

Knowing When Not to Act: Toward Capability-Aware Healthcare AI

Marzia Hoque Tania^[0000–0002–4496–1896]

Centre for Big Data Research in Health, University of New South Wales, Sydney,
Australia

`m.hoque_tania@unsw.edu.au`

Abstract. The growing adoption of AI in healthcare has increasingly shifted research from improving predictive performance toward ensuring trustworthy clinical deployment. Advances in uncertainty quantification, confidence calibration, out-of-distribution detection, selective prediction, and human–AI collaboration have substantially improved the ability to determine whether a prediction should be trusted. A complementary question, however, has received comparatively less attention: when should an AI system participate in clinical decision making? Addressing this question shifts the focus from evaluating prediction reliability to estimating the operational capability of an AI–clinical interaction. In this paper, a conceptual formulation based on four levels of abstraction—interaction, capability, competence, and operational admissibility—is developed to motivate the notion of an Operational Capability Set, representing the evolving domain over which AI participation is operationally admissible. This formulation provides a common perspective for interpreting uncertainty estimation, confidence calibration, out-of-distribution detection, selective prediction, and related approaches as complementary sources of evidence for capability estimation. It also highlights connections between trustworthy healthcare AI, computational metacognition, and meta-learning, providing a conceptual basis for future research on adaptive and trustworthy healthcare AI.

Keywords: Metacognition · Trustworthy AI · Clinical Decision Support

1 Introduction

Safe clinical practice depends not only on making accurate decisions but also on recognizing when a decision should not be made. Every day, clinicians assess whether they have sufficient expertise, information, and contextual understanding to diagnose a patient, recommend treatment, or manage a complex clinical situation. When these conditions are not met, they may seek additional investigations, consult colleagues, or refer the patient to another specialist. In this sense, good clinical judgment involves recognizing the boundaries of one’s own operational role as much as exercising one’s expertise.

As artificial intelligence (AI) becomes increasingly integrated into healthcare, AI systems are expected to support a growing range of clinical tasks,

including diagnosis, medical image interpretation, prognosis, treatment recommendation, patient monitoring, and clinical decision support [14, 3, 2]. Over the past decade, considerable progress has been made in improving predictive performance while simultaneously advancing methods that promote trustworthy deployment, including uncertainty quantification, confidence calibration, out-of-distribution (OOD) detection, selective prediction, and human–AI collaboration. These developments have substantially strengthened the reliability and safety of AI-assisted clinical decision making.

Complementary research has emphasised explainability, privacy, inclusivity, personalisation, and stakeholder-centred system design as important requirements for translating predictive models into trustworthy healthcare systems [6, 10, 11, 15].

Despite this progress, a fundamental question remains less explicitly addressed: When should an AI system participate in a clinical decision?

This question differs subtly from asking whether a particular prediction is accurate, confident, or well calibrated. An AI system may generate a highly confident prediction for a patient whose presentation lies beyond the conditions under which the model has been adequately validated. Conversely, a system may encounter incomplete or noisy clinical information while still operating within a context in which it has demonstrated reliable performance. Likewise, mechanisms for selective prediction or abstention determine *when* a model should withhold a recommendation, but they do not necessarily provide a broader understanding of *whether* the AI system is an appropriate participant in the clinical decision itself.

These observations suggest a shift in perspective. Much of trustworthy healthcare AI has focused on improving the trustworthiness of predictions. Increasingly, however, trustworthy deployment may also require considering the trustworthiness of participation. In other words, beyond evaluating whether a prediction should be trusted, future healthcare AI may also need to determine whether it should participate in a particular clinical decision under the prevailing patient, task, and deployment conditions.

In this perspective, we explore this broader question through the notion of *capability awareness*. We use the term as a conceptual lens for interpreting existing research on trustworthy healthcare AI [9]. From this perspective, uncertainty estimation, calibration, distributional awareness, selective prediction, and human oversight can be viewed as complementary sources of evidence informing whether an AI system is appropriately positioned to contribute to clinical decision making.

We aim to encourage a broader conversation about the role of AI in clinical care. We suggest that trustworthy healthcare AI may increasingly be viewed not only as a problem of producing reliable predictions, but also as a problem of understanding the boundaries of appropriate participation. The remainder of this paper develops this perspective.

2 Existing Approaches Address Different Questions

The increasing emphasis on trustworthy healthcare AI has stimulated substantial research into methods that improve the safety, reliability, and robustness of clinical decision support systems [13]. Over the past decade, advances in uncertainty quantification, confidence calibration, OOD detection, selective prediction, conformal prediction, and human–AI collaboration have collectively strengthened the foundations of trustworthy AI in healthcare [4, 1].

Although these research areas often overlap in practice, they have largely evolved independently, each addressing a different question about the behavior of AI systems. Viewing them through the questions they seek to answer, rather than through their underlying methodologies, reveals both their complementarity and their conceptual boundaries.

2.1 Can This Prediction Be Trusted?

One of the central questions in trustworthy AI concerns the reliability of individual predictions.

Confidence estimation, uncertainty quantification, Bayesian learning, ensemble methods, and related approaches seek to characterize how strongly a model supports a particular prediction and how much uncertainty surrounds that prediction. Similarly, calibration methods examine whether predicted confidence corresponds to observed performance, allowing clinicians to interpret probabilistic outputs more appropriately.

These approaches [12] have become essential components of trustworthy healthcare AI because they improve the interpretability and reliability of individual predictions. In clinical AI, explainability has similarly been used to expose the features and multimodal evidence contributing to predictions, thereby supporting transparency and clinical interpretation [6, 7, 17]. Their primary focus, however, remains the prediction itself rather than the broader clinical context in which that prediction is generated.

2.2 Have We Seen Cases Like This Before?

Another important line of research considers whether a patient resembles the data on which an AI system was developed [5].

Methods for out-of-distribution detection, dataset shift analysis, and domain generalization attempt to identify situations in which deployment conditions differ from those encountered during training. These methods have particular importance in healthcare, where patient populations, imaging devices, institutional practices, and disease prevalence often vary considerably across clinical settings.

This perspective addresses an important aspect of trustworthy deployment by asking whether current clinical data fall within the model’s previous experience. Nevertheless, similarity to prior data represents only one dimension of clinical appropriateness. Patients may resemble the training population while presenting

exceptionally complex clinical situations, whereas some forms of distributional shift may have relatively little practical effect on clinical performance.

2.3 Should the Model Abstain?

Recognizing that predictions are not always appropriate, another body of research has explored selective prediction, reject-option classification, conformal prediction, and human-in-the-loop decision support [16].

These approaches address an important practical question:

Under what conditions should an AI system withhold a recommendation or defer to human expertise?

Rather than requiring AI systems to produce predictions for every patient, these methods acknowledge that abstention, escalation, or referral may sometimes represent the safest clinical response.

In this sense, selective prediction shifts attention from generating predictions to managing clinical risk. However, these approaches primarily specify *how* an AI system should behave once predefined conditions have been identified, rather than addressing the broader question of how the appropriateness of participation should be assessed.

3 A Different Way to Think About Trustworthy Healthcare AI

The existing approaches discussed in the previous section share a common computational objective: determining whether a prediction should be trusted. We consider a different computational problem: determining whether an AI system should participate in clinical decision making before prediction itself is performed.

The proposed capability-aware perspective distinguishes four levels of abstraction: *interaction*, *capability*, *competence*, and *operational admissibility*. The primitive object is the interaction between an AI system and a clinical situation. Each interaction possesses a latent operational capability that cannot be observed directly. The AI system therefore maintains an evolving estimate of this capability, referred to as competence. Operational admissibility is subsequently determined from this estimate rather than assumed *a priori*. The mathematical formulation below follows directly from this decomposition.

Hypothesis. Trustworthy healthcare AI can be formulated as the sequential estimation of an evolving Operational Capability Set over the interaction space between AI systems and clinical situations.

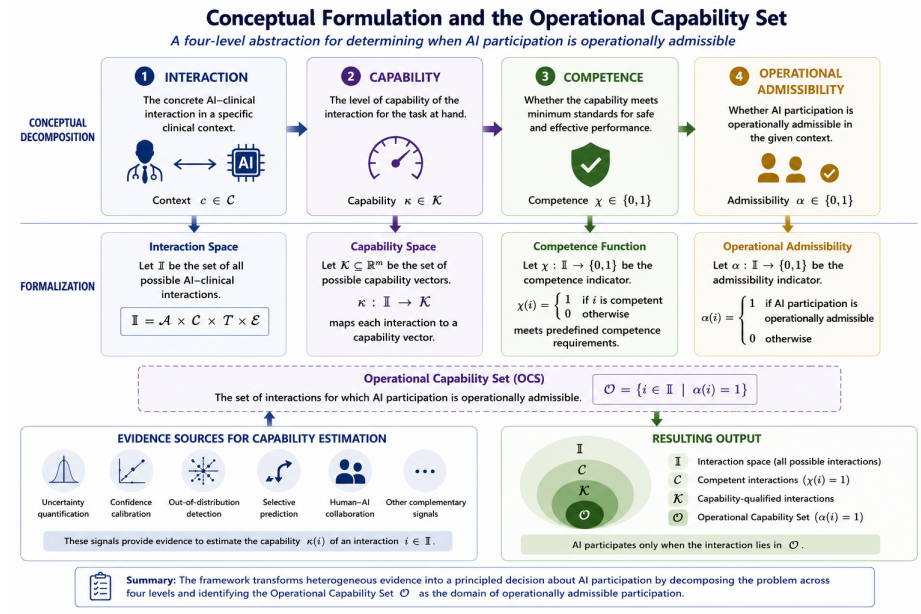


Fig. 1. Ontology of capability-aware healthcare AI.

3.1 Interaction Space

Let

$$\mathcal{X} = \mathcal{P} \times \mathcal{T} \times \mathcal{E}, \quad (1)$$

denote the space of clinical situations, where $\mathcal{P}$, $\mathcal{T}$, and $\mathcal{E}$ represent the patient, clinical task, and deployment environment, respectively. Let $\mathcal{M}$ denote the space of AI systems and define the interaction space

$$\mathcal{I} = \mathcal{M} \times \mathcal{X}, \quad (2)$$

whose elements $i = (m, x) \in \mathcal{I}$ represent interactions between an AI system and a clinical situation. Conventional healthcare AI learns a predictive mapping

$$f : \mathcal{I} \rightarrow \mathcal{Y}, \quad (3)$$

where $\mathcal{Y}$ denotes the space of clinical outputs. The present framework instead takes the interaction itself as the fundamental object of analysis.

3.2 Capability and Competence

Operational capability is represented by the latent mapping

$$C_t^* : \mathcal{I} \rightarrow \mathbb{R}, \quad (4)$$

where $C_t^*(i)$ denotes the latent operational capability index associated with interaction i . Larger values indicate greater operational admissibility for AI participation. The scalar representation is adopted for convenience; vector-valued or latent representations constitute equally valid realizations of the framework.

Operational capability is relational. It characterizes the interaction between an AI system and a clinical situation and implicitly aggregates all factors influencing operational admissibility, including the capabilities of the AI system, the available clinical evidence, the deployment environment, and the current state of medical knowledge. This relational view is consistent with recent healthcare-AI work in which model behaviour is considered jointly with patient heterogeneity, multimodal evidence, stakeholder requirements, personalisation, and deployment context [10, 11, 15, 2].

Since operational capability is not directly observable, the AI system maintains an estimate

$$\hat{C}_t : \mathcal{I} \rightarrow \mathbb{R}, \quad (5)$$

which we refer to as the system's *competence*. Capability is the latent operational property of an interaction, whereas competence is its evolving estimate. Competence may be refined through operational evidence including data-quality assessment, stakeholder or clinician feedback, patient monitoring, and observations accumulated during deployment [8, 15, 2]. Continual and online learning provide additional mechanisms through which such evidence may subsequently modify the underlying model.

3.3 Operational Capability Set

For an operational threshold $\tau \in \mathbb{R}$, define the Operational Capability Set

$$\mathcal{C}_{\tau,t} = \{i \in \mathcal{I} : C_t^*(i) \geq \tau\}, \quad (6)$$

and its estimated counterpart

$$\hat{\mathcal{C}}_{\tau,t} = \{i \in \mathcal{I} : \hat{C}_t(i) \geq \tau\}. \quad (7)$$

The Operational Capability Set characterizes the interactions for which AI participation is operationally admissible under the intended deployment conditions. The framework intentionally does not prescribe how operational admissibility should be translated into clinical workflow; autonomous prediction, clinician review, referral, or additional data acquisition are implementation-specific decisions.

Within this formulation, uncertainty estimation, confidence calibration, out-of-distribution detection, selective prediction, deployment monitoring, and clinician feedback are interpreted as complementary sources of evidence for refining the competence estimate $\hat{C}_t$ and, consequently, the estimated Operational Capability Set $\hat{\mathcal{C}}_{\tau,t}$. The Operational Capability Set thus plays a role analogous to feasible sets in constrained optimization, reachable sets in control theory, viability kernels in viability theory, and safe sets in reinforcement learning by defining the operational domain over which AI participation is admissible.

Table 1. Comparison of existing trustworthy AI concepts and the proposed capability-aware perspective.

Concept	Primary question
Prediction	What is the most likely clinical outcome?
Confidence	How certain is the predicted outcome?
Calibration	Is the reported confidence reliable?
Uncertainty quantification	How much uncertainty is associated with the prediction?
Out-of-distribution detection	Does the current interaction differ from prior experience?
Selective prediction	Should the model abstain from making a prediction?
Capability (latent)	What is the AI system’s true operational capability for this interaction?
Competence (proposed)	What is the AI system’s estimated operational capability?
Operational admissibility (proposed)	Should the AI participate in this clinical decision?

Accordingly, the capability-aware perspective shifts the primary object of inquiry from the trustworthiness of predictions to the estimation of the operational domain within which AI participation is admissible.

4 Discussion and Concluding Perspective

The capability-aware perspective formulates trustworthy healthcare AI as an estimation problem (Table 1). Uncertainty quantification, confidence calibration, out-of-distribution detection, selective prediction, and related methods characterize different aspects of prediction reliability. Capability estimation concerns a different object: the operational capability associated with an AI–clinical interaction.

Capability and competence separate the latent operational property of an interaction from its estimate. This formulation is closely related to computational metacognition, where performance on a task is accompanied by an assessment of the system’s ability to perform that task. Selective prediction, by comparison, determines whether a prediction should be accepted or withheld. Abstention is therefore a possible downstream response to limited estimated competence, whereas competence estimation concerns the operational domain of the AI–clinical interaction.

Consider two similar pneumonia predictions. The first is produced for an adult chest radiograph acquired under conditions represented during validation; the second concerns a neonatal intensive-care patient imaged using an unfamiliar protocol. Similar outputs need not imply similar operational admissibility. Evidence supporting capability may differ because of the patient, task, avail-

able information, or deployment conditions. Operational admissibility therefore cannot be determined from the prediction alone.

Comparable distinctions arise in other clinical applications. For example, the reliability of physiological monitoring can depend on whether an acquired PPG signal is itself of sufficient quality for downstream analysis [8], while multimodal diagnostic systems must additionally determine how heterogeneous clinical evidence should be integrated and interpreted [7, 17].

Clinical deployment also changes over time. Patient populations, clinical practice, data-generating processes, available evidence, and AI systems may differ from those encountered during initial validation. Current practice addresses these changes through intended-use specification, external and subgroup validation, quality control, uncertainty estimation, calibration, shift detection, monitoring, domain adaptation, transfer learning, retraining, and online or continual learning. Their roles within capability estimation are summarized in Table 2.

These mechanisms affect capability in different ways. Uncertainty, calibration, distributional shift, observed outcomes, monitoring, and clinician feedback provide evidence for updating the competence estimate $\hat{C}_t$. Domain adaptation, transfer learning, retraining, and online or continual learning can modify the model itself and thereby change the underlying capability C_t^* . A change in C_t^* requires corresponding reassessment of $\hat{C}_t$ and the estimated Operational Capability Set.

Competence estimates could be evaluated prospectively against observed operational outcomes. Interactions assigned higher competence should, under a useful estimator, exhibit greater reliability, while interactions near or beyond the estimated capability boundary should exhibit greater rates of failure or require additional support. The relevant outcome depends on the clinical task and intended use, precluding a single universal validation criterion.

Misestimation introduces its own risks. Overestimated competence can produce false admissibility, whereas underestimated competence can lead to unnecessary deferral. Sparse, delayed, or biased feedback can leave $\hat{C}_t$ inconsistent with current deployment conditions. Model adaptation can similarly alter C_t^* before the corresponding competence estimate has been updated. Expansion of $\hat{C}_{\tau,t}$ therefore requires evidence sufficient to support the revised operational boundary.

Capability estimation complements intended-use specification, uncertainty estimation, selective prediction, adaptation, and post-deployment monitoring. Intended use defines the nominal scope of deployment; reliability methods provide evidence about capability; adaptation can alter capability; and competence represents the evolving estimate used to characterize operational admissibility. This decomposition also provides a connection to computational metacognition and meta-learning, separating learning about capability from learning that changes capability.

Table 2: Positioning capability awareness within current clinical and machine-learning practice.

Capability factor	Illustrative limitation	Current clinical/ML practice	Capability-aware interpretation
Intended scope	The interaction falls outside the population, task, or setting for which the system was developed.	Research questions, eligibility criteria, intended-use specifications, task definitions, and validation establish the nominal deployment scope.	Known scope constraints provide prior evidence about the operational capability domain; capability estimation complements rather than replaces intended-use specification.
Model capability	The model lacks sufficient predictive, representational, or reasoning ability for the interaction.	Model development, benchmarking, subgroup analysis, external validation, and error analysis characterise model performance [14, 3].	Observed performance provides evidence about the latent operational capability $C_t^*(i)$.
Available evidence	Required clinical information is missing, incomplete, noisy, or of poor quality.	Clinicians obtain additional history, investigations, or repeat measurements. ML systems use data-quality checks, signal-quality assessment, missing-data methods, uncertainty estimation, or abstention [8].	Evidence sufficiency informs the competence estimate $\hat{C}_t(i)$ for the specific interaction.
Patient-task interaction	The case is unusual, complex, or poorly represented despite falling within the nominal intended-use population.	Clinical judgement, specialist consultation, subgroup analysis, multimodal evidence integration, uncertainty estimation, and selective prediction provide safeguards [7, 17, 15].	Capability is assessed at the interaction level rather than inferred solely from population-level validation or prediction confidence.
Deployment conditions	The data-generating process, workflow, equipment, prevalence, or institution differs from validation conditions.	External validation, OOD and shift detection, drift monitoring, recalibration, and local performance surveillance address deployment differences.	These signals provide evidence for updating $\hat{C}_t(i)$ and the estimated operational capability boundary.

Continued on next page

Table 2 continued from previous page

Capability factor	Illustrative limitation	Current practice	clinical/ML	Capability-aware interpretation
Medical knowledge	Scientific evidence relevant to the interaction is limited, conflicting, or evolving.	Clinical guidelines, evidence synthesis, expert consensus, and multidisciplinary review manage limitations in the knowledge base.		Operational capability may remain limited even for a technically strong model when the knowledge supporting reliable participation is insufficient.
Capability adaptation	The current model is insufficient for a new population, task, or deployment environment.	Domain adaptation, transfer learning, re-training, and online or continual learning modify the model using new data or experience.		Adaptation may change the underlying capability C_t^* and reshape the Operational Capability Set.
Capability estimation	The current operational boundary is uncertain or changes over time.	Calibration, uncertainty estimation, OOD detection, validation, monitoring, outcome assessment, and clinician feedback provide reliability evidence.		These signals refine competence $\hat{C}_t$ and the estimated Operational Capability Set $\hat{C}_{\tau,t}$ without necessarily changing the underlying capability.
Operational decision	Evidence does not sufficiently support AI participation for the current interaction.	Selective prediction, abstention, human review, referral, and requests for additional evidence are incorporated according to the clinical workflow.		The estimated capability domain characterises operational admissibility; the downstream response remains workflow dependent.

Capability estimation remains an open computational problem. Its value for healthcare AI will depend on whether operational capability can be estimated, calibrated, updated, and validated under real deployment conditions, and whether doing so improves decisions about when AI participation is warranted.

The contribution of this perspective is conceptual rather than algorithmic. By distinguishing interaction, capability, competence, and operational admissibility, it identifies capability estimation as a distinct problem in trustworthy healthcare AI. This perspective may stimulate further research on computational models, evaluation methodologies, and deployment strategies for estimating, updating, and validating operational capability boundaries under evolving clinical conditions.

Acknowledgements. The concept has been motivated by Maryam S. Ahmed. Generative AI (OpenAI ChatGPT) was used to assist in drafting and refining the conceptual figure.

References

1. Alelyani, T.: A validated framework for responsible AI in healthcare autonomous systems. *Scientific Reports* **15**, 44432 (2025)

2. Ali, M.S., Mahmud, M.: Towards a digital twin for critical care through explainable digital human modeling of icu patients. *Lecture Notes in Computer Science* **16720 LNCS**, 466 – 485 (2026). https://doi.org/10.1007/978-3-032-29842-3_30
3. Ali, M.S., Mahmud, M.: Towards explainable hospital mortality prediction using critical care data. *Lecture Notes in Computer Science* **16720 LNCS**, 449 – 465 (2026). https://doi.org/10.1007/978-3-032-29842-3_29
4. Everett, J.A.C., Claessens, S., Knöchel, T.D., Reinecke, M.G.: Principles for understanding trust in artificial intelligence. *Nature Reviews Psychology* **5**(6), 388–401 (2026)
5. Finlayson, S.G., Subbaswamy, A., Kohane, I.S.: The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine* **385**(3), 283–286 (2021)
6. Hassija, V., et al.: Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation* **16**(1), 45 – 74 (2024). <https://doi.org/10.1007/s12559-023-10179-8>
7. Jahan, S., Taher, K.A., Kaiser, M.S., Mahmud, M., Rahman, M.S., Hosen, A.S., Ra, I.H.: Explainable ai-based alzheimer’s prediction and management using multimodal data. *PLoS ONE* **18**(11 November) (2023). <https://doi.org/10.1371/journal.pone.0294253>
8. Khan, Z., Sabarimalai Manikandan, M., Mahmud, M.: Light weight graph neural network models for ppg signal quality assessment. *Communications in Computer and Information Science* **2607 CCIS**, 200 – 215 (2025). https://doi.org/10.1007/978-3-032-04657-4_14
9. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., et al.: FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025). <https://doi.org/10.1136/bmj-2024-081554>
10. Mahmud, M., et al.: Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **13309 LNCS**, 356 – 370 (2022). https://doi.org/10.1007/978-3-031-05039-8_26
11. Mahmud, M., et al.: Preventive healthcare through privacy-preserving, explainable and inclusive artificial intelligence. *Lecture Notes in Computer Science* **16340 LNCS**, 359 – 374 (2026). https://doi.org/10.1007/978-3-032-13022-8_25
12. Matos, J., Van Calster, B., Celi, L.A., Dhiman, P., Gichoya, J.W., Riley, R.D., Russell, C., Khalid, S., Collins, G.S.: Critical appraisal of fairness metrics for artificial intelligence-based clinical prediction models: a scoping review. *The Lancet Digital Health* p. 101001 (2026). <https://doi.org/10.1016/j.landig.2026.101001>
13. Palmieri, S., Declerck, J., Priou, S., Lea, N., Kalra, D.: AI literacy in healthcare organisations: implementing article 4 of the EU AI Act with M-SHALF. *npj Health Systems* **3**(1), 42 (2026)
14. Rahman, M.S., et al.: Machine learning-based prognostic model for 30-day mortality prediction in sepsis-3. *BMC Medical Informatics and Decision Making* **24**(1) (2024). <https://doi.org/10.1186/s12911-024-02655-4>
15. Shopland, N., et al.: Towards inclusive ai system development for disease risk prediction: Collecting, prioritising and incorporating user stories from heterogeneous stakeholders. *Lecture Notes in Computer Science* **16340 LNCS**, 386 – 403 (2026). https://doi.org/10.1007/978-3-032-13022-8_27
16. Skibinska, E.: Ethical governance of artificial intelligence in clinical decision-making: A systematic review and implementation framework.

- Journal of Health Policy Outcomes Research (JHPOR) (1), 2 (2026).
<https://doi.org/10.7365/JHPOR.2026.1.2>
17. Viswan, V., Shaffi, N., Malathy, E., Chemmalar Selvi, G., Kavitha, B., Abdesselam, A., Wang, S., Suganthan, P.N., Shezawi, I.A., Mahmud, M.: Multimodal fusion and explainability of artificial intelligence models in alzheimer's disease detection. *Brain Informatics* **13**(1) (2026). <https://doi.org/10.1186/s40708-025-00291-w>