Biology Methods and Protocols, 2025, **10(1)**, bpaf017
https://doi.org/10.1093/biomethods/bpaf017
Advance Access Publication Date: 25 February 2025
Methods Article

Retrieval Augmented Medical Diagnosis System

Ethan Thomas Johnson[1], Jathin Koushal Bande[1], Johnson Thomas[2,*]

[1]Central High School, 423 E Central St, Springfield, Missouri, 65802, United States
[2]Department of Endocrinology, Mercy Hospital, Springfield, Missouri, 65807, United States

*Corresponding author. Department of Endocrinology, Mercy Hospital, 3231 S National Ave, Springfield, Missouri, 65807, United States. Tel: (417) 888 5660; Fax: (417) 888-6793; E-mail: johnson.thomas@mercy.net

Abstract

Subjective variability in human interpretation of diagnostic imaging presents significant clinical limitations, potentially resulting in diagnostic errors and increased healthcare costs. While artificial intelligence (AI) algorithms offer promising solutions to reduce interpreter subjectivity, they frequently demonstrate poor generalizability across different healthcare settings. To address these issues, we introduce Retrieval Augmented Medical Diagnosis System (RAMDS), which integrates an AI classification model with a similar image model. This approach retrieves historical cases and their diagnoses to provide context for the AI predictions. By weighing similar image diagnoses alongside AI predictions, RAMDS produces a final weighted prediction, aiding physicians in understanding the diagnosis process. Moreover, RAMDS does not require complete retraining when applied to new datasets; rather, it simply necessitates re-calibration of the weighing system. When RAMDS fine-tuned for negative predictive value was evaluated on breast ultrasounds for cancer classification, RAMDS improved sensitivity by 21% and negative predictive value by 9% compared to ResNet-34. Offering enhanced metrics, explainability, and adaptability, RAMDS represents a notable advancement in medical AI. RAMDS is a new approach in medical AI that has the potential for pan-pathological uses, though further research is needed to optimize its performance and integrate multimodal data.

Keywords: explainable artificial intelligence (XAI); retrieval augmented generation (RAG); computer vision; physician-in-the-loop; case-based reasoning

Introduction

Subjectivity poses a significant challenge in human analysis of images, particularly in medical contexts. Image interpretation often suffers from both inter-observer and intra-observer [1] variability when diagnosing diseases from medical images. Inter-observer variability is defined as the discrepancy in interpretations when multiple physicians analyze the same medical image independently. Conversely, intra-observer variability describes the phenomenon in which a single physician produces inconsistent readings of an identical medical image across different time points [2]. Additionally, analysis of these images is prone to error and is time consuming. These errors can result in over-diagnosis or under-diagnosis. Over-diagnosis can lead to unnecessary invasive procedures and additional costs. Under-diagnosis can lead to physicians not detecting a disease early on, which could lead to increased morbidity and mortality.

Artificial intelligence (AI) models have emerged as a tool for reducing subjectivity, errors, and reading time [3]. Despite the benefits that medical AI models bring, they have an apparent lack of explainability. The challenge of explainability in AI-based medical diagnosis systems is a topic of considerable interest and concern in the medical community [4]. Traditional deep learning models, considered as black boxes [5].

Moreover, contemporary medical AI models do not adhere to the principles of case-based reasoning in medical diagnosis [6]. Instead, these models are designed to recognize specific characteristics of various pathologies solely based on the data used during their training.

Another key challenge that medical AI faces is reduced performance when testing on new datasets from different health systems [7]. This is particularly relevant in medical imaging, where datasets can vary significantly in quality and characteristics across different institutions and regions. Furthermore, transferring AI models to new medical imaging datasets is also challenging and performance might degrade [8].

To address these issues, we aim to create an explainable AI model that integrates a standard AI prediction model, and an image similarity model paired with a weighing equation called Retrieval Augmented Medical Diagnosis System (RAMDS). To evaluate the efficacy of RAMDS, we conducted experiments on breast ultrasound images.

Materials and methods

Our study's goal was to test whether RAMDS, can outperform a standard AI model in critical aspects of medical diagnostics. The focus is not on basic accuracy, but on sensitivity (correctly identifying disease when it is present), negative predictive value (probability that a person with a negative test result truly does not have the disease) and explainability (providing clear, understandable reasoning behind its diagnoses). By comparing the RAMDS to

Received: 10 January 2025; **Revised:** 14 February 2025; **Editorial decision:** 17 February 2025; **Accepted:** 21 February 2025

the base ResNet 34 model, the study seeks to quantify the added value of the retrieval-augmented mechanism in improving diagnostics through AI.

Datasets

All data used in the article are available in the public domain. All images were resized to 256 × 256 pixels and converted to greyscale with padding to maintain aspect ratio. Multiple image augmentation techniques were used to generate more images on the fly for training.

Training data consisted of images from BUSC [9], BUSI_corrected [10], and QAMEBI [11] datasets. RODTOOK [12] dataset was used for testing. These datasets had breast ultrasound images with corresponding diagnoses. Features of each dataset are listed in Table 1.

Model creation

A pretrained ResNet 34 model was used as the base model. FastAI library [13] was used to find the optimal learning rate and fine tune this model on the training dataset to predict the diagnosis. This fine-tuned ResNet model was used to predict on the test set. After this, the model was used to generate embeddings for the training images, and these embeddings were stored in a file. Then, using a cosine similarity function, the test images are compared to the previously stored embeddings. After comparing the input embedding to the stored embeddings, N number of similar images (determined by the user) and corresponding diagnosis are retrieved.

Then, we created a model that combined the deep learning model's predictions with a similarity-based weighted adjustment mechanism; refining the diagnostic threshold based on the concordance with the top similar historical cases. The core of our method lies in the adjusted prediction function, which operates in several steps:

- **Base Prediction and Probability**: The function begins by obtaining a base prediction (*base_pred*) and its associated probability (P_{score}) from the fine-tuned ResNet34 model.
- **Similarity Assessment**: It then retrieves diagnoses and similarity scores of the top N images closely resembling the input image. These similarities are quantified using cosine similarity measures in the embedding space generated by the same ResNet34 model.
- **Weighted Agreement Calculation**: The function calculates a weighted agreement rate between the base prediction and the retrieved similar cases. If the base prediction from the fine-tuned ResNet-34 is cancer and all retrieved similar images are cancer, then there will be a higher agreement rate. Higher agreement rate translates to higher weightage being assigned to similar image diagnosis in the final prediction.
- **Threshold Adjustment:** The base threshold, initially set at a specific value, is dynamically adjusted based on the weighted agreement rate. An adjustment factor, determined through a grid search algorithm on the validation test, modulates how much the threshold should shift. The rationale behind this adjustment is to increase the prediction's confidence when there is a high agreement with similar cases and decrease it when there is a discrepancy.
- **Final Prediction**: The adjusted threshold is then applied to the base prediction probability to yield the final diagnosis.
- The core equation can be represented as

$$T_{adjusted} = base\_threshold - adjustment\_factor \times \left(2 \times \sum_{i=1}^{N} w_i \cdot I(D_i = base\_pred) - 1\right)$$

where:

$$w_i = \frac{S_i}{\sum_{j=1}^{N} S_j}$$

The final prediction is then determined as

$$Final\ prediction = \{1\ if\ P_{score} \geq T_{adjusted}\ \ 0\ if\ P_{score} < T_{adjusted}$$

given:

- P_{score} is the prediction score from the AI model.
- S_i is the similarity score from the ith similar image.
- D_i is the diagnosis for the ith similar image.
- *base_pred* is the base prediction from the model.
- N is the number of similar images.

here:

- $I(D_i = base\_pred)$ is an indicator function that equals 1 if $D_i = base\_pred$, *and* 0 *otherwise*.
- w_i represents the normalized weight for each similar image based on its similarity score.

This equation can be fine-tuned to new datasets without retraining the original model. Grid search algorithms can be used on a sample of the new dataset to find the optimal parameters for a desired performance metric. Figure 1 depicts the workflow of RAMDS.

Statistical analysis

We used Python programming language and packages to conduct statistical analysis. Accuracy,sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), were calculated to quantify and compare the results.

Table 1. Breast ultrasound datasets.

Dataset	Total images	Benign Images	Malignant Images	Ultrasound Device	Country
BUSC	250	100	150	Voluson 730 with Voluson small part transducer S-VNW5-10	Brazil
BUSI_corrected	589	410	179	GE LOGIQ E9, Agile with ML6-15-D Matrix linear probe	Egypt
QAMEBI	232	109	123	AixPlorer Ultimate ultrasound machine with linear transducer	Iran
RODTOOK	278	131	147	Philips iU22 US scanner	Thailand
Total	1349	750	599		

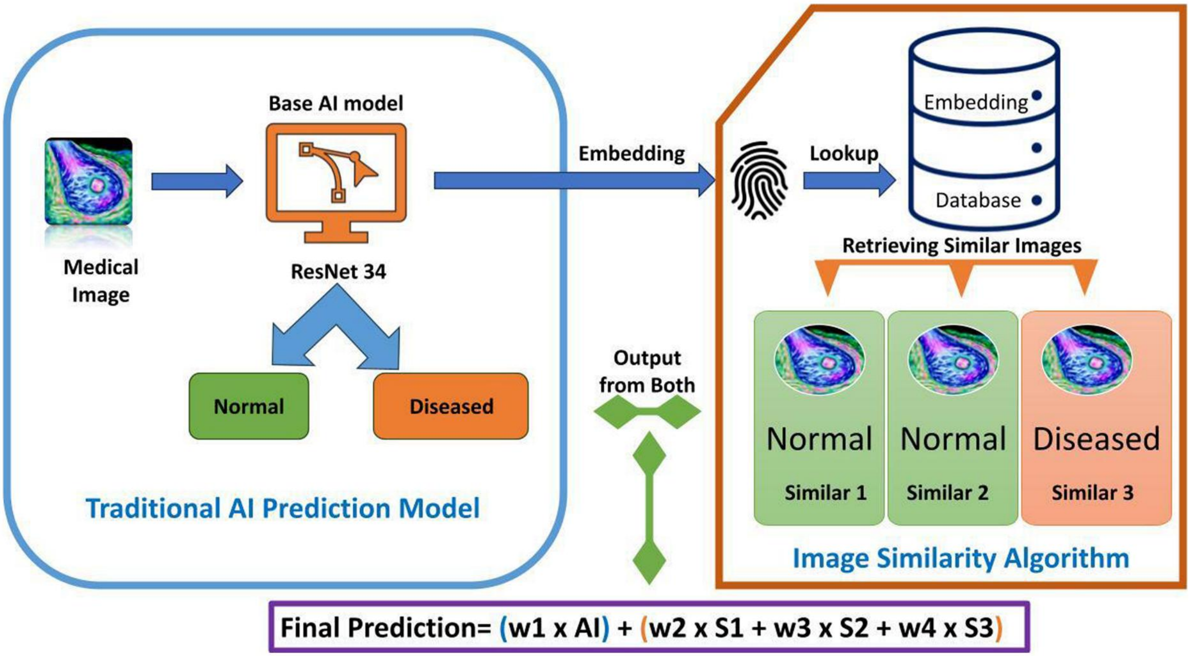

Figure 1. Workflow of the RAMDS hybrid Model for medical image diagnosis. The process begins with the input of a medical image into the traditional ResNet prediction Model, which generates an initial prediction of either "Diseased" or "Normal." Then the embedding from the base AI model is utilized to perform a lookup in an image database to retrieve visually similar images through an Image Similarity Algorithm. These similar images are then analyzed, with their diagnostic labels ("Diseased" or "Normal") considered. The final prediction from the RAMDS model is derived by combining the traditional AI model's output with the weighted agreement from the similar images, leading to a more refined and accurate diagnosis

Table 2. Comparison of ResNet-34 and RAMDS models performance on breast ultrasound images.

	ResNet-34 (%)	RAMDS (%)
Sensitivity	79	90
Specificity	59	49
Accuracy	69	70
Positive Predictive Value	67	65
Negative Predictive Value	73	82

Results

Table 2 compares the performance of the fine-tuned ResNet model and RAMDS for breast cancer classification from ultrasounds. When tuned for negative predictive value, sensitivity increased by 11% and NPV increased by 9%. Specificity decreased by 10% and there was a decrease in PPV by 2%.

Discussion

Explainability in AI diagnostics

The challenge of explainability in AI-based medical diagnosis systems is a topic of considerable interest and concern in the medical community. Traditional deep learning models often operate as black boxes, offering little insight into the reasoning behind their predictions [5]. In contrast, the retrieval-augmented approach in this study enhances explainability by relating the AI's diagnosis to similar historical cases. This approach aligns with the findings of Holzinger *et al.* [14], who emphasized the importance of making AI decisions transparent, understandable, and explainable, especially in healthcare. RAMDS is also helpful in teaching less experienced operators using similar images.

The system's reliance on similarity assessment with historical cases provides clinicians with a reference point, thereby facilitating a better understanding of the model's decision-making process. This is very similar to the principles of case-based reasoning in medical diagnosis, as discussed by Bichindaritz and Marling [6], where past cases are often used to inform the diagnosis of new cases.

Clinicians can also review similar images and corresponding diagnosis using RAMDS graphical user interface. After reviewing this information, they can decide whether to accept or reject the provided predictions based on how similar the images are. This makes clinicians an active participant and a partner in the AI assisted diagnostic model rather than a passive receiver of information.

Adaptability to local imaging contexts

Another important aspect of this system is its adaptability to different imaging contexts without the need for retraining the entire model. This is particularly relevant in medical imaging, where datasets can vary significantly in quality and characteristics across different institutions and regions [7]. By creating embeddings for the new local images and fine-tuning the retrieval augmentation system, the model can be adapted to new contexts, without retraining the base model, enhancing its utility and scalability. This approach tries to alleviate the concerns of Cheplygina *et al.* [8], who highlighted the challenges of transferring AI models to new medical imaging datasets. Our proposed RAMDS framework is composed of three synergistic components: a fine-tuned convolutional neural network (CNN), a similarity finding algorithm, and a weight adjustment algorithm. First, the fine-tuned CNN is designed to extract discriminative features from images. Next, the similarity finding algorithm quantifies the relationships among images, effectively clustering those with shared characteristics. Finally, the weight adjustment algorithm dynamically calibrates the contribution of each feature, mitigating overfitting and reinforcing the model's ability to adapt to new data. This modular design not only enhances the model's performance on the testing data but also underpins its generalization capability, enabling the framework to be effectively applied to any binary image classification problem.

Limitations of the study

RAMDS was not tested in real world clinical setting, nor has it undergone regulatory scrutiny or approval. Prospective real-world testing is preferred when compared to retrospective testing on historical cases. The feasibility of incorporating RAMDS in the radiology workflow is critical and this has not been done. When making a decision, radiologists review multiple images of the same lesion, but RAMDS makes predictions based on a single image. When RAMDS was used specificity decreased by 10%. This is because of the inherent tradeoff between sensitivity and specificity [15]. We could not find other studies using the same datasets in a similar fashion, hence we are not able to provide comparison to other studies. We have not compared it to other models like U-Net with classification encoder, Vision Transformer, etc. Since RAMDS is a framework and any classification model which can be used to generate embeddings can be substituted for the ResNet in RAMDS. In this article, we explored the feasibility of the framework. In the future other models could be substituted for ResNet.

Future directions and improvements

Future research could focus on enhancing the specificity of the retrieval-augmented model without compromising its sensitivity. Additionally, incorporating video clips and multimodal data could potentially improve diagnostic accuracy. A unique direction for exploration for RAMDS is possibly using it to teach medical students about the fundamentals of Case-Based Reasoning.

Conclusion

RAMDS introduces a novel method of using retrieval augmentation in image based medical AI diagnosis. This methodology addresses critical limitations in current medical AI imaging models, particularly the lack of explainability and adaptability, which are possible barriers to clinical implementation. Leveraging both physician expertise and AI efficiency, RAMDS represents a promising advancement in medical image diagnostics. Further research and development are necessary to refine this system and fully realize its potential in clinical settings.

Acknowledgements

We would like to thank all the researchers who made their data publicly available so that we could conduct this study. ChatGPT was used for language and grammar cleanup.

Author contributions

Ethan Thomas Johnson (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Visualization [equal]), Jathin Koushal Bande (Investigation [supporting], Software [supporting], Visualization [supporting]), and Johnson Thomas (Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal])

Conflict of interest statement. No conlict of interest declared.

Funding

No funding received.

Data availability

All datasets used in this study are open-source and are publicly available at the following links. BUSC_Mendeley—https://data.mendeley.com/datasets/wmy84gzngw/1—Rodrigues, Paulo Sergio (2017), Breast Ultrasound Image, Mendeley Data, V1, doi: 10.17632/wmy84gzngw.1. BUSI_corrected—https://www.kaggle.com/datasets/jarintasnim090/busi-corrected—Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data in Brief. 2020 Feb; 28:104863. DOI: 10.1016/j.dib.2019.104863. RODTOOk—origin—http://www.onlinemedicalimages.com/index.php/en/site-map—Rodtook, A., Kirimasthong, K., Lohitvisate, W., Makhanov, S.S. (2018) Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities. Pattern Recognition, Vol 79, pp 172-182. QAMEBI—origin—https://qamebi.com/breast-ultrasound-images-database/—[1] A. Abbasian Ardakani, A. Mohammadi, M. Mirza-Aghazadeh-Attari, U.R. Acharya, An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies, Computers in Biology and Medicine, 152 (2023) 106438. https://doi.org/10.1016/j.compbiomed.2022.106438—[2] H. Hamyoon, W. Yee Chan, A. Mohammadi, T. Yusuf Kuzan, M. Mirza-Aghazadeh-Attari, W.L. Leong, K. Murzoglu Altintoprak, A. Vijayananthan, K. Rahmat, N. Ab Mumin, S. Sam Leong, S. Ejtehadifar, F. Faeghi, J. Abolghasemi, E.J. Ciaccio, U. Rajendra Acharya, A. Abbasian Ardakani, Artificial intelligence, BI-RADS evaluation and morphometry: A novel combination to diagnose breast cancer using ultrasonography, results from multi-center cohorts, European Journal of Radiology, 157 (2022) 110591. https://doi.org/10.1016/j.ejrad.2022.110591. Algorithm blueprint and schematics is available at https://github.com/ZappyFountain/similarity-search/blob/main/similarity-search.

References

1. Berg WA, Campassi C, Langenberg P *et al.* Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. *AJR Am J Roentgenol* 2000;**174**:1769–77. https://doi.org/10.2214/ajr.174.6.1741769
2. Popović ZB, Thomas JD. Assessing observer variability: a user's guide. *Cardiovasc Diagn Ther* 2017;**7**:317–24. https://doi.org/10.21037/cdt.2017.03.12
3. Alowais SA, Alghamdi SS, Alsuhebany N *et al.* Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 2023;**23**:689. https://doi.org/10.1186/s12909-023-04698-z
4. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of 'do no harm'. *Intelligent Medicine* 2024;**4**:52–7. https://doi.org/10.1016/j.imed.2023.08.001
5. Castelvecchi D. Can we open the black box of AI? *Nature News* 2016;**538**:20–3. https://doi.org/10.1038/538020a
6. Bichindaritz I, Marling C. Case-based reasoning in the health sciences: what's next?". *Artif Intell Med* 2006;**36**:127–35. https://doi.org/10.1016/j.artmed.2005.10.008
7. Naddaf M. Medical AI falters when assessing patients it hasn't seen". *Nature* 2024. https://doi.org/10.1038/d41586-024-00094-9
8. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer

learning in medical image analysis. *Med Image Anal* 2019;**54**: 280–96. https://doi.org/10.1016/j.media.2019.03.009

9. Rodrigues PSS *et al.* CAD system for breast US images with speckle noise reduction and bio-inspired segmentation. In: *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2019;68–75. https://doi.org/10.1109/SIBGRAPI.2019.00018
10. Tasnim J, Hasan MK. CAM-QUS guided self-tuning modular CNNs with multi-loss functions for fully automated breast lesion classification in ultrasound images. *Phys Med Biol* 2023;**69**. https://doi.org/10.1088/1361-6560/ad1319
11. Abbasian Ardakani A, Mohammadi A, Mirza-Aghazadeh-Attari M *et al.* An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. *Comput Biol Med* 2023;**152**: 106438. https://doi.org/10.1016/j.compbiomed.2022.106438
12. Rodtook A, Kirimasthong K, Lohitvisate W *et al.* Automatic initialization of active contours and level set method in ultrasound images of breast abnormalities. *Pattern Recognition* 2018;**79**: 172–82. https://doi.org/10.1016/j.patcog.2018.01.032
13. Howard J, Gugger S. *Deep Learning for Coders with Fastai and PyTorch*. Japan: O'Reilly Media, Inc., 2020.
14. Holzinger A, Langs G, Denk H *et al.* Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019;**9**:e1312. https://doi.org/10.1002/widm.1312
15. Monaghan TF, Rahman SN, Agudelo CW *et al.* Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina (Kaunas)* 2021; **57**:503. https://doi.org/10.3390/medicina57050503