applied sciences

MDPI

Article

Deep Learning Enhances Radiologists' Detection of Potential Spinal Malignancies in CT Scans

Leonard Gilberg [1,*,†], Bianca Teodorescu [1,†], Leander Maerkisch [1], Andre Baumgart [2], Rishi Ramaesh [3], Elmer Jeto Gomes Ataide [1] and Ali Murat Koç [1,4]

1 Floy GmbH, 80335 Munich, Germany
2 UMM Universitätsmedizin Mannheim, 68305 Mannheim, Germany
3 University of Edinburgh, Midlothian EH25 9RG, UK
4 Department of Radiology, Ataturk Education and Research Hospital, Izmir Katip Celebi University, 35620 Izmir, Turkey
* Correspondence: leonard.gilberg@floy.com; Tel.: +49-15170826793
† These authors contributed equally to this work.

Featured Application: This work presents an AI-based second reader application tailored for computed tomography (CT) scans in Radiology. Its primary objective is to detect overlooked potential malignant cases in the vertebral body during routine radiological reporting.

Abstract: Incidental spinal bone lesions, potential indicators of malignancies, are frequently underreported in abdominal and thoracic CT imaging due to scan focus and diagnostic bias towards patient complaints. Here, we evaluate a deep-learning algorithm (DLA) designed to support radiologists' reporting of incidental lesions during routine clinical practice. The present study is structured into two phases: unaided and AI-assisted. A total of 32 scans from multiple radiology centers were selected randomly and independently annotated by two experts. The U-Net-like architecture-based DLA used for the AI-assisted phase showed a sensitivity of 75.0% in identifying potentially malignant spinal bone lesions. Six radiologists of varying experience levels participated in this observational study. During routine reporting, the DLA helped improve the radiologists' sensitivity by 20.8 percentage points. Notably, DLA-generated false-positive predictions did not significantly bias radiologists in their final diagnosis. These observations clearly indicate that using a suitable DLA improves the detection of otherwise missed potentially malignant spinal cases. Our results further emphasize the potential of artificial intelligence as a second reader in the clinical setting.

Keywords: deep learning; computed tomography; malignancies; AI detection; second reader; spine; vertebral lesions

Citation: Gilberg, L.; Teodorescu, B.; Maerkisch, L.; Baumgart, A.; Ramaesh, R.; Gomes Ataide, E.J.; Koç, A.M. Deep Learning Enhances Radiologists' Detection of Potential Spinal Malignancies in CT Scans. *Appl. Sci.* **2023**, *13*, 8140. https://doi.org/10.3390/app13148140

Academic Editor: Cosimo Nardi

Received: 22 June 2023
Revised: 10 July 2023
Accepted: 12 July 2023
Published: 13 July 2023

1. Introduction

Discrepancies in radiology are a long-known issue and—due to the extended workload and limited evaluation time—have remained the same despite continuously improving imaging techniques over the last decades [1,2]. A perceptual error, or false negative, is an abnormality present in a diagnostic image but not described by the interpreter. Such overlooked findings constitute the vast majority of human error in image interpretation [2–4].

Spinal bone lesions that present themselves as a conglomerate are frequently an indicator of malignancy, with the vertebrae being the most prevalent hotspot for bone metastasis [5]. On the other side of the spectrum, solitary lesions are more challenging and can indicate both malignant and benign processes [6,7], creating uncertainty within the diagnostic procedure, which may require further investigation [8]. In this case, if missed or initially overlooked (perceptual error), they can exhibit major negative consequences on a patient's quality of life and, subsequently, their morbidity and mortality [9,10].

With CT being a reliable imaging modality for assessing osseous involvement and the destruction degree of spine abnormalities [6,11], the past years have shown a rising interest in automatizing the detection and classification of spinal lesions to a large extent [12–14]. Artificial intelligence (AI) is now an active part of various medical diagnostic procedures within real-life hospital workflows, with deep-learning (DL)-based analysis of radiologic images as one of its key applications. AI as a screening tool or a second reader for abnormality detection already shows promising results in various fields, such as chest X-ray reporting and lung nodule detection [15,16]. However, reliable deep-learning-based algorithms for spinal lesion detection are still sparse as their development has proven to be more challenging, mostly due to the overlapping image features of degenerative and neoplastic events [6]. To our knowledge, there are currently no EU MDR-certified or FDA-approved AI second reader software that detect incidental spinal lesions in CT scans of unrelated indications. A reliable algorithm with such capacities could assist the reporting physician with accurate supplementary information, reduce the rate of missed potentially malignant lesions, and streamline the diagnostic pathway.

This work examines the clinical impact of a deep-learning algorithm (DLA) that assists radiologists in their day-to-day workflow within a simulated hospital setting. The algorithm was developed to detect potentially malignant cases within the vertebrae and act as a second reader, using native and contrast-enhanced abdomen and thoracic CT imaging sequences. Its clinical efficacy was evaluated in an observational cross-over study design, where the algorithm's performance and the effect on the decision-making of six subspecialty radiologists with and without the intervention of the DLA were assessed. The distinct feature of this study design is its emphasis on reducing incidental findings during the reporting of other main underlying diseases that the patient has. This approach, further accentuated by limiting the scope to only CT abdomen and thoracic scans, shifts the focus away from solely detecting vertebral malignancy as the primary objective of the responsible radiologist. Such "background acting" tools open new avenues in how one can correctly integrate AI in the medical sector and underline the crucial role of human involvement in the overall process.

2. Materials and Methods

This section details the materials and methods used for the study.

2.1. Data Acquisition

All clinical and imaging information was obtained retrospectively from multiple outpatient radiology centers in Germany. The data selection process is detailed in Figure 1.

We included studies of native or contrast-enhanced thoracic and abdominal CT examinations with multiplanar bone and soft tissue reconstructions. Incomplete or broken studies, individuals with prior spinal surgery, and individuals under 18 were excluded from the cohort. Once the data was filtered based on the chosen inclusion criteria, 32 randomly sampled studies were picked. Data contracts are signed with the data providers, and the studies were anonymized before being included in the study. Additionally, the data is retrospective, with CT scans from a multicenter data provider collected over 12 months from January 2022 until January 2023. Due to these factors, the need for informed consent is waived. The anonymization process strips away all identifying tags such as name, contact details, and address. Demographic details such as sex are preserved, and age is rounded to the nearest whole number.

2.2. Establishment of the Ground Truth

Images were pre-processed and stored in a Digital Imaging and Communications in Medicine (DICOM) format before the expert annotation. The ground truth labels were then established via manual segmentation by two board-certified radiologists with expertise in the field (MK: Associate Professor, 14 years of experience; and RR: Senior Lecturer, ten years of experience). The studies were annotated on an object level by drawing bounding boxes

around all regions of interest (ROIs) and subsequently classified as positive or negative. We consider a positive case in which at least one finding is indicative of a potentially malignant vertebral lesion (lytic, sclerotic, or mixed with a circumscribed boundary) and has been manually segmented and evaluated by our two experts. This labeling process was performed on the Encord platform (© 2022 Encord), and a consensus was reached in case of divergent opinions.

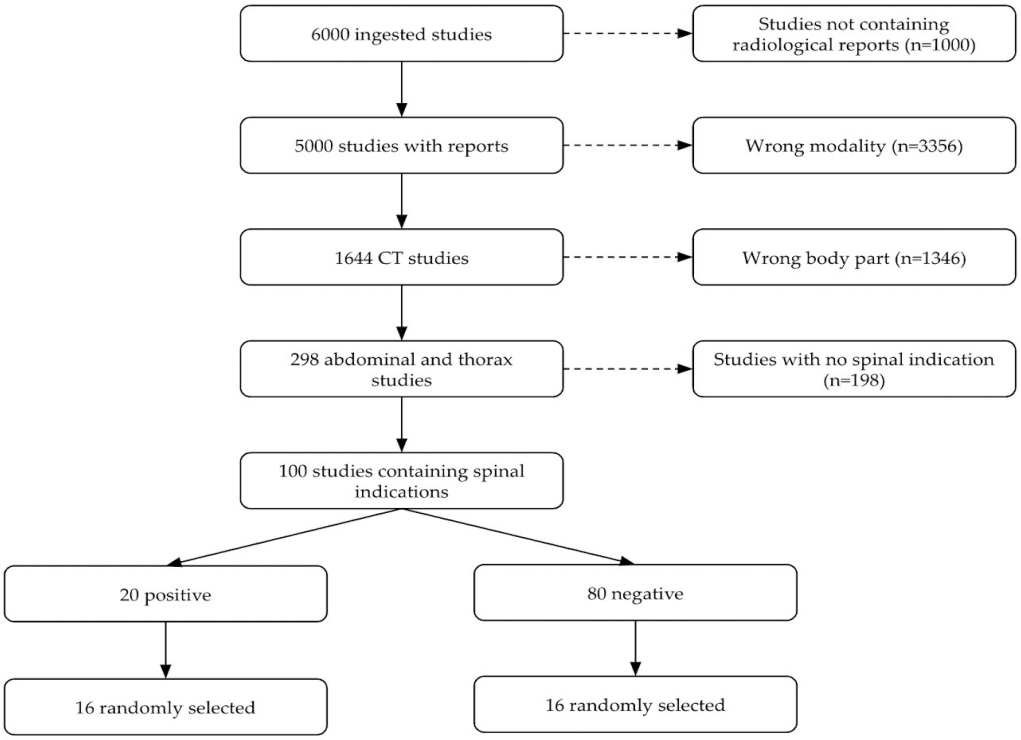

Figure 1. Flowchart of data collection and distribution. All CT studies were collected retrospectively from the clinical databases of multiple institutions.

2.3. AI Algorithm Development

The deep-learning algorithm used in this study was developed using native and contrast-enhanced CT studies of the spine, abdomen, and thorax. It is meant to be used as a medical device in a clinical setting. The DLA consists of two deep-learning models that work together. One model uses a U-Net-like segmentation approach coupled with volumetric analysis to determine the presence of potentially malignant lesions in the spinal vertebrae. The second model employs a vertebral localization component that enables the proper selection of the region of interest. It is trained using a training set of 224 cases and tested on 735 cases. Additionally, the DLA is evaluated on an external dataset of 420 cases. Statistical analysis of the datasets was performed to verify the demographic distribution of the data.

2.4. Experimental Setup

In this study, our primary goal was to investigate the effectiveness of a deep-learning algorithm (DLA) in assisting radiologists in identifying previously overlooked potentially malignant cases in the spine through abdomen and thoracic CT imaging. A total of 32 studies were selected from our data pool which contained both positive and negative cases in equal numbers. Details of the selection process are given in Figure 1.

The study consisted of two phases. In the first phase, six radiologists independently reviewed the 32 studies without DLA assistance, following their routine reading process. In this case, a routine reading process is defined as the reading of the radiological images by

the radiologist based on the patient's complaint. This phase aimed to establish a baseline for their diagnostic performance.

The six radiologists (participants) we recruited have varying experience levels, ranging from 1 to 11 years. Each participant received the same set of 32 studies but presented in random order. The participants were blinded to the gold standard and any patient-specific information to ensure unbiased assessments, except for a brief description of the patient's complaint. These complaints were not focused on spine-related issues but were general complaints of the patients who visited the clinics. The details of this information can be found in Supplementary Table S1.

After a break of 10 days, the second phase was conducted. In this phase, the participants reevaluated the same set of studies, but this time, they had access to the DLA predictions to assist them in their evaluations. The DLA provided predictions about the presence or absence of abnormalities in the spinal vertebrae in the CT scans.

To ensure accurate documentation of their assessments, the participants were asked to use screen-recording software [17] to capture their computer screens. They were also instructed to describe their findings verbally while indicating them with their mouse. A DICOM visualization tool [18] was made available to aid in interpreting the scans and the predictions of the DLA.

To evaluate the effectiveness of the AI intervention, the results obtained during the reading sessions were manually compared to the gold standard. This comparison enabled the determination of the accuracy of the participants' diagnoses with and without the DLA's assistance.

2.5. Statistical Analysis

Statistical analysis was performed using GraphPad Prism v8.4.2. The primary measured outcome was based on assessing case and object level (per-patient and per-lesion) sensitivity, specificity, and average false positive (AvgFP). Because of the infinite number of possible locations for a spinal lesion, we could not define the true negative and thus did not calculate the per-lesion specificity. To assess the significance between the two reading sessions, we conducted the McNemar test in Python (statsmodels v 0.14.0).

3. Results

3.1. Demographics of the Dataset and Spinal Lesion Assessment in the Reference Standard

No significant differences or inhomogeneities were noted concerning the demographic qualities of our cohort. Supplementary Table S1 shows the background information of the 32 patients involved in this study, including patient demographics (sex, age). 53.1% of the patients are female. The average age was noted to be 56.6 years. An overview of the scan conditions and other image acquisition details can be found in Supplementary Table S2. As per the gold standard, there were 16 positive and 16 negative scans. There were a total of 27 annotated suspicious lesions. The annotators exhibited a mutual agreement for 75.0% of the cases while determining the gold standard. Both annotators initially disagreed on the remaining 25.0%, but these conflicts were resolved through discussions leading to a consensus.

3.2. Algorithm Performance

Figure 2 exemplifies two true-positive predictions on different vertebral sites for lytic (Figure 2a,b) and sclerotic (Figure 2c,d) lesions. The deep-learning algorithm (DLA) for spinal lesion detection was tested on the same patient studies and correctly detected 12 out of 27 lesions in 16 patients. It also falsely indicated 13 spinal findings (false positives) that were not considered true findings following the gold standard. The overall outcome of the DLA for the established dataset is shown in Table 1. On a case level, compared to the gold standard, the DLA performance had a sensitivity and specificity of 75.0% and 56.3%, respectively. Regarding the results on an object level, the sensitivity was 44.4%, as shown in Table 2.

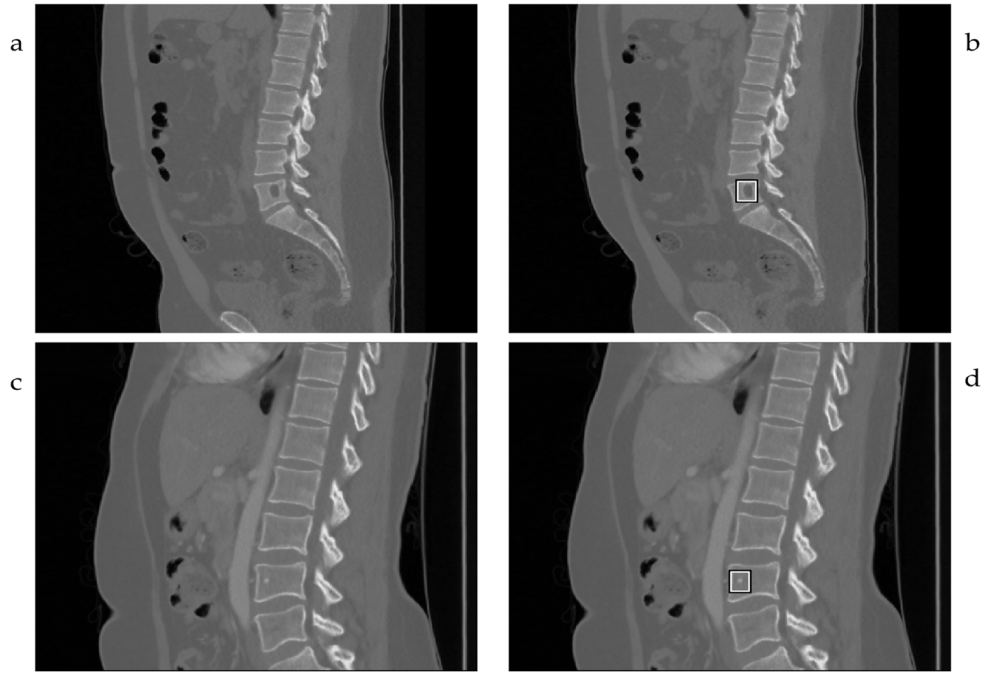

Figure 2. DLA predictions of osteolytic (**a**) and oseoblastic (**c**) lesions are shown in corresponding images (**b**,**d**).

Table 1. Algorithm performance in comparison to the gold standard (count).

	Positive Cases	Total Number of Detected Objects	Number of Detected Spinal Lesions (TP)	Number of Undetected Spinal Lesions (FN)	Number of Falsely Detected Spinal Lesions (FP)
Gold Standard	16	27	27	N/A	N/A
DLA	11	40	12	15	13

Table 2. Algorithm performance in comparison to the gold standard (metrics).

Sensitivity (TP Rate)		Specificity (TN Rate)		Accuracy	
Case Level	**Object Level**	**Case Level**	**Object Level**	**Case Level**	**Object Level**
75.00%	44.44%	56.25%	N/A	65.63%	N/A

3.3. Intra-Observer Agreement without and with the Aid of the DLA

The observers' performance results are summarized in Table 3, with visual exemplification in Figure 3 showcasing sensitivity, specificity, and the true-positive rate on a case level.

Considering that the participants were asked to perform routine reporting based on general complaints (such as acute abdominal pain or elevated liver enzymes), only one out of six radiologists did not include spinal findings in the first round (without DLA). Another radiologist solely identified degenerative changes without detecting any significant or suspicious findings that aligned with the ground truth. The participant with the highest experience level (11 years) initially reported two true positives in the first round. However, when aided by the DLA, this number increased to 14 true-positive and two false-positive findings. The two least experienced radiologists (one year) did not report any spinal findings without relying on the algorithm. Following the predictions in round 2, only one included spinal findings in the report.

Considering a potential maximum of 162 true positives (27 lesions multiplied by 6 participants), the radiologists reported 11 true-positive objects alone. However, when assisted by the DLA, this number increased to more than three times that value, with 35 true-positive findings in the study's second phase.

The DLA predicted 13 false-positive objects. When interpreting these predictions, on average, the radiologists included less than one false positive in the report, resulting in four false-positive reports across all participants.

Table 3. Intra-observer agreement depicted for the two study phases according to the gold standard (phase 1—without the DLA and phase 2—with the DLA support).

	Participants						
	1	**2**	**3**	**4**	**5**	**6**	
Experience (years)	6	7	5	1	11	1	
	Phase 1—no support from DLA						**Mean**
Suspicious lesions reported (n)	2.00	4.00	3.00	0.00	2.00	0.00	1.83
False positives (n)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sensitivity (case level)	12.50%	25.00%	18.75%	0.00%	12.50%	0.00%	11.46%
Sensitivity (object level)	7.41%	14.81%	11.11%	0.00%	7.41%	0.00%	6.79%
Specificity (case level)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Accuracy (case level)	56.25%	62.50%	59.38%	50.00%	56.25%	50.00%	55.73%
Accuracy (object level)	53.70%	57.41%	55.56%	50.00%	53.70%	50.00%	53.40%
False positive rate	0.00%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
	Phase 2—with support from DLA						**Mean**
Suspicious lesions reported (n)	5.00	5.00	8.00	0.00	14.00	3.00	5.83
False positives (n)	1.00	0.00	1.00	0.00	2.00	1.00	0.83
Sensitivity (case level)	31.25%	31.25%	43.75%	0.00%	75.00%	12.50%	32.29%
Sensitivity (object level)	14.81%	18.52%	25.93%	0.00%	44.44%	7.41%	18.52%
Specificity (case level)	93.75%	100.00%	93.75%	100.00%	87.50%	93.75%	94.79%
Accuracy (case level)	62.25%	65.62%	68.75%	50.00%	81.25%	53.12%	63.54%
Accuracy (object level)	55.56%	59.26%	68.12%	50.00%	68.52%	51.85%	58.89%
False-positive rate	3.70%	0.00%	3.57%	0.00%	7.41%	3.70%	3.06%

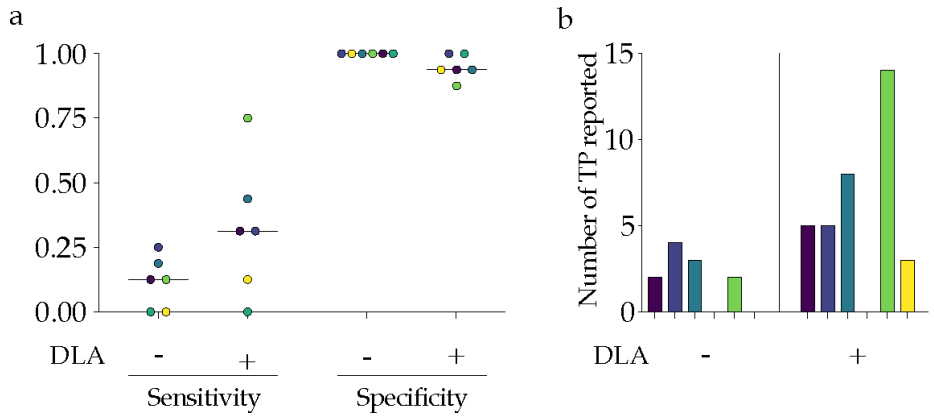

Figure 3. Performance metrics of the six participants on case level, with each participant color coordinated; (**a**) Sensitivity and specificity of reporting potentially malignant spinal lesions without (−) and with (+) the aid of an AI tool as a second reader; (**b**) Number of reported true-positive findings without (−) and with (+) the aid of an AI tool as a second reader.

Based on the AI's indication of true-positive spinal conditions, four participants felt requesting further follow-up or diagnostic tests was necessary. These requests were made to investigate potential abnormalities or conditions related to the spine, as suggested by the AI. None of the participants requested additional examinations due to possible FP findings.

On a case level, participants' average sensitivity in detecting spinal lesion(s) (which translates into reporting at least one correct lesion in a positive case) increased from 11.5% to 32.3% (20.08 percentage points) when using the DLA tool. The sensitivity increased from 6.8% to 18.5% (11.70 percentage points) on an object level. The average FP rate increased from 0.0% (no primarily reported false spinal findings) to 3.1% on a case and object level. The mean accuracy value increased from 55.7% to 63.5% on a case level and from 53.4% to 58.9% on an object level.

There was no clear trend that could link the participants' clinical experience and their responsiveness to the AI predictions, with both junior and more experienced radiologists having heterogeneous behaviors toward the presented algorithm results. It should be noted that, in less-experienced participants, we observed changes in their reports between reading sessions unrelated to the AI findings.

4. Discussion

A complete and comprehensive review of a CT scan is crucial for a patient's health and has a significant impact on decision-making. CT imaging can assess spinal bone metastatic lesions up to 6 months before plain radiographs [19]. However, smaller lesions or those that do not have significant cortical destruction are often underreported or missed during CT image reporting [20]. In a systematic review, Bartalena et al. reported that radiologists' recognition of incidental vertebral findings (in this case, fractures) was low, with a mean reporting rate of just 27.4% [21]. A major fraction of false negatives are significant bone lesions that could indicate potential malignancy [3]. J Donald et al. showed in an internal department analysis that these types of spinal lesions were most frequently misinterpreted on CT images, with 14 out of 16 missed findings being metastatic [22].

False-negative cases are the most common perceptual errors [4], with CT imaging being especially susceptible [22]. Errors made in previous radiology reports can lead to the tendency of radiologists to replicate the error in subsequent reports, which is referred to as 'alliterative bias' [23]. This concept reiterates the importance of some confirmation protocols in medical practice. While double-reading practices significantly impact the quality of radiological reports, clinical workload and staff shortages make a routine human double-reader scheme hard to implement [24,25].

Computer-aided detection (CAD) systems have supported radiologists in their workflow even before the era of deep-learning tools, with some of the best examples being

small functions and add-ons for DICOM viewers, such as contrast enhancers or manual annotation support. With the use of AI increasing rapidly in the fields of medicine, with radiology as a leading candidate, future deep-learning-based CAD systems will not only optimize the users' workflow and improve their diagnostic abilities but also weigh in on their medical judgment and decision-making process. CTs are not just a series of images; they contain extensive information about the pathology in question that sometimes cannot be interpreted by the bare human eye [26]. There are models developed for almost every disease that can be assessed radiologically. More specifically, detection systems have shown an emerging potential in reducing missed radiological spinal lesions as a second reader [13,27–30].

In our study, we investigated the effects of a DLA when implemented in a routine CT reporting process performed by six radiologists having different experience levels. Our results show that DL-based spinal lesion detection can improve inter-observer agreement and overall increase performance in detecting these radiological findings, regardless of the training level of our participants. The case level sensitivity increased by 20.83 percentage points when the participants were aided by the DLA. However, this improvement should only be interpreted in the context of a rather low baseline of reported spinal findings in the first round. Another observational study by Noguchi et al. [13] showed that the sensitivity of radiologists in detecting bone metastases could be elevated with the help of a DLA by 15.3 percentage points. In a similar study, Kato et al. [31] reported an improvement in the performance of less-experienced radiologists in brain metastasis detection by 4.90 percentage points. Like our own study, Kato et al. observed no significant increase in false-positive findings when utilizing a CAD system. While these previous studies have primarily focused on DLAs improving performance in explicit detection tasks, our study provides a novel perspective by highlighting the potential benefits of a tool for reducing incidental findings during routine reporting. This observation helps to explain the relatively lower baseline performance of participants in our study. Since a complete radiological report includes information regarding all body parts that can be seen on the scan, we expected insights on all findings that the six radiologists encountered while analyzing the images. Our findings support the already existing issue of missed spinal findings in clinical practice, which might rely on the aforementioned reasons for perceptual errors in radiology. It is worth noting that the sensitivities and specificities are calculated only for potentially malignant spinal lesion findings and not findings for all other organs (lungs, liver, kidneys). Hence, the lowered sensitivities and specificities due to many spinal lesions being missed during the initial diagnostic process without the assistance of the DLA.

Indeed, the DLA reported false-positive findings. It is important to mention that the radiologists' assessments were not solely based on the AI's predictions for false-positive lesions. They considered various factors, including other findings identified as true positives by the AI. Furthermore, based on the AI's indication of true-positive spinal conditions, four participants requested further follow-up or diagnostic tests. These requests were made to investigate potential abnormalities or conditions related to the spine, as suggested by the AI.

Several studies have already investigated the potential of deep-learning systems for pathology assessment of the spine [13,28,32–35]. Although deep-learning algorithms (DLAs) can demonstrate accuracy in lesion detection on par with radiologists, it is crucial to consider their real-world implementation in datasets that differ significantly from the training data. DLAs may face challenges in such scenarios and are more prone to producing incorrect results. However, by addressing these challenges through ongoing research and fine-tuning, we can further optimize the effectiveness of DLAs in practical settings. It is, therefore, crucial to focus on the effects and performance of the human reader when in conjunction with this emerging technology. While our DLA's performance may not have demonstrated superiority over expert human radiologists, there is potential for radiologists of all experience levels to benefit from its second reader function. It is essential to conduct future studies with larger cohorts and greater sample sizes to validate any hypothetical

benefits and potential risks associated with AI. These studies will ensure the safe and effective integration of machine learning software into the clinical setting.

This study has several limitations. Since the allocated time between the two reading sessions was set for only ten days, one could argue that the first reading could have biased the participants' performance in the second session. We conducted the McNemar test to assess significance. The test ($p = 0.25$) confirmed that the study did not yield statistically significant results ($p > 0.05$). This outcome is linked to the small number of radiologist participants used for this study and probably could not prove generalizable effects. However, it is important to note that this finding presents an opportunity for improvement in future studies. The number of patient studies investigated by the participants was also small and should be increased for further studies. Moreover, the training that radiologists receive, their reporting style, and thus their attention to detail differ between countries and subspecialties.

5. Conclusions

We show that the implementation of a DLA as a second reader in routine reporting of CT scans can increase radiologists' true-positive rate for spinal lesion detection while at the same time having close to no impact on the false-positive rate. These findings showcase the potential that an AI-based technology could have in the hospital setting, particularly in detecting missed potential malignancies during routine reporting. However, it is important to consider that the impact of AI in medical imaging may vary depending on the interpreter's background and training. While current trends and discoveries indicate improvements in diagnostic performance, further comprehensive studies are needed to validate these results on a larger scale and gain a deeper understanding of the implications this technology may have in the clinical setting. Nevertheless, the enhanced metrics observed in this study provide evidence of AI's prospect as a valuable detection tool.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app13148140/s1, Table S1: Demographics, Complaints and Lesion Characteristics. Table S2: Scanner and Technical Specifications.

Author Contributions: L.G., B.T., A.M.K., L.M. and E.J.G.A. formulated the study concept, registered the review, contributed to all stages of the review process, performed the literature search, data extraction, and quality assessment, wrote the first draft of the manuscript, and acted as first reviewers. L.G., B.T. and E.J.G.A. analyzed the data and wrote the first draft of the manuscript. L.M., A.B. and A.M.K. oversaw the review conceptualization, registration, and execution and acted as second reviewers. A.M.K., E.J.G.A., R.R. and A.B. acted as third reviewers for discrepancies and revisions of the manuscript. All authors contributed equally in preparing the final version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Floy GmbH. Floy GmbH played no role in the study design, data collection or analysis, decision to publish, or manuscript preparation.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the usage of anonymized retrospective patient data.

Informed Consent Statement: Data contracts were signed by the data providers, and the studies were anonymized before being included in the study. Additionally, the data is retrospective. Due to these factors, the need for informed consent was waived.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Conflicts of Interest: L.G., B.T. and A.M.K. are Clinical Scientists at Floy GmbH. EJGA is the Head of Clinical Research at Floy GmbH. A.B. is a Regulatory Consultant specializing in medical technology translation activities. L.M. is the Co-Founder of Floy GmbH. Correspondence and requests for materials should be addressed to L.G.

References

1. Waite, S.; Scott, J.; Gale, B.; Fuchs, T.; Kolla, S.; Reede, D. Interpretive Error in Radiology. *AJR Am. J. Roentgenol.* **2017**, *208*, 739–749. [CrossRef]
2. Fitzgerald, R. Error in radiology. *Clin. Radiol.* **2001**, *56*, 938–946. [CrossRef]
3. McCreadie, G.; Oliver, T.B. Eight CT lessons that we learned the hard way: An analysis of current patterns of radiological error and discrepancy with particular emphasis on CT. *Clin. Radiol.* **2009**, *64*, 491–499; discussion 500-1. [CrossRef]
4. Terreblanche, O.D.; Andronikou, S.; Hlabangana, L.T.; Brown, T.; Boshoff, P.E. Should registrars be reporting after-hours CT scans? A calculation of error rate and the influencing factors in South Africa. *Acta Radiol.* **2012**, *53*, 61–68. [CrossRef] [PubMed]
5. Ziu, E.; Viswanathan, V.K.; Mesfin, F.B. *Spinal Metastasis*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
6. Rodallec, M.H.; Feydy, A.; Larousserie, F.; Anract, P.; Campagna, R.; Babinet, A.; Zins, M.; Drapé, J.-L. Diagnostic Imaging of Solitary Tumors of the Spine: What to Do and Say. *Radiographics* **2008**, *28*, 1019–1041. [CrossRef]
7. Kim, Y.S.; Han, I.H.; Lee, I.S.; Lee, J.S.; Choi, B.K. Imaging findings of solitary spinal bony lesions and the differential diagnosis of benign and malignant lesions. *J. Korean Neurosurg. Soc.* **2012**, *52*, 126–132. [CrossRef]
8. Nguyen, T.T.; Thelen, J.C.; Bhatt, A.A. Bone up on spinal osseous lesions: A case review series. *Insights Imaging* **2020**, *11*, 80. [CrossRef] [PubMed]
9. Coleman, R.E. Metastatic bone disease: Clinical features, pathophysiology and treatment strategies. *Cancer Treat. Rev.* **2001**, *27*, 165–176. [CrossRef] [PubMed]
10. Martin, M.; Bell, R.; Bourgeois, H.; Brufsky, A.; Diel, I.; Eniu, A.; Fallowfield, L.; Fujiwara, Y.; Jassem, J.; Paterson, A.H.; et al. Bone-related complications and quality of life in advanced breast cancer: Results from a randomized phase III trial of denosumab versus zoledronic acid. *Clin. Cancer Res.* **2012**, *18*, 4841–4849. [CrossRef]
11. Jarvik, J.G.; Deyo, R.A. Diagnostic evaluation of low back pain with emphasis on imaging. *Ann. Intern. Med.* **2002**, *137*, 586–597. [CrossRef]
12. Wang, J.; Fang, Z.; Lang, N.; Yuan, H.; Su, M.Y.; Baldi, P. A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. *Comput. Biol. Med.* **2017**, *84*, 137–146. [CrossRef] [PubMed]
13. Noguchi, S.; Nishio, M.; Sakamoto, R.; Yakami, M.; Fujimoto, K.; Emoto, Y.; Kubo, T.; Iizuka, Y.; Nakagomi, K.; Miyasa, K.; et al. Deep learning–based algorithm improved radiologists' performance in bone metastases detection on CT. *Eur. Radiol.* 2022. [CrossRef] [PubMed]
14. Li, Z.; Wu, F.; Hong, F.; Gai, X.; Cao, W.; Zhang, Z.; Yang, T.; Wang, J.; Gao, S.; Peng, C. Computer-Aided Diagnosis of Spinal Tuberculosis From CT Images Based on Deep Learning with Multimodal Feature Fusion. *Front. Microbiol.* **2022**, *13*, 823324. [CrossRef] [PubMed]
15. Liang, M.; Tang, W.; Xu, D.M.; Jirapatnakul, A.C.; Reeves, A.P.; Henschke, C.I.; Yankelevitz, D. Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers. *Radiology.* **2016**, *281*, 279–288. [CrossRef]
16. McKinney, S.M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G.S.; Darzi, A.; et al. International evaluation of an AI system for breast cancer screening. *Nature* **2020**, *577*, 89–94. [CrossRef]
17. Loom. Available online: https://www.loom.com/ (accessed on 21 June 2023).
18. Visian. Available online: https://visian.org/ (accessed on 21 June 2023).
19. Salvo, N.; Christakis, M.; Rubenstein, J.; de Sa, E.; Napolskikh, J.; Sinclair, E.; Ford, M.; Goh, P.; Chow, E. The Role of Plain Radiographs in Management of Bone Metastases. *J. Palliat Med.* **2009**, *12*, 195–198. [CrossRef]
20. Williams, A.L.; Al-Busaidi, A.; Sparrow, P.J.; Adams, J.E.; Whitehouse, R.W. Under-reporting of osteoporotic vertebral fractures on computed tomography. *Eur. J. Radiol.* **2009**, *69*, 179–183. [CrossRef]
21. Bartalena, T.; Rinaldi, M.F.; Modolon, C.; Braccaioli, L.; Sverzellati, N.; Rossi, G.; Rimondi, E.; Busacca, M.; Albisinni, U.; Resnick, D. Incidental vertebral compression fractures in imaging studies: Lessons not learned by radiologists. *World J. Radiol.* **2010**, *2*, 399–404. [CrossRef]
22. Donald, J.J.; Barnard, S.A. Common patterns in 558 diagnostic radiology errors. *J. Med. Imaging Radiat. Oncol.* **2012**, *56*, 173–178. [CrossRef]
23. Smith, M.J. *Error and Variation in Diagnostic Radiology*; C.C. Thomas: Springfield, IL, USA, 1967.
24. Markus, J.B.; Somers, S.; O'Malley, B.P.; Stevenson, G.W. Double-contrast barium enema studies: Effect of multiple reading on perception error. *Radiology* **1990**, *175*, 155–156. [CrossRef] [PubMed]
25. Wakeley, C.J.; Jones, A.M.; Kabala, J.E.; Prince, D.; Goddard, P.R. Audit of the value of double reading magnetic resonance imaging films. *Br. J. Radiol.* **1995**, *68*, 358–360. [CrossRef] [PubMed]
26. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577. [CrossRef] [PubMed]
27. Masoudi, S.; Mehralivand, S.; Harmon, S.; Walker, S.; Pinto, P.A.; Wood, B.J.; Citrin, D.E.; Karzai, F.; Gulley, J.L.; Madan, R.A.; et al. Artificial intelligence assisted bone lesion detection and classification in computed tomography scans of prostate cancer patients. *J. Clin. Orthod.* **2020**, *38*, e17567. [CrossRef]
28. Li, Y.; Zhang, Y.; Zhang, E.; Chen, Y.; Wang, Q.; Liu, K.; Yu, H.J.; Yuan, H.; Lang, N.; Su, M.-Y. Differential diagnosis of benign and malignant vertebral fracture on CT using deep learning. *Eur. Radiol.* **2021**, *31*, 9612–9619. [CrossRef]
29. Han, S.; Oh, J.S.; Lee, J.J. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. *Eur. J. Nucl. Med. Mol. Imaging* **2022**, *49*, 585–595. [CrossRef]

30. Ohno, Y.; Aoyagi, K.; Takenaka, D.; Yoshikawa, T.; Ikezaki, A.; Fujisawa, Y.; Murayama, K.; Hattori, H.; Toyama, H. Machine learning for lung CT texture analysis: Improvement of inter-observer agreement for radiological finding classification in patients with pulmonary diseases. *Eur. J. Radiol.* **2021**, *134*, 109410. [CrossRef]
31. Kato, S.; Amemiya, S.; Takao, H.; Yamashita, H.; Sakamoto, N.; Miki, S.; Watanabe, Y.; Suzuki, F.; Fujimoto, K.; Mizuki, M.; et al. Computer-aided detection improves brain metastasis identification on non-enhanced CT in less experienced radiologists. *Acta Radiol.* **2023**, *64*, 1958–1965. [CrossRef]
32. Yoda, T.; Maki, S.; Furuya, T.; Yokota, H.; Matsumoto, K.; Takaoka, H.; Miyamoto, T.; Okimatsu, S.; Shiga, Y.; Inage, K.; et al. Automated Differentiation between Osteoporotic Vertebral Fracture and Malignant Vertebral Fracture on MRI Using a Deep Convolutional Neural Network. *Spine* **2022**, *47*, E347–E352. [CrossRef]
33. Yeh, L.-R.; Zhang, Y.; Chen, J.-H.; Liu, Y.-L.; Wang, A.-C.; Yang, J.-Y.; Yeh, W.-C.; Cheng, C.-S.; Chen, L.-K.; Su, M.-Y. A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: Retrospective training and validation of ResNet. *Eur. Spine J.* **2022**, *31*, 2022–2030. [CrossRef]
34. Ouyang, H.; Meng, F.; Liu, J.; Song, X.; Li, Y.; Yuan, Y.; Wang, C.; Lang, N.; Tian, S.; Yao, M.; et al. Evaluation of Deep Learning-Based Automated Detection of Primary Spine Tumors on MRI Using the Turing Test. *Front. Oncol.* **2022**, *12*, 814667. [CrossRef]
35. Liu, H.; Jiao, M.; Yuan, Y.; Ouyang, H.; Liu, J.; Li, Y.; Wang, C.; Lang, N.; Qian, Y.; Jiang, L.; et al. Benign and malignant diagnosis of spinal tumors based on deep learning and weighted fusion framework on MRI. *Insights Imaging* **2022**, *13*, 87. [CrossRef] [PubMed]