Implementing Trustworthy AI in Real-World Medical Imaging Using the SimpleMind Software Environment

Matthew S. Brown[1], M. Wasil Wahi-Anwar[1], Youngwon Choi[1], Morgan Daly[1], Liza Shrestha[1], Koon-Pong Wong[1], Jonathan G. Goldin[1], Dieter R. Enzmann[1]

[1] *Center for Computer Vision and Imaging Biomarkers, Department of Radiological Sciences, David Geffen School of Medicine at UCLA, University of California, Los Angeles, USA*

Abstract

Deep neural networks (DNNs) have good overall performance in medical imaging, but they are susceptible to obvious mistakes that violate common sense concepts. Unexplained errors have reduced trust and prevented widespread adoption in real-world clinical practice. We introduce SimpleMind, an open-source Cognitive AI software environment for medical image understanding. It uses a hybrid Neurosymbolic AI approach that integrates both DNNs and machine reasoning from a knowledge base. We demonstrate its use in building trustworthy AI for checking endotracheal tube (ETT) placement on chest X-rays (CXRs). The AI was integrated into clinical practice and the correctness of the ETT misplacement alerts were compared with radiology reports as the reference. 214 CXRs were ordered by ICU physicians to check ETT placement with AI assistance. ETT alert messages had a positive predictive value (PPV) of 42% and a negative predictive value (NPV) of 98%. Physicians indicated that they agreed with the AI outputs, had increased confidence in their decisions, and were more effective with AI assistance.

Keywords

Trustworthy AI, Medical Imaging, Knowledge Base, Machine Reasoning, Deep Neural Networks

1. Introduction

Deep neural networks (DNNs) detect patterns in data and have shown versatility and strong performance in many computer vision applications. However, despite many research publications there has not been broad adoption of artificial intelligence (AI) in crucial tasks such as medical imaging. While DNNs have good overall performance metrics in medical imaging, they are susceptible to obvious mistakes that violate simple, common sense concepts and are limited in their ability to use explicit knowledge to guide their search and decision making. Obvious, unexplained errors have reduced trust and prevented widespread adoption in real-world clinical practice [1] and motivated research into trustworthy AI [2, 3].

To improve computer vision accuracy and reliability we embed deep neural networks within a Cognitive AI environment. Cognitive AI includes not only the learning of patterns in data, but also learning through teaching and concepts (declared knowledge) as well as reasoning to apply this knowledge to guide the interpretation of a specific image. It uses a hybrid Neurosymbolic AI approach that integrates both DNNs and machine reasoning from a knowledge base. We demonstrate its use and benefits in building reliable, trustworthy AI for checking endotracheal tube (ETT) placement on chest X- rays (CXRs) that has been adopted and evaluated in clinical practice.

Chest radiographs (CXRs) are used in the intensive care unit (ICU) to examine and monitor critically ill patients on life-supporting devices. Endotracheal tubes (ETTs) are used to maintain airway patency and lung ventilation. The desired ETT tip position is within the mid trachea, approximately 5±2 cm above the carina [4]. There is a risk of inefficient ventilation and vocal cord injury if the ETT is too high, and lung collapse, pneumothorax, and even death if the ETT tip is too low. Given the urgent need

NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy
EMAIL: mbrown@mednet.ucla.edu
ORCID: 0000-0001-6805-8151
© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

for intervention if the ETT is misplaced, ICU physicians often take a preliminary look at the CXR at the bedside and immediately adjust a misplaced tube. However, assessment of tube placement can be challenging, especially for non-radiologists.

In this paper we introduce SimpleMind, an open-source Cognitive AI software environment for medical image understanding [5]. We demonstrate its use to build trustworthy AI to assist in checking ETT placement on CXRs in clinical practice and evaluate its real-world performance and physician acceptance.

2. Methods

A SimpleMind application was developed to automatically identify the ETT, trachea, and carina in CXRs. In SimpleMind, an application is built by specifying a knowledge base that describes expected characteristics and relationships between image objects. To check the ETT tip placement, a "safe zone" is defined in the knowledge base as the region inside the trachea and 3 - 7 cm above the carina. SimpleMind computes this region using spatial inferencing for explainable decisions regarding ETT placement. During image understanding, SimpleMind uses the knowledge base to guide DNN segmentation agents and machine reasoning agents that evaluate the results. It enables reasoning on multiple detected objects to ensure consistency, providing cross-checking between DNN outputs. This machine reasoning improves the reliability and trustworthiness of DNNs through an interpretable model and explainable decisions. The CXR application was integrated and evaluated in the clinical imaging workflow at our institution.

2.1. Knowledge Representation

The knowledge base for a SimpleMind application is created as a semantic network (SN) [6] where each node represents an object, object component, or object state. The SimpleMind environment provides a human-readable intuitive language to specify a semantic network. Each node contains attributes that describe expected object characteristics relating to size, shape, pixel intensity, and relative position. Spatial relationships that can be described between objects, include part of, right of, left of, above, below, inside, etc. Attributes are derived from a vocabulary that defines the name of the attribute and its associated parameters. Relational attributes form the links between nodes in the semantic network. For example, the vocabulary defines "RightOf", which includes two parameters: (1) the related node (forming a relational link between nodes), and (2) the expected distance to the right. Fuzzy sets are used to represent prior expectations for object characteristics using a confidence function over the range of possible parameter values [7], so the expected distance to the right is represented in the knowledge base as a fuzzy membership function. The fuzzy functions can be set initially by a human expert and refined by learning from data.

The semantic network attributes can also represent procedural knowledge used by processing agents, including DNN architectures (e.g., U-Net, ResNet, or any user-defined architectures), learning hyper parameters, and image pre and post processing parameters. Crucially, all attribute parameters in the semantic network are exposed (separate from the processing code) and human readable, so they can be both specified by a human and auto optimized by SimpleMind. The SimpleMind environment allows a DNN agent to train weights using the above attribute parameters from a given SN node. The DNN weights are then stored with the node, embedding the DNN within the semantic network. Thus, a SimpleMind knowledge base can include both declared knowledge (that it is "taught") and learned knowledge from examples (acquired through machine learning), i.e., we can actively teach the Cognitive AI as well as have it learn passively from data.

The CXR knowledge base is derived from the medical literature that states that the tip of the endotracheal tube should be 5 ± 2 cm above the carina, where the trachea bifurcates into the two main stem bronchi. The semantic network shown in Fig 1 includes DNNs for the trachea (*trachea_cnn*), carina (*carina_cnn*), and ETT (*et_tube_1_cnn* and *et_tube_2_cnn*). It defines a "safe zone" for the ETT tip using spatial concepts:

- part of the trachea: Line 3 of the *et_zone_1* node (Fig 1B)
- 3 - 7 cm above the carina: Line 4 of the *et_zone_1* node - based on the y-coordinate of the centroid (Fig 1B)
- ETT tip must be inside the safe zone: *et_tip_correct* node describes this relative to the *et_zone* node and represents the state of the ETT tip (Fig 1C)
- the ETT path must be within the trachea (and thus not going into the esophagus): *et_path_incorrect* node describes this relative to the *trachea* node and represents the state of the ETT path
- for the ETT position to be correct the two criteria above must be met - this requirement is defined in the *et_tube_correct* node which represents the final decision of the system based on its machine reasoning (Fig 1D)

The knowledge base also demonstrates checking of DNN outputs for consistency. The *carina_cnn* outputs a coordinate for the position of the carina, represented by the *carina_1* node. The carina location can also be derived from the inferior portion of the trachea where it branches into the two main stem bronchi, represented by the *carina_2* node. The *carina_3* node indicates that these two should correspond and refines the final result. Accurate detection the carina is necessary for ETT position checking. Crucially, if the alternate carina locations do not correspond, then the system will report that it is unable to reliably identify the carina rather than outputting an incorrect result. Using knowledge to identify interpretation errors is an important benefit of machine reasoning that allows the system to determine when it is likely to be wrong rather than failing silently.

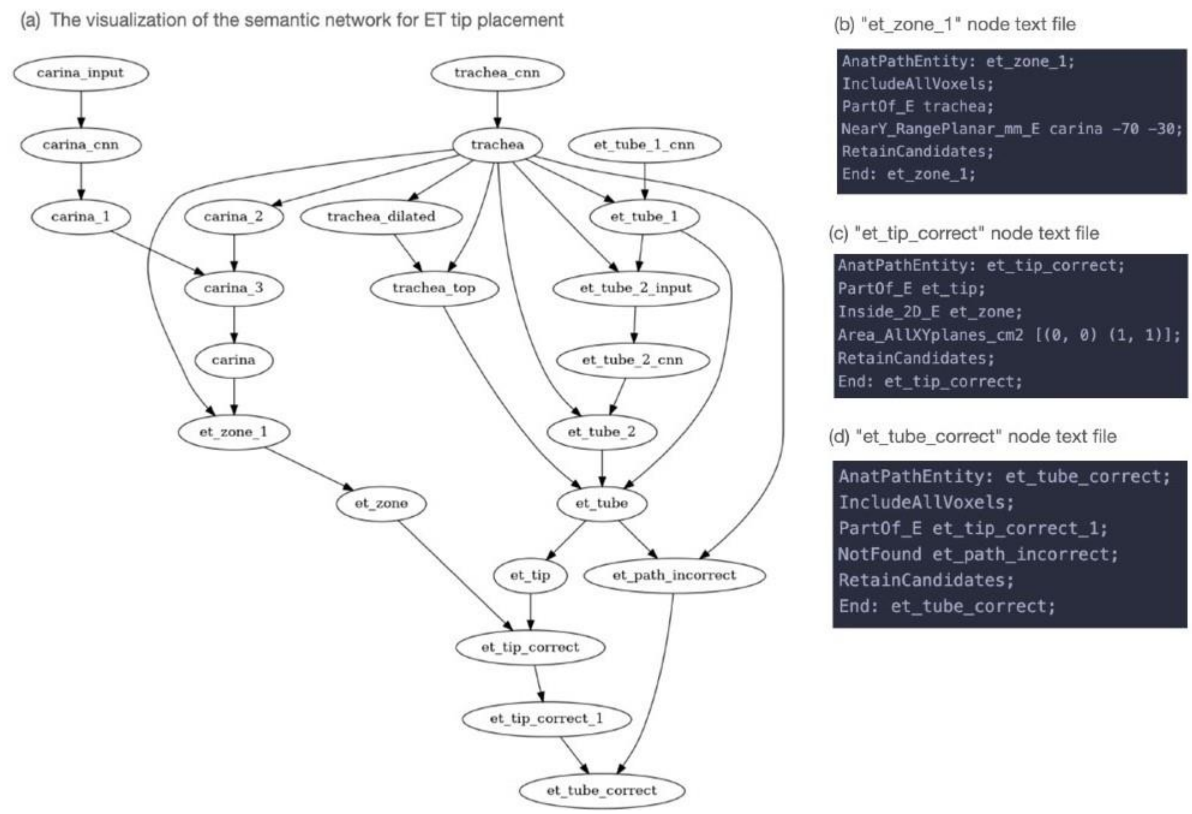

Figure 1: The SimpleMind semantic network (SN) for the ETT with selected node details.

2.2. Multi-Agent Thinking

SimpleMind provides a Think Module for computer vision, i.e., recognizing objects (nodes) from the knowledge base in a given image. Multiple software agents work together to segment the image into candidate regions, then select the best candidate based on object attributes described in the knowledge base. Software agents collaborate to solve the vision problem by reading from, and writing to, a global Blackboard data structure [8]. The Blackboard is the working space of SimpleMind during the "thinking" process, i.e., during comparison and matching of the image to a knowledge base for image understanding. An agent can read information from the Blackboard generated by other agents and add or update information. Agents operate independently and collaborate only via the Blackboard, giving them a degree of autonomy and making the system more flexible and scalable. Agent types provided in SimpleMind are shown in Fig 2.

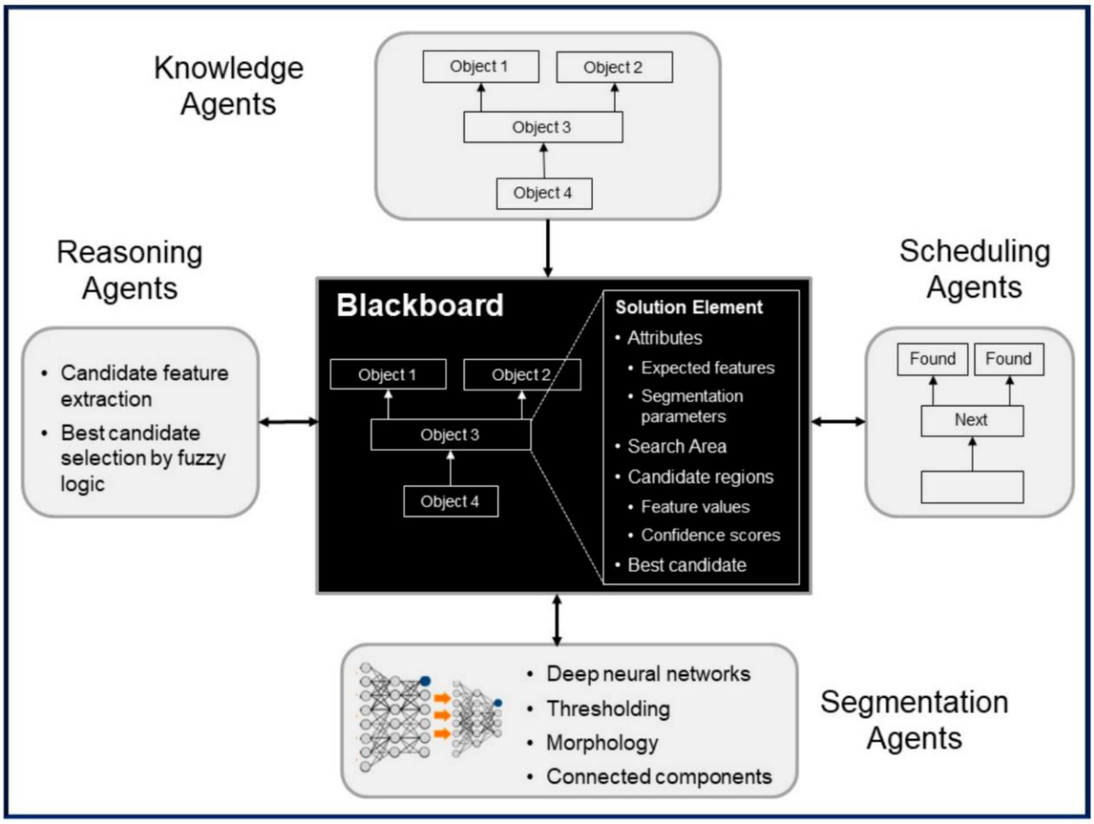

Figure 2: SimpleMind multi-agent thinking architecture. Schematic showing agent types and the Blackboard for sharing of results during image understanding.

For each node in the semantic network, a data structure called a Solution Element is created on the Blackboard, corresponding to an object to be recognized in the image. The Solution Element stores all agent contributions while recognizing the object. Knowledge base attributes and candidate image regions are transformed into a common feature space for selection of the best candidate. A Knowledge Agent accesses the knowledge base and creates each Solution Element, initializing it with prior expectations for object feature values. Thus, the objects and their relationships represented in the semantic network are transformed into a directed graph of Solution Elements on the Blackboard, with the direction of the link reflecting the dependency of an object's attribute upon another object.

Solution Elements are processed sequentially by agents. Objects are recognized in order based on the directed links between their Solution Elements. When a particular Solution Element is scheduled for processing (by a Scheduling Agent), a Reasoning Agent computes an image search area using spatial inferencing from the relationships to previously recognized objects. This search area is provided as a mask to guide a Segmentation Agent that generates candidate image regions for the object. Segmentation is typically performed by a DNN agent that generates multiple connected components as candidate regions. Feature values are computed for each candidate region and compared against the expected values by a Reasoning Agent. Feature values are computed for each candidate according to the attributes provided in the knowledge base and the corresponding fuzzy membership function yields a confidence value for that attribute. The overall confidence for a candidate region is then computed as the minimum confidence of any attribute. Thus, by pattern classification, the candidate that best matches these expectations from the knowledge base is selected.

Agents are activated iteratively, one at a time. At each iteration an activation score is computed for each registered agent. The agent with the highest score is activated and can contribute to the solution on the Blackboard. Each agent provides a function to compute its activation score based on the contents of the Blackboard, in particular whether the Solution Element being processed has the relevant attributes and necessary data required by the agent. The process repeats until all activation scores are zero and no further agents activate. The system control is simple, yet highly flexible, with agent priorities determined through their activation functions.

The contents of the Blackboard reflect what SimpleMind is thinking at any point in time and its current understanding of the image. Once all Solution Elements have been processed, the Blackboard contains an instantiation of the general knowledge base to a particular image. The object attributes from the general knowledge base are now instantiated with actual numerical feature values from the image, enabling further high-level reasoning.

2.3. Machine Learning

SimpleMind provides a Learn Module for machine learning within a knowledge base, in particular for training the weights of embedded DNNs. For the CXR AI, we used 2000 images collected retrospectively from ICU patients between April 2018 and September 2019. All of the DNN nodes are trained with 1488 images, and the application was initially tested experimentally on 512 images [9].

Although not used in this application, the SimpleMind environment also includes a Knowledge Network Learning and Optimization (KNoLO) method. It comprehensively co-optimizes all attribute parameters from all nodes simultaneously, including: object expected characteristics, DNN input channels and image preprocessing options, and DNN learning hyper parameters. The parameter optimization is performed using a genetic algorithm and details can be found in [5].

2.4. Implementation and Evaluation in Clinical Practice

The CXR AI is currently deployed within our institution's clinical workflow for investigational use only as a quality improvement (QI) tool. A new cloud-based computing infrastructure was designed to integrate the CXR AI system with the clinical Picture Archiving and Communications System (PACS). An image router was configured to push CXRs to an on-premise Azure AI/ML platform where the AI system is deployed. The CXR AI processes the image, detects tubes and anatomic landmarks on the image, and generates an enhanced CXR image with overlays and an alert/informational message that is pushed back to the PACS. Both the original and AI CXR images are available in PACS viewers for the radiologist or ICU physician. The total turnaround time from CXRs reaching PACS to the AI output being available in PACS is within 3 to 4 minutes, ensuring that AI outputs are available to ICU physicians at the point of care during their CXR review. A specific order code was set up for CXR with AI processing, providing a limited deployment on identifiable cases to be reviewed by a selected pool of ICU physicians and radiologists. From June 11, 2021 to November 3, 2022, 214 CXRs were ordered by ICU physicians through this specific order code for checking ETT placement with AI assistance.

The AI displays one of the three possible ETT messages: (1) "Found" (ETT tip was determined to be in the safe zone), (2) "Position Alert" (ETT tip was not in the safe zone or the AI could not determine the safe zone), (3) "Not Found" (no ETT was detected by the AI). The AI alerts were evaluated against the findings in the radiology report in which the radiologists were asked to include the following statement: "An investigational endotracheal tube AI overlay was available and was/was not consistent with my interpretation". We evaluated the AI performance by defining a positive output (alert) as messages (2) and (3), and a negative output (no alert) as message (1). When the AI output was positive, a true positive (TP) required that the ETT be misplaced per the radiology report or that the ETT was missing (since cases being routed to AI were expected to have an ETT), otherwise it was a false positive (FP). When the output was negative, a false negative (FN) required the ETT to be misplaced, otherwise it was a true negative (TN). When alerts were issued, follow-up CXRs were reviewed and radiology reports checked to confirm repositioning of the tube. Positive predictive value (PPV = TP/(TP+FP)) and negative predictive value (NPV=TN/(TN+FN)) metrics were computed to give a sense of trustworthiness of the AI from a physician perspective. In previous experimental testing, the PPV, NPV, and sensitivity to misplaced tubes were 42%, 99%, and 95%, respectively [8]. The AI system was designed to be highly sensitive to avoid missed alerts when ETTs were misplaced, thus higher NPV and lower PPV were considered sufficient for the system to be deployed in clinical practice and further evaluated as described in this paper.

A survey was conducted to qualitatively evaluate the ICU physicians and radiologists experience in using the CXR AI in their clinical workflow. They were asked to provide ratings for usefulness and satisfaction with the AI clinical application.

3. Results

For the 214 CXR images ordered to check ETT placement with AI assistance by ICU physicians, a confusion matrix is shown in Table 1. The AI alert messages had a positive predictive value (PPV) of 42% (21 / (21 + 29)) and a negative predictive value NPV of 98% (161 / (161 + 3)) based on the radiology reports. These performance metrics were consistent with clinical requirements and previous experimental testing.

Table 1: Confusion matrix of AI alert vs. actual ETT placement

	AI Alert	
Actual ETT placement	**Yes**	**No**
Incorrect	21	3
Correct	29	161

The AI generates CXR overlays, showing the ETT path and distance from the ETT tip to the carina as shown in Figs 3 and 4. Fig 3 is a case with correct ETT position showing the internal results of SimpleMind (Fig 3b,c) that explain why it thinks the position is correct (tip inside the safe zone) and the final output of the system as presented to the ICU physician (Fig 3D). Fig 4 shows an example of incorrect ETT placement, with the tip too low relative to the carina (outside of the safe zone).

Seven clinicians completed the user survey: three were radiologists with 9 - 26 years of experience, four were physicians with 1 - 10 years of experience in critical care medicine. Five of the seven clinicians had reviewed over 20 CXRs with AI, and two had reviewed over 50. Table 2 summarizes the frequency and median ratings. Users indicated that they agreed with the AI outputs, had increased confidence in their decisions, and were more effective with AI assistance. The trust and willingness to adopt the system was further confirmed in weekly user group meetings.

Table 2: Frequency of user ratings for survey questions: 1 = “Strongly disagree”; 5 = “Strongly agree”.

System Outputs	**1**	**2**	**3**	**4**	**5**	**Median**
The system output agrees with my assessment	0	0	1	4	2	4
Tube annotations are helpful/appropriate	0	0	2	3	2	4
Informational/alert messages are helpful/appropriate	0	0	3	2	2	4
The system output increases my confidence	0	0	2	3	2	4

Usefulness and Satisfaction	**1**	**2**	**3**	**4**	**5**	**Median**
It helps me be more effective	0	0	3	4	0	4
It helps me be more productive	0	1	4	2	0	3
It works the way I want it to work	0	0	3	4	0	4
I am satisfied with it	0	0	2	4	0	4

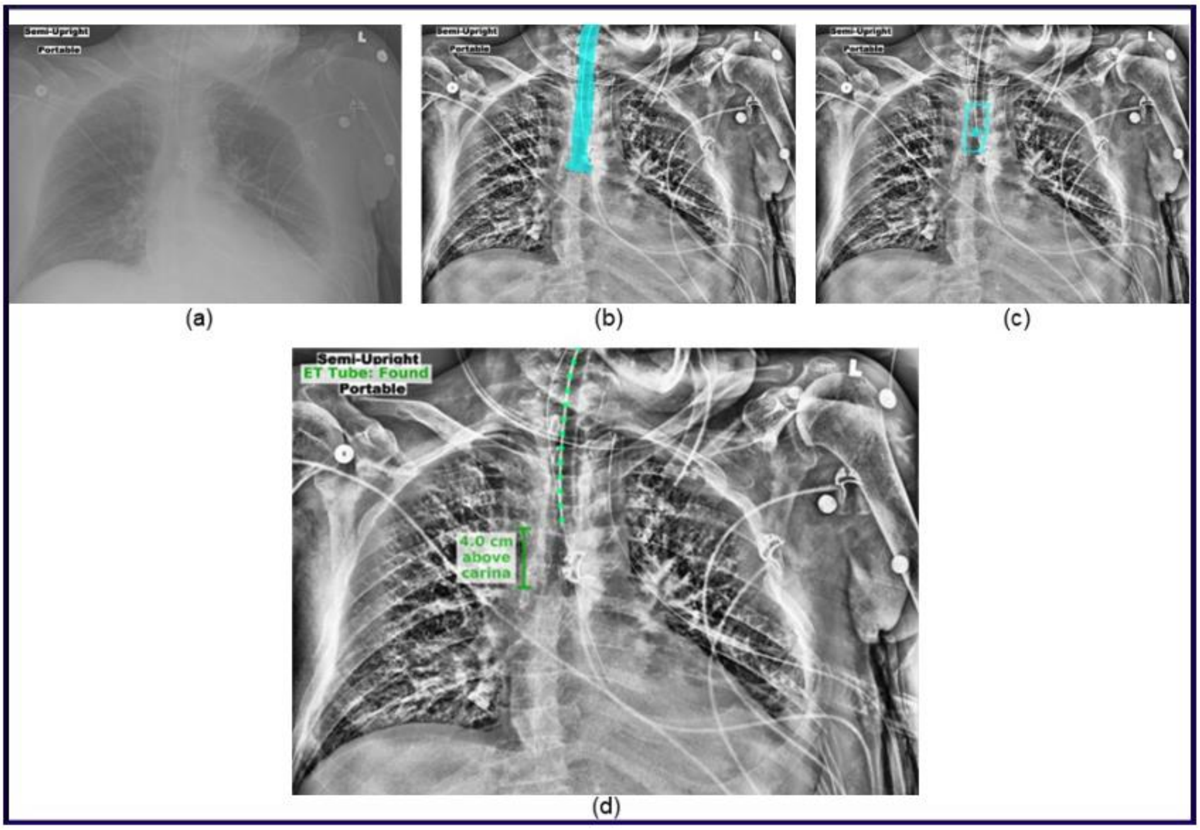

Figure 3: Visualization of SimpleMind nodes for correct ETT placement. (A) Original CXR; (B) *trachea* region on an enhanced image; (C) ETT safe zone (*et_zone*) and tube tip location (*et_tip*) showing the tip within the safe zone; (D) output of the system as presented to the physician with the green overlay indicating correct tube placement.

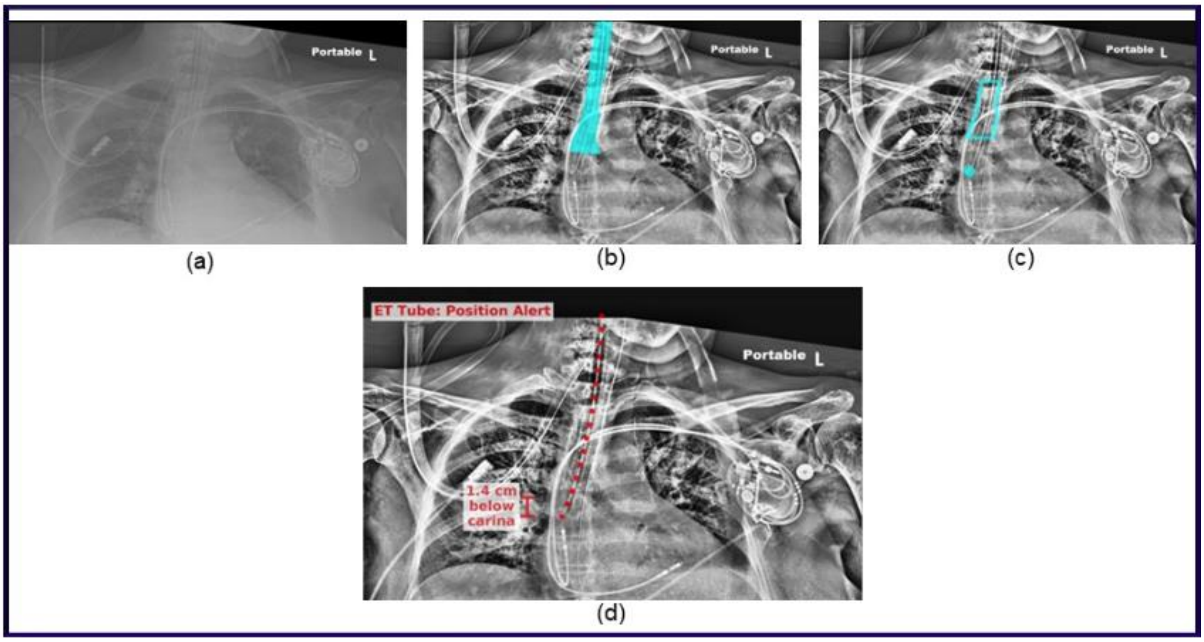

Figure 4: Visualization of SimpleMind nodes for incorrect ETT placement. (A) Original CXR; (B) *trachea* region on an enhanced image; (C) ETT safe zone (*et_zone*) and tube tip location (*et_tip*) showing the tip outside the safe zone; (D) output of the system as presented to the physician with the red overlay indicating incorrect tube placement (tip too low relative to the carina).

4. Discussion

SimpleMind brings explainability and trustworthiness to ETT placement checking on CXRs using a knowledge base that describes not only the ETT but also relevant anatomic landmarks and includes relational attributes to cross-check multiple DNNs and ensure consistency and overall reliability of the system. Rather than attempting to learn misplacement of the ETT indirectly from examples, the SimpleMind knowledge base can directly describe when an alert should be given based on the tip location relative to the carina.

SimpleMind is a Cognitive AI software environment that enables users to build applications for image understanding by specifying a knowledge base in the human-readable language of SimpleMind and then tuning its parameters. Developing a SimpleMind application is like teaching or instructing a human at a cognitive level, it allows non-programmers to build a medical application directly and completely using their domain knowledge without knowing the details of the processing code. At runtime, the knowledge base is applied to recognize objects. SimpleMind is open source (https://gitlab.com/sm-ai-team/simplemind) and the environment can be extended through application programming interfaces (APIs) whereby developers can expand the vocabulary and implement new processing algorithms as agents. SimpleMind automatically handles the aggregation and chaining of many processing agents, enabling a multi-DNN Cognitive AI system. It has also been applied to segmentation of the kidney on CT [10, 11] and the prostate on MRI [12].

SimpleMind can be considered a "hybrid learning system" that brings together features from connectionism and symbolic AI. Four key advantages have been suggested as arising from this combined approach [13]: (1) interpretability, (2) error recovery, (3) out of distribution (OOD) handling, and (4) learning from small data; and SimpleMind supports DNNs accordingly:

- It allows explicit knowledge to be applied systematically to improve performance and reliability.
  - Computing a search area in which to apply DNN segmentation using spatial relationships in the semantic network.
  - Selection of the best candidate image region outputted by the DNN based on expected characteristics defined in the knowledge base, or conversely, rejection of the output if it does not meet expectations. It enables reasoning on multiple detected objects, providing cross-checking between DNN outputs for more robust the image interpretation. Rejecting candidates for an object does not preclude recognition of subsequent objects based on other knowledge and avoids propagating errors. This gives SimpleMind applications more resilience in handling OOD cases and error recovery.
- It provides a high degree of interpretability and explainability.
  - The knowledge base makes explicit the knowledge that was previously implicit in pre and post processing code and makes it easier to apply more knowledge intuitively.
  - The thinking of SimpleMind as it processes an image is captured in the Blackboard. A human can know what it was thinking by reviewing the Blackboard contents.
- Using a human-provided knowledge base, SimpleMind can perform object recognition with little or no training data.
  - When there is insufficient data to train a DNN, other segmentation agents (e.g., intensity thresholding or edge detection) can use the knowledge base to generate initial segmentation results. Little or no training data is needed since the initial semantic network can be constructed using declarative knowledge rather than machine learning. These initial results can be used with manual editing to generate training sets for DNN learning. When an OOD situation arises it can be added to the knowledge base and handled without training data being initially available.

These benefits are also consistent with goals of trustworthy AI according to the High-Level Expert Group on AI from the European Commission [14], in particular the following guidelines:

- Transparency: AI systems and their decisions should be explained.
- Technical Robustness and safety: AI systems need to be resilient with a fall back plan in case something goes wrong.
- Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions with proper oversight mechanisms.

5. Conclusion

SimpleMind is a Neurosymbolic AI environment for medical imaging that supports DNNs with a knowledge base and machine reasoning. It was used to build an AI for checking ETT placement on CXR that was adopted and evaluated as trustworthy in real-world clinical practice. We believe that there is strong potential utility for broader research and commercial applications in building trustworthy AI. The open source software allows for knowledge base expansion and agent aggregation by a community of developers.

6. Acknowledgements

The authors wish to thank their friends and colleagues at the UCLA Center for Computer Vision and Imaging Biomarkers.

7. References

[1] T. Panch, H. Mattie, L.A. Celi. The "inconvenient truth" about AI in healthcare. NPJ digital medicine. 2019;2(1):1–3.
[2] N. Hasani, M.A. Morris, A. Rahmim, R.M. Summers, E. Jones, E. Siegel, et al. Trustworthy artificial intelligence in medical imaging. PET clinics. 2022;17(1):1–12.
[3] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion. 2022: 79, (3), 263--278, doi:10.1016/j.inffus.2021.10.007
[4] L.R. Goodman, P.A. Conrardy, F. Laing, M.M. Singer. Radiographic evaluation of endotracheal tube position. AJR Am J Roentgenol. 1976;127(3):433-434. doi:10.2214/ajr.127.3.433.
[5] Y. Choi, M.W. Wahi-Anwar, M.S. Brown. SimpleMind adds thinking to deep neural networks. 2022. arXiv:2212.00951 [cs.AI].
[6] M.R. Quillian. Semantic Networks. In: Minsky ML, editor. Semantic Information Processing. MIT Press; 1968.
[7] L.A. Zadeh. Fuzzy sets. Information and Control. 1965;8(3):338–353. doi:https://doi.org/10.1016/S0019-9958(65)90241-X.
[8] D. Corkill. Collaborating Software. In: International Lisp Conference, New York. vol. 44; 2003.
[9] M.S. Brown, K.P. Wong, L. Shrestha, M. Wahi-Anwar, M. Daly, G. Foster, et al. Automated Endotracheal Tube Placement Check Using Semantically Embedded Deep Neural Networks. Academic Radiology Vol. 30, No. 3, March 2023 pp 412-430. PMID: 35644754 DOI: 10.1016/j.acra.2022.04.022.
[10] G. Melendez-Corres, M.W. Wahi-Anwar, Y. Choi, H. Coy, S.S. Raman, M.S. Brown. Accelerating training data annotation via a continuous ai-assisted, human-supervised feedback loop in kidney segmentation in ct.; 2021. Available from: http://archive.rsna.org/2021/704158.html.
[11] G. Melendez-Corres, Y. Choi, M.W. Wahi-Anwar, H. Coy, S.S. Raman, M.S. Brown. Machine reasoning for segmentation of the kidneys on CT images: improving CNN performance by incorporating anatomical knowledge in post-processing. Available from: https://spie.org/medical-imaging/presentation/Machine-reasoning-for-segmentation-of-the-kidneys-on-CT-images/12465-63.
[12] Y. Choi, M. Garcia, S.S. Raman, D.R. Enzmann, M.S. Brown. AI-human interactive pipeline with feedback to accelerate medical image annotation. In: Medical Imaging 2022: Computer-Aided Diagnosis. vol. 12033. SPIE; 2022. p. 741–747.
[13] M.K. Sarker, L. Zhou, A. Eberhart, P. Hitzler. Neuro-Symbolic Artificial Intelligence: Current Trends; 2021. Available from: https://arxiv.org/abs/2105.05330.
[14] P. Ala-Pietila, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer W, et al. The assessment list for trustworthy artificial intelligence (ALTAI). European Commission; 2020.