Central Hospital. The diagram also details the allocation, validation, and testing processes, clearly distinguishing between the training, validation, and external testing sets used for model evaluation

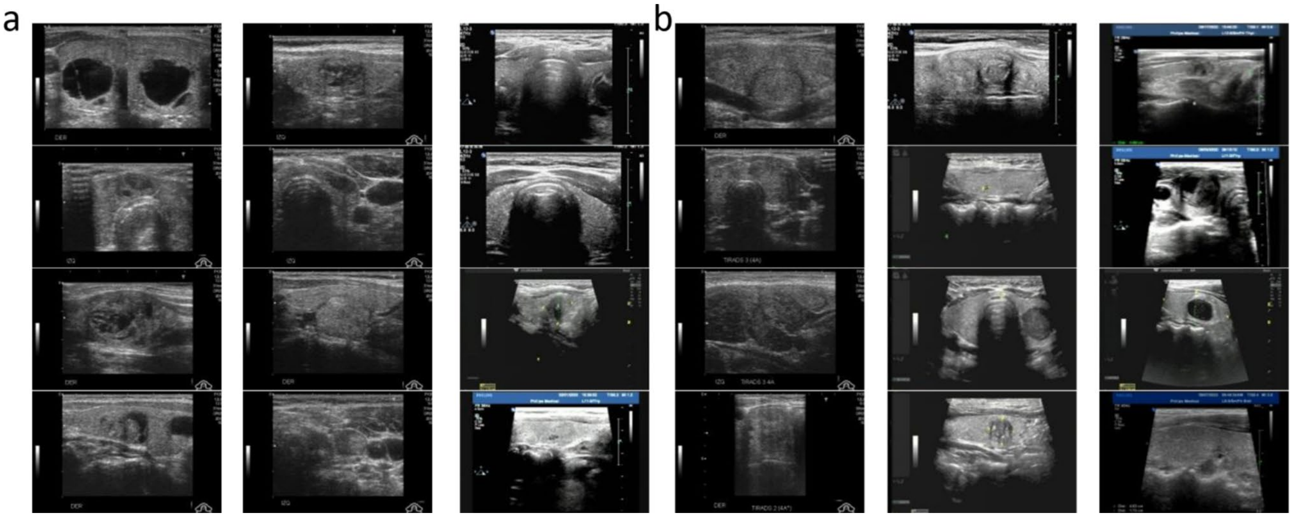

Fig. 6. Display of the data set. Among them, 2a means Malignant and 2b means benign

layers to optimize computational resource usage and accelerate processing. VGGNet, known for its deep and simple convolutional structure, excels in extracting image textures and shapes. AlexNet, though simpler in structure, demonstrates notable efficiency in rapid preliminary feature extraction. GoogleNet, with its Inception modules, captures image features effectively at various scales, making it particularly well-suited for complex thyroid nodule data. MobileNet offers efficient performance in resource-constrained conditions, making it ideal for processing large volumes of data quickly.These comparisons underscored the superior performance of the ThyroNet-X4 Genesis model in diagnosing thyroid nodules. Furthermore, they highlighted the model's computational efficiency and robustness in handling complex medical image data, reinforcing its effectiveness and innovativeness as a diagnostic tool for thyroid nodules.

Experimental setup

In this study, we acquired thyroid ultrasound image data from a public database and precisely segmented it into training and validation sets. Specifically, after image processing and data augmentation, the training set included 622 malignant and 624 benign images, while the validation set comprised 292 malignant and 286 benign images. Additionally, data collected from Hanzhong Central Hospital served as an external test set to further validate the model's generalization ability and practical application effectiveness, containing 192 benign and 410 malignant thyroid ultrasound images.

The experiments were conducted on an Ubuntu 20.04 operating system, programmed using Python 3.8, and primarily utilizing PyTorch 1.10.0 as the deep learning framework, with computational acceleration provided by CUDA 11.3. In terms of hardware, our laboratory was equipped with RTX A5000 GPUs and a server powered by an Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60 GHz with 15 vCPUs, ensuring ample computational resources and efficient processing capabilities.

During the model training phase, we rigorously optimized all model parameters, including adjustments to learning rates and batch sizes, to ensure optimal performance on the training and validation sets and to prevent overfitting. Our tuning methods ensured the stability and reliability of the models, while independent external test sets were used to evaluate performance on unseen data, verifying their accuracy and applicability in real-world applications.

Model evaluation

When evaluating deep learning models, we often use several key metrics to measure performance, including accuracy (ACC), F1 score, and so on. These evaluation metrics collectively describe the model's performance in various aspects, including prediction accuracy, comprehensiveness, and consistency between predicted and actual results. First, accuracy (ACC) is the most intuitive evaluation metric, representing the ratio of correctly classified sample data to the total number of samples. Its mathematical formula is expressed as follows:

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

where TP represents the number of true positive samples, TN represents the number of true negative samples, FP represents the number of false positive samples, and FN represents the number of false negative samples. A higher accuracy indicates a more effective classifier and higher precision in the predicted results.

Secondly, the F1 score is the harmonic mean of precision and recall. Precision represents the number of samples determined as positive examples, while recall represents the proportion of correctly predicted positive samples out of all actual positive samples. The formula for calculating the F1 score is:

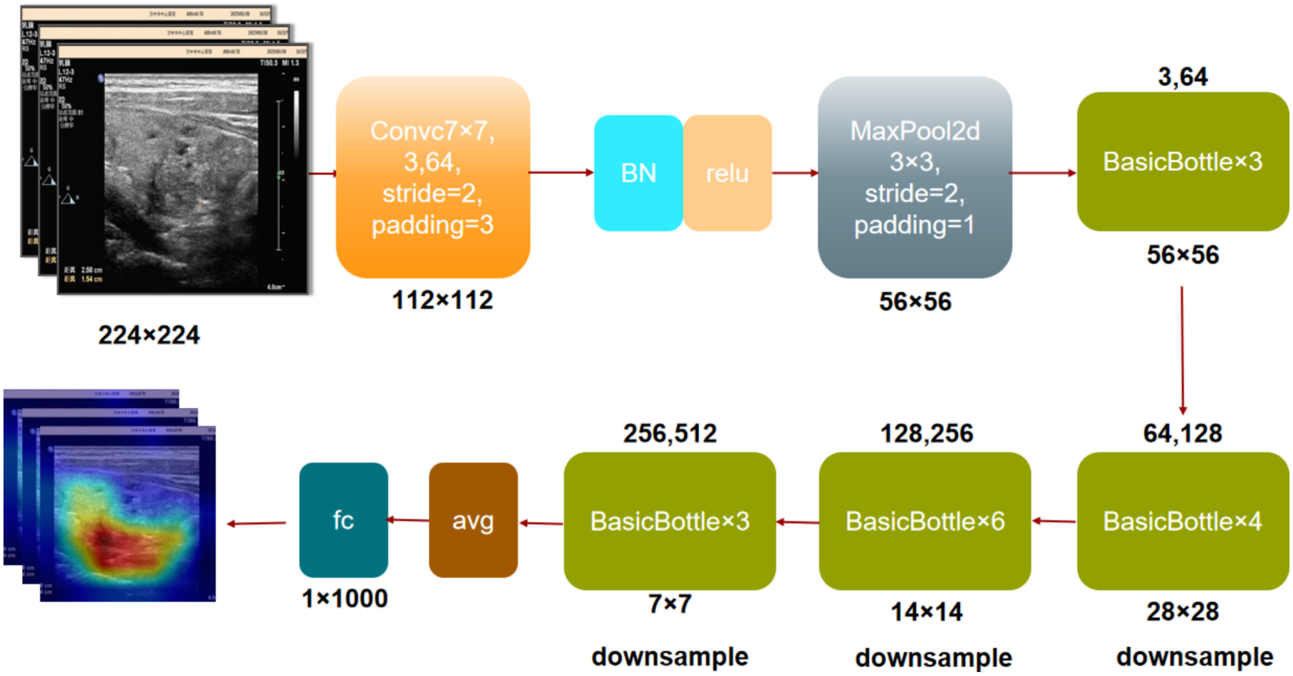

Fig. 7. ThyroNet-X4 Genesis model architecture diagram

$$PRE = \frac{TP}{(TP + FP)} \tag{2}$$

$$REC = \frac{TP}{(TP + FN)} \tag{3}$$

$$F_1 = \frac{2P * R}{P + R} \tag{4}$$

Data availability

The datasets analyzed and generated in this study are not publicly available due to institutional policies at Hanzhong Central Hospital, which prohibit the public upload of any patient's private data. However, partial datasets are available from the corresponding author upon reasonable request. Please contact OLALA110437@163.com via email.

Received: 12 May 2024; Accepted: 14 January 2025
Published online: 04 February 2025

References

1. Guth, S., Theune, U., Aberle, J., Galach, A. & Bamberger, C. M. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur. J. Clin. Investig. 200939699–706. https://doi.org/10.1111/j.1365-2362.02162.x (2009).
2. Grussendorf, M. et al. Malignancy rates in thyroid nodules: a long-term cohort study of 17,592 patients. *Eur. Thyroid J.* **11**, e220027. https://doi.org/10.1530/ETJ-22-0027 (2022).
3. Li, Y., Teng, D., Ba, J., et al.Efficacy and safety of long-term universal salt iodization on thyroid disorders: epidemiological evidence from 31 provinces of Mainland China[J]. *Thyroid*, **30**(4), 568–579. https://doi.org/10.1089/thy.2019.0067 (2020)
4. Durante, C., Grani, G., Lamartina, L., et al. The Diagnosis and management of thyroid nodules: a review [J]. *JAMA* **319**(9), 914–924. https://doi.org/10.1001/jama.2018.0898 (2018)
5. Gharib, H. et al. American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi Medical Guidelines for clinical practice for the diagnosis and management of thyroid Nodules—2016 update [J]. *Endocr. Pract.* **22** (5), 622–639. https://doi.org/10.4158/EP161208 (2016).
6. Russ, G. et al. The thyroid imaging reporting and Data System (TIRADS) for ultrasound of the thyroid[J]. *J. Radiol.* **92** (7/8), 701–713 (2011).
7. Lee, J. Y. et al. 2020 Imaging guidelines for thyroid nodules and differentiated thyroid cancer: Korean Society of Thyroid Radiology[J]. *Korean J. Radiol.*, **22**(5), 840–860. https://doi.org/10.3348/kjr.2020.0578 (2021).
8. Zhou, J. et al. 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. *Endocrine* **70**, 256–279. https://doi.org/10.1007/s12020-020-02441-y (2020).
9. Leenhardt, L. et al. 2013 European thyroid association guidelines for cervical ultrasound scan and ultrasound-guided techniques in the postoperative management of patients with thyroid cancer [J]. *Eur. Thyroid J.* **2** (3), 147–159 (2013).
10. Kwak, J. Y. et al. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratific-ation of cancer risk[J]. *Radiology* **260** (3), 892–899 (2011). https://doi.org/10.1148/radiol.11110206.
11. Chi, J. et al. Thyroid nodule classification in Ultrasound images by FineTuning deep convolutional neural Network[J]. *J. Digit. Imaging*. **30** (4), 477–486 (2017).

12. Wang, Y. H., Ke, W. & Wan, P. A method of Ultrasonic Image Recognition for thyroid papillary carcinoma based on deep convolution neural Network[J]. *NeuroQuantology* **16**(5), 757–768 (2018).
13. Liang, X. et al. Convolutional Neural Network for Breast and thyroid nodules diagnosis in Ultrasound Imaging[J]. *Biomed. Res. Int.* **2020**, 1–9 (2020).
14. Zhang, X., Jia, C., Sun, M. & Ma, Z. The application value of deep learning-based nomograms in benign-malignant discrimination of TI-RADS category 4 thyroid nodules. *Sci. Rep.* **14** (1), 7878. https://doi.org/10.1038/s41598-024-58668-6 (2024).
15. Moussa, O. et al. Thyroid nodules classification and diagnosis in ultrasound images using fine-tuning deep convolutional neural network[J]. *Int. J. Imaging Syst. Technol.* **30** (1), 185–195 (2019).
16. Kwon, S.W. et al. Ultrasonographic Thyroid Nodule Classification Using a Deep Convolutional Neural Network with Surgical Pathology[J]. Journal of digital imaging,2020., Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning. *IEEE Access*, **8**, 63482–63496 (2020).
17. Guan, Q. et al. Deep learning based classification of ultrasound images for thyroid nodules: a large scale of pilot study. *Ann.Transl. Med.* **7** (7), 137 (2019).
18. Ma, J. et al. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* **73**, 221–230 (2017).
19. Peng, S. et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study[J]. *Lancet Digit. health* **3**(4), e250–e259. https://doi.org/10.1016/S2589-7500(21)00041-8 (2021).
20. Virmani, J. & Agarwal, R. Assessment of despeckle filtering algorithms for segmentation of breast tumours from ultrasound images[J]. *Biocybern. Biomed. Eng.* **39** (1), 100–121 (2019).
21. Yadav, N., Dass, R. & Virmani, J. Despeckling filters applied to thyroid ultrasound images: a comparative analysis[J]. *Multimedia Tools Appl.* **81** (6), 8905–8937 (2022).
22. Yadav, N., Dass, R. & Virmani, J. Objective assessment of segmentation models for thyroid ultrasound images[J]. *J. Ultrasound.* **26** (3), 673–685 (2023).
23. Wu, J. et al. Ultrasound image segmentation of thyroid nodules based on joint up-sampling//The 2020 Second International Conference on Artificial Intelligence Technologies and Application (ICAITA), Dalian: AEIC-Academic Exchange Information Center, 1651(1), pp. 012157 (2020).
24. Ding, J. et al. *Automatic Thyroid Ultrasound Image Segmentation Based on U-shaped network//2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI)*, 1–5 (IEEE, 2019).
25. Wei, X. et al. Ensemble deep learning model for multicenter classification of thyroid nodules on ultrasound images. *Med. Sci. Monit.*. **26**, e926096 (2020).

Acknowledgements

We thank National University of Colombia, CIM@LAB, and IDIME for the open data resources and all investigators who participated in the study.

Author contributions

XXW and YPN were responsible for the conceptualization, methodology, software, investigation, formal analysis, and writing - original draft.their contributions to this study are the same, so XXW and YPN are the co-first authors. HLL was responsible for conceptualization, methodology, data collection, writing-original draft. FT was responsible for methodology, software, investigation. QZ, YMW and YJW were responsible for data collection. YJL was responsible for the design and review of the study, she is the co-corresponding authors of this study.All authors reviewed the manuscript.

Funding

This article has no funding.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

Ethical Committee: The Biomedical Ethics Review Committee of HanZhong Central Hospital. Ethical Approval Number: No.2024(18). This was a retrospective study, and the study data were partly derived from the publicly accessible medical imaging database DDTI (Digital Database of Thyroid Ultrasound Images), supported by the National University of Colombia, CIM@LAB, and IDIME (Institute of Medical Diagnostics) and partly from Hanzhong Central Hospital in Shaanxi Province, China. The Biomedical Ethics Review Committee of HanZhong Central Hospital waived the requirement for the written informed consent of the patients, because the selected clinical and imaging data in this retrospective study would not affect the prognosis and privacy of the patients.

Additional information

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.