Original papers

Med Ultrason 2023, Vol. 25, no. 4, 375-383
DOI: 10.11152/mu-4306

ChatGPT-assisted deep learning model for thyroid nodule analysis: beyond artificial intelligence

Ismail Mese[1], Neslihan Gokmen Inan[2], Ozan Kocadagli[2], Artur Salmaslioglu[3], Duzgun Yildirim[4]

[1]Department of Radiology, Health Sciences University, Erenkoy Mental Health and Neurology Training and Research Hospital, [2]Department of Statistics, Mimar Sinan Fine Arts University, [3]Department of Radiology, Istanbul Medical Faculty, Istanbul University, [4]Department of Radiology, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey

Abstract

Aims: To develop a deep learning model, with the aid of ChatGPT, for thyroid nodules, utilizing ultrasound images. The cytopathology of the fine needle aspiration biopsy (FNAB) serves as the baseline. **Material and methods:** After securing IRB approval, a retrospective study was conducted, analyzing thyroid ultrasound images and FNAB results from 1,061 patients between January 2017 and January 2022. Detailed examinations of their demographic profiles, imaging characteristics, and cytological features were conducted. The images were used for training a deep learning model to identify various thyroid pathologies. ChatGPT assisted in developing this model by aiding in code writing, preprocessing, model optimization, and troubleshooting. **Results:** The model demonstrated an accuracy of 0.81 on the testing set, within a 95% confidence interval of 0.76 to 0.87. It presented remarkable results across thyroid subgroups, particularly in the benign category, with high precision (0.78) and recall (0.96), yielding a balanced F1-score of 0.86. The malignant category also displayed high precision (0.82) and recall (0.92), with an F1-score of 0.87. **Conclusions:** The study demonstrates the potential of artificial intelligence, particularly ChatGPT, in aiding the creation of robust deep learning models for medical image analysis.

Keywords: artificial ıntelligence; deep learning; medical informatics applications; thyroid nodule; ultrasonography

Introduction

Thyroid nodules are common clinical findings, affecting up to 50% of the general population. Women and the elderly exhibit a higher prevalence. While most of these nodules are benign, a small percentage can be malignant [1]. Ultrasound (US) imaging is a vital tool for thyroid nodule assessment. It offers a noninvasive, cost-effective, and widely accessible means to determine nodule size, shape, echogenicity, and other features pivotal for malignancy risk stratification [2]. The Thyroid Imaging Reporting and Data System (TIRADS) provides a standardized classification for thyroid nodules based on US characteristics. This aids in predicting the risk of malignancy and guiding subsequent interventions [3]. Low-risk thyroid nodules, such as those that are benign, asymptomatic, or small, are generally managed through monitoring. This approach often includes regular check-ups that encompass clinical and US examinations, and thyroid function tests [4,5]. However, nodules that display uncertain results, exhibit growth, or have suspicious features might require a more invasive diagnostic procedure like fine-needle aspiration biopsy (FNAB) [6]. US guidance also ensures the accuracy of FNAB by confirming that the tissue sampled genuinely represents the nodule in question [2,6].

Received 03.10.2023 Accepted 27.10.2023
Med Ultrason
2023, Vol. 25, No 4, 375-383
Corresponding author: Ismail Mese
Department of Radiology,
Health Sciences University,
Erenkoy Mental Health and Neurology
Training and Research Hospital
19 Mayıs, Sinan Ercan Cd. No:23,
34736 Kadıköy/İstanbul
E-mail: ismail_mese@yahoo.com
Phone/fax: +90 (506) 895 16 53
+90 (216) 356 04 96

The interpretation and communication of findings from FNAB of thyroid nodules are facilitated by a standardized system known as The Bethesda System for Reporting Thyroid Cytopathology [7]. The system comprises six categories, ranging from 'Non-diagnostic or Unsatisfactory' to 'Malignant.' Each category correlates with a specific risk of malignancy and suggests a particular management approach [2].

Recently, deep learning techniques have revolutionized medical imaging analysis. These techniques utilize convolutional neural networks (CNN) or other neural networks to learn complex patterns and extract valuable information from large datasets [8]. By analyzing these datasets and discerning intricate patterns, artifical intelligence (AI) algorithms can assist in differentiating benign and malignant nodules, thereby improving clinical decision-making and potentially reducing the need for invasive procedures such as surgery [9]. Moreover, deep learning models have demonstrated remarkable results in diverse medical imaging applications, including tumor classification, anatomical structure segmentation, and abnormality detection [10].

The practical application of these advanced algorithms and models requires not only an intricate understanding of the underlying theory but also the ability to write an effective code [11,12]. Researchers lacking a multidisciplinary team face a challenge in developing a deep learning model using medical data, as it necessitates a blend of medical and programming proficiency. Python, a widely used programming language for AI and machine learning, can be difficult for individuals lacking extensive programming experience or expertise in the field [11-13]. Such obstacles can hinder the adoption and implementation of AI solutions, as well as the training and education of professionals interested in these advanced technologies [11,12]. ChatGPT, a large language model developed by OpenAI, has capabilities in programming and coding that have been highlighted [14,15]. Currently, there are no studies that explore the utilization of ChatGPT for the development of deep learning models.

The current study seeks to address AI adoption-implementation challenges by exploring the use of ChatGPT, as a tool to facilitate the coding process and enhance learning for individuals working with Python. This study primarily aims to develop a deep learning model with the assistance of ChatGPT for thyroid nodules, using US images, with cytopathology of the FNAB serving as a reference. The hope is that this approach will streamline the development process and enable researchers and professionals to leverage the capabilities of deep learning and AI technologies.

Material and methods

Data collection

In this retrospective study, we included patients over 18 years old who underwent thyroid US and subsequent FNAB for suspicious nodules between January 2017 and January 2022. The inclusion criteria consisted of the availability of high-quality US images in the PACS system, a corresponding FNAB diagnosis in the pathology database, and no history of previous thyroid surgery or ongoing thyroid-related treatments. The exclusion criteria encompassed patients with inadequate image quality, and instances with missing or incomplete clinical data.

We gathered thyroid US images of the chosen patients from the PACS system, which contained information on nodule size, shape, echogenicity, and other relevant features critical for deep learning model training and analysis. The images were saved in a suitable format to ensure compliance with CNN requirements for image processing and analysis.

In our analysis, we examined the demographic profiles, imaging characteristics, and cytological features of 1079 patients, each presenting with a single thyroid nodule. However, due to incomplete clinical, pathological, or radiological data, we excluded 18 patients from the study, resulting in a final count of 1061 patients for analysis.

Firstly, based on the FNAB cytology findings, we removed any samples that fell under the non-diagnostic category of the Bethesda classification from our dataset (n=180). For clarity and to remain consistent with the original Bethesda classification, we then divided the patients into five categories: benign, atypia of undetermined significance (AUS) follicular lesion of undetermined significance (FLUS), follicular neoplasm, and malignant.

The enrolled patients had an average age of 41.61 years, ranging from 18 to 82 years. Regarding gender distribution, 190 nodules were associated with male patients (21.6%), and 691 nodules were associated with female patients (78.4%).

As for the distribution of thyroid pathologies in the dataset, there were 428 images of benign (48.6%), 125 images of AUS (14.2%), 79 images of FLUS (9.0%), 63 images of follicular neoplasm (7.2%), and 186 images of malignant (21.1%) pathologies.

Due to the retrospective nature of the study the requirement for patient consent was waived. The local ethics committee and IRB approval for this study were obtained.

ChatGPT assistance

We proceeded with the development of the deep learning model, employing ChatGPT to aid in code writ-

ing. ChatGPT provided guidance on preprocessing steps, such as resizing, normalizing pixel values, and applying data augmentation techniques. It also assisted in choosing the optimal network architecture, algorithm selection, model optimization, suggesting suitable hyperparameters, and helped troubleshoot code-related issues, thereby ensuring an efficient and effective development process.

Data organization, software installation, and Python library setup

For implementing deep learning applications, our research utilized Python [13]. We used PyCharm, a free and open-source integrated development environment (IDE) with features such as code completion, debugging, and version control integration [16]. We used an NVIDIA GeForce GTX 1630 Ti laptop GPU (VRAM: 8 GB) for the execution.

The successful implementation of a deep learning algorithm necessitates the installation and configuration of various Python libraries and packages [17]. We followed ChatGPT's recommendations for a model creation regarding the necessary libraries and packages for our deep learning model.

ChatGPT assisted-deep learning model development and implementation

With the assistance of ChatGPT, we chose a CNN as the deep learning architecture for our study, given its proven effectiveness in medical image analysis [9]. We employed a step-by-step approach, using various Python libraries and choosing specific parameters based on our study's requirements with ChatGPT's help (supplementary file, fig 1).

We imported the collected data into our Python environment using the pandas library. The images and labels were loaded from subfolders. Next, we preprocessed the images using the OpenCV library by resizing them to a uniform size. Images with dimensions of 440x440 were resized to 128x128 using the bilinear interpolation method. We also normalized their pixel values, and applied data augmentation techniques to enhance the dataset's size and improve the model's generalization capabilities. For the train-test split, we used the train_test_split function of the scikit-learn library, allocating 70% of the data for training, 15% for validation, and 15% for testing (supplementary file, fig 2). All test sets were independent of the training sets. This allocation is common practice in machine learning to ensure sufficient data for model evaluation and prevent overfitting [18].

In the model creation stage, we used TensorFlow and Keras libraries to build a CNN architecture. The model architecture consisted of two convolutional layers (each followed by a max pooling layer) and two fully connected layers. The output layer had five neurons, corresponding to the five classes of thyroid nodules, with a softmax activation function for outputting class probabilities. To prevent overfitting, dropout layers were added with a rate of 0.5. The input shape for the first convolutional layer was set to (128, 128, 3) to accommodate the resized and preprocessed images. During the training process, we used the Adam optimizer, a popular and effective algorithm for deep learning tasks. We compiled the model with the Adam optimizer, sparse categorical cross-entropy loss, and accuracy as the performance metric [19]. We implemented early stopping to monitor the validation loss and halt training if it did not improve after a specified number of epochs, also restoring the best weights obtained during training. We then trained the model for 100 epochs, using the training and validation data, and applied the early stopping callback (supplementary file, fig 3). The training process stopped at the 29th epoch, utilizing the early stopping mechanism to prevent overfitting and optimize model performance. This setup ensured effective learning while avoiding overfitting of the training data.

After training the model, we evaluated its performance using various metrics, including accuracy, precision, recall, and F1-score. We also calculated the 95% confidence intervals for each metric to determine the range of possible values. In addition, we used scikit-learn's roc_curve and auc functions to generate ROC curves and AUC values for each thyroid pathology. We plotted loss and accuracy curves to visualize the training process and identify any signs of overfitting or underfitting.

Finally, we utilized the Gradient Weighted Class Activation Mapping (Grad-CAM) visualization technique to discern the areas of the image that the model deemed significant for making its predictions (Supplementary file, fig 4).

Statistical Analysis

We conducted descriptive statistics to summarize and describe the key characteristics of our dataset, including demographic data and thyroid pathology distribution. We assessed our deep learning model's performance using metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). In our study, we calculated 95% confidence intervals for each performance metric (accuracy, precision, recall, F1-score, and AUC) to quantify the uncertainty around our model's performance.

Results

Model performance-training and validation set

During the training process, the model demonstrated a consistent decrease in loss and a steady increase in ac-

curacy (fig 1). The exact values for loss and accuracy at the end of training were 0.42 and 0.84, respectively, for the training set, and 0.82 and 0.82 for the validation set. We observed potential overfitting, evidenced by the continuous decrease in training loss, while validation loss stagnated. We addressed this through the implementation of regularization techniques such as dropout and weight decay, which improved the model's generalizability.

Model performance-testing set

The model's performance showcases significant results across various metrics. The accuracy, a fundamental measure of the model's overall correctness, is recorded at 0.81. This indicates that our model correctly predicted 81% of the test data instances, making it considerably reliable. This high level of accuracy falls within a 95% confidence interval of 0.76 to 0.87, offering further assurance of the model's performance. Table I provides a comprehensive view of the performance metrics on the testing set.

In terms of the deep learning model's performance metrics on thyroid subgroups, the benign category is particularly noteworthy. This category exhibits high precision (0.78) and an exceptional recall (0.96), illustrating the model's robust ability to accurately identify this category while effectively minimizing false positives. This balance between precision and recall is mirrored in the F1-score of 0.86. Additionally, it's worth noting that this category had the highest number of instances, with a total of 94.

The follicular neoplasm category also presents significant performance, with a perfect precision score of 1.00. This indicates that every prediction made by the model for this category was accurate. However, the recall for this category was lower at 0.71, suggesting the model may not have detected all true positives. This balance is reflected in the F1-score of 0.83.

Similarly, the malignant category demonstrated strong results, with high precision (0.82) and recall (0.92). The F1-score for this category is 0.87, indicating a balanced performance between precision and recall. Overall, these results illustrate the deep learning model's effective performance across the different thyroid subgroups. Table III provides a comprehensive classification report, breaking down the performance metrics for each category under study. Figure 2 visually represents the ROC curves for each subgroup, along with their respective AUC values, further detailing the model's performance on subgroup classification.

Table I. Performance metrics on the testing set

Metric	Performance	95% Confidence Interval
Accuracy	0.81	0.76 - 0.87
Precision	0.83	0.73 - 0.86
Recall	0.81	0.78 - 0.88
F1-score	0.80	0.75 - 0.86
AUC	0.89	

AUC: The Area Under the Curve

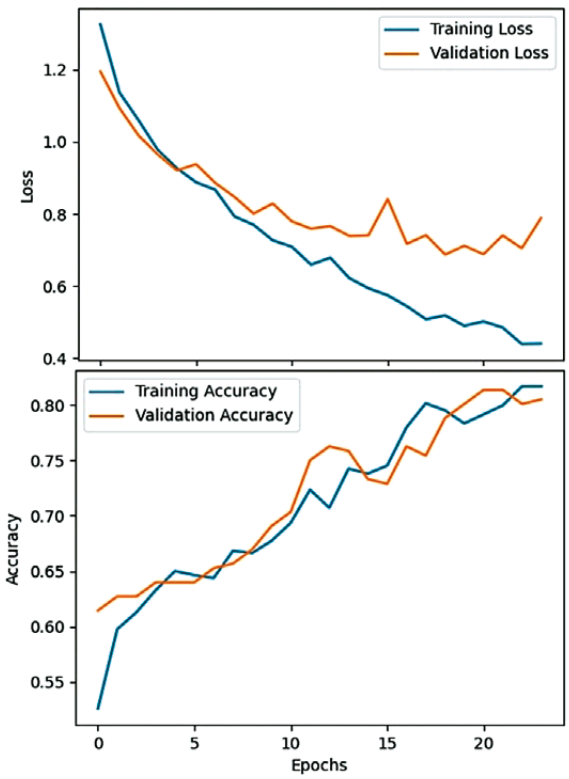

Fig 1. Loss and accuracy curves for training and validation sets

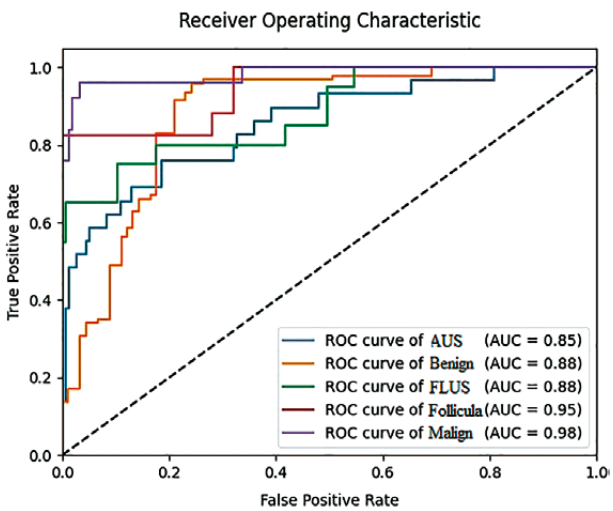

Fig 2. Subgroup categorization evaluation through ROC Curve and AUC values

Grad-CAM heatmaps were utilized to visualize the features and patterns learnt by the CNN. These heatmaps serve as validation, demonstrating that the algorithm correctly identified nodules, particularly the solid areas within these nodules, for feature extraction, classification and subsequent decision-making (fig 3).

The classification report in Table II presents the performance metrics for each subcategory.

Discussion

Our study aimed to develop a robust, accurate deep learning model with the assistance of ChatGPT for the assessment of thyroid nodules using ultrasound images. The application of AI in thyroid nodule evaluation could potentially improve diagnostic accuracy, aid in decision-making, reduce unnecessary invasive procedures, and contribute to personalized therapeutic strategies.

This study's unique proposition was the application of the ChatGPT language model to assist in the development of the deep learning model. By leveraging ChatGPT's capabilities, we streamlined the process of model development. The high accuracy achieved by our model, 81%, suggests that the ChatGPT-assisted development process was effective. The model demonstrated strong performance across various thyroid pathology subgroups, particularly benign category, follicular neoplasm category, and malignant category. These results underline the model's ability to distinguish different thyroid pathologies, highlighting its potential usefulness in clinical decision-making. By offering a non-invasive technique for thyroid pathology classification via ultrasound, it presents a patient-centric alternative to more invasive diagnostics such as biopsies. Furthermore, integrating ChatGPT into the developmental phase fosters transparency and replicability, possibly mitigating radiologists' workload and amplifying diagnostic precision.

The accuracy of our model (81%) is comparable to other studies that have employed deep learning techniques for thyroid nodule classification. For instance, Ma et al [20] reported an accuracy of 83.7%, while Song et al [21] achieved an accuracy of 85.6%. The slight differences in performance could be attributed to variations in dataset size, data quality, model architecture, and training parameters.

The demographic characteristics of the patient sample in our study provide crucial insights into the prevalence and distribution of thyroid nodules. The gender distribution in our sample, with 21.6% male and 78.4% female patients, aligns with the broader literature, reflecting the known predilection of thyroid diseases for women [20-24].

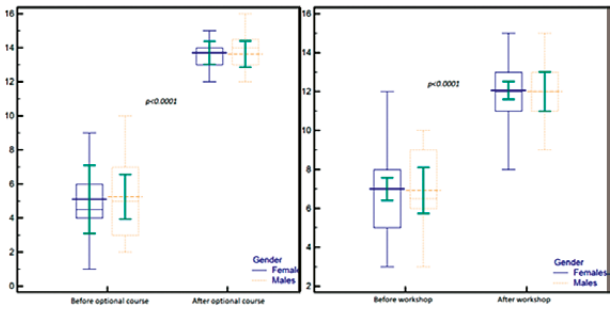

Fig 3. Thyroid nodule ımages and corresponding Grad-CAM heatmaps

Table II. A classification report for the different categories under examination

Category	Precision	Recall	F1-score	Support
AUS	0.82	0.48	0.61	29
Benign	0.78	0.96	0.86	94
FLUS	0.92	0.55	0.69	20
Follicular Neoplasm	1.00	0.71	0.83	17
Malignant	0.82	0.92	0.87	25

AUS: Atypia of Undetermined Significance, FLUS: Follicular Lesion of Undetermined Significance

The originality and strength of this study lie in its innovative approach to merging traditional medical imaging techniques with cutting-edge artificial intelligence. With this methodology used for the first time in the literature for thyroid nodules, the integration of ChatGPT into our deep learning model development brought about profound enhancements. ChatGPT facilitated our decision-making, especially when it came to the pivotal step of algorithm selection [25]. It not only shed light on the best network architecture for our study but also elaborated on various options with comprehensive explanations. The study's meticulous data selection, involving comprehensive inclusion and exclusion criteria, combined with the detailed preprocessing of images with ChatGPT, highlights its methodological rigor. The utilization of various evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC curves, provides a holistic assessment of the model's performance. Additionally, the adoption of Grad-CAM heatmaps to elucidate the model's decision-making process embodies a commitment to transparency and interpretability, critical in fostering trust among clinicians and patients. The integration not only simplified the model creation process but also enhanced efficiency.

We incorporated a CNN as the deep learning architecture for our study, following ChatGPT's advices. Compared to other classification algorithms, a CNN requires significantly less pre-processing [26]. Whereas conventional methods require manually engineered filters, with

sufficient training, CNNs can learn these filters or features autonomously [26,27]. To address the overfitting problem, we employed regularization techniques such as dropout and weight decay and implemented early stopping callback [28]. Early Stopping is a regularization technique for deep neural networks that stops training when parameter updates no longer begin to yield improve on a validation set. In essence, technique stores and updates the current best parameters during training, and when parameter updates no longer yield an improvement (after a set number of iterations) it stops training and use the last best parameters. It works as a regularizer by restricting the optimization procedure to a smaller volume of parameter space [28]. ChatGPT offered valuable assistance by providing suggestions in this manner. With optimization of functions dropout, weight decay, L2 regularization, it has significantly improved the generalization capability of our model [29,30].

We employed a suite of evaluation metrics to thoroughly assess the performance of our deep learning model, namely accuracy, precision, recall, and the F1 score. The 95% confidence intervals provided for each metric reinforced the model's reliability and its readiness for deployment in real-world scenarios [31,32]. AUS demonstrates high precision but low recall, leading to a low F1 score. This indicates that while the model is precise in identifying AUS cases, it tends to miss a significant number of true positive cases. In contrast, benign category shows average precision and high recall, resulting in a high F1 score. This suggests that the model accurately identifies the majority of true positive cases in this category, with a slight trade-off in precision. FLUS has high precision but low recall, yielding a moderate F1 score, implying that while the model is precise in classifying FLUS cases, it misses several true positive cases. Follicular neoplasm exhibits perfect precision and average recall, leading to a high F1 score, signifying that the model correctly identifies all true positive cases, but its recall suggests that some cases are misclassified. Finally, the malignant category demonstrates high precision and recall, resulting in a high F1 score, indicating that the model performs well in this category, accurately identifying most true positive cases while maintaining a good balance between precision and recall.

Understanding the decision-making process of deep learning models is crucial in clinical applications to build trust and facilitate the adoption of AI-assisted models. Methods such as Grad-CAM, LIME, or SHAP can be employed to enhance model interpretability and explainability, ensuring that the model's predictions align with human expert knowledge [33]. As suggested by ChatGPT, we used Grad-CAM heatmaps, which provided valuable insights into our model's decision-making process. Selvaraju et al introduced Grad-CAM technique, which offers a visual explanation of deep learning models [34]. This method allows for enhanced understanding of models during detection or prediction tasks. In its operation, Grad-CAM takes an image, processes it using the designated model, and after predicting a label, applies the Grad-CAM technique to one of the Convolutional layers, typically the last one. Radiologists can then use a color visualization feature of Grad-CAM to view clearer images, facilitating more informed and confident decisions [35]. By visualizing the regions of the images that the model found most informative, we could verify that the model was focusing on relevant features. This technique assists in making deep learning models more interpretable and explainable [35].

While ChatGPT's capabilities in programming and coding have been highlighted, a noticeable gap exists in the literature [14,15]. Specifically, no studies focus on using ChatGPT's coding potential to create deep learning models. Our research successfully addressed this void. For coding context, in a study with undergraduate students, ChatGPT proved instrumental in providing quick solutions, enhancing thinking abilities, simplifying debugging, and boosting self-assurance during programming tasks [14]. However, it was not without raising some concerns: there were instances of promoting laziness and occasional inaccurate responses. In software bug resolution, ChatGPT's performance stood on par with other deep learning methods. Its unique strength lies in integrating additional information, such as anticipated outputs or error signals [14]. This underscores ChatGPT's potential in programming, but it is essential to approach its use with an awareness of its limitations.

In addition to assisting with the technical aspects of deep learning model development, ChatGPT also played a significant role in troubleshooting code-related issues throughout our study. These issues ranged from debugging syntax errors to addressing problems related to data processing, data formatting, and the implementation of various deep learning libraries and frameworks. For instance, ChatGPT was instrumental in resolving problems associated with the integration of TensorFlow and Keras into our deep learning model. It helped us identify and correct errors in our model's architecture, the implementation of optimization algorithms, and the setting of hyperparameters. ChatGPT also guided us in the identification of the most effective methods for data augmentation, normalization, and encoding.

AUS/FLUS remain a challenge, at both the diagnostic and therapeutic level [36]. In this category, the risk of malignancy varies between 18% and 81% [37]. Within

this group, discerning malign samples from benign ones with the human eye can be challenging. In this context, we believe that artificial intelligence, especially with ChatGPT's guidance, can be of assistance in differentiating between malign and benign cases within those categories.

In our study which focused on the development of a deep learning model for the assessment of thyroid nodules using US images, several limitations and potential areas for improvement were noted. Starting with ChatGPT's limitations: firstly, our reliance on ChatGPT for guidance in algorithm development is worth highlighting. While ChatGPT proved instrumental, offering insights into the knowledge available at its last update, the rapidly evolving field of AI could introduce newer methodologies or paradigms postdating this knowledge. If there have been groundbreaking advancements in CNN architectures or optimization techniques after the last update, ChatGPT would not be aware of them. Secondly, it is crucial to understand the potential biases embedded in ChatGPT's recommendations. Derived from its training data, these biases, if unaddressed, can permeate its suggestions [38,39]. For instance, while CNN is well-known, there might be niche architectures specifically designed for certain types of medical images that the model does not emphasize. The simplicity of the CNN architecture chosen by ChatGPT might not recognize certain intricate features in ultrasound images, which could potentially undermine its diagnostic proficiency for specific pathologies [40,41]. Investigating more sophisticated architectures, such as DenseNets, ResNets, or Inception networks, might bolster performance. These have demonstrated heightened aptitude for discerning complex image features, subsequently elevating classification accuracy [42,43]. Leveraging transfer learning techniques could enhance the model's adaptive prowess [44]. Furthermore, a fundamental issue is the limited transparency into ChatGPT's training data [38]. This training enables ChatGPT to generate code snippets that may seem innovative and useful at first glance, but it also raises critical concerns. Specifically, there is no way to ascertain whether the code it generates is an exact or approximate replica of a code previously written by someone else. Utilizing such code in a commercial product or service could lead to complicated copyright disputes or even legal actions [38]. The ethical implications are equally significant. Therefore, while ChatGPT offers exciting possibilities, it is crucial to approach its output with caution and due diligence. It is imperative to scrutinize ChatGPT's advice against expert opinion and industry knowledge [38]. Adopting a multidisciplinary approach, including both domain experts and machine learning specialists, can ensure the development process remains comprehensive and devoid of undue biases [39].

Regarding human experts or data limitations: the primary constraint is the quality and volume of our training dataset. If inadequate, this could compromise the model's accuracy and generalizability [45]. If the dataset was collected from a single institution or geographic location such as ours, the findings might not generalize well to other settings or populations. Certain demographic or clinical subgroups might be underrepresented in the study, making the model less effective for those populations. Additionally, in our retrospective study, we utilized the Bethesda 2017 criteria because our database includes patients from January 2017 to January 2022; however, it is important to note that this may limit the generalizability of our findings, as the updated Bethesda 2023 criteria are now available.

Future perspectives

With the rapid advancements in the field of artificial intelligence and deep learning, the future holds promising opportunities to enhance our current methodologies and models for diagnosing thyroid nodules. Building upon our ChatGPT-assisted deep learning model, several modifications can be considered for future implementations. First, incorporating a broader spectrum of data from diverse sources and demographics can bolster the model's robustness and generalizability. Second, the implementation of transfer learning using state-of-the-art architectures such as ResNet, Inception, or VGG can potentially improve accuracy and reduce training time [44]. Furthermore, as ultrasound images can sometimes have inherent noise, incorporating advanced image processing and enhancement techniques before training can ensure cleaner and more focused data input. This will help in the extraction of more relevant features, thus improving diagnosis accuracy [46]. To reduce errors, future implementations can consider ensemble techniques, where multiple models' predictions are combined to arrive at a final decision [47]. Lastly, continuous feedback loops can be established where expert radiologists review and correct the model's predictions, creating an ever-evolving system that learns from its mistakes [48]. As we move forward, a synergy between human expertise and artificial intelligence, like the one we established with ChatGPT, will be pivotal in pushing the boundaries of medical diagnostics and patient care. To truly gauge the model's viability in practical scenarios, external validations using distinct datasets and forward-looking clinical trials are essential. These investigations can validate the model's versatility across varied populations, measure its influence on clinical choices, and spotlight potential enhancement areas [49].

Conclusion

Our study highlights the potential of AI-assisted deep learning models, such as those developed with ChatGPT, in medical image analysis. As an initial study, we were able to achieve accuracy close to the success of more advanced neural networks in the literature.

Conflict of interest: none

References

1. Dean DS, Gharib H. Epidemiology of thyroid nodules. Best Pract Res Clin Endocrinol Metab 2008;22:901-911.
2. Bomeli SR, LeBeau SO, Ferris RL. Evaluation of a thyroid nodule. Otolaryngol Clin North Am 2010;43:229-238.
3. Tessler FN, Middleton WD, Grant EG, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. J Am Coll Radiol 2017;14:587-595.
4. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The Diagnosis and Management of Thyroid Nodules: A Review. JAMA 2018;319:914-924.
5. Ito Y, Miyauchi A, Inoue, H, et al. An observational trial for papillary thyroid microcarcinoma in Japanese patients. World J Surg 2010;34:28-35.
6. Todsen T, Bennedbaek FN, Kiss K, Hegedüs L. Ultrasound-guided fine-needle aspiration biopsy of thyroid nodules. Head Neck 2021;43:1009-1013.
7. Cibas ES, Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. Thyroid 2017;27:1341–1346.
8. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D Deep Learning on Medical Images: A Review. Sensors (Basel) 2020;20:5097.
9. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017;73:221-230.
10. Tsuneki M. Deep learning models in medical image analysis. J Oral Biosci 2022;64:312-320.
11. Tejani AS. Identifying and addressing barriers to an artificial intelligence curriculum. J Am Coll Radiol 2021;18:605–607.
12. Wiggins WF, Caton MT Jr, Magudia K, Rosenthal MH, Andriole KP. A Conference-Friendly, Hands-on Introduction to Deep Learning for Radiology Trainees. J Digit Imaging 2021;34:1026-1033.
13. Open-source community. Python programming language. Version 3.9.6. Python software foundation. Available from: https://www.python.org/. Published 2021. Accessed August 10, 2023.
14. Surameery NMS, Shako MY. Use Chat GPT to solve programming bugs. Int J Inform Tech Comput Eng 2022;3:17-22.
15. Perkel JM. Six tips for better coding with Chat GPT. Nature 2023;618:422-423.
16. JetBrains. PyCharm professional. Version 2021.2. JetBrains. Available from: https://www.jetbrains.com/pycharm/. Published 2021. Accessed August 10, 2023.
17. Raschka S, Patterson J, Nolet C. Machine learning in python: main developments and technology trends in data science, machine learning, and artificial intelligence. Information 2020;11:193.
18. El Naqa I, Ruan D, Valdes G, et al. Machine learning and modeling: data, validation, communication challenges. Med Phys 2018;45:e834-e840.
19. Hassan E, Shams MY, Hikal NA, Elmougy S. The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. Multimed Tools Appl 2023;82:16591-16633.
20. Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. Ultrasonics 2017;73:221-230.
21. Song W, Li S, Liu J, et al. Multitask Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition. IEEE J Biomed Health Inform 2019;23:1215-1224.
22. Kim DH, Chung SR, Choi SH, Kim KW. Accuracy of thyroid imaging reporting and data system category 4 or 5 for diagnosing malignancy: a systematic review and meta-analysis. Eur Radiol 2020;30:5611-5624.
23. Suteau V, Munier M, Briet C, Rodien P. Sex Bias in Differentiated Thyroid Cancer. Int J Mol Sci 2021;22:12992.
24. Ling J, Li W, Lalwani N. Atypia of undetermined significance/follicular lesions of undetermined significance: What radiologists need to know. Neuroradiol J 2021;34:70–79.
25. Yu AC, Eng J. One algorithm may not fit all: how selection bias affects machine learning performance. Radiographics. 2020;40:1932–1937.
26. Zang B, Ding L, Feng Z, et al. CNN-LRP: Understanding convolutional neural networks performance for target recognition in SAR images. Sensors (Basel) 2021;21:4536.
27. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging 2018;9:611-629.
28. Charilaou P, Battat R. Machine learning models and overfitting considerations. World J Gastroenterol 2022;28:605-607.
29. Rodríguez-Muñiz LJ, Bernardo AB, Esteban M, Díaz I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? PLoS One 2019;14:e0218796.
30. Kumar U, Bhar A. Studying and analysing the effect of weight norm penalties and dropout as regularizers for small convolutional neural networks. Int J Eng Res Technol (IJERT) 2021;10:47-51.
31. Hazra A. Using the confidence interval confidently. J Thorac Dis 2017;9:4125-4130.
32. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022;12:5979.
33. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy (Basel) 2020;23:18.
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep net-

works via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. Venice, Italy, 2017;618-626.
35. Hung SC, Wu HC, Tseng MH. Integrating image quality enhancement methods and deep learning techniques for remote sensing scene classification. Appl Sci 2021;11: 11659.
36. Nishino M, Wang HH. Should the thyroid AUS/FLUS category be further stratified by malignancy risk? Cancer Cytopathol 2014;122:481-483.
37. Kim TH, Krane JF. The evolution of "atypia" in thyroid fine-needle aspiration specimens. Diagn Cytopathol 2022;50:146-153.
38. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 2023;15:e37281.
39. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. J Med Internet Res 2023;25:e43251.
40. Nirthika R, Manivannan S, Ramanan A, Wang R. Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. Neural Comput Appl 2022;34:5321-5347.
41. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell 2022;15:1-22.
42. Girdhar N, Sinha A, Gupta S. DenseNet-II: an improved deep convolutional neural network for melanoma cancer detection. Soft comput 2022;24:1-20.
43. Pan Y, Liu J, Cai Y, et al. Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases. Front Physiol 2023;14:1126780.
44. Gupta P, Malhotra P, Narwariya J, Vig L, Shroff G. Transfer learning for clinical time series analysis using Deep neural networks. J Healthc Inform Res 2019;4:112-137.
45. Holland L, Wei D, Olson KA, et al. Limited number of cases may yield generalizable models, a proof of concept in deep learning for colon histology. J Pathol Inform 2020;11:5.
46. Panwar H, Gupta PK, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. Chaos Solitons Fractals 2020;140:110190.
47. Westphal M, Brannath W. Evaluation of multiple prediction models: A novel view on model selection and performance assessment. Stat Methods Med Res 2020;29:1728-1745.
48. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 2023;56:3005–3054.
49. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022;4:e210064.