Review Article

AO SPINE

Global Spine Journal
2025, Vol. 15(4) 2547–2558

DOI: 10.1177/21925682251314379
journals.sagepub.com/home/gsj
Sage

Artificial Intelligence for Cervical Spine Fracture Detection: A Systematic Review of Diagnostic Performance and Clinical Potential

Wongthawat Liawrungrueang, MD[1], Watcharaporn Cholamjiak, PhD[2], Arunee Promsri, PhD[3], Khanathip Jitpakdee, MD[4], Sompoom Sunpaweravong[5], Vit Kotheeranurak, MD[6,7], and Peem Sarasombath, MD[8]

Abstract

Study Design: Systematic review.

Objective: Artificial intelligence (AI) and deep learning (DL) models have recently emerged as tools to improve fracture detection, mainly through imaging modalities such as computed tomography (CT) and radiographs. This systematic review evaluates the diagnostic performance of AI and DL models in detecting cervical spine fractures and assesses their potential role in clinical practice.

Methods: A systematic search of PubMed/Medline, Embase, Scopus, and Web of Science was conducted for studies published between January 2000 and July 2024. Studies that evaluated AI models for cervical spine fracture detection were included. Diagnostic performance metrics were extracted and included sensitivity, specificity, accuracy, and area under the curve. The PROBAST tool assessed bias, and PRISMA criteria were used for study selection and reporting.

Results: Eleven studies published between 2021 and 2024 were included in the review. AI models demonstrated variable performance, with sensitivity ranging from 54.9% to 100% and specificity from 72% to 98.6%. Models applied to CT imaging generally outperformed those applied to radiographs, with convolutional neural networks (CNN) and advanced architectures such as MobileNetV2 and Vision Transformer (ViT) achieving the highest accuracy. However, most studies lacked external validation, raising concerns about the generalizability of their findings.

Conclusions: AI and DL models show significant potential in improving fracture detection, particularly in CT imaging. While these models offer high diagnostic accuracy, further validation and refinement are necessary before they can be widely integrated into clinical practice. AI should complement, rather than replace, human expertise in diagnostic workflows.

Keywords

artificial intelligence, deep learning, cervical spine fracture, diagnostic accuracy, convolutional neural networks, spine trauma

[1] Department of Orthopaedics, School of Medicine, University of Phayao, Phayao, Thailand
[2] Department of Mathematics, School of Science, University of Phayao, Phayao, Thailand
[3] Department of Physical Therapy, School of Allied Health Sciences, University of Phayao, Phayao, Thailand
[4] Department of Orthopedics, Queen Savang Vadhana Memorial Hospital, Sriracha, Chonburi, Thailand
[5] Faculty of Medicine, Chulalongkorn University, and King Chulalongkorn Memorial Hospital, Bangkok, Thailand
[6] Department of Orthopaedics, Faculty of Medicine, Chulalongkorn University, and King Chulalongkorn Memorial Hospital, Bangkok, Thailand
[7] Center of Excellence in Biomechanics and Innovative Spine Surgery, Chulalongkorn University, Bangkok, Thailand
[8] Department of Orthopaedics, Phramongkutklao Hospital and College of Medicine, Bangkok, Thailand

Corresponding Author:
Peem Sarasombath, MD, Department of Orthopaedics, Phramongkutklao Hospital and College of Medicine, Bangkok 10400, Thailand.
Email: peems13063@gmail.com

Introduction

Cervical spine fractures are critical injuries that can result in significant morbidity and mortality.[1,2] Clinical practice of the accurate and timely detection of these fractures is essential to prevent complications such as spinal cord injury, paralysis, or death.[3] Radiological imaging, particularly computed tomography (CT) and radiographs, plays a pivotal role in diagnosing cervical spine fractures. However, the complexity of cervical spine anatomy, coupled with subtle fracture patterns, can make detection challenging, even for experienced radiologists.[4] In recent years, artificial intelligence (AI) and deep learning (DL) have emerged as promising tools to enhance diagnostic accuracy and reduce human error in medical imaging.[5,6]

Deep learning, a subset of machine learning, involves training convolutional neural networks (CNNs) and other advanced models to automatically learn features from large datasets (Figure 1). These models have demonstrated remarkable success in image classification tasks across various medical fields, including radiology.[7] The application of AI models, specifically CNNs, in cervical spine fracture detection has shown a potential to improve diagnostic accuracy, streamline workflows, and assist in clinical decision-making. Despite these advancements, there remains variability in the performance of AI models across different datasets, imaging modalities, and clinical settings, highlighting the need for systematic evaluation.[8,9]

This systematic review aims to comprehensively analyse the current literature on using AI and deep learning models for cervical spine fracture detection. By evaluating these models' diagnostic accuracy, object detection performance, and clinical applicability, we seek to elucidate their strengths, limitations, and potential for integration into routine clinical practice. Additionally, we will address the challenges associated with adopting AI in this domain, including the need for external validation, standardized evaluation metrics, and larger datasets to ensure robust model performance across diverse patient populations.

Material and Methods

Literature Search Strategy

A systematic search of four databases included PubMed/Medline, Embase, Scopus, and Web of Science. It was conducted to identify studies on applying AI and DL models for detecting cervical spine fractures. The search strategy included a combination of Medical Subject Headings (MeSH) terms and keywords, such as "artificial intelligence," "deep learning," "cervical spine fracture," "fracture detection," and "diagnostic accuracy." Boolean operators (AND, OR) were

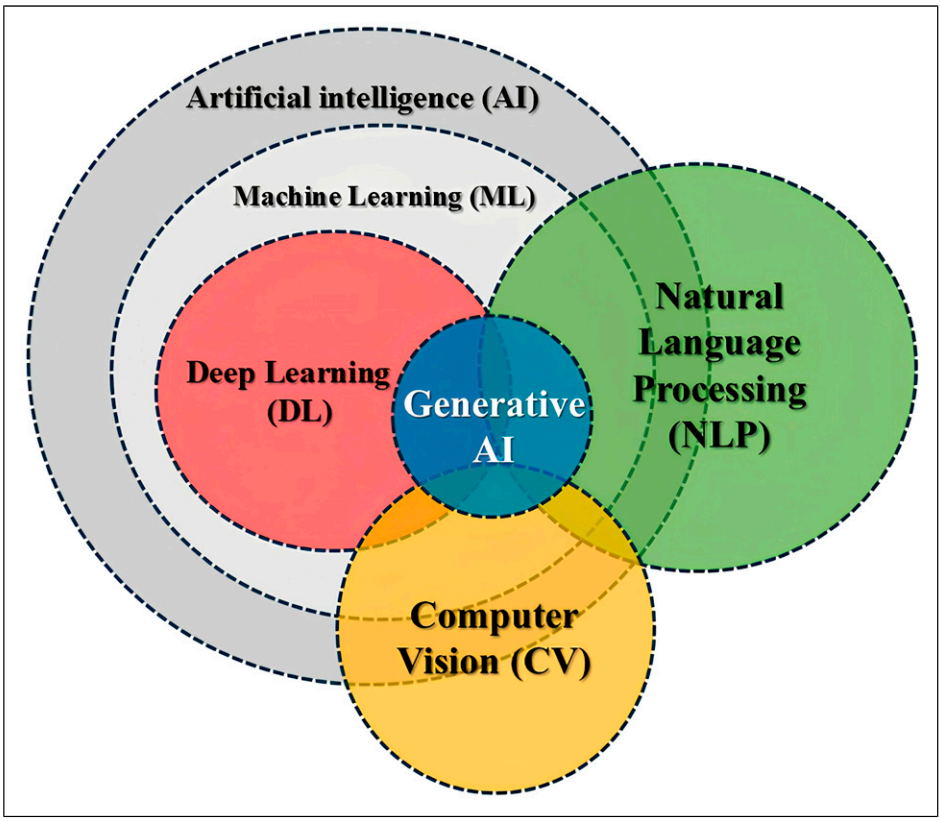

Figure 1. Synergistic relationships: AI is the overarching field, ML provides foundational learning algorithms within AI, DL enhances ML with deep neural networks, NLP and CV apply ML/DL to language and vision, and generative AI uses DL to create new information.

applied to refine the search, and reference lists of relevant articles were manually screened for additional studies. The literature search from January 2000 to July 2024 adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline.[10]

Study Selection

The initial search yielded 893 records (Medline/Mesh: 286, Embase: 182, Scopus: 365, Web of Science: 60). After removing 201 duplicate records, 692 studies were screened based on titles and abstracts. A total of 423 records were excluded, leaving 269 reports for further evaluation. These reports were reviewed for eligibility, and 113 full-text articles were assessed. Studies were excluded if they did not meet the following inclusion criteria: The population consisted of patients with suspected or confirmed cervical spine fractures. The intervention used AI or DL models for cervical spine fracture detection. The performance of AI models compared to radiologists or expert labels. The outcome was diagnostic accuracy metrics. The study design was a diagnosis study, prospective, retrospective and randomised controlled trials studies evaluating the performance of AI models. The excluded were those that did not involve cervical spine fracture detection, discussed generative AI models (ChatGPT) unrelated to radiology and focused solely on image dataset algorithms, data science, or mathematical models. Studies excluded from the review were case reports, reviews, meta-analyses, and studies without available abstracts or full texts. Non-English language articles were also excluded. Two reviewers conducted the study selection process independently, with discrepancies resolved through discussion and, if necessary, by consultation with a third reviewer. The inclusion and exclusion criteria were included in this systematic review. The study selection process is visualized in the PRISMA flow diagram.

Data Extraction

Data were extracted by two independent reviewers using a standardized form. The data included: Study characteristics: Author, year, country of origin, sample size, and population demographics. Imaging modality: CT, plain radiograph x-ray. AI model type: Models such as CNN, YOLO, Vision Transformer (ViT), and others. Performance metrics: sensitivity, specificity, accuracy, recall, F-Score, predictability and AUC. Any discrepancies between reviewers were resolved through discussion or by consulting a third reviewer.

Risk of Bias Assessment

The quality of included studies was assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST).[11,12] The PROBAST tool evaluated the risk of bias across four domains: participants, predictors, outcomes, and analysis. Each study was rated as having low, high, or unclear risk of bias for each domain. The risk of bias assessment was conducted independently by two reviewers, with discrepancies resolved through discussion or, when necessary, by involving a third reviewer.

Results

Study Selection

The literature search yielded a total of 893 records from four databases: Medline/Mesh (286), Embase (182), Scopus (365), and Web of Science (60). After removing 201 duplicate records, 692 studies remained for screening. Title and abstract screening excluded 423 articles due to irrelevance, leaving 269 studies for full-text review. Following the application of inclusion and exclusion criteria, 113 full manuscripts were further assessed for eligibility, and 102 studies were excluded for reasons such as a focus on non-radiological imaging, lack of relevance to cervical spine fractures, or dealing with generative AI models unrelated to radiological diagnosis. A total of 11 studies were included in this systematic review. The selection process is shown in the PRISMA flow diagram (Figure 2).

Study Characteristics

The selected studies were published between 2020 and 2024, with sample sizes ranging from 229 to 4200 participants. Most studies (7/11) utilized CT-scan as the primary imaging modality, while three studies focused on lateral cervical spine radiographs and one study focused on open-mouth odontoid radiographs. The studies spanned various geographical regions, including the USA, Thailand, Egypt, Bangladesh, and the Netherlands, reflecting the global interest in applying AI to cervical spine fracture detection. The AI models evaluated in these studies ranged from CNN to more complex architectures like ViT and Realtime Object Detection Model. Table 1 provides detailed demographic data for each study, including the year, nationality, sample size, population characteristics, and imaging modality used.

Performance of AI Models

The AI models assessed in this review demonstrated a wide range of diagnostic performance with variability in sensitivity, specificity, accuracy, recall, F-Score, predictability and AUC. The performance of AI Models was summary for the detection cervical spine fracture in Table 2

Sensitivity. Sensitivity, a measure of the model's ability to correctly identify fractures, ranged widely from 54.9% to 100%. Aidoc DSS showed the lowest sensitivity, underperforming in identifying fractures on non-contrast CT scans, particularly in chronic cases and fractures with complex

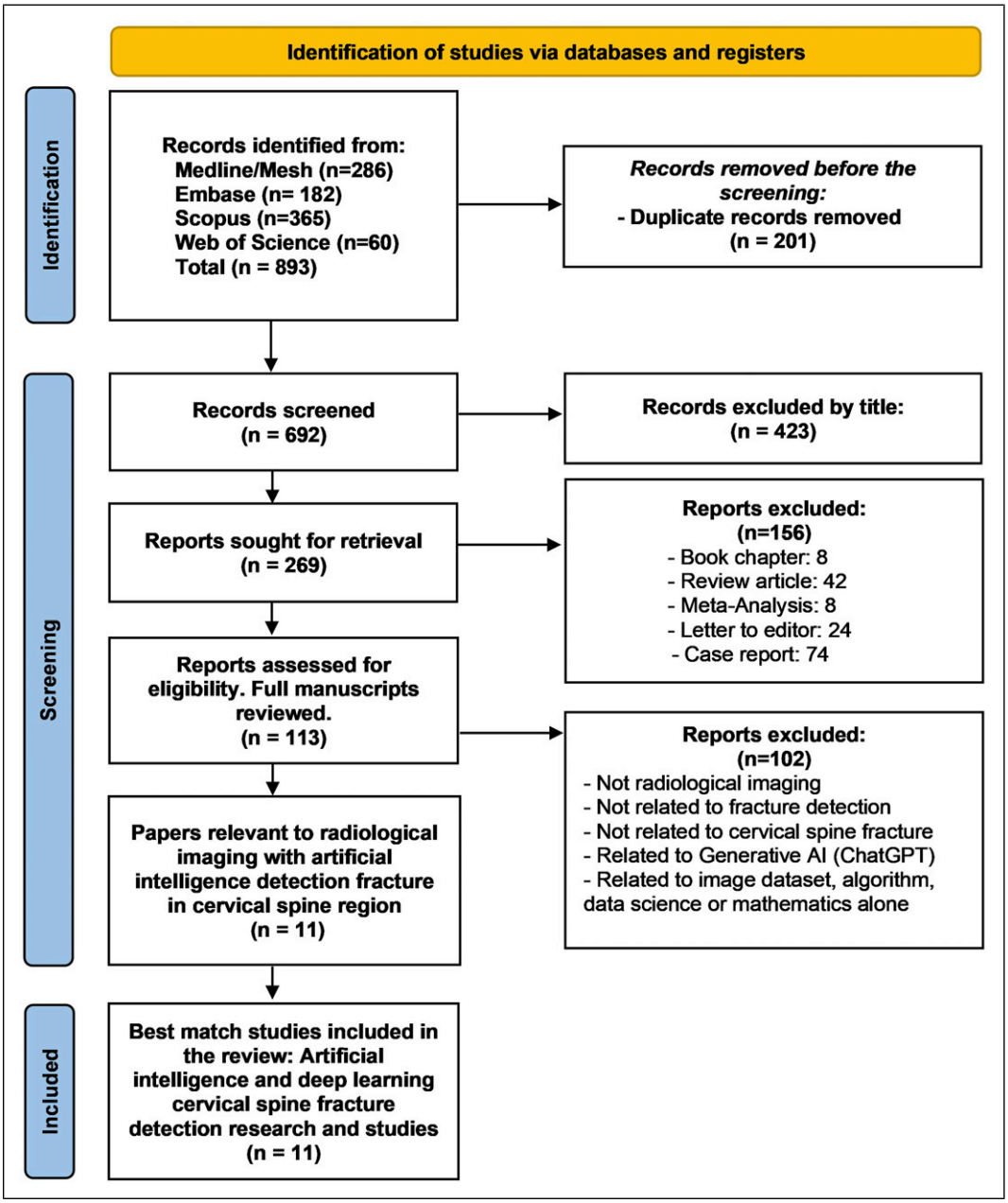

Figure 2. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) flow diagram in this systematic study.

anatomical variations. Conversely, CNN-based models for CT imaging and odontoid fracture detection achieved 100% sensitivity, showcasing their ability to identify all positive cases without missing fractures. Advanced architectures like GoogleNet and MobileNetV2 consistently reported sensitivity values above 98%, demonstrating their robustness for detecting a variety of cervical spine injuries.

Specificity. Specificity, which measures the ability to correctly identify non-fracture cases, was uniformly high across studies, ranging from 72% (YOLO V4) to 98.6% (AIDOC Medical AI). Models such as AIDOC Medical AI excelled in avoiding false positives, particularly for CT-based fracture detection, while YOLO V4, used on radiographs, faced challenges with specificity due to limited dataset diversity and radiographic artifacts.

Accuracy. Overall accuracy values ranged from 75% to 100%, with models like YOLO V4 at the lower end, reflecting its limited performance on radiographs, and CNN-based approaches achieving perfect scores on CT imaging. Models like MobileNetV2 and Vision Transformers consistently delivered accuracy above 98%, demonstrating their capability to generalize effectively across diverse datasets and imaging scenarios.

Table 1. Demographic Data of the Studies in this Systematic Review.

Study	Year	Nationality	Sample Size (n)	Population	Age Range	Imaging Modality	Deep Learning Model
Small et al[13]	2021	USA	665	Cervical spine injury patients	16-98	CT-scan	Aidoc (FDA-approved CNN)
Boonrod et al[14]	2022	Thailand	229	Cervical spine injury patients	≥21	Radiographs (lateral view)	Realtime object detection model (YOLO V4)
Paul et al[15]	2023	Bangladesh	4200	Cervical spine fracture patients	Not specified	CT-scan	MobileNetV2, InceptionV3, ResNet50V2
Naguib et al[16]	2023	Egypt	2009	Cervical spine injury patients	Not specified	Radiographs (lateral view)	AlexNet, GoogleNet
Chład et al[17]	2023	Poland	2019	Cervical spine fracture patients	Not specified	CT-scan	Vision transformer (ViT)
Riazi Esfahani et al[18]	2023	USA, Egypt	4050	Cervical spine fracture patients	Not specified	CT-scan	CNN (google collab platform)
Liawrungrueang et al[19]	2024	Thailand, Korea, USA	500	Cervical spine fracture patients	Not specified	Lateral cervical spine radiograph	CNN (KNIME-based)
Voter et al[20]	2024	USA	1904	Emergent non-contrast CT scans	18-80	CT-scan	Aidoc AI decision support system (DSS)
van den Wittenboer et al[21]	2024	Netherlands	2368	Cervical spine fracture patients	18-97	CT-scan	Aidoc AI algorithm
Ruitenbeek et al[22]	2024	Netherlands	2974	Cervical spine CT patients	18-101	CT-scan	Aidoc AI algorithm
Liawrungrueang et al[23]	2024	Thailand, Korea, Australia	432	Axis(C2) fracture	Not specified	AP open-mouth radiograph	CNN (KNIME-based)

Abbreviations: CT – computed tomography, CNN – convolutional neural network, YOLO – you only look once (realtime object detection model), ANN – artificial neural network, KNIME – konstanz information miner (analytics platform), ViT – vision transformer, AP – anteroposterior.

Table 2. Summary of the Current Performance of Artificial Intelligence Detecting Cervical Spine Fracture.

Study	Deep Learning Model	Object Detection Task	Sensitivity (%)	Specificity (%)	Accuracy (%)	Recall (%)	F-Score	Predictability	AUC	Key Findings	Limitations
Small et al[13]	Aidoc CNN	Cervical spine fracture detection on CT	76	97	92	76	Not reported	Not reported	Not reported	High specificity; effective in worklist prioritization	Limited sensitivity in the lower cervical spine
Boonrod et al[14]	YOLO V4	C-spine injury detection on radiographs	80	72	75	80	Not reported	Not reported	0.743	YOLO V4 outperforms earlier versions	Small dataset, further validation needed
Paul et al[15]	MobileNetV2, InceptionV3, ResNet50V2	Cervical spine fracture detection on CT	99.75	Not reported	99.75	99.75	0.9975	High	High	MobileNetV2 with augmentation performs best	No web-based deployment; focused on android app
Naguib et al[16]	AlexNet, GoogleNet	Classification of fractures, dislocations, and normal on X-rays	99.33	99.66	99.55	99.33	99.67	High	Not reported	GoogleNet performs best with high sensitivity and specificity	Small dataset, no external validation, focus on X-ray only
Chład et al[17]	Vision transformer (ViT)	Cervical spine fracture detection on CT	98	Not reported	98	98	Not reported	High	Not reported	ViT is explainable and fast with high accuracy	No external validation, cloud-based system still under testing
Riazi Esfahani et al[18]	CNN (google collab platform)	Cervical spine fracture detection and classification on CT	100	100	100	100	1.0	High	1.0	Perfect precision, recall, and F1 scores; exceptional accuracy	Limited dataset diversity; further validation needed in real-world settings
Liawrungrueang et al[19]	CNN (KNIME-based)	Cervical spine fracture detection on lateral radiographs	88.60 (fractures), 95.70 (normal)	95.70 (fractures), 88.60 (normal)	92.14	88.6	91.9	High	0.921 (fractures), 0.942 (normal)	Balanced performance, strong in true positive and negative identification	It is a relatively small dataset; further validation on diverse datasets is needed
Voter et al[20]	Aidoc DSS	Cervical spine fracture detection on non-contrast CT	54.9	94.1	Not reported	54.9	38.7	Low	Not reported	High specificity, but poor sensitivity in fracture detection; failure mode analysis identifies chronic fractures as problematic	Limited performance in chronic fractures, false positives due to degeneration
van den Wittenboer et al[21]	Aidoc AI algorithm	Cervical spine fracture detection on CT	71.5	98.6	96.1	71.5	84.5	High	Not reported	AI performed well on transverse process fractures missed by radiologists but missed many fractures requiring stabilizing therapy	AI missed 48% of fractures needing stabilizing therapy
Ruitenbeek et al[22]	Aidoc AI algorithm	Cervical spine fracture detection on non-contrast CT	89.8	95.3	94.8	89.8	63.0	High	Not reported	Reduced time to diagnosis by 16 minutes for fracture cases, with a high NPV.	22 fractures missed, 5 required stabilizing therapy; some false positives caused by anatomical variations and degeneration
Liawrungrueang et al[23]	CNN (KNIME-based)	Odontoid fracture detection on open-mouth X-rays	100	95.4	97	100	97.77	High	0.97	High accuracy, sensitivity, and specificity for detecting odontoid fractures	Small dataset with no external validation

Abbreviations: AI: artificial intelligence, CNN: convolutional neural network, ViT: vision transformer, AUC: area under the curve, NPV: negative predictive value, PPV: positive predictive value, F-Score: harmonic mean of precision and recall, DNT: detection and notification time.

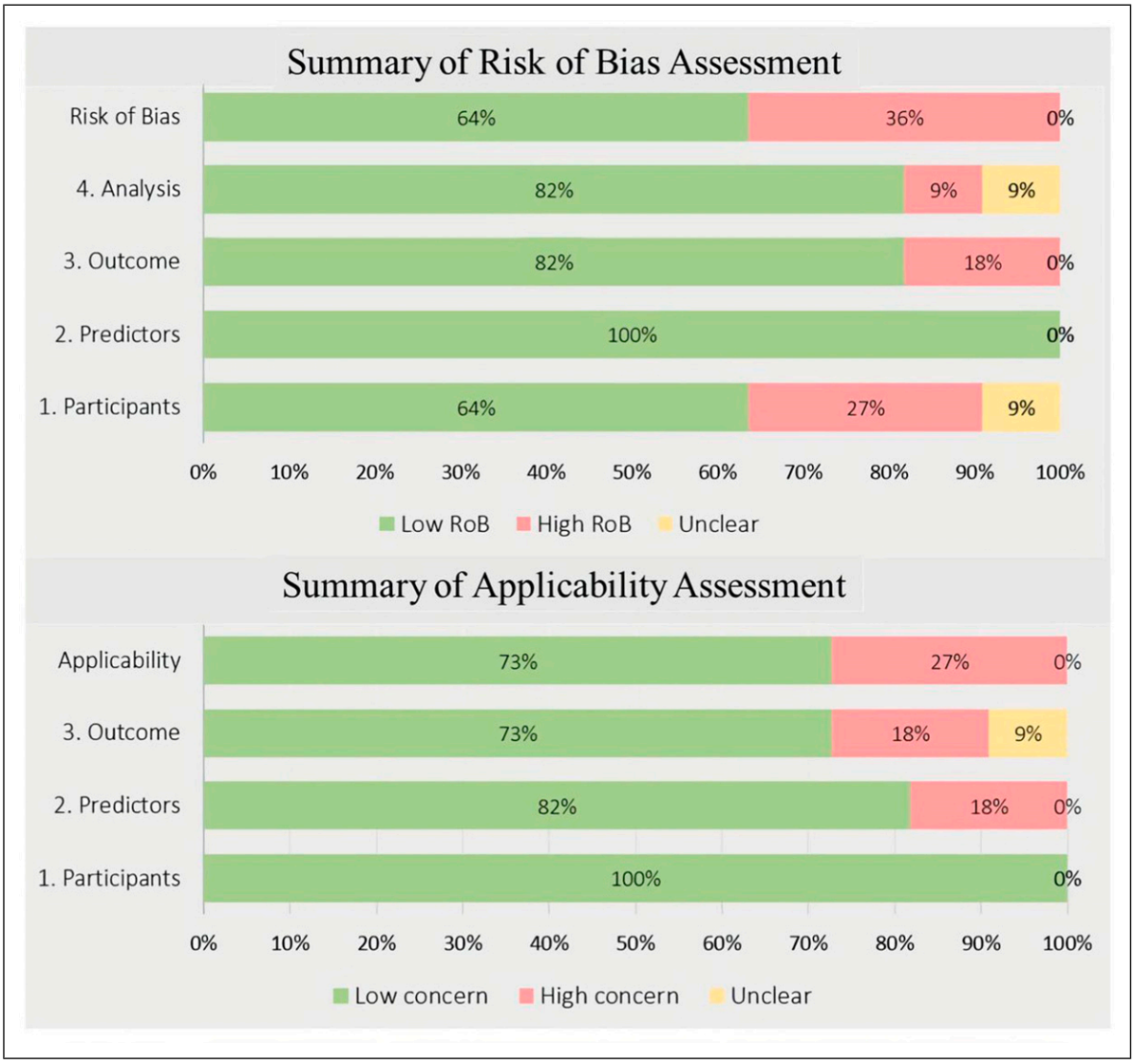

Figure 3. Summary of the risk of bias assessment and applicability using prediction model risk of bias assessment tool (PROBAST).

Recall. Recall, synonymous with sensitivity in this context, varied similarly across studies. High recall values were evident in top-performing models like GoogleNet, MobileNetV2, and CNNs trained for specific fracture detection tasks, indicating their reliability in minimizing false negatives.

F-Score. F-Score, the harmonic mean of precision and recall, offered a balanced perspective on model performance. Models like CNN-based approaches for odontoid fracture detection achieved F-Scores as high as 97.77, reflecting their ability to maintain strong performance across both precision and recall metrics. This is critical for clinical applications where both true positive identification and minimizing false positives are essential.

Predictability. Most studies reported high predictability, indicating consistent model performance across datasets. Models like MobileNetV2 and GoogleNet demonstrated strong generalizability, crucial for deployment in clinical settings where robustness across diverse patient populations is required.

AUC. AUC values, which measure the overall discriminative ability of the models, ranged up to 1.0 in CNN-based studies, highlighting their capacity to differentiate between fracture and non-fracture cases effectively. This metric underscores the potential of AI tools to support clinical decision-making by reducing diagnostic uncertainty.

CT vs Radiographs. AI models applied to CT scans consistently outperformed radiographs with higher accuracy, sensitivity, and specificity. The detailed anatomical imaging provided by CT likely contributed to these better outcomes. This was particularly evident in studies like Riazi Esfahani et al[18] and Paul et al,[15] where CT-based models achieved near-perfect diagnostic performance.

Table 3. The Prediction Model Studies the Risk of Bias Assessment Tool (PROBAST) Guidelines for Assessing the Risk of Bias in this Systematic Review.

	Risk of Bias				Applicability			Overall	
Author, Year	1. Participants	2. Predictors	3. Outcome	4. Analysis	1. Participants	2. Predictors	3. Outcome	Risk of Bias	Applicability
Small et al[13]	+	+	+	+	+	+	+	+	+
Boonrod et al[14]	–	+	–	+	+	–	?	–	–
Paul et al[15]	+	+	+	+	+	+	+	+	+
Naguib et al[16]	+	+	+	+	+	+	+	+	+
Chład et al[17]	+	+	+	+	+	+	+	+	+
Riazi Esfahani et al[18]	?	+	–	?	+	+	–	–	–
Liawrungrueang et al[19]	+	+	+	+	+	+	+	+	+
Voter et al[20]	–	+	+	–	+	–	–	–	–
van den Wittenboer et al[21]	+	+	+	+	+	+	+	+	+
Ruitenbeek et al[22]	+	+	+	+	+	+	+	+	+
Liawrungrueang et al[23]	–	+	+	+	+	+	+	–	+

+ indicates low risk of bias/low concern regarding applicability; - indicates high risk of bias/high concern regarding applicability; ? Indicates unclear risk of bias/unclear concern regarding applicability.

Model Variability. The review found significant variability in the performance of different AI models. While some models, such as MobileNetV2 and ViT, demonstrated strong performance metrics, others, such as YOLO V4, had lower specificity and struggled with the complexity of detecting fractures in lateral radiographs. This variability suggests that the choice of model architecture and training dataset greatly influences performance.

Training Dataset Size and Quality. Studies with larger datasets, such as those by Paul et al[15] with 4200 patients and Riazi Esfahani et al[18] with 4050 patients, reported better performance metrics across all evaluation criteria. In contrast, studies with smaller sample sizes, such as Boonrod et al[14] with 229 participants, showed reduced performance, emphasizing the importance of large, diverse datasets for training AI models.

Clinical Implications. Several studies, including Voter et al[20] and van den Wittenboer et al,[21] highlighted the potential for AI models to assist radiologists in fracture detection, particularly in high-volume trauma settings. These models can prioritize urgent cases and reduce diagnostic delays. However, concerns remain about the clinical deployment of these models, especially those with lower sensitivity.

Risk of Bias

The risk of bias was assessed using the PROBAST tool.[11,12] Most studies were rated as having a low risk of bias. However, three studies (Boonrod et al,[14] Voter et al,[20] and Riazi Esfahani et al[18]) exhibited a higher risk of bias in participant selection and predictor domains. These studies used small datasets or datasets lacking diversity, which may limit the generalizability of their findings. Boonrod et al[14] relied heavily on a single-site dataset, while Voter et al[20] focused on emergent CT scans, which may not fully represent the broader patient population. Results from the PROBAST assessment are presented in Figure 3 and Table 3.

Discussion

This systematic review assesses the diagnostic performance of AI and DL models in detecting cervical spine fractures based on 11 studies published between 2021 and 2024. These studies demonstrated significant variation in diagnostic outcomes, with sensitivity ranging from 54.9% to 100% and specificity between 72% and 98.6%. The review highlights that AI models applied to CT imaging generally performed better than those applied to radiographs, likely due to the superior anatomical detail provided by CT imaging. The superior performance of CNN and more advanced models like MobileNetV[24] and ViT[25] demonstrates that AI has the potential to improve diagnostic accuracy and efficiency in clinical settings significantly. Studies with larger datasets, such as Paul et al[15] and Riazi Esfahani et al,[18] reported near-perfect accuracy and sensitivity, highlighting the importance of comprehensive and diverse training datasets in optimizing model performance. However, variability in model outcomes across studies indicates the need for further refinement, particularly in terms of external validation and standardized evaluation metrics, to ensure the generalizability and reliability of these models in clinical practice.

AI integration into clinical workflows offers several potential advantages. In particular, the application of AI in radiology can augment diagnostic processes by providing automated decision support, improving diagnostic speed, and reducing human error. One of the most immediate and impactful ways AI can be integrated into clinical practice is through worklist prioritisation. AI models, such as the Aidoc AI Decision Support System (DSS),[26] can be embedded in radiology information systems (RIS) and picture archiving and communication systems (PACS) to triage cases based on the likelihood of fractures.[27] This approach allows urgent cases to be flagged for immediate review, improving the efficiency of radiological workflows, especially in high-volume trauma centers where quick identification of fractures is critical. Another key area for AI integration is real-time decision support. AI models can provide instant feedback to radiologists by highlighting regions of interest (ROIs) or suspicious areas on imaging scans. This serves as a second check, helping radiologists focus on potential fracture sites. This assistance is particularly valuable in complex cases, where subtle fractures may be easily overcome, or in high-stress environments such as emergency departments, where rapid decision-making is paramount. Integrating AI directly into PACS would minimize workflow disruptions by allowing radiologists to interact with AI-driven insights without needing to switch between different systems.

AI can also play an essential role in training and education. Radiology trainees and early-career clinicians could benefit from AI-assisted tools that provide instant feedback on image interpretation. AI models could help trainees learn to identify subtle fracture patterns, especially in less commonly encountered cases. This could be particularly valuable in low-resource settings or hospitals with limited access to advanced imaging modalities like CT or MRI, where clinicians might rely more heavily on radiographs.[28] AI can help bridge the gap between experienced and less experienced radiologists by continuously learning from diverse datasets and improving over time as more data are fed into these models. Another promising area for AI integration is interdisciplinary use. AI can assist radiologists and other clinicians involved in trauma care, such as emergency room physicians, trauma surgeons, and orthopaedic specialists. For instance, non-radiologists could use AI models as preliminary screening tools to identify patients at high risk of cervical spine fractures in emergency situations. This is particularly valuable in settings where access to radiologists may be delayed, such as rural or under-resourced hospitals. By providing AI-driven preliminary assessments, these models could assist in triaging patients and guiding immediate clinical decision-making before radiological confirmation is available.

The type of study is critical for optimizing AI performance in fracture detection. Prospective, multi-center studies using large, diverse datasets are essential for ensuring that AI models can generalize across varied patient populations and clinical settings. Thus, the AI research detection of the fracture should focus on larger, more diverse datasets, particularly those from multi-center studies, to improve the reliability. The importance of CT scans, which consistently demonstrate superior performance compared to radiographs in fracture detection. The detailed anatomical information provided by CT imaging allows for more accurate and reliable fracture identification, making it the preferred modality for AI-based fracture detection in cervical spine trauma.

AI to be fully integrated into routine clinical practice, there are important considerations regarding regulatory approval and ethical deployment. AI systems that influence clinical decision-making must undergo rigorous validation and be approved by regulatory agencies such as the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA).[29] The development and deployment of AI systems must be transparent, ensuring that they are explainable to clinicians and that their decision-making processes are interpretable. One concern with AI is the “black box” problem, where the underlying reasoning behind AI outputs is unclear.[30] Clinicians must be able to trust and understand the outputs provided by AI models to incorporate them into their practice confidently. In addition to these considerations, AI models should be continuously monitored and updated to ensure ongoing accuracy and relevance. Models trained on historical datasets may become outdated as new diagnostic criteria, imaging techniques, or population health trends emerge. Establishing a feedback loop that allows AI systems to learn from newly acquired data can improve the long-term effectiveness of these tools. Furthermore, AI integration should include mechanisms for real-time performance monitoring, ensuring that these systems maintain high diagnostic accuracy as they encounter new and diverse patient populations.

Despite these promising developments, AI models are not ready to fully replace radiologists. The variability in model performance, particularly in cases where sensitivity is suboptimal, such as the Aidoc AI Decision Support System’s (DSS) lower sensitivity for detecting chronic fractures, highlights the need for further refinement before these models can be deployed confidently in routine clinical practice.[26] Several limitations were identified in the studies included in this review. Most studies were retrospective which may introduce selection bias and limit the real-world applicability of AI models. Additionally, a lack of external validation was noted in several studies, such as Boonrod et al[14] and Liawrungrueang et al,[23] raising concerns about the generalizability of these models to broader patient populations. Furthermore, many included studies relied on single-centre datasets, which may not adequately represent the diversity of imaging conditions in everyday clinical practice.

The limitation of this study is the relatively limited application of AI models to radiographs. Radiographs are frequently used in initial trauma assessments due to their accessibility and cost-effectiveness. However, the suboptimal performance of AI models on this modality suggests that

Table 4. Future Directions for AI in Cervical Spine Fracture Detection.

Future Direction	Description	Challenges	Potential Impact
External validation	Validate AI models on larger, multi-center datasets to ensure generalizability across different patient populations and clinical settings	Limited access to diverse datasets; need for collaboration	Increased confidence in the robustness of AI models
Prospective studies	Conduct real-time studies in clinical settings to evaluate the performance and utility of AI models beyond retrospective data analysis	Logistical complexities; time and resource intensive	Better understanding of AI's real-world effectiveness
Improved performance on radiographs	Enhance AI accuracy in interpreting radiographs, particularly for initial trauma assessments in resource-limited environments	Lower resolution images; subtle fracture patterns missed	Broader applicability of AI in low-resource settings
Integration into clinical workflows	Seamless integration of AI into existing radiology systems (RIS/PACS) for worklist prioritization and real-time decision support	Compatibility with current systems; ensuring user adoption	Improved diagnostic efficiency and reduced workload
Continuous model updates	Regular updates to AI models with new data to ensure that they evolve with changes in imaging techniques and diagnostic standards	Risk of outdated models; need for continuous data input	Ensures long-term relevance and accuracy of AI systems
Interdisciplinary collaboration	Foster collaborations between clinicians, AI developers, and engineers to design AI systems tailored to real-world diagnostic challenges	Coordinating cross-disciplinary teams; differing priorities	AI models optimized for clinical practice and user needs
Ethical deployment and regulatory approval	Ensure AI models meet regulatory standards (FDA, EMA) and are explainable to clinicians for ethical deployment in clinical settings	Regulatory hurdles; transparency in decision-making	Safe, reliable use of AI in healthcare environments
Education and training	Use AI as a training tool for radiologists to improve diagnostic accuracy, especially for subtle or difficult-to-detect fractures	Need for proper educational frameworks and integration	Enhanced diagnostic skills and experience for trainees

further improvements are necessary to enhance their utility in resource-limited settings or in the initial evaluation of trauma patients. Future research must address several critical areas to fully integrate AI models into clinical practice for cervical spine fracture detection. First, external validation using larger, multi-center datasets is necessary to ensure that these models can generalize across different patient populations and clinical environments. Second, prospective studies evaluating AI models in real-time clinical workflows are essential to validate their utility beyond retrospective analyses. Third, enhancing AI performance on radiographs is vital, particularly in low-resource settings where CT imaging may not always be available. Fourth, various classifications used from studies fail to include essential patient information, such as neurologic status, polytrauma, and other clinical modifiers, which are crucial for therapeutic decisions. Focusing on imaging data alone may be insufficient for comprehensive decision-making in cervical spine trauma. Future AI training should incorporate datasets that include these essential clinical parameters. Finally, ongoing collaboration between clinicians and AI developers is crucial to ensure that AI models are optimized for real-world diagnostic challenges, with a particular focus on improving sensitivity and minimizing false positives. This systematic review provides a comprehensive assessment of the current landscape of AI models for cervical spine fracture detection, highlighting both the potential and limitations of these technologies. By adhering to the PRISMA guidelines and employing the PROBAST tool for risk of bias assessment, this review ensures a rigorous evaluation of the available literature, offering valuable insights into the future integration of AI into clinical practice. The author summarizes the future directions for AI in cervical spine fracture detection in Table 4.

Conclusion

This systematic review demonstrates that AI and deep learning models, particularly CNN and advanced architectures like MobileNetV2 and ViT, show significant potential in detecting cervical spine fractures, especially in CT imaging. These models can enhance diagnostic accuracy and streamline clinical workflows, offering valuable support to radiologists. However, challenges remain, including variability in performance across imaging modalities and a lack of external validation in many studies. Future work must focus on refining AI models, validating them in diverse clinical settings, and ensuring they complement human expertise. With further development, AI has the potential to play a critical role in improving the detection of cervical spine fractures.

Acknowledgments

The authors would like to thank the Thailand Science Research and Innovation Fund (Fundamental Fund 2025, Grant No. 5025/2567) and the School of Medicine, University of Phayao.

Author Contributions

Conceptualization: W Liawrungruean, Data curation: W Liawrungrueang, W Cholamjiak, A Promsri, P Sarasombath, Formal analysis: W Liawrungrueang, P Sarasombath, Funding acquisition: W Liawrungrueang, Methodology: W Liawrungrueang, Project administration: W Liawrungrueang, Visualization: W Liawrungrueang, W Cholamjiak, A Promsri, V Kotheeranurak, K Jitpakdee, S Sunpaweravong, P Sarasombath, Writing - original draft: W Liawrungrueang, P Sarasombath, Writing - review & editing: W Liawrungrueang, P Sarasombath.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

IRB Statement

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Phayao University Hospital (Approval No. HREC-UP-HSST 1.1/046/67).

ORCID iDs

Wongthawat Liawrungrueang https://orcid.org/0000-0002-4491-6569
Peem Sarasombath https://orcid.org/0000-0003-0513-238X

References

1. Nemani VM, Kim HJ. The management of unstable cervical spine injuries. *Clin Med Insights Trauma Intensive Med*. 2014;5:CMTIM.S12263.
2. McMordie JH, Viswanathan VK, Gillis CC. Cervical spine fractures overview. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2024. https://www.ncbi.nlm.nih.gov/books/NBK448129/. accessed 18 July 2024.
3. Zileli M, Osorio-Fonseca E, Konovalov N, et al. Early management of cervical spine trauma: WFNS spine committee recommendations. *Neurospine*. 2020;17:710-722.
4. Pinto A, Berritto D, Russo A, et al. Traumatic fractures in adults: missed diagnosis on plain radiographs in the emergency department. *Acta Biomed*. 2018;89:111-123.
5. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. *Bioengineering*. 2023;10:1435.
6. Khalifa M, Albadawy M. AI in diagnostic imaging: revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*. 2024;5:100146.
7. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data*. 2021;8:53.
8. Kutbi M. Artificial intelligence-based applications for bone fracture detection using medical images: a systematic review. *Diagnostics*. 2024;14:1879.
9. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. *Front Public Health*. 2023;11:1273253.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151:264-269.
11. Fernandez-Felix BM, López-Alcalde J, Roqué M, Muriel A, Zamora J. CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models. *BMC Med Res Methodol*. 2023;23:44.
12. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51-58.
13. Small JE, Osler P, Paul AB, Kunst M. CT cervical spine fracture detection using a convolutional neural network. *AJNR Am J Neuroradiol*. 2021;42:1341-1347.
14. Boonrod A, Boonrod A, Meethawolgul A, Twinprai P. Diagnostic accuracy of deep learning for evaluation of C-spine injury from lateral neck radiographs. *Heliyon*. 2022;8:e10372.
15. Guha Paul S, Saha A, Assaduzzaman M. A real-time deep learning approach for classifying cervical spine fractures. *Healthcare Analytics*. 2023;4:100265.
16. Naguib SM, Hamza HM, Hosny KM, Saleh MK, Kassem MA. Classification of cervical spine fracture and dislocation using refined pre-trained deep model and saliency map. *Diagnostics*. 2023;13:1273.
17. Chłąd P, Ogiela MR. Deep learning and cloud-based computation for cervical spine fracture detection system. *Electronics*. 2023;12:2056. doi:10.3390/electronics12092056.
18. Riazi Esfahani P, Guirgus M, Maalouf M, et al. Development of a machine learning-based model for accurate detection and classification of cervical spine fractures using CT imaging. *Cureus*. 2023;15:e47328.
19. Liawrungrueang W, Han I, Cholamjiak W, Sarasombath P, Riew KD. Artificial intelligence detection of cervical spine fractures using convolutional neural network models. *Neurospine*. 2024;21:833-841.
20. Voter AF, Larson ME, Garrett JW, Yu JPJ. Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of cervical spine fractures. *AJNR Am J Neuroradiol*. 2021;42:1550-1556.
21. van den Wittenboer GJ, van der Kolk BYM, Nijholt IM, et al. Diagnostic accuracy of an artificial intelligence algorithm versus radiologists for fracture detection on cervical spine CT. *Eur Radiol*. 2024;34:5041-5048.
22. Ruitenbeek HC, Oei EHG, Schmahl BL, Bos EM, Verdonschot RJCG, Visser JJ. Towards clinical implementation of an AI-algorithm for detection of cervical spine fractures on computed tomography. *Eur J Radiol*. 2024;173:111375.

23. Liawrungrueang W, Cho ST, Kotheeranurak V, Pun A, Jitpakdee K, Sarasombath P. Artificial neural networks for the detection of odontoid fractures using the Konstanz information miner analytics platform. *Asian Spine J*. 2024;18:407-414.
24. Zaidi SSA, Ansari MS, Aslam A, Kanwal N, Asghar M, Lee B. A survey of modern deep learning based object detection models. *Digit Signal Process*. 2022;126:103514.
25. Papa L, Russo P, Amerini I, Zhou L. A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking. *IEEE Trans Pattern Anal Mach Intell*. 2024 Dec; 46(12):7682-7700. doi:10.1109/TPAMI.2024.3392941
26. Shiang T, Garwood E, Debenedectis CM. Artificial intelligence-based decision support system (AI-DSS) implementation in radiology residency: introducing residents to AI in the clinical setting. *Clin Imag*. 2022;92:32-37.
27. Honeyman JC. Information systems integration in radiology. *J Digit Imag*. 1999;12:218-222.
28. Duong MT, Rauschecker AM, Rudie JD, et al. Artificial intelligence for precision education in radiology. *Br J Radiol*. 2019;92:20190389.
29. Karalis VD. The integration of artificial intelligence into clinical practice. *Applied Biosciences*. 2024;3:14-44.
30. 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: 52-57.