Machine Versus Man: Artificial Intelligence Diagnostic Accuracy in Fracture Diagnosis

Máquina *Versus* Homem: A Inteligência Artificial no Diagnóstico de Fraturas

Fábia Silva[1]*, Diogo Tomaz[1], Micaela Gonçalves[1], Jorge Lopes[1], Miguel Relvas Silva[1], Vítor Vidinha[1], António Sousa[1]

1. Departamento de Ortopedia e Traumatologia, Centro Hospitalar Universitário de São João, Porto, Portugal

https://doi.org/

ABSTRACT

Introduction: Yearly around 21 thousand adult patients visit our tertiary hospital's emergency department after suffering from high or low energy trauma. Skeletal radiographs, being inexpensive and widely available, are the first-line imaging modality. Recent studies are showing encouraging results of the use of artificial intelligence in the detection of bone fractures. The main objective of this study is to compare the diagnostic accuracy between a medical-grade artificial intelligence (AI) software (BoneView®, Gleamer) and orthopaedic surgeons of various levels of expertise for the detection of bone fractures in a tertiary hospital's emergency department.

Methods: Retrospective analysis of a series of posttraumatic radiographic examinations, including only adult patients with plain radiographs of limbs or pelvis obtained after a recent trauma. Exclusion criteria were patients with cast control radiographs, images with inadequate radiographic quality, and examinations showing only obvious fractures. The diagnostic performance of the AI software and six orthopaedic surgeons was measured by sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

Results: The AI software had 91.3% sensitivity (95% CI: 82.03-96.74) and 97.3% specificity (95% CI: 93.22-99.26), with 0.95 AUC (95% CI: 91.3-98.8; p <0.001). All six readers had inferior results in every measure obtained, with slight differences between them.

Conclusion: Our study demonstrated that the BoneView® software has a high diagnostic capacity for fractures and, in this regard, can be considered a useful tool in the emergency department.

Keywords: Artificial Intelligence; Emergency Service, Hospital; Fractures, Bone/diagnostic imaging;

RESUMO

Introdução: Anualmente cerca de 21 mil adultos recorrem à urgência do nosso hospital terciário por traumatismos de pequena ou elevada energia. A radiografia, de acessibilidade imediata, é geralmente o primeiro exame diagnóstico

Autor Correspondente/Corresponding Author: Fábia Silva [fabia.natacha@gmail.com], Alameda Prof. Hernâni Monteiro 4200–319 Porto, Portugal

Recebido/Received: 2024/08/27 **Aceite/Accepted:** 2024/10/06 **Publicado online/Published online:** 2024/10/16 **Publicado/Published:** -

realizado. Recentemente a inteligência artificial (IA) na saúde expandiu-se para o diagnóstico de fraturas, com resultados encorajadores.
O objetivo deste trabalho é comparar a acuidade diagnóstica de fraturas de um *software* de IA (BoneView®, Gleamer) com ortopedistas de diferentes graus de diferenciação no serviço de urgência de um hospital terciário.
Métodos: Análise retrospetiva de uma amostra de radiografias realizadas no contexto de trauma, incluindo adultos que realizaram pelo menos 1 incidência radiográfica do esqueleto apendicular. Excluídas imagens radiográficas de seguimento, baixa qualidade ou com apenas fraturas óbvias. O desempenho diagnóstico do sistema IA e ortopedistas incluídos foi avaliado através da sensibilidade, especificidade e área sob a curva ROC (AUC).
Resultados: O sistema IA demonstrou uma sensibilidade de 91,3% (95% CI: 82,03-96,74) e especificidade de 97,3% (95% CI: 93,22-99,26), com uma AUC de 0,95 (95% CI: 91,3-98,8; *p* <0,001). Em média estes valores foram ligeiramente inferiores para todos os ortopedistas, com pequenas diferenças entre eles.
Conclusão: O nosso estudo mostrou que o *software* BoneView® tem uma elevada capacidade diagnóstica para fraturas e, nesse sentido, poderá assumir-se como uma ferramenta útil no serviço de urgência.

Palavras-chave: Fracturas Ósseas/diagnóstico por imagem; Inteligência Artificial; Serviço de Urgência Hospitalar

INTRODUCTION

Yearly, around 21 thousand adult patients visit the emergency department (ED) of our tertiary referral hospital after suffering from either high or low-energy trauma. Skeletal radiographs are the first-line imaging modality to diagnose traumatic skeletal injuries as they are inexpensive, widely available, and expeditious.[1,2] However, diagnostic errors in the emergency department are the highest when interpreting trauma radiographs.[1,3] This can be explained by long or overnight shifts due to doctors' fatigue, distracting injury effect in multiple trauma patients, lack of radiologic expertise or simply by the high volume of patients and exiguous time to consult each of them.[4] Missed fractures can have a serious impact on patients' outcomes resulting in malunion or arthritis, for instance. Delayed treatment can have a similar effect on patients' morbidity.[2,5] Computer-aided detection software has been around for a couple of decades now and is widely used for breast cancer screening, among other pathologies. Recent studies are showing encouraging results of the use of artificial intelligence in the detection of bone fractures as well as dislocations or joint effusions.[2,6] Providing emergency department physicians with artificial intelligence (AI) fracture detection tools could help reduce diagnostic error rates in trauma settings. Consequently, this could aid reduce malpractice claims against physicians, a prominent concern in developed countries nowadays.

The main objective of this study is to compare the diagnostic accuracy between medical-grade AI software (BoneView®, Gleamer) and orthopedic surgeons of various levels of expertise for the detection of bone fractures in a tertiary hospital's emergency department.

MATERIAL AND METHODS

Study Design and Population

This is a retrospective analysis of a series of posttraumatic radiographic examinations performed in the ED of a tertiary hospital, during the first week of the BoneView® experimental period in the trauma setting.

Inclusion criteria were age 18 years or older and at least one digital plain radiograph of limbs or pelvis obtained after a recent trauma, with or without lesion. Exclusion criteria were patients with cast control radiographs, images with inadequate radiographic quality, and examinations showing only obvious fractures (displaced, dislocated, or multiple fragments). Examinations where the AI software was not able to ascertain a diagnosis, either positive or negative, were excluded as well.

All data was treated anonymously, and the entire study protocol was approved by the hospital's ethics committee.

Reading of the Data Set and Ground Truth Definition

A total of six readers, two orthopaedic surgery specialists (Specialist 1 and 2) with about 15 years of experience and four orthopaedic residents (two sixth year residents – Resident 1 and 2; two third year residents – Resident 3 and 4) interpreted every view of patient's radiographs. The readers did not have access to patients' clinical history nor to the ground truth, they were also blinded to one another. Each fracture was identified by the orthopaedic surgeons in at least one view of every singular case. The reading time of

each examination was not registered. The BoneView® software analyzed the same sets of radiographs and when a fracture was identified it would be highlighted with a box on each radiographic view (Fig. 1).The ground truth was defined by a complete agreement between AI and the two senior elements composing the orthopedics team assessing trauma patients at the time of presentation in the ED. Disagreements were resolved by majority consensus with a third senior orthopaedics specialist (A.S., with 25 years of experience).

Figure 1. Example of examination of a true positive 5th metacarpal base fracture by the AI software.

Statistical Analysis

The main outcomes measured were the sensitivity and specificity of the AI software and every orthopedic surgeon. Positive predictive value (PPV) and negative predictive value (NPV) were also calculated. The performance of the BoneView® software was compared with the assessments made by six independent orthopedic surgeons by using receiver operating characteristic curves (ROC). IBM SPSS Statistics version 29.0 software was utilized for statistical analysis.

RESULTS

Data Set Characteristics

A total of 217 sets of radiographs were included in this study with 67 (30.87%) identified fractures. All anatomic regions were represented, apart from the spine, as previously stated. A summary of the examination location and presence of fracture per anatomical area is given in Table 1.

Tabela 1. Summary of examination location and presence of fracture.

Anatomical area	Fractures, no. (% total)	Examinations, no. (%)
Shoulder girdle and Elbow	9 (23.07)	39 (17.97)
Forearm, Wrist and Hand	22 (40.00)	55 (25.34)
Pelvis and Thigh	12 (46.15)	26 (10.59)
Knee and Leg	6 (16.66)	36 (16.58)
Ankle and Foot	18 (29.50)	61 (29.03)

Performance Analysis

The AI software was calculated to have a sensitivity of 91.3% (95% CI: 82.03-96.74) and a specificity of 97.3% (95% CI: 93.22-99.26). When analyzing the six readers' individual performances, sensitivity varied from 77.4% (95% CI: 66.00-86.54) to 86.3% (95% CI: 75.69-93.57) with an overall value 10 points lower (81.3%; 95% CI: 77.17-85.07) than the BoneView® software. A less than 6 points difference (5.4) is seen in the readers obtained specificity with an overall value of 91.9% (95% CI: 89.96-93.62) in comparison with the AI's 97.3% (95% CI: 93.22-99.26). The PPV and NPV were highest for the AI software as well, being 94.0% (95% CI: 85.66%-97.65) and 96.0% (95% CI: 91.78-98.10), respectively. Results of the AI software and readers performance are detailed in Table 2.

Area under the receiver operating characteristic curve (AUC) of the BoneView® software (0.95; 95% CI: 0.913--0.988; p <0.001) was larger than that of any reader to diagnose patient's fractures. It was followed by one of the most experienced orthopedic surgeons (Specialist 1) with an AUC of 0.89 (95% CI: 0.842-0.950; p <0.001). All four residents showed substantial agreement in diagnostic performance, with the less experienced ones (Resident 3 and 4) having slightly smaller AUCs. Lowest-performing with an AUC of 0.83 (95% CI: 0.765-0.900; p <0.001) was the remaining orthopedic surgery specialist. Fig. 2 shows the diagnostic performance of the AI software and each reader. Table 3 includes the data presented in the figure.

Table 2. Diagnostic performances of AI software, orthopaedic residents, and specialists.

	Sensibility (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
AI software	0.913 (82.03-96.74)	0.973 (93.22-99.26)	0.940 (85.66%-97.65)	0.960 (91.78-98.10)
Resident 1	0.797 (68,31-88.44)	0.918 (86.27-95.74)	0.820 (72.45-88.87)	0.906 (85.86-93.96)
Resident 2	0.846 (73.52-92.37)	0.921 (86.62-95.85)	0.820 (72.50-88.85)	0.933 (88.77-96.12)
Resident 3	0.774 (66.00-86.54)	0.917 (86.08-95.68)	0.820 (72.43-88.88)	0.893 (84.44-92.82)
Resident 4	0.776 (65.78-86.89)	0.940 (88.92-97.22)	0.852 (75.17-91.69)	0.903 (85.73-93.63)
Specialist 1	0.863 (75.69-93.57)	0.933 (88.16-96.78)	0.850 (75.66-91.27)	0.940 (89.50-96.64)
Specialist 2	0.830 (71.03-91.56)	0.933 (88.08-96.76)	0.830 (72.70-90.02)	0.933 (88.82-96.10)
All readers	0.813 (77.17-85.07)	0.919 (89.96-93.62)	0.815 (77.94-84.71)	0.918 (90.14-93.25)
All Residents	0.797 (74.51-84.39)	0.924 (90.03-94.44)	0.828 (78.35-86.54)	0.909 (88.77-92.70)
All Specialists	0.848 (77.29-90.59)	0.909 (87.17-93.89)	0.791 (72.52-84.45)	0.936 (90.71-95.73)

PPV: positive predictive value, NPV: negative predictive value; CI: confidence interval

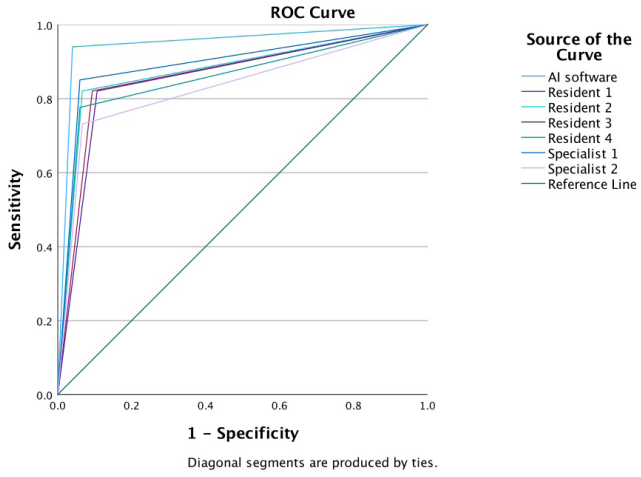

Figure 2. Receiver operating characteristic curves show artificial intelligence and readers' performance.

Table 3. Area under the ROC curve.

	Area	*P*-value[a]	95% Confidence Interval	
			Lower Bound	Upper Bound
AI software	.950	.000	.913	.988
Resident 1	.864	.000	.804	.924
Resident 2	.877	.000	.818	.936
Resident 3	.857	.000	.796	.918
Resident 4	.858	.000	.795	.921
Specialist 1	.895	.000	.841	.950
Specialist 2	.832	.000	.765	.900

a. Null hypothesis: true area = 0.5

DISCUSSION

To our knowledge, this was the first study to assess the performance of an AI software for the diagnosis of appendicular and pelvic bone fractures in a Portuguese hospital. Our results are comparable with several other mono and multicentric studies with high sensitivity (91.3%) and specificity (97.3%) of the AI software, but also comparable results for the orthopaedic surgeons and residents who were part of the study.[1-3,5-7] However, not including obvious fractures in the data set might have resulted in underestimating the overall diagnostic capability of the readers.

The receiver operating curves (Fig. 2) allow us to assess the success rate of each subject, by calculating the corresponding AUC. A subject with a higher AUC is better at distinguishing between patients with a fracture and the ones without. Comparable recently published research articles have documented AUCs higher than 0.90 for various AI software. Likewise, we report a BoneView® software AUC

of 0.950. All six readers had lower AUCs, ranging from 0.832 to 0.895. As expected, there was performance variation across readers with different experience levels. However, the lowest performing reader was one of the most experienced ones, contrary to the expected. This could be related, for instance, to a difference in allotted reading time during the examination, which was not measured. One strength of this study is the assessment of all appendicular radiographs, as opposed to other published studies that focus on single body parts.

Our study has several limitations. Firstly, the sample size of the study was relatively small, when compared to other published studies with several hundred patients.

Another limitation was that the reading time of each examination was not recorded, even though we recognize that time constraints can be a relevant factor in a frequently busy ED. In everyday clinical practice, the sparse time to interpret conventional plain radiographs may partially account for the failed diagnoses. We note several studies that report an improvement in reading time when readers have an AI aid,[3,4] and we recognize that it might make the best of the limited time available to professionals.

As previously detailed in the literature, in more visually challenging anatomical areas, like the hand and foot, fractures are more commonly missed.[3,5,8] A subgroup analysis, by anatomical area, would be interesting in future work with larger data sets. Furthermore, the readers not having access to patient's clinical history and making a diagnosis solely with radiographic images creates a context bias, and the same can be applied to the BoneView® software. Additionally, the context of the research could have influenced the readers to do a more rigorous review of the radiographs, creating a Hawthorne effect.

Radiologists were not included in the establishment of ground truth, nor were they part of the evaluated readers. The purpose here was to emulate a real ED setting in our tertiary hospital, where all appendicular and spine trauma is assessed exclusively by orthopedic surgeons.

Previous studies suggest that the highest diagnostic performance might be attained when clinicians have access to AI reports.[1,4-6,9] Subsequent research could benefit from including an AI-aided assessment of radiographs by the orthopedic surgeons, as well as analysis of imaging reading time and inclusion of background clinical information.

CONCLUSION

Our study demonstrated that the AI system has a high diagnostic capacity for fractures and, in this regard, can be considered a useful tool in the emergency department. By reducing diagnostic errors, it assists orthopedic surgeons in decision-making, thereby enhancing care for the population.

AI software like BoneView®, despite not having yet well-defined regulations established for their use, can be of great value in the ED for physicians of different backgrounds assessing trauma patients. Legal and ethical issues may come to light, but the seeming benefit of AI fracture detection tools cannot be ignored. In the near future, we see AI tools becoming a complement in the assessment of traumatic skeletal injuries, improving healthcare quality for the population.

Prémios e Apresentações prévias

Este trabalho foi apresentado na forma de Comunicação Oral no 42º Congresso Nacional de Ortopedia e Traumatologia em Novembro de 2023, tendo recebido as distinções de "Melhor Comunicação Livre" e "Melhor Trabalho de Traumatologia".

Responsabilidades Éticas

Conflitos de Interesse: Os autores declaram a inexistência de conflitos de interesse na realização do presente trabalho.
Fontes de Financiamento: Não existiram fontes externas de financiamento para a realização deste artigo.
Confidencialidade dos Dados: Os autores declaram ter seguido os protocolos da sua instituição acerca da publicação dos dados de doentes.
Proteção de Pessoas e Animais: Os autores declaram que os procedimentos seguidos estavam de acordo com os regulamentos estabelecidos pela Comissão de Ética responsável e de acordo com a Declaração de Helsínquia revista em 2013 e da Associação Médica Mundial.
Proveniência e Revisão por Pares: Não comissionado; revisão externa por pares.

Ethical Disclosures

Conflicts of Interest: The authors have no conflicts of interest to declare.
Financing Support: This work has not received any contribution, grant or scholarship
Confidentiality of Data: The authors declare that they have followed the protocols of their work center on the publication of data from patients.

Protection of Human and Animal Subjects: The authors declare that the procedures followed were in accordance with the regulations of the relevant clinical research ethics committee and with those of the Code of Ethics of the World Medical Association (Declaration of Helsinki as revised in 2013).

Provenance and Peer Review: Not commissioned; externally peer reviewed.

Declaração de Contribuição

FS: Conceção e desenho do trabalho; aquisição de dados, análise, interpretação e redação.

DT, MG, MG, JL e MRS: Aquisição, análise e interpretação dos dados.

VV e AS: Revisão crítica do manuscrito.

Todos os autores aprovaram a versão final a ser publicada.

Contributorship Statement

FS: Conception and design of the work; data acquisition, analysis, interpretation and writing.

DT, MG, MG, JL and MRS: Data acquisition, analysis and interpretation.

VV and AS: Critical review of the manuscript.

All authors approved the final version to bel published.

References

1. Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, et al. Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study. Radiology. 2021;300:120-9. doi: 10.1148/radiol.2021203886.
2. Zhang X, Yang Y, Shen YW, Zhang KR, Jiang ZK, Ma LT, et al. Diagnostic accuracy and potential covariates of artificial intelligence for diagnosing orthopedic fractures: a systematic literature review and meta-analysis. Eur Radiol. 2022;32:7196-216. doi: 10.1007/s00330--022-08956-4.
3. Canoni-Meynet L, Verdot P, Danner A, Calame P, Aubry S. Added value of an artificial intelligence solution for fracture detection in the radiologist's daily trauma emergencies workflow. Diagn Interv Imaging. 2022;103:594-600. doi: 10.1016/j.diii.2022.06.004.
4. Guermazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology. 2022;302:627--36. doi: 10.1148/radiol.210937.
5. Jones RM, Sharma A, Hotchkiss R, Sperling JW, Hamburger J, Ledig C, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med. 2020;3:144. doi: 10.1038/s41746-020-00352-w.
6. Kuo RY, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, Stewart M, Collins GS, Furniss D. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. 2022;304:50-62. doi: 10.1148/radiol.211785.
7. Harrison W, Newton AW, Cheung G. The litigation cost of negligent scaphoid fracture management. Eur J Emerg Med. 2015;22:142-3. doi: 10.1097/MEJ.0000000000000152.
8. Jassar S, Adams SJ, Zarzeczny A, Burbridge BE. The future of artificial intelligence in medicine: Medical-legal considerations for health leaders. Healthc Manage Forum. 2022;35:185-9. doi: 10.1177/08404704221082069.
9. Naik N, Hameed BM, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? Front Surg. 2022;9:862322. doi: 10.3389/fsurg.2022.862322.
10. Regnard NE, Lanseur B, Ventre J, Ducarouge A, Clovis L, Lassalle L, et al. Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays. Eur J Radiol. 2022;154:110447. doi: 10.1016/j.ejrad.2022.110447.