

POROVNÁNÍ DIAGNOSTICKÉ PŘESNOSTI SYSTÉMŮ UMĚLÉ INTELIGENCE A JUNIORNÍCH RADIOLOGŮ PŘI DETEKCI FRAKTUR: RETROSPEKTIVNÍ, MULTI-READER STUDIE

Šimon Kubov, Daniel Kvak, Jakub Dandár

Abstrakt

Retrospektivní, monocentrická, multi-reader, multi-case studie hodnotila diagnostickou výkonnost dvou systémů umělé inteligence (AI 1, AI 2) při detekci prvodiagnostických fraktur na rentgenových snímcích muskuloskeletálního aparátu a porovnávala ji s výkonem tří klinicky aktivních radiologů (RAD 1–3). Z archivu PACS Fakultní nemocnice Olomouc bylo vybráno 1 137 anonymizovaných snímků pořízených mezi 5. a 11. březnem 2025; prevalence fraktur činila 10,4 %. Referenční standard byl stanoven majority-vote konsensem tří nezávislých radiologů, přičemž substanciální shodu potvrdila Fleissova $\kappa = 0,739$. Power analýza požadovala minimálně 110 fraktur a 900 negativních případů; skutečný soubor tyto limity komfortně překročil. Statistické hodnocení vycházelo ze čtyřpolních kontingenčních tabulek s Wilsonovými 95% intervaly spolehlivosti (CI); rozdíly mezi čtenáři byly testovány McNemarovým testem s Holmovou korekcí. AI 1 dosáhla senzitivity 0,941 (95 % CI 0,883–0,971) a specifity 0,930 (0,913–0,944), AI 2 senzitivity 0,932 (0,872–0,965) a specifity 0,935 (0,918–0,949). Senzitivita obou algoritmů se statisticky nelišila ($p = 0,79$); AI 2 však vykázala vyšší specifitu než AI 1 ($p < 0,05$). Radiologové dosáhli senzitivity 0,915–0,941 a specifity 0,944–0,977. Jejich senzitivita byla srovnatelná s oběma AI, zatímco specifita překonala AI 1 a byla obdobná či mírně vyšší než AI 2. Negativní prediktivní hodnota všech hodnocených komparátorů přesáhla 0,990. Subanalýzy šesti anatomických regionů a čtyř věkových kohort potvrdily konzistentně vysokou citlivost obou AI, s nejvyšší přesností v oblasti ramene/klíční kosti a největším poklesem specifity AI 1 u ruky/zápěstí. Studie ukazuje, že oba systémy poskytují detekci fraktur se senzitivitou srovnatelnou se zapojenými radiology.

Klíčová slova

umělá inteligence, fraktury, diagnostická přesnost, muskuloskeletální radiografie, multi-reader studie

1 Úvod

Rentgenová diagnostika skeletárních traumat zůstává navzdory rozvoji pokročilých modalit klíčovou metodou akutní radiologie. Přehlédnuté zlomeniny jsou přitom častým zdrojem diagnostických pochybení a mohou vést ke zpoždění léčby [1,2]. V posledních letech proto roste zájem o využití systémů umělé inteligence (AI) založených na hlubokém učení, které slibují zvýšit spolehlivost detekce fraktur na běžných rentgenových snímcích [3]. Dosavadní práce většinou ukazují nárůst senzitivity, vliv na specifitu je však proměnlivý a mezi studiemi panuje výrazná heterogenita daná rozdíly v datových souborech, anatomických lokalitách i klinickém workflow [4]. To podtrhuje nutnost nezávislého, transparentního a metodicky ukotveného ověřování výkonnosti komerčních AI řešení v reálné praxi.

Mnohé publikace se zaměřují na jediný anatomický segment nebo specifickou populaci (např. dětské pacienty) a probíhají v kontrolovaném prostředí [5]. Komplexní validace napříč více částmi pohybového aparátu a širokým věkovým spektrem je však klíčová: dětské snímky představují zvláštní výzvu pro interpretaci, zatímco u seniorů se častěji vyskytují degenerativní změny a komorbidita komplikující hodnocení [6].

Tato práce představuje retrospektivní, monocentrickou studii typu multi-reader, multi-case realizovanou ve FN Olomouc, která na souboru 1 137 rentgenových snímků hodnotí diagnostickou přesnost dvou nezávislých AI modelů (AI 1, AI 2) pro detekci fraktur a porovnává ji s výkonem tří klinicky aktivních radiologů (RAD 1–3; 1–3 roky praxe). Referenční standard byl stanoven majority-vote konsensem trojice nezávislých radiologů, kteří se neúčastnili srovnávacího čtení. Primárním cílem je ověřit, zda oba AI systémy dosahují na úrovni snímku senzitivity i specifity $\geq 0,90$, přičemž 95% intervaly spolehlivosti nesmí tyto prahové hodnoty vylučovat. Sekundárním cílem je posoudit robustnost výkonu v šesti předem definovaných anatomických oblastech a ve čtyřech věkových kohortách. Výsledky jsou kontextualizovány vůči lidským čtenářům s cílem poskytnout důkazy o možnostech využití AI v diagnostice skeletálních traumat.

2 Materiál a metody

2.1 Design studie

Studie má retrospektivní, monocentrický design typu multi-reader, multi-case (MRMC), ve kterém několik hodnotitelů posuzovalo shodnou sadu případů. Hodnocena byla diagnostická přesnost dvou nezávislých AI systémů (označených zde jako AI 1 a AI 2) při detekci nově diagnostikovaných (akutních) fraktur na rentgenových snímcích pohybového aparátu. V rámci přípravy byla provedena power analýza k odhadu potřebného počtu snímků. Následně byl retrospektivně získán soubor snímků z archivu nemocnice za definované období jednoho týdne (5.–11. březen 2025).

Inkluzní kritéria zahrnovala:

- snímky pacientů všech věkových kategorií s podezřením na akutní traumatickou zlomeninu
- snímky pořízené při prvotním vyšetření (tj. před zahájením léčby dané zlomeniny)
- digitální rentgenové snímky adekvátní diagnostické kvality

Exkluzní kritéria:

- snímky jiné anatomické oblasti než pohybového aparátu (např. páteř, lebka, žebra)
- snímky s nedostatečnou technickou kvalitou (artefakty, neúplné zobrazení)
- opakované snímky téhož úrazu po provedení repozice či jiné terapie
- případy, u nichž nebylo spolehlivě možné stanovit referenční pravdu (neshoda mezi hodnotiteli)

Po aplikaci kritérií bylo do další analýzy zařazeno 1 201 snímků splňujících všechny podmínky. Každý snímek byl považován za samostatný případ (jednotku hodnocení). Pokud měl jeden pacient více projekcí či snímků různých anatomických lokalit, hodnotily se tyto samostatně. Před zpracováním byla obrazová data anonymizována a exportována z PACS ve formátu DICOM; nebylo prováděno žádné speciální předzpracování kromě případné konverze formátu podle potřeb AI.

2.2 Identifikace software

V pilotním nasazení ve Fakultní nemocnici Olomouc byly testovány dvě softwarové aplikace umělé inteligence pro detekci fraktur na rentgenových snímcích: Carebot AI Bones (Carebot s.r.o., ČR), verze 1.11 a Gleamer BoneView (Gleamer, Francie), verze 2.5.4. Dále v textu jsou tyto systémy označeny neutrálně jako AI 1 a AI 2.

Pohlaví	n_{GT}	$n_{FRACTURE}$	n_{NORMAL}	Prevalence (%)	Věk (roky)
Celkem	1 137	118	1 019	10,4	44,0 ± 22,8
Muži	560	66	494	11,8	38,6 ± 21,5
Ženy	577	52	525	9,0	49,1 ± 22,8

Tabulka 1 – Demografie

Prahová hodnota obou systémů pro rozhodnutí nebyla během hodnocení upravována – byl použit pevný práh nastavený výrobcem na základě interní validace. Studie probíhala podle předem definovaného harmonogramu: sběr primárních snímků proběhl 5.–12. března 2025 souběžně s analýzou AI, následoval export dat (časový bod T_0), čištění dat a nezávislé radiologické hodnocení. Statistické vyhodnocení proběhlo na konci června 2025 a studie byla uzavřena 4. července 2025.

2.3 Akvizice dat

Obrazová data byla získána retrospektivně z nemocničního systému PACS (Marie PACS, OR-CZ spol. s r.o., ČR) Fakultní nemocnice Olomouc. Do vyhledávacího dotazu byly zahrnuty všechny rentgenové studie muskuloskeletálního aparátu provedené v období od 5. března 2025 00:00 do 11. března 2025 23:59. Z PACS bylo původně extrahováno 3 547 souborů ve formátu DICOM.

Následné manuální čištění dat provedl radiolog s 3 lety praxe. Ze souboru byly vyřazeny:

- snímky nepodporovaných anatomických oblastí (žebra, páteř, lebka, abdomen)
- kontrolní („follow-up“) studie
- snímky s osteosyntetickým materiálem či fixačními dlahami
- duplicitní záznamy vzniklé opakovaným ukládáním totožných sérií

Tento postup byl zvolen záměrně, protože studie je zaměřena výhradně na prvotní diagnostické zobrazení, kde má rozhodnutí o přítomnosti či nepřítomnosti fraktury největší klinický dopad. Po aplikaci vylučovacích kritérií zůstalo 1 201 unikátních snímků, které splňovaly požadavky na původní diagnostické zobrazení a mohly být dále hodnoceny.

Snímky byly před zpracováním nevratně anonymizovány v souladu s nařízením (EU) 2016/679 (GDPR) a zákonem 110/2019 Sb. Byly odstraněny všechny přímé i nepřímé identifikátory pacienta a zdravotnického personálu, přičemž technické parametry expozice a informace o použitém zařízení zůstaly zachovány.

Studie byla provedena v souladu s principy správné klinické praxe pro retrospektivní analýzy (ISO 14155:2020). [7]

2.4 Cíle studie

Primární cíl: Kvantitativně prokázat, že oba zkoumané systémy umělé inteligence (AI 1 a AI 2) dosahují na prvodiagnostických rentgenových snímcích muskuloskeletálního systému senzitivity a specifity alespoň $Se \geq 0,90$ a $Sp \geq 0,90$. U obou metrik zároveň dolní mez 95% intervalu spolehlivosti nesmí klesnout pod tyto prahové hodnoty. Referenčním standardem byl konsenzus tří radiologů stanovený metodou majority vote. Uvedené prahy vycházejí z požadavků urgentního příjmu – senzitivita pod 0,90 by znamenala nepřijatelný počet přehlédnutých fraktur, zatímco specifita pod 0,90 by generovala nadměrné množství falešně pozitivních nálezů.

Sekundární cíle: Posoudit robustnost algoritmu v klinicky relevantních podskupinách. Hodnotí se, zda oba systémy zachovávají (případně s akceptovatelnou odchylkou dosahují) výše definovaných mezních hodnot senzitivity a specifity

v šesti anatomických oblastech (chodidlo/kotník, koleno/dolní končetina, ruka/zápěstí, loket/paže, rameno/klíční kost, pánev/kyčelní kloub) a ve čtyřech věkových kohortách pacientů (0–17 let, 18–40 let, 41–65 let, ≥ 66 let).

Explorativní cíl: Porovnat výsledky obou algoritmů se zapojenými radiology. Toto srovnání slouží k dokumentaci stavu současné technologie, nikoli k testování superiority či non-inferiority.

2.5 Referenční standard

Referenční standard byl stanoven trojicí nezávislých radiologů (GT 1: 7 let praxe; GT 2: 4 roky praxe; GT 3: 2 roky praxe). Všichni posuzovali snímky zaslepeně, bez přístupu k výstupům AI či klinickým údajům o pacientech. Hodnocení probíhalo ve specializovaném anotačním softwaru s běžně dostupnými nástroji pro zobrazení (např. lupa, nastavení jasu a kontrastu).

Pro klasifikaci fraktura-pozitivních případů (FRACTURE) bylo použito pravidlo většinového hlasování (alespoň 2 ze 3 hodnotitelů). Naopak pro označení snímku jako fraktura-negativního (NORMAL) byla vyžadována plná shoda všech tří hodnotitelů. Snímky s neshodou (1/3 proti 2/3) byly z analýzy vyřazeny.

Z původního počtu 1 201 snímků byl referenční standard stanoven u 1 137 snímků muskuloskeletálního systému (94,7 %). Z toho bylo 1 019 klasifikováno jako fraktura-negativní (84,3 %) a 118 jako fraktura-pozitivní (10,4 %). Celkem 64 snímků (~5,3 %) vykazovalo neshodu mezi hodnotiteli a bylo vyřazeno. Tento nízký podíl neměl vliv na robustnost datového souboru.

Co se týče použitých rentgenových přístrojů, největší podíl snímků byl pořízen detektorem Samsung GC85A (583 snímků, ~ 51 %). Další zastoupené platformy zahrnovaly Philips Medical Systems PCR (16 snímků, ~ 1,4 %), Philips DigitalDiagnost (3 snímky, ~ 0,3 %) a DRGEM GXR-32SD (2 snímky, ~ 0,2 %). Zbývající přístroje jednotlivě nepřesáhly 1 % z celkového souboru snímků.

Rentgenový přístroj	Počet snímků	Podíl (%)
Samsung Electronics GC85A	563	49,5
Agfa – Pixium 4343E CSI	326	28,7
Agfa 3543EZE	204	17,9
Siemens Fluorospot Compact FD	24	2,1
Philips DigitalDiagnost	8	0,7
DRGEM GXR System	4	0,4
GE Healthcare DR-F	3	0,3
DRGEM GXR-32SD System	2	0,2
Canon CXDI Control Software NE	2	0,2
Agfa XD 17	1	0,1

Tabulka 2 – Počty snímků dle použitého přístroje

2.6 Multi-reader studie

Do studie byli zapojeni tři klinicky aktivní radiologové (RAD 1–3) s 1–3 lety praxe, kteří se nepodíleli na stanovení referenčního standardu. Všichni běžně vyšetřují akutní traumata pohybového aparátu a jejich profil odpovídá složení služeb na urgentním příjmu.

Hodnocení probíhalo zaslepeně v anotační aplikaci, snímky byly anonymizovány a prezentovány v náhodném pořadí. Nebyly poskytovány žádné klinické ani demografické údaje a nebyla poskytnuta asistence AI. Každý čtenář provedl binární klasifikaci („FRACTURE“ / „NORMAL“) u všech snímků, bez časového limitu a bez možnosti konzultace s ostatními.

2.7 Statistická analýza

Primární hodnocení výkonu AI 1 a AI 2 vycházelo ze čtyřpolního kontingenčního uspořádání. Senzitivita (Se) a specifita (Sp) jsou vyjádřeny s 95 % intervalem spolehlivosti (CI). Pro střední až vysoké četnosti byla použita Wilsonova metoda bez korekce kontinuity:

$$\hat{p}_{\text{Wilson}} = \frac{p + \frac{z^2}{2n}}{1 + \frac{z^2}{n}}, \text{CI} = \hat{p}_{\text{Wilson}} \pm z \sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}},$$

kde p je bodový odhad, n odpovídající jmenovatel a $z=1,96$. Pokud počet nesplnil kritérium $np(1-p) < 5$, byl automaticky použit exaktní Clopper–Pearsonův interval.

Z téhož kontingenčního stolu byly odvozeny pozitivní a negativní prediktivní hodnoty (PPV, NPV). Jejich 95 % CI byly stanoveny bootstrapem s 10 000 re-samply, neboť aproximace pro prediktivní hodnoty je silně závislá na prevalenci. Metriky byly počítány i pro šest předem definovaných anatomických oblastí a čtyři věkové kohorty. V podskupinách s méně než 20 frakturami jsou CI vypočteny exaktní Fisherovou metodou.

Rozdíl senzitivity a specifity (ΔSe , ΔSp) mezi AI 1, AI 2 a radiology byl hodnocen pomocí párových Newcombeových intervalů a McNemarova testu; p -hodnoty byly korigovány metodou Holm–Bonferroni. Formální testování superiority či non-inferiority nebylo primárním cílem. Protože modely vracejí pouze binární výstup, nebylo možné konstruovat klasickou ROC křivku.

Shoda mezi třemi referenčními radiology byla kvantifikována pomocí Fleissovy κ včetně 95% CI získaného bootstrapem.

K posouzení, zda se výkonnost algoritmů významně liší od průměru lidských hodnotitelů, byl použit generalizovaný lineární smíšený model (GLMM) s logitovou vazbou. Do modelu vstoupil každý snímek jako náhodný intercept a typ hodnotitele (AI vs. radiolog) jako fixní efekt. Parametry byly odhadnuty metodou Laplaceovy aproximace maxima likelihood a konzervativní robustní (Huber–White) standardní chyby byly použity pro intervaly.

Základní deskriptivní statistika charakterizovala soubor absolutními a relativními četnostmi, průměrem $\pm$ směrodatnou odchylkou (věk), případně mediánem [IQR], pokud distribuce neprošla testem normality (Shapiro–Wilk $p < 0,05$). Všechny výsledky a intervaly jsou uváděny na tři desetinná místa; CI v kulatých závorkách. Kritická hladina významnosti $\alpha = 0,05$ se uváděla pouze tam, kde byla potřeba pro konstrukci CI nebo testu.

Veškeré výpočty byly provedeny v prostředí Python 3.11 (pandas 2.2, numpy 1.26, statsmodels 0.15).

$$H_0 : p \leq 0.90 - \Delta \text{ vs. } H_1 : p > 0.90 - \Delta$$

kde p zastupuje Se nebo Sp a $\Delta=0,05$. Non-inferiorita byla považována za prokázanou, pokud dolní mez jednostranného 95% Wilsonova intervalu spolehlivosti (LCL) překročila hodnotu 0,85.

Pro binomální model bylo iterativně zjištěno, že k zaručení dolní meze Wilsonova intervalu nad 0,85 při skutečné Se = 0,93 je zapotřebí nejméně 110 fraktur. Při očekávané Sp = 0,93 posta-

čuje 900 negativních snímků, aby analýza dosáhla požadované síly; tato potřeba narůstá zejména při poklesu specifity pod 0,90. Při předpokládané prevalenci fraktur přibližně 10 % tak vyšel cílový rozsah souboru mezi 1 100 a 1 200 snímky. Tento rozsah zajišťuje minimálně 110 pozitivních a přibližně 990 negativních případů a zároveň obsahuje rezervu okolo 5 % pro možné ztráty způsobené duplicitami, nízkou obrazovou kvalitou či absencí referenčního standardu.

Třída	Minimální počet	Plánované rozpětí	Důvod
FRACTURE	≥ 110	110–130	Síla 80 % při Se $\approx 0,93$
NORMAL	≥ 900	950–1 050	Síla 80 % při Sp $\approx 0,93$
Celkem	$\geq 1 050$	1 100–1 200	Rezerva 5 % + rovnoměrné pokrytí podskupin

Tabulka 3 – Počty snímků dle power analýzy

Vedle primární analýzy bylo rovněž nutné zajistit orientační odhady výkonnosti v šesti anatomických oblastech a čtyřech věkových kohortách. Abychom v nejmenší z těchto podskupin dosáhli Wilsonova intervalu přibližně $\pm 0,20$ pro senzitivitu, byl stanoven minimální cíl deseti fraktur na každou oblast. Formální inferenční testy v podskupinách se neplánovaly; údaje sloužily pouze pro deskriptivní robustnostní analýzu.

3 Výsledky

3.1 Demografie

Analýzovaný soubor tvořilo 1 137 digitálních rentgenových snímků, z nichž 118 (10,4 %) bylo klasifikováno jako fraktura-pozitivní podle referenčního standardu. Průměrný věk všech pacientů činil $44,0 \pm 22,8$ roku; pacienti s frakturou byli v průměru o dva roky starší ($46,1 \pm 27,9$) než pacienti bez fraktury ($43,7 \pm 22,1$).

Distribuce snímků podle anatomických oblastí byla nerovnoměrná, s převahou kolene/dolní končetiny, chodidla/kotníku a ruky/zápěstí. Nejvyšší prevalence fraktur byla zaznamenána u lokte/paže a ramene/klíční kosti, zatímco koleno vykazovalo nejnížší podíl pozitivních nálezů (viz tabulka 4). Věkové rozložení ukázalo, že poranění lokte/paže se častěji vyskytovalo u mladších pacientů, zatímco fraktury pánve a kyčelního kloubu byly typické pro starší věkové skupiny.

Rozdělení souboru podle věkových skupin ukázalo největší zastoupení ve skupině 41–65 let, následované skupinou 18–40 let. Prevalence fraktur byla nejvyšší u dětí a adolescentů (0–17 let) a seniorů (≥ 66 let), zatímco u dospělých ve věku 18–65 let zůstala nižší a relativně vyrovnaná (viz tabulka 5).

3.2 Výkonnostní test

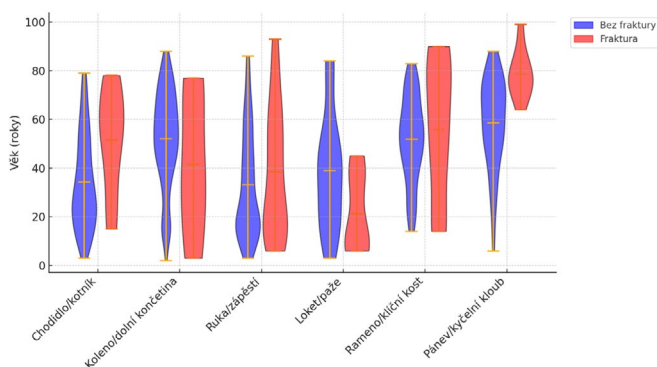
Ve validačním souboru dosáhl systém AI 1 senzitivity a specifity, které bezpečně splňují předem stanovené prahové hodnoty. Hodnoty senzitivity se statisticky nelišily od AI 2 ani od hodnocených radiologů, zatímco ve specifitě byl AI 1 překonán jak druhým algoritmem, tak všemi třemi radiology (viz tabulka 6).

Analýza prediktivních hodnot ukázala, že negativní prediktivní hodnota (NPV) zůstává u všech systémů i radiologů nad 0,99, což prakticky eliminuje riziko přehlédnutí fraktury. Pozitivní prediktivní hodnota (PPV) byla u AI 1 a AI 2 nižší než u dvou radiologů s vyšší mírou zkušeností (viz tabulka 7).

Podrobné výsledky senzitivity a specifity v jednotlivých anatomických oblastech ukazuje tabulka 8. Oba AI systémy vykáža-

Část těla	n_{GT}	Věk celkem (roky)	$n_{FRACTURE}$	Věk FRACTURE (roky)	n_{NORMAL}	Věk NORMAL (roky)
Celkem	1 137	44,0 ± 22,8	118	46,1 ± 27,9	1 019	43,7 ± 22,1
Chodidlo/kotník	279	35,4 ± 19,7	19	51,4 ± 21,5	260	34,3 ± 19,1
Koleno/dolní končetina	372	51,7 ± 20,2	11	41,5 ± 28,7	361	52,0 ± 19,9
Ruka/zápěstí	251	34,1 ± 22,6	44	38,5 ± 26,5	207	33,1 ± 21,6
Loket/paže	45	34,9 ± 23,4	10	21,2 ± 16,2	35	38,8 ± 23,8
Rameno/klíční kost	103	52,9 ± 20,7	25	55,8 ± 28,6	78	51,9 ± 17,6
Pánev/kyčelní kloub	87	60,7 ± 19,0	9	78,6 ± 10,1	78	58,6 ± 18,7

Tabulka 4 – Distribuce pacientů podle části těla: počet a věk u fraktur a normálních nálezů



Graf 1 – Violin ploty demografie

Věková skupina	n_{GT}	$n_{FRACTURE}$	n_{NORMAL}
Celkem	1 137	118	1 019
0–17 let	205	35	170
18–40 let	293	19	274
41–65 let	409	28	381
≥ 66 let	230	36	194

Tabulka 5 – Demografie, rozpad věkových skupin

ly konzistentně vysokou senzitivitu, nejvyšší v oblasti ramene/klíční kosti (AI 1 dosáhl senzitivity 1,000; AI 2 vykázal obdobně vysokou hodnotu) a nejnižší v oblasti pánve/kyčelního kloubu. Specifita AI 1 byla nižší zejména v oblasti ruky/zápěstí a pánve, zatímco AI 2 si zde udržel hodnoty blízké radiologům. V ostatních oblastech byly výsledky obou algoritmů srovnatelné a ve většině případů se statisticky významně nelišily od výkonu lidských čtenářů. Statisticky významné rozdíly se objevily pouze v několika případech, například nižší specifita AI 1 u ruky/zápěstí vůči všem třem radiologům.

3.3 Power analýza

Před zahájením studie byla stanovena minimální velikost souboru 1 000 snímků (≥ 110 fraktur a ≥ 900 negativních případů) s rezervou pro ztráty a rovnoměrné pokrytí podskupin. Vlastní akvizice v PACS FN Olomouc přinesla 1 137 plně vyhodnotitelných snímků, z nichž 118 (10,4 %) bylo fraktur-pozitivních a 1 019 (89,6 %) fraktur-negativních. Tím byly požadavky power analýzy splněny s rezervou a prevalence fraktur (~ 10 %) odpovídá předpokladu.

ID	Se (95 % CI)	p-Se	Sp (95 % CI)	p-Sp
AI 1	0,941 (0,883 – 0,971)	—	0,930 (0,913 – 0,944)	—
AI 2	0,932 (0,872 – 0,965)	0,790	0,935 (0,918 – 0,949)	< 0,05
RAD 1	0,924 (0,861 – 0,959)	0,605	0,944 (0,928 – 0,957)	< 0,05
RAD 2	0,915 (0,851 – 0,953)	0,450	0,977 (0,966 – 0,985)	< 0,05
RAD 3	0,941 (0,883 – 0,971)	1,000	0,975 (0,964 – 0,983)	< 0,05

Tabulka 6 – Výsledky výkonostního testu (Se, Sp)

ID	PPV (95 % CI)	p-PPV	NPV (95 % CI)	p-NPV
AI 1	0,610 (0,537 – 0,678)	—	0,993 (0,985 – 0,996)	—
AI 2	0,625 (0,552 – 0,693)	0,769	0,992 (0,984 – 0,996)	0,805
RAD 1	0,657 (0,582 – 0,725)	0,365	0,991 (0,982 – 0,995)	0,639
RAD 2	0,824 (0,750 – 0,880)	< 0,05	0,990 (0,982 – 0,995)	0,531
RAD 3	0,816 (0,743 – 0,872)	< 0,05	0,993 (0,986 – 0,997)	0,930

Tabulka 7 – Výsledky výkonostního testu (PPV, NPV)

Plán počítal s prokázáním non-inferiority vůči klinickému prahu 0,90, přičemž dolní mez jednostranného 95 % Wilsonova intervalu musela překročit 0,85. Skutečná výkonnost systému AI 1 dosáhla senzitivity 0,941 (95 % CI 0,883–0,971) a specifity 0,930 (0,913–0,944); AI 2 dosáhla senzitivity 0,932 (0,872–0,965) a specifity 0,935 (0,918–0,949), obě dolní meze bezpečně převyšovaly hranici 0,85.

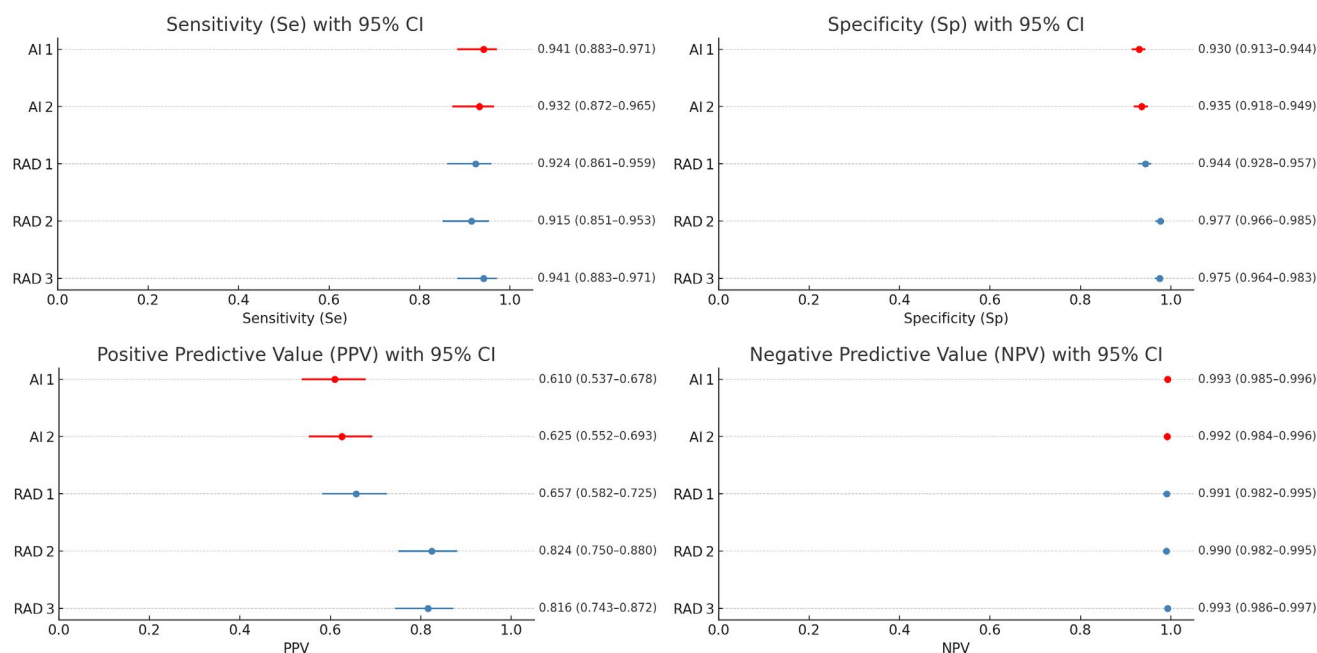
Power analýza požadovala minimálně 10 fraktur v každé ze šesti anatomických oblastí. Tento cíl byl splněn ve všech segmentech kromě pánve/kyčelního kloubu, kde bylo k dispozici pouze devět fraktur. Ve věkových kohortách byl cílový počet ≥ 10 fraktur splněn ve všech čtyřech skupinách (rozmezí 19–36), takže věková robustnost algoritmu mohla být vyhodnocena podle plánu.

POROVNÁNÍ DIAGNOSTICKÉ PŘESNOSTI SYSTÉMŮ UMĚLÉ INTELIGENCE A JUNIORNÍCH RADIOLOGŮ PŘI DETEKCI FRAKTUR: RETROSPEKTIVNÍ, MULTI-READER STUDIE

Šimon Kubov, Daniel Kvak, Jakub Dandár

Část těla	ID	Se (95 % CI)	p-Se	Sp (95 % CI)	p-Sp
Chodidlo / kotník	AI 1	0,947 (0,754 – 0,991)	—	0,946 (0,912 – 0,968)	—
	AI 2	0,947 (0,754 – 0,991)	1,000	0,904 (0,862 – 0,934)	0,067
	RAD 1	0,947 (0,754 – 0,991)	1,000	0,900 (0,858 – 0,931)	0,048
	RAD 2	0,737 (0,512 – 0,882)	0,075	0,973 (0,945 – 0,987)	0,119
	RAD 3	0,895 (0,686 – 0,971)	0,547	0,969 (0,940 – 0,984)	0,191
Koleno / dolní končetina	AI 1	0,909 (0,623 – 0,984)	—	0,972 (0,950 – 0,985)	—
	AI 2	0,909 (0,623 – 0,984)	1,000	0,964 (0,939 – 0,979)	0,525
	RAD 1	0,818 (0,523 – 0,949)	0,534	0,983 (0,964 – 0,992)	0,312
	RAD 2	0,909 (0,623 – 0,984)	1,000	0,994 (0,980 – 0,998)	0,020
	RAD 3	0,909 (0,623 – 0,984)	1,000	0,997 (0,984 – 1,000)	0,006
Ruka / zápěstí	AI 1	0,932 (0,818 – 0,977)	—	0,855 (0,801 – 0,897)	—
	AI 2	0,955 (0,849 – 0,987)	0,645	0,899 (0,850 – 0,933)	0,178
	RAD 1	0,886 (0,760 – 0,950)	0,458	0,928 (0,884 – 0,956)	0,018
	RAD 2	0,932 (0,818 – 0,977)	1,000	0,976 (0,945 – 0,990)	< 0,001
	RAD 3	0,909 (0,788 – 0,964)	0,694	0,947 (0,907 – 0,970)	0,002
Loket / paže	AI 1	0,900 (0,596 – 0,982)	—	0,914 (0,776 – 0,970)	—
	AI 2	1,000 (0,722 – 1,000)	0,303	0,914 (0,776 – 0,970)	1,000
	RAD 1	1,000 (0,722 – 1,000)	0,303	0,914 (0,776 – 0,970)	1,000
	RAD 2	0,900 (0,596 – 0,982)	1,000	0,829 (0,673 – 0,919)	0,642
	RAD 3	1,000 (0,722 – 1,000)	0,305	0,943 (0,814 – 0,984)	0,643
Rameno / klíční kost	AI 1	1,000 (0,867 – 1,000)	—	0,936 (0,859 – 0,972)	—
	AI 2	0,840 (0,653 – 0,936)	0,037	0,962 (0,893 – 0,987)	0,468
	RAD 1	0,960 (0,805 – 0,993)	0,312	0,974 (0,911 – 0,993)	0,246
	RAD 2	1,000 (0,867 – 1,000)	1,000	0,974 (0,911 – 0,993)	0,246
	RAD 3	1,000 (0,867 – 1,000)	1,000	0,974 (0,911 – 0,993)	0,246
Pánev / kyčelní kloub	AI 1	0,889 (0,565 – 0,980)	—	0,885 (0,795 – 0,938)	—
	AI 2	1,000 (0,701 – 1,000)	0,303	0,987 (0,931 – 0,998)	0,009
	RAD 1	1,000 (0,701 – 1,000)	0,303	0,936 (0,859 – 0,972)	0,262
	RAD 2	1,000 (0,701 – 1,000)	0,303	0,987 (0,931 – 0,998)	0,009
	RAD 3	1,000 (0,701 – 1,000)	0,303	0,987 (0,931 – 0,998)	0,009

Tabulka 8 – Výsledky výkonnostního testu (Se, Sp) dle anatomické oblasti



Graf 2 – Forest plot, výsledky výkonnostního testu (Se, Sp, PV)

Kategorie	Plán (min.)	Skutečnost	Rezerva	Kritérium splněno?
FRACTURE (pozitivní)	≥ 110	118	+8	ANO
NORMAL (negativní)	≥ 900	1 019	+119	ANO
Celkem	≥ 1 050	1 137	+87	ANO

Tabulka 9 – Výsledky power analýzy

3.4 Hodnocení robustnosti referenčního standardu

Na 1 137 vyhodnocených snímcích byla vypočtena Fleissova kappa = 0,739, což odpovídá „substanční“ úrovni shody. Interval spolehlivosti získaný bootstrapem (2 000 vzorků) měl dolní mez 0,687, což potvrzuje stabilitu konsenzu mezi radiology.

4 Diskuze

Výsledky retrospektivní, monocentrické MRMC studie ukazují, že oba hodnocené systémy umělé inteligence (AI 1 a AI2) dosahují na prvotně diagnostických rentgenových snímcích pohybového aparátu vysoké úrovně senzitivity i specifity. Oba modely splnily předem stanovená kritéria non-inferiority (Se a Sp ≥ 0,90 s dolní mezí 95% CI nad 0,85), čímž prokázaly diagnostickou přijatelnost pro klinické použití.

Při srovnání s třemi klinickými radiology se ukázalo, že senzitivita AI 1 i AI 2 byla srovnatelná s výkony lékařů, zatímco ve specifitě se projevila rozdílnost mezi algoritmy – AI 2 dosáhl hodnot blízkých radiologům, zatímco AI 1 vykazoval mírně nižší specifitu. Tento rozdíl se projevil i v prediktivních hodnotách – negativní prediktivní hodnota přesahovala 0,99 u všech metod, což znamená minimální riziko přehlédnutí fraktury. Naopak pozitivní prediktivní hodnota byla u AI 1 a AI 2 nižší než u dvou radiologů s vyšší mírou zkušeností, což odráží častější výskyt falešně pozitivních nálezů u algoritmů.

Subanalýzy ukázaly relativně stabilní výkonnost obou systémů napříč šesti anatomickými oblastmi i čtyřmi věkovými kohortami, přestože prevalence fraktur se mezi jednotlivými

Anatomická oblast	Požadavek	Skutečnost	Kritérium splněno?
Chodidlo / kotník	≥ 10	19	ANO
Koleno / dolní končetina	≥ 10	11	ANO
Ruka / zápěstí	≥ 10	44	ANO
Loket / paže	≥ 10	10	ANO
Rameno / klíční kost	≥ 10	25	ANO
Pánev / kyčelní kloub	≥ 10	9	NE (hraničně)

Tabulka 10 – Výsledky power analýzy dle anatomické oblasti

Metrika	Bodový odhad	Dolní 95 % CI	Hranice NI (0,85)	Kritérium splněno?
Senzitivita (Se)	0,941	0,883	0,85	ANO
Specifita (Sp)	0,930	0,913	0,85	ANO

Tabulka 11 – Vyhodnocení power analýzy

Ukazatel	Odhad	95 % CI
Fleissova κ	0,739	0,687 – 0,790
GLMM	log-OR = 0,612	0,337 – 0,887

Tabulka 12 – Výsledky Fleissova κ a GLMM

lokalitami lišila. Typické bylo bimodální rozložení výskytu fraktur se zvýšeným podílem u pediatrické a geriatrické populace. Nejvyšší senzitivitu oba algoritmy dosáhly v oblasti ramenního pletence, zatímco oblast ruky a zápěstí byla spojena se zvýšeným výskytem falešně pozitivních nálezů, zejména u AI 1.

Studie využila rozsáhlý soubor 1 137 snímků, který překročil požadavky power analýzy a zahrnoval dostatečnou věkovou i technickou variabilitu pro základní zobecnitelnost výsledků. Přibližně polovina snímků pocházela z přístroje Samsung GC85A a další významná část z detektorů Agfa, což dokládá technickou heterogenitu souboru

Limitace

Interpretaci výsledků je nutné chápat v kontextu omezení. Především jde o monocentrický, retrospektivní design, který – i přes široké věkové spektrum a různé typy zobrazovacích systémů – může podléhat lokálním specifickým a ovlivnit míru falešně pozitivních i falešně negativních nálezů. U obou systémů se falešně pozitivní nálezy koncentrovaly zejména v oblasti ruky a zápěstí, což poukazuje na obtížnost interpretace této anatomické lokality vzhledem k projekční variabilitě a jemné anatomické struktuře.

Dalším omezením je skutečnost, že hodnocení bylo založeno výhradně na binární klasifikaci na úrovni snímku. Nebyla zkoumána přesnost lokalizace zlomenin ani časová efektivita čtení s využitím AI. Tyto aspekty zůstávají předmětem dalších analýz.

Navzdory uvedeným limitům poskytují výsledky této studie robustní důkaz, že systémy AI 1 a AI 2 dosahují výkonnosti srovnatelné s radiology při detekci fraktur na standardních rentgenových snímcích, a že jejich výkon zůstává stabilní napříč různými podskupinami dat.

5 Závěr

Oba testované systémy umělé inteligence (AI 1 i AI 2) prokázaly na rozsáhlém retrospektivním souboru rentgenových snímků schopnost detekovat akutní fraktury s vysokou přesností. Jejich senzitivita při zachytu zlomenin dosahovala ~93 % a byla srovnatelná s výkony mladých radiologů – AI tedy dokázala odhalit obdobný podíl fraktur jako lékař. Specifita algoritmů se pohybovala kolem 93–94 %. Zejména AI 2 se v této metrice blížil výkonnosti radiologů, zatímco AI 1 generoval o něco více falešných nálezů.

Velmi vysoká negativní prediktivní hodnota (> 99 %) u obou AI i radiologů navíc naznačuje, že negativní nálezy lze považovat za téměř jisté (minimální riziko přehlédnutí úrazu). Výsledky studie tedy ukazují, že oba hodnocené AI systémy mohou v klinické praxi spolehlivě asistovat při diagnostice skeletálních traumat – dosahují úrovně senzitivity srovnatelné s lidskými čtenáři a specifity pouze o málo nižší. Tyto poznatky podporují další integraci AI do akutní muskuloskeletální radiologie s cílem snížit počet přehlédnutých zlomenin při zachování přijatelného množství falešných poplachů.

COMPARISON OF DIAGNOSTIC ACCURACY BETWEEN ARTIFICIAL INTELLIGENCE SYSTEMS AND JUNIOR RADIOLOGISTS IN FRACTURE DETECTION: A RETROSPECTIVE, MULTI-READER STUDY

Abstract

A retrospective, single-center, multi-reader, multi-case study evaluated the diagnostic performance of two artificial intelligence systems (AI 1, AI 2) in detecting primary fractures on musculoskeletal X-ray images and compared it with the performance of three clinically active radiologists (RAD 1–3). A total of 1,137 anonymized images taken between March 5 and 11, 2025, were selected from the PACS archive of the Olomouc University Hospital; the prevalence of fractures was 10.4%. The reference standard was established by majority vote consensus of three independent radiologists, with substantial agreement confirmed by Fleiss' $\kappa = 0.739$. The power analysis required a minimum of 110 fractures and 900 negative cases; the actual sample comfortably exceeded these limits. The statistical evaluation was based on four-field contingency tables with Wilson's 95% confidence intervals (CI); differences between readers were tested using McNemar's test with Holm's correction. AI 1 achieved a sensitivity of 0.941 (95% CI 0.883–0.971) and a specificity of 0.930 (0.913–0.944), AI 2 achieved a sensitivity of 0.932 (0.872–0.965) and specificity of 0.935 (0.918–0.949). The sensitivity of both algorithms did not differ statistically ($p = 0.79$); However, AI 2 showed higher specificity than AI 1 ($p < 0.05$). Radiologists achieved sensitivity of 0.915–0.941 and specificity of 0.944–0.977. Their sensitivity was comparable to both AIs, while their specificity exceeded AI 1 and was similar to or slightly higher than AI 2. The negative predictive value of all evaluated entities exceeded 0.990. Subanalyses of six anatomical regions and four age cohorts confirmed the consistently high sensitivity of both AIs, with the highest accuracy in the shoulder/clavicle region and the greatest decrease in specificity of AI 1 in the hand/wrist region. The study shows that both systems provide fracture detection with sensitivity comparable to that of the radiologists involved.

Keywords

artificial intelligence, fractures, diagnostic accuracy, musculoskeletal radiography, multi-reader study

Literatura

- [1.] Berlin, L. (2007). Radiologic errors and malpractice: a blurry distinction. *American Journal of Roentgenology*, 189(3), 517–522.
- [2.] Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., & Reede, D. (2017). Interpretive error in radiology. *American Journal of Roentgenology*, 208(4), 739–749.
- [3.] Kuo, R. Y., Harrison, C., Curran, T. A., Jones, B., Freethy, A., Cussons, D., ... & Furniss, D. (2022). Artificial intelligence in fracture detection: a systematic review and meta-analysis. *Radiology*, 304(1), 50–62.
- [4.] Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3), 199–217.
- [5.] Thian, Y. L., Li, Y., Jagmohan, P., Sia, D., Chan, V. E. Y., & Tan, R. T. (2019). Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence*, 1(1), e180001.
- [6.] Shelmerdine, S. C., White, R. D., Liu, H., Arthurs, O. J., & Sebire, N. J. (2022). Artificial intelligence for radiological paediatric fracture assessment: a systematic review. *Insights into Imaging*, 13(1), 94.
- [7.] International Organization for Standardization. (2020). ISO 14155:2020 – Clinical investigation of medical devices for human subjects – Good clinical practice (3rd ed.). <https://www.iso.org/standard/71690.html>.

Kontakt

Mgr. Daniel Kvak
Carebot s. r. o.
Rašínovo nábřeží 71/10
128 00 Praha 2
+420 739 174 316
daniel.kvak@carebot.com
<https://www.carebot.com/>

MUDr. Jakub Dandár
Carebot s. r. o.
jakub.dandar@carebot.com

MUDr. Šimon Kubov
Fakultní Thomayerova nemocnice
kubov.simon@gmail.com