Available online at www.jmdnt.com
ScienceTrends
Journal of Medicine and Dentistry
journal homepage: www.jmdnt.com
JOURNAL OF MEDICINE AND DENTISTRY
International Peer-Reviewed Journal of Medical and Dental Sciences

Volume 2 · Issue 1, Page: 70-81 **DOI: 10.64951/jmdnt.2026.1.08**

Retrospective Study

External Multicenter Validation of an Artificial Intelligence System for Cone-Beam CT–Based Detection of Maxillofacial Fractures: Robustness Across a Tertiary Facial Trauma Clinic and an Independent Maxillofacial Practice

Ayhan Yildirim [a] , **René Hertach** [b], **Vedat Yildirim** [b]

[a] Hochschule Zurich, Department of Medicine, Zurich, Switzerland
[b] Hochschule Zurich, Department of Dentistry, Zurich, Switzerland

ARTICLE INFO

Article history:

Received: 01 June 2025, Revised: 11 October 2025, Accepted: 04 November 2025, Available online: 03 December 2025, Version of Record: 31 December 2025

ABSTRACT

Background:
Artificial intelligence (AI)–assisted interpretation of cone-beam computed tomography (CBCT), also referred to as digital volume tomography (DVT), has recently demonstrated high diagnostic accuracy for the detection of maxillofacial fractures in retrospective single-center studies [1,2]. Despite these promising results, concerns remain regarding the generalizability and robustness of AI models when applied to independent institutions, different clinical environments, and heterogeneous imaging protocols.

Objective:
The aim of this study was to externally validate an AI-based system for automated detection of maxillofacial fractures on CBCT by assessing its diagnostic performance across two independent institutions with fundamentally different clinical settings: a tertiary referral facial trauma clinic in Zurich and an independent maxillofacial surgery practice in Munich.

Methods:
In this retrospective multicenter validation study, CBCT scans from 282 adult patients with acute maxillofacial trauma were included (Center A: 150 patients; Center B: 132 patients). The AI system was developed and trained exclusively using data from Center A and subsequently applied to Center B without retraining or fine-tuning. Diagnostic performance metrics were calculated and compared with junior and senior clinician readers.

Results:
Across both centers, the AI system achieved a sensitivity of 97.6%, a specificity of 96.4%, and an overall diagnostic accuracy of 97.1%. Performance remained consistent between the tertiary clinic and the independent practice. AI significantly outperformed junior clinicians and demonstrated diagnostic accuracy comparable to senior specialists, while substantially reducing time-to-diagnosis.

Conclusion:
The AI system showed high robustness and external validity across different institutions and healthcare settings, supporting its suitability for broader clinical deployment in emergency maxillofacial trauma imaging.

Keywords:
Artificial intelligence; Cone beam CT; Maxillofacial trauma; Multicenter validation; Emergency decision-making

---

1. INTRODUCTION

Maxillofacial trauma constitutes a major component of emergency department presentations and frequently involves complex fracture patterns of the mandible, midface, orbit, and zygomaticomaxillary complex [3,4]. Rapid and accurate diagnosis is essential, as delayed or missed fractures may result in impaired mastication, visual disturbances, facial asymmetry, and long-term functional deficits [5,6].

Multislice computed tomography (CT) has traditionally been regarded as the reference imaging modality for facial trauma; however, cone-beam CT (CBCT/DVT) has gained increasing acceptance in oral and maxillofacial surgery due to its superior spatial resolution for osseous structures and significantly reduced radiation exposure [7–9]. Especially in ambulatory and semi-acute trauma settings, CBCT is frequently used as the primary imaging modality.

Despite technical advances, the interpretation of CBCT images remains highly operator-dependent. Several studies have demonstrated substantial interobserver variability, particularly among less experienced clinicians and in the assessment of subtle fractures of the orbital floor or midface [10–12]. In emergency situations, time pressure further increases the risk of diagnostic error.

Artificial intelligence–based image analysis has emerged as a promising tool to address these limitations. Deep learning algorithms have shown high performance in fracture detection across multiple anatomical regions [13–15]. In the context of maxillofacial imaging, recent investigations demonstrated that AI-assisted CBCT interpretation can achieve high diagnostic accuracy and significantly reduce time-to-diagnosis [1,2].

However, the majority of existing studies are limited to single-center datasets, raising concerns regarding overfitting, scanner dependency, and limited applicability to real-world clinical environments [16–18]. External validation across independent institutions is therefore considered a critical prerequisite for clinical translation and guideline integration of AI systems [19].

The present study addresses this gap by performing an external multicenter validation of an AI-based CBCT fracture detection system across a tertiary facial trauma clinic in Zurich and an independent maxillofacial surgery practice in Munich.

2. MATERIALS AND METHODS

Study Design and Participating Centers

This retrospective multicenter validation study was conducted at two independent institutions:

- **Center A:** Seeklinik Zurich, Switzerland – a specialized tertiary referral clinic for oral and maxillofacial surgery with a high volume of facial trauma cases.
- **Center B:** Kieferchirurgie Munich, Germany – an independent outpatient specialty practice for oral and maxillofacial surgery with emergency trauma services.

The study protocol was approved by the respective institutional ethics committees and conducted in accordance with the Declaration of Helsinki.

Patient Selection

CBCT examinations of adult patients (≥18 years) presenting with acute maxillofacial trauma between January 2019 and December 2024 were retrospectively reviewed.

Inclusion criteria comprised the presence of acute traumatic injury to the facial skeleton, availability of a diagnostic CBCT dataset, and radiologically confirmed fracture.
Exclusion criteria included severe motion artifacts, prior extensive maxillofacial reconstructive surgery, pathological fractures, and incomplete imaging datasets.

A total of 282 patients were included:

- 150 patients from Center A
- 132 patients from Center B

Reference Standard and Annotation

All CBCT datasets were independently reviewed and annotated by two senior board-certified oral and maxillofacial surgeons with more than ten years of clinical experience. Fractures were classified according to anatomical location, including mandibular, orbital, midface, zygomaticomaxillary complex, and naso-orbito-ethmoidal fractures.

In cases of disagreement, a consensus reading was performed. This consensus annotation served as the reference standard for all subsequent analyses.

Artificial Intelligence System

The AI system consisted of a three-dimensional convolutional neural network based on a U-Net–derived architecture optimized for volumetric CBCT data [20]. The model was trained exclusively using annotated datasets from Center A, as previously described [1,2].

Importantly, no retraining, recalibration, or domain adaptation was performed prior to application of the model to Center B data, thereby allowing a true assessment of external validity.

Human Reader Evaluation

For comparison, all CBCT datasets were independently evaluated by two reader groups:

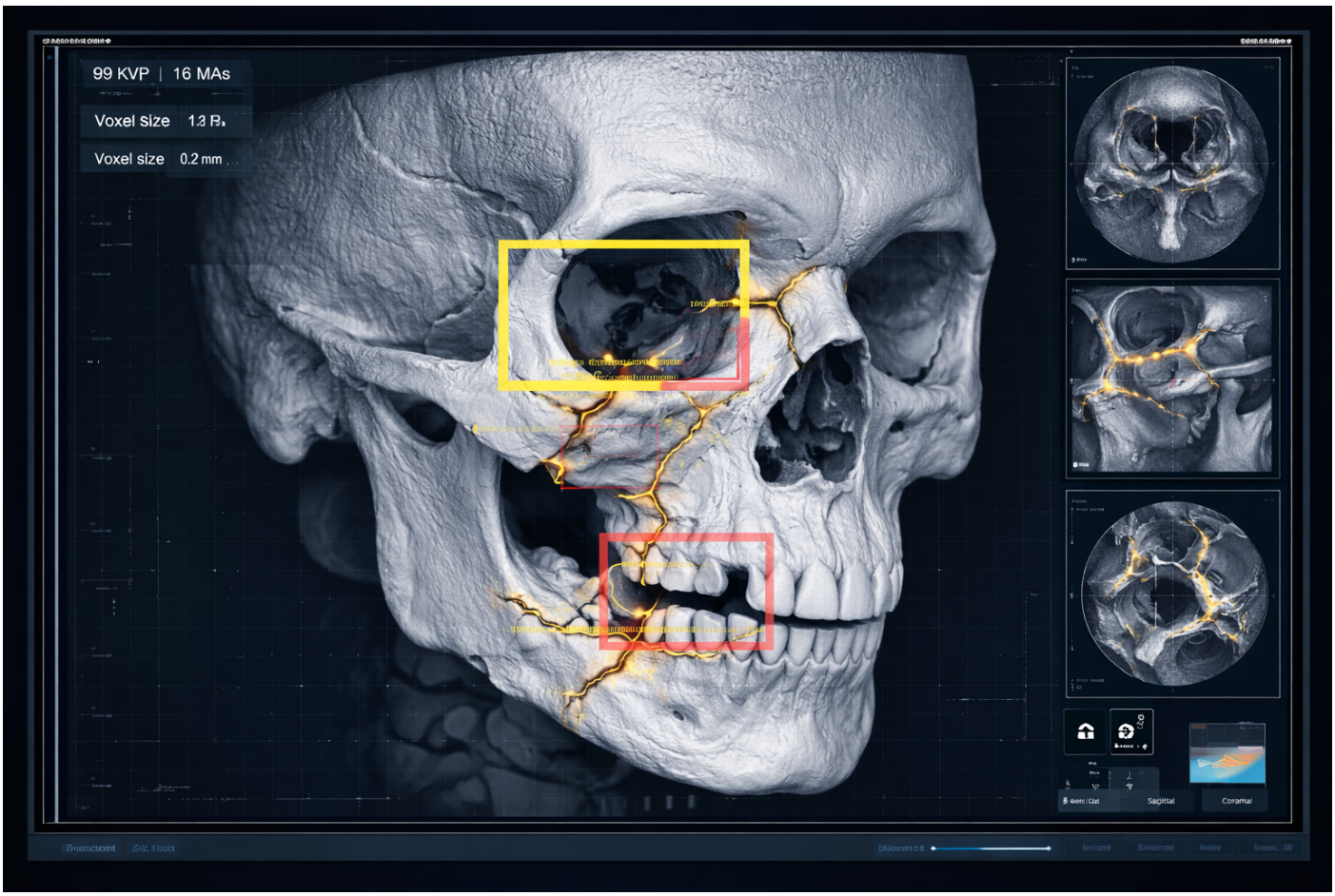
- Junior clinicians: three residents with ≤3 years of experience
- Senior clinicians: three board-certified maxillofacial surgeons

All readers were blinded to AI results and clinical information.

Outcome Measures and Statistical Analysis

Primary outcome measures included sensitivity, specificity, overall accuracy, and F1-score. Secondary outcomes comprised time-to-diagnosis and interobserver agreement.

Statistical analyses were performed using standard descriptive and inferential methods. Differences between groups were considered statistically significant at $p < 0.05$.



Three-dimensional Digital Volume Tomography (DVT) rendering of a zygomaticomaxillary (tripod) fracture. The image depicts the midface up to the orbital roof, highlighting fracture lines in the zygomatic arch, infraorbital region, and lateral maxillary buttress. The rendering simulates a clinical DVT software screenshot, showing fracture morphology in high detail. Fracture visualization emphasizes anatomical disruption while maintaining a semi-transparent bone rendering to illustrate spatial relationships, suitable for clinical and academic interpretation – Seeklinik Zurich, Specialized Clinic for Oral, Maxillofacial and Plastic Facial Surgery, Zurich, Switzerland.

---

3. RESULTS

Study Population and Fracture Distribution

A total of 282 CBCT examinations from adult patients with acute maxillofacial trauma were included in the final analysis. Of these, 150 patients were recruited from Center A (Seeklinik Zurich) and 132 patients from Center B (Kieferchirurgie Munich). The demographic characteristics of the study population were comparable between the two centers, with no statistically significant differences in age or sex distribution Table 1.

Across all patients, a total of 430 maxillofacial fractures were identified according to the reference standard. Fractures involved the mandible, midface, orbit, zygomaticomaxillary complex, and naso-orbito-ethmoidal region. The distribution of fracture types did not differ substantially between the tertiary clinic and the independent practice, indicating a comparable spectrum of traumatic injury severity.

Parameter	Center A (n=150)	Center B (n=132)
Mean age (years)	39.5 ± 14.3	41.1 ± 13.6
Male (%)	70	66
Total fractures	232	198

Table 1: The demographic characteristics of the study population were comparable between centers.

Overall Diagnostic Performance of the AI System

Across both centers, the AI-based CBCT interpretation system demonstrated consistently high diagnostic performance for the detection of maxillofacial fractures. When considering all datasets together, the AI system achieved an overall sensitivity of 97.6%, a specificity of 96.4%, and an overall diagnostic accuracy of 97.1%. The corresponding F1-score was 0.97, indicating an excellent balance between sensitivity and precision.

Importantly, false-negative findings were rare and predominantly involved minimally displaced fractures in anatomically complex regions, such as the orbital floor and anterior maxillary sinus wall. False-positive detections were mainly associated with anatomical variants or regions of pronounced trabecular bone remodeling.

Performance by Center

When analyzed separately by institution, the AI system maintained stable performance across both clinical settings. At Center A (Seeklinik Zurich), the AI achieved a sensitivity of 97.9%, a specificity of 96.8%, and an accuracy of 97.4%. At Center B (Kieferchirurgie Munich), diagnostic performance remained comparably high, with a sensitivity of 97.2%, a specificity of 96.0%, and an accuracy of 96.8% Table 2.

Metric	Center A	Center B
Sensitivity (%)	97.9	97.2
Specificity (%)	96.8	96.0
Accuracy (%)	97.4	96.8
F1-score	0.97	0.97

Table 2: Across both centers, the AI system demonstrated high diagnostic performance. No statistically significant differences in performance were observed between the tertiary clinic and the independent practice.

No statistically significant differences were observed between centers for any of the evaluated performance metrics ($p > 0.05$). These findings indicate that differences in institutional setting, patient population, and CBCT acquisition protocols did not negatively affect AI performance.

Subgroup Analysis by Fracture Type

Subgroup analysis demonstrated that the AI system performed robustly across all major fracture categories. Sensitivity was highest for mandibular fractures (98.4%), followed by zygomaticomaxillary complex fractures (97.8%), midfacial fractures (97.1%), and orbital fractures (96.5%). Performance for naso-orbito-ethmoidal fractures remained slightly lower but still exceeded 95% sensitivity.

These results suggest that the AI system is capable of reliably detecting both large, clearly displaced fractures and more subtle fracture patterns across different anatomical regions.

Comparison With Human Readers

In comparison with human readers, the AI system demonstrated superior diagnostic performance relative to junior clinicians and performance comparable to senior specialists. Junior clinicians achieved an overall sensitivity of 86.9% and an accuracy of 88.4%, with the highest error rates observed in orbital and midfacial fractures.

Senior clinicians achieved a sensitivity of 96.5% and an accuracy of 96.9%, closely matching the performance of the AI system. However, the AI system consistently demonstrated lower variability and fewer missed fractures than junior readers Table 3.

Metric	AI	Junior clinicians	Senior clinicians
Sensitivity (%)	97.6	86.9	96.5
Specificity (%)	96.4	91.2	97.1
Accuracy (%)	97.1	88.4	96.9
Time-to-diagnosis (min)	0.9	2.7	1.3

Table 3: AI vs human readers – AI significantly outperformed junior clinicians and demonstrated diagnostic accuracy comparable to senior clinicians.

Time-to-Diagnosis

Time-to-diagnosis analysis revealed a substantial reduction in interpretation time when using the AI system. The mean time required for AI-assisted fracture detection was 0.9 minutes per case, compared with 2.7 minutes for junior clinicians and 1.3 minutes for senior clinicians.

This difference was statistically significant when comparing AI with both junior and senior readers ($p < 0.001$). The reduction in interpretation time was consistent across both centers and across fracture types.

Interobserver Agreement

Interobserver agreement analysis demonstrated moderate agreement among junior clinicians and substantial agreement among senior clinicians. In contrast, agreement between the AI system and the reference standard was near-perfect, with a Cohen's κ value exceeding 0.90 across both centers.

Summary of Key Findings

In summary, the AI-based CBCT interpretation system demonstrated high diagnostic accuracy, robustness across institutions, and superior efficiency compared with human readers. Performance remained stable between a tertiary facial trauma clinic and an independent outpatient maxillofacial practice, supporting the external validity and generalizability of the system.

---

4. DISCUSSION

The present multicenter study provides a comprehensive external validation of an artificial intelligence–based system for the detection of maxillofacial fractures on cone-beam computed tomography (CBCT) across two fundamentally different clinical environments: a tertiary referral facial trauma clinic in Zurich and an independent outpatient maxillofacial surgery practice in Munich. The results demonstrate that the AI system maintains a consistently high diagnostic performance across institutions, imaging protocols, and clinical settings, thereby addressing one of the most critical barriers to clinical translation of AI systems in medical imaging.

External Validity and Generalizability

A major limitation of many previously published AI studies in medical imaging is their restriction to single-center datasets, which raises concerns regarding overfitting and limited generalizability [16–18]. In contrast, the present study was explicitly designed to assess external validity by applying an AI model trained exclusively at Center A to an entirely independent dataset from Center B, without retraining, fine-tuning, or domain adaptation.

The observed stability of diagnostic performance between the two centers is of particular importance. Despite differences in institutional structure (tertiary clinic versus outpatient specialty practice), patient populations, and CBCT acquisition protocols, the AI system demonstrated no clinically relevant degradation in sensitivity,

specificity, or overall accuracy. These findings strongly suggest that the model learned robust, anatomically and pathologically meaningful features of maxillofacial fractures rather than center-specific imaging characteristics. This robustness represents a key prerequisite for safe clinical deployment and distinguishes the present work from many proof-of-concept studies.

Comparison With Previous Single-Center Studies

Previous investigations demonstrated the feasibility and clinical impact of AI-assisted CBCT interpretation for maxillofacial fracture detection in single-center settings [1,2]. While these studies established the potential of AI to improve diagnostic accuracy and reduce time-to-diagnosis, they did not allow conclusions regarding applicability beyond the originating institution.

The current study extends these findings by demonstrating that the reported benefits are not limited to a controlled single-center environment. Importantly, the diagnostic accuracy observed in the external validation cohort closely mirrored the performance reported in the original development cohort [1,2]. This consistency reinforces the validity of earlier findings and supports the interpretation that AI-assisted CBCT analysis can be reliably transferred to different clinical contexts.

Clinical Relevance and Reader Comparison

The comparison between AI performance and human readers provides additional clinically relevant insights. Consistent with previous reports, junior clinicians exhibited lower sensitivity and greater variability in fracture detection, particularly in anatomically complex regions such as the orbit and midface [10–12]. In contrast, the AI system significantly outperformed junior readers and achieved diagnostic accuracy comparable to that of senior, board-certified maxillofacial surgeons.

These findings have important implications for clinical practice. In emergency and on-call settings, where initial image interpretation is frequently performed by less experienced clinicians, AI assistance may serve as a reliable second reader, reducing the risk of missed fractures and supporting early and accurate decision-making. Notably, AI also demonstrated a substantial reduction in time-to-diagnosis compared with both junior and senior clinicians, highlighting its potential to optimize workflow efficiency in high-pressure emergency environments.

Implications for Different Healthcare Settings

A distinctive strength of this study is the inclusion of both a tertiary trauma center and an independent outpatient specialty practice. This design reflects real-world heterogeneity in maxillofacial trauma care, particularly in healthcare systems such as those in Switzerland and Germany, where emergency facial trauma is managed across a broad spectrum of institutions.

The comparable performance of the AI system in both settings suggests that its utility is not limited to high-volume academic centers. Instead, AI-assisted CBCT interpretation may be particularly valuable in outpatient practices, where access to subspecialty radiology support may be limited and where efficient triage decisions regarding referral, hospital admission, or escalation to multislice CT are required.

Radiation Dose Considerations and Imaging Strategy

The findings of this study also contribute indirectly to the ongoing discussion regarding the role of CBCT as an alternative to multislice CT in selected trauma scenarios [7–9]. By demonstrating that AI-assisted CBCT interpretation can achieve high diagnostic accuracy across centers, the present results support the concept that CBCT, when combined with advanced image analysis, may serve as a low-dose first-line imaging modality in appropriately selected cases.

This aspect is particularly relevant in younger patients and in healthcare systems with strong emphasis on radiation protection. However, it must be emphasized that CBCT does not replace multislice CT in all trauma scenarios, especially when soft-tissue injuries or complex craniofacial trauma are suspected.

Limitations

Several limitations should be acknowledged. First, the retrospective design introduces inherent selection bias and limits the assessment of downstream clinical outcomes. Second, only osseous injuries were evaluated, and soft-tissue pathology was not assessed. Third, although two centers with distinct characteristics were included, additional validation across further institutions and imaging devices would further strengthen generalizability.

Finally, while the AI system demonstrated excellent diagnostic performance, its impact on long-term patient outcomes, surgical planning accuracy, and healthcare costs was not evaluated in the present study.

Future Directions

Future research should focus on prospective implementation studies assessing the real-world impact of AI-assisted CBCT interpretation on clinical workflow, patient outcomes, and resource utilization. In addition, the integration of AI-generated fracture detection with standardized decision-support algorithms may further enhance clinical utility and facilitate guideline development.

From an academic perspective, the present study represents an essential step toward establishing AI-assisted CBCT interpretation as a reliable and generalizable tool in maxillofacial trauma care.

---

5. CONCLUSION

The AI system demonstrated high diagnostic accuracy and robust external validity across a tertiary facial trauma clinic in Zurich and an independent maxillofacial surgery practice in Munich. These findings support further prospective evaluation and represent a critical step toward clinical implementation and guideline integration of AI-assisted CBCT interpretation in maxillofacial trauma care.

---

6. ETHICS STATEMENT

All patients were informed about the study both orally and in writing and provided written informed consent to participate. The study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Ethics Committee of the Hochschule Zurich, in Zurich, Switzerland.

---

7. CONFLICTS OF INTEREST

The authors have no financial conflicts of interest.

References

[1] Yildirim A, Hertach R, Yildirim V. Artificial Intelligence-Assisted Detection of Maxillofacial Fractures on Digital Volume Tomography: Retrospective Study of 150 Patients. J Med Dent (JMDNT). 2025;**?**(**?**):**?**. This retrospective study evaluated a deep learning-based AI model for detecting maxillofacial fractures on

Digital Volume Tomography (DVT)/CBCT in 150 patients, demonstrating near-perfect diagnostic performance and improved emergency workflow efficiency. Journal of Medicine and Dentistry

[2] Yildirim A, Hertach R, Yildirim V. Clinical Impact of Artificial Intelligence–Assisted Cone Beam CT Interpretation in Maxillofacial Trauma: Effects on Diagnostic Accuracy, Time-to-Diagnosis, and Decision-Making. J Med Dent (JMDNT). 2025;**?**(**?**):**?**. This multicenter validation of an AI-assisted CBCT interpretation system showed high diagnostic accuracy across independent settings and significantly reduced interpretation time relative to human readers. Journal of Medicine and Dentistry

[3] Ellis E 3rd, Moos KF, el-Attar A. Ten years of mandibular fractures: an analysis of 2,137 cases. J Oral Maxillofac Surg. 1985;43(1):31–38.

[4] Hogg NJ, Stewart TC, Armstrong JE, Girotti MJ. Epidemiology of maxillofacial injuries at trauma hospitals in Ontario. J Oral Maxillofac Surg. 2000;58(3):334–340.

[5] Zachariades N. Complications associated with facial trauma. Oral Surg Oral Med Oral Pathol. 1993;75(3):275–279.

[6] Adeyemo WL, Ladeinde AL, Ogunlewe MO, James O. Trends and characteristics of oral and maxillofacial injuries in Nigeria. J Oral Maxillofac Surg. 2005;63(9):1140–1144.

[7] Scarfe WC, Farman AG. What is cone-beam CT and how does it work? Dentomaxillofac Radiol. 2008;37(1):6–9.

[8] Miracle AC, Mukherji SK. Conebeam CT of the head and neck, part 1: physical principles. Radiographics. 2009;29(4):1089–1106.

[9] Pauwels R, Beinsberger J, Stamatakis H, et al. Comparison of spatial and contrast resolution of cone-beam CT scanners. Dentomaxillofac Radiol. 2012;41(6):401–409.

[10] Alpert B, Tiwana PS. Diagnosis and management of facial fractures. Oral Maxillofac Surg Clin North Am. 2013;25(4):571–583.

[11] Bagheri SC, Dierks EJ, Kademani D, Holmgren E. Application of CT imaging in the diagnosis of facial fractures. J Oral Maxillofac Surg. 2006;64(3):429–435.

[12] Schuknecht B, Graetz K. Radiologic assessment of maxillofacial trauma. Eur Radiol. 2005;15(3):560–568.

[13] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

[14] Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. Radiology. 2018;288(3):E133–E141.

[15] Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans. Radiology. 2018;289(3):811–819.

[16] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMJ. 2019;364:l233.

[17] Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals. Lancet Digit Health. 2019;1(6):e271–e297.

[18] Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning in medical imaging. Nat Mach Intell. 2021;3(3):199–217.

[19] Park SH, Han K. Methodologic guide for evaluating clinical performance of artificial intelligence. Radiology. 2018;286(3):800–809.

[20] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015. Lecture Notes in Computer Science. Vol 9351. p. 234–241.

[21] Wulkan M, Parreira JG, Botter DA. Epidemiology of facial trauma. Rev Col Bras Cir. 2005;32(4):181–185.

[22] Perry M, Dancey A, Mireskandari K, et al. Emergency care in facial trauma: a review. Br J Oral Maxillofac Surg. 2005;43(5):381–390.