J BONE ONCOL 44 (2024) 100525

Contents lists available at ScienceDirect

Journal of Bone Oncology

journal homepage: www.elsevier.com/locate/jbo

Research Paper

Large-Scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications

Fan Yang [a], Dong Yan [a], Zhixiang Wang [b,*]

[a] Department of Radiation, Beijing Jishuitan Hospital,Capital Medical University, Beijing 100035, China
[b] Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China

HIGHLIGHTS

- ChatGPT shows potential in malignant bone tumor diagnosis with valuable advantages such as increasedefficiency and reduced missed diagnoses.
- ChatGPT struggled with complex cases due to ambiguous symptoms and the overlap of imaging features in bone lesions.
- Collaboration between physicians and ChatGPT is crucial in real-world settings.
- AI has potential in the diagnosis of malignant bone tumors and lays the foundation for the future development of medical AI.

ARTICLE INFO

Keywords:
ChatGPT
Bone tumors
Artificial intelligence
Diagnosis

ABSTRACT

Objective: This study was designed to delve into the complexities involved in diagnosing of benign and malignant bone tumors and to assess the potential of AI technologies like ChatGPT in improving diagnostic accuracy and efficiency. The study also explores the few-shot learning as a method to optimize ChatGPT's performance in specialized medical domains such as benign and malignant bone tumors diagnosis.

Methods: A total of 1366 benign and malignant bone tumors-related imaging reports were collected and diagnosed by 25 experienced physicians. The gold standard of diagnosis was established by combining clinical, imaging and pathological principles.These reports were then input into the ChatGPT model which underwent a few-shot learning method to generate diagnostic results. The diagnostic results of the physicians and the AI model were compared to evaluate the performance of ChatGPT. An experiment was conducted to assess the influence of different radiologist's reporting styles on the model's diagnostic performance. Furthermore, in-depth analysis of misdiagnosed cases was carried out, categorizing diagnostic errors and exploring possible causes.

Results: The diagnostic results generated by ChatGPT showed an accuracy of 0.73, sensitivity of 0.95, and specificity of 0.58. After few-shot learning, ChatGPT demonstrated significant improvement, achieving an accuracy of 0.87, sensitivity of 0.99, and specificity of 0.73, bringing it much closer to the level of physician diagnostics. In an experiment analyzing the influence of the radiologist's reporting style, the model demonstrated higher sensitivity when interpreting reports written by high-level radiologists. In 56 benign cases, ChatGPT misdiagnosed them as malignant. Among these, 35 benign lesions- fibrous dysplasia and osteofibrous dysplasia-were incorrectly identified as metastatic tumors or osteosarcomas; 8 cases of myositis ossificans were wrongly diagnosed as extraosseous osteosarcoma. 7 cases of giant cell tumor of bone at the end of long bone were misdiagnosed as osteosarcoma by intermediate doctors. Chondroblastoma was misdiagnosed as malignant tumor in 6 cases −2 osteosarcoma and 4 chondrosarcoma-In this study, 23 osteosarcoma cases were misdiagnosed by ChatGPT as osteomyelitis; Chondrosarcoma was misdiagnosed as fibrous dysplasia or aneurysmal bone cyst in 8 cases. Four cases of spinal chordoma were misdiagnosed as spinal tuberculosis.

Conclusion: Our findings highlight the potential of ChatGPT in the diagnosis of benign and malignant bone tumors, offering advantages like enhanced efficiency and a reduction in missed diagnoses. However, the necessity of collaborative interactions between physicians and ChatGPT in practical settings was underscored. With an examination into AI's capacity in benign and malignant bone tumors diagnosis, this study lays the groundwork

* Corresponding author at: Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing 100050, China.
E-mail address: zhixiang.wang@maastro.nl (Z. Wang).

https://doi.org/10.1016/j.jbo.2024.100525
Received 26 October 2023; Received in revised form 3 January 2024; Accepted 7 January 2024
Available online 22 January 2024

for future AI advancements in medicine. Additionally, the benefits of few-shot learning in fine-tuning ChatGPT applications in specialized fields were also demonstrated.

1. Introduction

In daily clinical practice, localized bone lesions are commonly encountered. While some of these lesions are true tumors, the majority present as benign abnormalities. Deciding which lesions require evaluation and which can be left untreated can be a challenging task. At times, radiological findings possess disease-specific characteristics or strongly suggest a particular condition, making radiology crucial in clinical practice [1]. Although the ultimate goal of clinical work is to determine the correct diagnosis, this objective is often unattainable based on existing clinical and radiological data. In clinical practice, it is essential to succinctly and reasonably list a series of relevant diagnoses to ensure that malignant tumors are not overlooked and benign abnormalities are not overtreated. To effectively achieve this purpose, standardized radiology reports and descriptions of radiological features are of utmost importance. The radiology diagnostic report serves as both the culmination of the comprehensive medical imaging process and the most significant source of information for disease diagnosis. It vividly illustrates the correlations among medical imaging findings, clinical presentations, and imaging-based diagnosis.

Artificial intelligence (AI) technology, including natural language processing models such as ChatGPT developed by OpenAI, holds significant potential for medical imaging report diagnostics [2]. Among the various AI models, ChatGPT capable of analyzing large amounts of textual data in a short time, greatly improving the efficiency of diagnostic processes [3].

The advanced AI model ChatGPT, built on the foundation of the GPT series, possesses the ability to understand and generate human-like text [4]. By identifying patterns in large data sets, it can generate comprehensive responses, thereby enhancing its usability in complex tasks that require language comprehension, such as medical imaging report diagnostics [5].

ChatGPT's robust language understanding and generation capabilities reduce the risk of misdiagnosis and missed diagnoses due to human factors [6]. In addition, its ability to uncover potential correlations and patterns within the text of imaging reports can provide valuable insights for doctors, helping them better understand the disease's development mechanisms and optimize treatment plans [7].

As AI technology rapidly advances, the use of ChatGPT in the field of benign and malignant bone tumors diagnosis is garnering increasing attention. Despite AI technology facing certain challenges in practical applications, such as issues with accuracy, professionalism, and usability [8], ChatGPT shows promising potential. Its impressive natural language processing capabilities enable it to effectively navigate the complexities and nuances of medical imaging reports, making it a powerful tool for this application [9]. However, as ChatGPT is a general, large-scale model, its performance in specialized fields can still exhibit shortcomings [10]. Additionally, language differences significantly impact diagnostic outcomes. Variations in medical terminology and patient data across languages can lead to discrepancies in AI analyses. Therefore, it's necessary to explore optimization strategies for enhancing ChatGPT's performance in niche domains, such as employing methods based on few-shot learning. Few-shot learning is a machine learning paradigm designed to allow AI models to make accurate predictions with limited sample data [11]. By providing the model with a small number of training examples ("shots"), it can learn to generalize to new, unseen cases. In the context of benign and malignant bone tumors diagnosis, few-shot learning can significantly enhance the model's ability to differentiate between complex cases, thereby potentially improving diagnostic accuracy and reducing the risk of misdiagnoses.

The significance of this study is twofold. Firstly, it provides a thorough evaluation of ChatGPT's performance in diagnosing benign and malignant bone tumors, compared to physicians, laying a foundation for its potential clinical applications [12]. Secondly, it pioneers the use of few-shot learning to fine-tune ChatGPT, demonstrating its enhanced performance in specialized medical domains such as benign and malignant bone tumors diagnosis [13].

Moreover, our research rigorously verifies the reliability of ChatGPT through repeated experiments, crucial for any potential clinical diagnostic tool [14]. We also propose strategies for refining ChatGPT's application in benign and malignant bone tumors diagnosis, contributing to the broader advancement of AI in medicine. In essence, this study underscores the potential of AI as a customizable and reliable tool in medical diagnostics, particularly for conditions like malignant bone tumors.

2. Materials and method

2.1. Data collection and clinician participation

For this study, 1366 imaging reports of bone tumors diagnosed by CT in our hospital from January 2020 to June 2023 were collected. Patients meeting the following inclusion criteria were included:(1) bone tumors confirmed by puncture or histology after surgery; (2) no frequent puncture and clinical treatment before CT examination; (3)CT examination includes CT plain scan and enhanced scan; (4) In line with the 2020 WHO classification of bone tumors; Exclusion criteria:(1) Receiving clinical treatment; (2) Incomplete clinical data. Finally, 1366 imaging reports of patients with bone tumors were included. These reports contain basic information of patients, such as demographics, CT radiological findings, and puncture biopsy or surgical pathological features, providing a rich data source for our research [15,16]. All imaging reports were from PACS, This study has been approved by the Ethics Committee of our hospital(approval number 201512–02), exempting patients from informed consent.

To ensure the validity of the study, we invited 25 experienced physicians-including 14 intermediate doctors and 11 senior doctors-to manually diagnose the collected imaging reports [17]. Based on the information within the reports and their own professional knowledge and experience, the physicians provided diagnostic opinions for each patient.The gold standard of diagnosis was established by combining clinical, imaging and pathological principles (multidisciplinary team pattern in diagnosis and treatment) , allowing us to compare them with the AI model's diagnostic results and assess the model's performance in benign and malignant bone tumors diagnosis [18]. The research workflow is shown in Fig. 1.

Diagnosis Result Generation of ChatGPT Based on Few-shot Learning. The base model in this research is ChatGPT 3.5. In this research, the few-shot learning method is adopted instead of directly using the ChatGPT model to generate diagnostic results. Because the ChatGPT model has been deeply trained to understand and generate language texts, it may need further improvement in understanding professional medical imaging reports [10]. By adopting few-shot learning and providing specific benign and malignant bone tumors diagnostic samples, the model can more accurately grasp the characteristics and rules of such imaging reports. Therefore, in the experiment, the collected imaging reports were input into the ChatGPT model, and combined with the few-shot learning method, a series of benign and malignant bone tumors samples were provided to the model. These samples include imaging reports and their corresponding diagnostic results. This method fully utilizes the ability of ChatGPT, allowing it to learn and generalize the rules of tasks through a few examples. After

few-shot learning, the GPT model can understand the key information in the report more deeply. The prompt for the experiment is "Here are two examples of imaging reports, which contain the inspection results and corresponding diagnostic results: Patient 1: The radiological findings of the left tibia and fibula CT scan + enhancement examination are as follows: localized cortical bone defect in the anterior part of the upper segment of the left tibia, with clear boundaries, hardened inner edge, medium density center, uniformly enhanced on enhanced scan; slightly swollen nearby sartorius and gracilis tendon. The rest of the left tibia and fibula have smooth and continuous cortical bone. The bone joint surface of the left knee and ankle joints corresponds well. The diagnosis result is benign. Patient 2: The radiological findings of the right femur CT scan + 3D reconstruction examination are as follows: bone destruction at the distal metaphyseal end of the right femur, discontinuous periosteal reaction visible, local Codman triangle visible, peripheral soft tissue swelling. The diagnosis result is malignant. Please learn the information of the above two imaging reports, and then diagnose the benign and malignant of the following patients: The generated response examples will be shown in the results section. Fig. 2 exhibits example of generated responses.

2.2. Comparison analysis of few-shot learning method and direct answer generation

To evaluate the performance of the few-shot learning method and the direct answer generation method in diagnosing benign and malignant bone tumors, we compared the diagnostic results generated by these two methods. We evaluated the accuracy of the diagnosis, such as whether the diagnostic result is consistent with the multidisciplinary team pattern in diagnosis, and the reliability of the diagnostic result, for example, whether the model can stably produce accurate diagnoses. Through this comparative analysis, we can better understand the advantages and limitations of the few-shot learning method relative to direct answer generation in diagnosing benign and malignant bone tumors.

2.3. Report evaluation

In order to thoroughly assess the diagnostic outcomes produced by ChatGPT, they are juxtaposed with gold standard diagnoses determined by multidisciplinary team pattern [19]. Throughout the evaluation process, attention is concentrated on the following aspects: Accuracy: By comparing the concurrence between the model's diagnostic outcomes and the gold standard diagnoses, its precision is gauged [20]. This allows for a deeper understanding of ChatGPT's performance in diagnosing benign and malignant bone tumors. Misdiagnosis and missed diagnosis: Cases of misdiagnosis and missed diagnosis are meticulously scrutinized to pinpoint possible issues within the model's diagnostic procedure [21]. This aids in refining the model further and enhancing its accuracy for future implementations. Model performance analysis: The model's performance is further explored across various imaging report categories, such as its capacity to render accurate judgments in intricate and nuanced cases [3]. This helps appraise the practical application potential and constraints of the model. By employing these assessment techniques, a comprehensive understanding of ChatGPT's capabilities in benign and malignant bone tumors diagnosis can be obtained, offering robust backing for its potential worth and clinical utilization.

2.4. *Impact of radiologist's reporting style experiment*

To investigate the potential impact of individual radiologist's reporting style on the model's performance, Simple random sampling method was adopted,50 radiology reports — 25 authored by senior radiologists and 25 by mid-level radiologists — were input into the model for analysis. The reports were carefully selected to encompass a range of benign and malignant bone tumors presentations and were anonymized to prevent any potential bias. By comparing the diagnostic results provided by the model for each group, we aimed to evaluate if the model's performance varied depending on the reporting style and experience level of the radiologist.

2.5. Analysis of misdiagnosed cases

To gain further insight into ChatGPT's performance in benign and malignant bone tumors diagnosis, we have conducted an in-depth analysis of misdiagnosed cases.By categorizing error cases, we can better understand the model's diagnostic issues in various aspects. Analyzing Error Causes: For each erroneous case, we examine its imaging report and relevant medical data to explore possible causes for the diagnostic error [22]. These causes may include the model's insufficient recognition of certain pathological features, inaccurate understanding of medical terminology, and more.

3. Results*Patient information*

Below is a three-line table summarizing the key demographic and clinical characteristics of the patients included in this study(Table 1).

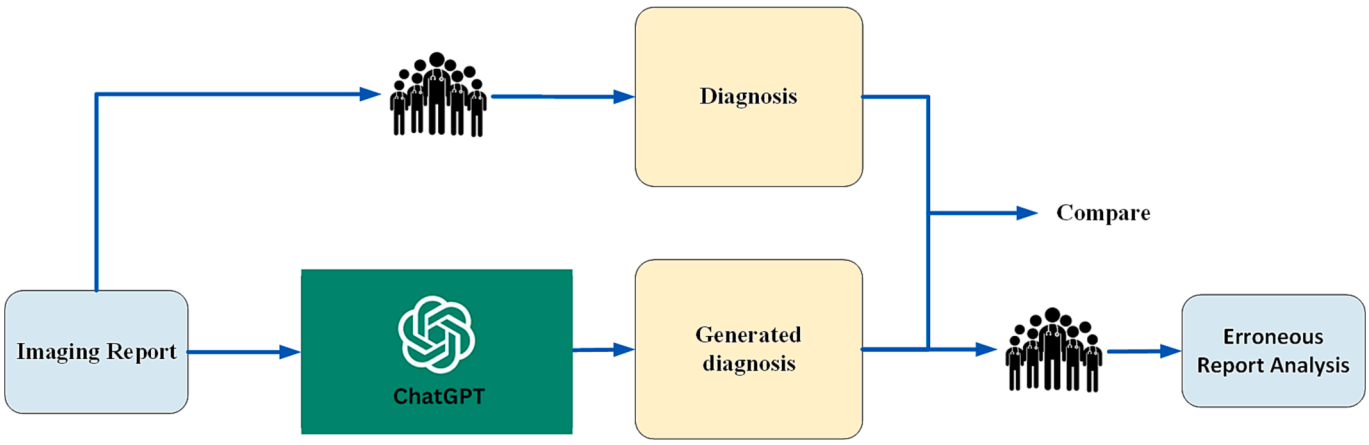

Fig. 1. Experimental workflow diagram illustrating the process, which begins with data collection and acquisition of imaging reports. Next, ChatGPT generates diagnostic results, which are then compared with clinical diagnoses and few-shot learning based method. Concurrently, misdiagnosed reports and results undergo analysis by clinicians. Finally, the impact of radiologist's reporting style was evaluated.

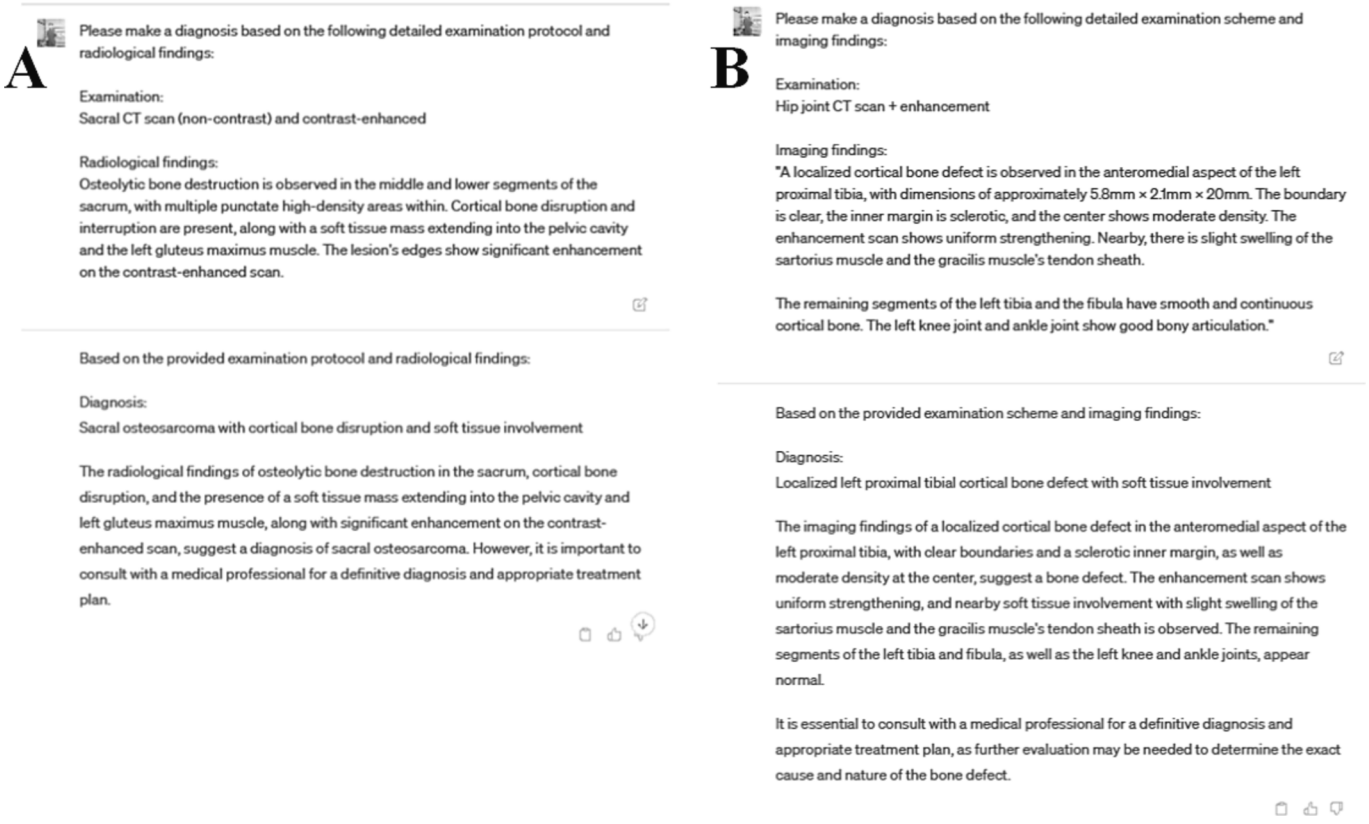

Fig. 2. Example of generated responses. A represents osteosarcoma diagnosis, while B represents benign tumor diagnosis.

Table 1
Clinical information of benign and malignant bone tumors.

Group	Age	Sex		N
	(mean ± SD, years)	Male	Female	
Benign bone tumor group	18.18 ± 11.05	484	345	829
Malignant bone tumor group	19.03 ± 11.15	299	238	537

3.1. *Statistical evaluation metrics*

Based on the collected data, we compared the diagnostic outcomes of physicians with both the direct and few-shot learning based outputs of ChatGPT. The physician diagnostic results showed an accuracy of 0.88, sensitivity of 0.99, and specificity of 0.81 with the AUC 0.92. In contrast, the direct results from ChatGPT showed an accuracy of 0.73, sensitivity of 0.95, and specificity of 0.58 with the AUC 0.72. Notably, after applying few-shot learning for fine-tuning, ChatGPT's performance significantly improved, achieving an accuracy of 0.87, sensitivity of 0.99, and specificity of 0.73 with the AUC 0.83. The T-statistic between ChatGPT's initial and few-shot learning-enhanced outputs was 2.58, yielding a significant P-value of less than 0.01, indicating a notable enhancement in ChatGPT's diagnostic performance post-fine-tuning. Additionally, the comparison between physicians and few-shot enhanced ChatGPT showed a T-statistic of 3.42 with a P-value of 0.0008, underscoring the physicians' statistically superior performance. (Table 2). Our findings underscore that few-shot learning can significantly enhance the specificity of ChatGPT, bringing it much closer to the level of physician diagnostics. The average time cost for ChatGPT is 4.28 s, which demonstrates its ability to almost achieve real-time analysis.

3.2. *Influence of radiologist's reporting style on model performance*

The experiment aimed to evaluate the impact of different radiologist's reporting styles on the model's diagnostic performance yielded interesting results. For the group of reports written by mid-level radiologists, the model achieved an accuracy of 0.76, a sensitivity of 0.83, and a specificity of 0.69. In contrast, the model demonstrated slightly enhanced performance when analysing the reports authored by high-level radiologists, with an accuracy of 0.80, a sensitivity of 0.91, and a specificity of 0.69(Table 3). These findings suggest that the model's performance may slightly vary depending on the radiologist's level of experience and potentially, their unique reporting style. Notably, the model demonstrated higher sensitivity when interpreting reports written by high-level radiologists, which could be attributed to the more detailed and precise language generally used by experienced radiologists. However, the specificity remained consistent for both groups, indicating that the model's ability to correctly rule out benign bone tumors cases was unaffected by the level of the radiologist's experience.

Table 2
Comparison of diagnostic performance in each group.

Group	Accuracy	Sensitivity	Specificity	AUC
Physician	0.88	0.99	0.81	0.92
ChatGPT(directly)	0.73	0.95	0.58	0.72
ChatGPT(few-shot)	0.87	0.99	0.73	0.83

Table 3
Impact of Radiologist's Reporting Style Experiment result.

Group	Accuracy	Sensitivity	Specificity	AUC
middle	0.76	0.83	0.69	0.78
high	0.8	0.91	0.69	0.84

3.3. Analysis of misdiagnosis cases

In 56 benign cases, ChatGPT misdiagnosed them as malignant. Among these, 35 benign lesions- fibrous dysplasia and osteofibrous dysplasia- were incorrectly identified as metastatic tumors or osteosarcomas.;8 cases of myositis ossificans were wrongly diagnosed as extraosseous osteosarcoma. 7 cases of giant cell tumor of bone at the end of long bone were misdiagnosed as osteosarcoma by intermediate doctors. Chondroblastoma was misdiagnosed as malignant tumor in 6 cases −2 osteosarcoma and 4 chondrosarcoma. In this study, 23 osteosarcoma cases were misdiagnosed by ChatGPT as osteomyelitis. Chondrosarcoma was misdiagnosed as fibrous dysplasia or aneurysmal bone cyst in 8 cases. Four cases of spinal chordoma were misdiagnosed as spinal tuberculosis. Notably, 70 benign cases were consistently misdiagnosed as malignant by both ChatGPT and radiologists.

4. Discussion

In this study, our attention converges on the intricacies of benign and malignant bone tumors diagnosis, especially regarding the utilization of artificial intelligence technologies like the ChatGPT model, with the aim to enhance diagnostic accuracy and efficiency [12]. Malignant bone tumors, especially osteosarcoma, a malignant tumor, poses a significant diagnostic challenge owing to its complexity. Radiology reports play a vital role in malignant bone tumors diagnosis in clinical practice, yet the risk of misdiagnosis and missed diagnoses still lingers. This study, therefore, seeks to assess the diagnostic proficiency of the ChatGPT model in the context of benign and malignant bone tumors by analyzing imaging reports, and in doing so, we aim to unearth its potential value and real-world clinical applications [23].

The significance of this study is multifaceted. Firstly, it comprehensive evaluation of ChatGPT's diagnostic proficiency in benign and malignant bone tumors, compare against physician diagnoses, thereby illuminating its potential for real-world clinical applications. Secondly, our study underscores the efficacy of few-shot learning in fine-tuning ChatGPT, which results in improved performance in specialized medical domains such as benign and malignant bone tumors diagnosis. In doing so, it offers a pioneering exploration of machine learning optimization strategies to enhance the specificity of large-scale language models like ChatGPT in the field of medical diagnostics. Lastly, by analyzing the impact of varying physician reports, we gauge the robustness of ChatGPT, emphasizing its reliability and consistency in clinical applications.

Undertaking these objectives, we aimed to gain a holistic understanding of ChatGPT's abilities in benign and malignant bone tumors diagnosis. This could provide empirical evidence of its potential utility in clinical environments, and shed light on areas where diagnostic errors occur, offering crucial insights for refining AI applications in medical diagnostics [18].

Based on the data collected, our study unveils a comprehensive comparison of diagnostic outcomes between physicians and the ChatGPT model, both in its direct and few-shot learning tuned forms. The study results highlight the potential of fine-tuning approaches such as few-shot learning in improving AI diagnostics. Notably, we observed significant improvements in both accuracy and specificity after fine-tuning the ChatGPT model.

The stark contrast, however, lies in the specificity. Physicians' diagnostics demonstrate high specificity of 0.81, substantially outperforming the direct output of ChatGPT, which yielded 0.58. Nevertheless, the application of few-shot learning presented a remarkable enhancement in ChatGPT's specificity, elevating it to 0.73. This not only indicates the effectiveness of few-shot learning but also signals that the ChatGPT model can be effectively customized for medical applications.

Building upon the collected data, our study presents a detailed comparison of diagnostic outcomes derived from physicians and the ChatGPT model in both its original and few-shot learning tuned forms. Initial observations indicated that physicians had a superior diagnostic accuracy of 0.88, notably higher than the direct output of ChatGPT, which registered 0.73. Intriguingly, upon the application of few-shot learning for fine-tuning, the diagnostic accuracy of ChatGPT surged to 0.87, indicating a notable improvement and closing in on the performance of human physicians [24].

In the case of sensitivity, the diagnostics from physicians outperformed slightly with a value of 0.99, as compared to ChatGPT's initial output at 0.95. Yet, the employment of few-shot learning once again elevated ChatGPT's sensitivity to match that of the physicians at 0.99, thus demonstrating its efficacious capability to detect benign and malignant bone tumors [25].

However, the gap was more pronounced when it came to specificity. The physicians' diagnostics exhibited high specificity at 0.81, significantly surpassing the initial output from ChatGPT, recorded at 0.58. The application of few-shot learning introduced a striking enhancement in ChatGPT's specificity, propelling it to 0.73. This ascension signifies that the fine-tuning process, like few-shot learning, can considerably improve ChatGPT's proficiency in accurately distinguishing benign bone tumors cases [3].

The upturn in specificity, upon the implementation of few-shot learning to ChatGPT, warrants particular attention. Few-shot learning essentially involves learning from a limited number of samples and extrapolating that knowledge to novel cases [26].In this context, the ChatGPT model enhanced its ability to differentiate benign bone tumors from malignant bone tumors cases, following exposure to a handful of example cases, which resulted in the dramatic boost in specificity.

This observation reinforces the potential of few-shot learning in adapting broad AI models like ChatGPT for distinct medical applications. The rise in specificity translates to a diminished probability of false-positive results. In practical medical terms, this reduction signifies that fewer patients would face the distressing prospect of being inaccurately identified as having malignant bone tumors, thus preventing undue stress and economizing healthcare resources by avoiding unnecessary follow-up testing and treatment [27].

Our research not only demonstrates the feasibility of few-shot learning in the medical domain but also highlights its revolutionary potential. Unlike traditional machine learning models that typically rely on extensive datasets for effective training, few-shot learning allows models like ChatGPT to quickly adapt to specialized tasks with minimal data. This is especially pertinent in healthcare, where data can be scarce or highly specific to individual patients [28]. The adaptability introduced by few-shot prompt fine-tuning positions ChatGPT as a distinctly valuable tool in healthcare applications, offering a new paradigm in AI-driven medical diagnostics and decision-making [29].

Furthermore, our results underline ChatGPT's promise for near-real-time analysis, boasting an average processing time of a mere 4.28 s. This processing efficiency illuminates the immense advantages of integrating AI-based tools like ChatGPT into clinical practice. Especially in situations where time is of the essence or quick decision-making is crucial, AI tools like ChatGPT can considerably expedite the diagnostic process.

In addition to its reproducibility, our findings suggest that the performance of ChatGPT may be influenced by the reporting style and experience level of the radiologist whose reports it analyzes. ChatGPT demonstrated a slightly improved accuracy and sensitivity when interpreting reports written by high-level radiologists, potentially due to the more detailed and precise language used by these experienced professionals. This observation suggests that the utility of ChatGPT as a diagnostic tool might be optimized when used in conjunction with reports authored by more experienced radiologists. However, regardless of the radiologist's experience, the specificity of ChatGPT remained consistent, indicating its reliable ability to correctly rule out benign bone tumors cases. This reliability, in combination with its potential for fine-tuning through few-shot learning, suggests that ChatGPT could serve as a robust support tool for physicians. It can provide consistent

diagnostic suggestions, potentially alleviating the workload and diagnostic pressure on healthcare professionals [30].

From the perspective of clinical diagnostic principles, the overlap of certain features among benign and malignant bone lesions, such as radiographic appearance and growth patterns, may pose a challenge for both physicians and AI models like ChatGPT. Without considering additional clinical information, this overlap could result in difficulty distinguishing between these conditions [31].

In 56 benign cases, ChatGPT was misdiagnosed as malignant. Of these, 35 benign lesions- fibrous dysplasia and osteofibrous dysplasia- were incorrectly identified as metastatic tumors or osteosarcomas. The underlying reason for this misunderstanding may be related to the simultaneous occurrence of "multiple bone destruction", "periosteal reaction" and "ground glass density" in the diagnostic report. 8 cases of myositis ossificans were wrongly diagnosed as extraosseous osteosarcoma. The reason was that the diagnostic doctor did not correctly understand "ossification" and "tumor bone". The ossification shadow was lamellar or indefinite shape, with trabecular bone structure in the middle, and the ossification was centrifugal mature from the outside to the inside, and there was a clear gap between the bone cortex and the bone cortex. The calcification or ossification of the tumor is mostly distributed in the tumor. In addition, combined with clinical history and CT image analysis, 4 cases were localized lesions, most of the ossification was far from the bone, and there was no clear history of trauma, which was also one of the reasons for misdiagnosis[32,33].

Seven cases of giant cell tumor of bone occurring at the end of long bone were misdiagnosed as osteosarcoma (mainly osteolytic) by intermediate doctors. This may be due to the fact that intermediate doctors did not analyze the image carefully enough, over-considered the location of the disease at the end of bone, and did not show high-density tumor bone in the lesion, and the patients were over 30 years old, so the possibility of adult osteosarcoma was ignored. 6 cases of chondroblastoma were misdiagnosed as malignant tumors (2 cases of osteosarcoma and 4 cases of chondrosarcoma), one of the reasons for misdiagnosis was insufficient understanding of chondroblastoma in rare sites. In addition, the lesions were large, so it was considered malignant.

Both osteomyelitis and osteosarcoma commonly occur in the metaphyseal region of bones, originating from the medullary cavity, extending to the surrounding bone, and spreading up and down along the medullary canal. Due to similarities in age of onset, location, and clinical presentation, distinguishing between the two poses a challenge for clinical diagnosis[34,35]. In this study, 23 cases of osteosarcoma were misdiagnosed as osteomyelitis by ChatGPT. The reasons for this misdiagnosis may be as follows: 1. It is related to their similar report descriptions, such as "focal osteosclerosis", "periosteal triangle" and "soft tissue mass" in both cases; 2. 2. Failure to correctly understand "dead bone" and "residual bone", osteosarcoma, like other bone malignancies, does not form the ischemic mechanism of normal bone, so it rarely forms dead bone, but the uneven infiltration, growth and bone destruction and absorption of the tumor can make part of normal bone remain in the tumor tissue, and can also enter the soft tissue mass with the growth of the tumor, forming similar signs of dead bone. It is worth noting that 4 patients diagnosed with osteomyelitis by pathology and ChatGPT were misdiagnosed as osteosarcoma by radiologists. Combined with text analysis, this may be related to the simultaneous occurrence of the fields of "round or circular bone destruction", "with clear sclerotic edge", "peripheral bone defect", "broad band periosteum hyperplasia" and "cortical thickening". Of course, this hypothesis needs a large sample of data to verify.

Chondrosarcoma was misdiagnosed as fibrous dysplasia and aneurysmal bone cyst in 8 cases. CT findings of cases misdiagnosed as fibrous dysplasia showed more typical loofah sac-like changes in the near middle of trunk bone, which are common in fibrous dysplasia. Retrospective analysis showed speckle high-density calcification shadow and short acicular periosteum reaction, and the peripheral soft tissue swelling was more obvious, which should be considered in the direction of malignant tumor. In cases misdiagnosed as aneurysmal bone cysts, CT showed obvious dilatant growth, with more fine compartments and higher penetration, and periosteum and soft tissue were negative. Postoperative analysis showed that misdiagnosis was due to the ignorance of the blurred local edges of the lesion, and bone destruction with blurred boundaries is a common sign of malignant bone tumors. Although the incidence of clear cell chondrosarcoma is very low, the concurrent appearance of the above CT signs should be considered as a possibility of malignancy. Four cases of vertebral chordoma were misdiagnosed as spinal tuberculosis. On the one hand, the clinical manifestations of the patients were very similar to tuberculosis; on the other hand, due to the lack of rich clinical experience, the diagnostic doctor only paid attention to the patient's history of tuberculosis, did not take the initiative to check the clinical details, and did not take into account the image manifestations of "different diseases" when making diagnosis, which may also lead to misdiagnosis.

This complicates the diagnostic process for both physicians and AI models, emphasizing the importance of taking additional clinical information into account to enhance diagnostic accuracy. Therefore, factors like patient demographics, medical history, laboratory test results, and biopsy findings should all be considered alongside the radiographic appearance [36–38].

ChatGPT presents several strengths in the diagnosis of malignant bone tumors. As an AI-based technology, it can quickly process a large volume of cases, enhancing diagnostic efficiency. Additionally, ChatGPT's high sensitivity improves its ability to detect malignant bone tumors accurately, aiding physicians in making preliminary judgments and minimizing the chance of missed diagnoses.

Despite these strengths, ChatGPT encounters some practical challenges. However, this process is further complicated by the impact of language differences, such as between English and Chinese. Variations in medical terminologies and expressions can significantly affect AI analysis. The disparity in the size of language corpora used for training also contributes to differences in analytical abilities. A more comprehensive corpus in one language might lead to more accurate analysis than in a language with a smaller, less diverse corpus [39]. Ambiguities or inaccuracies in diagnostic reports may lead to incorrect diagnoses. Therefore, the indispensable role of physicians in the diagnostic process should not be underestimated, and their synergistic collaboration with ChatGPT is vital to ensure effective diagnoses [40].

5. Conclusions

Our investigation of benign and malignant bone tumors diagnosis underscored the challenges and complexities of the process while revealing the potential of AI tools like ChatGPT. Although high sensitivity was observed in ChatGPT's initial output, the model grappled with differentiating complex and ambiguous cases. The overlapping imaging features between benign and malignant bone lesions further emphasized the necessity for additional clinical information. However, the application of few-shot learning demonstrated a notable enhancement in ChatGPT's diagnostic accuracy and specificity. This fine-tuning not only increased efficiency but also reduced the likelihood of missed diagnoses and false positives. Despite these improvements, the study reinforced the essential role of human physicians in the diagnostic process and the importance of their synergistic collaboration with AI tools like ChatGPT. In sum, our study illuminates the potential role of ChatGPT in benign and malignant bone tumors diagnosis and paves the way for further advancements in the integration of AI into the medical field, aiming for more accurate, efficient, and patient-centered diagnoses.

CRediT authorship contribution statement

Fan Yang: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Dong Yan:** Conceptualization, Data curation. **Zhixiang Wang:** Conceptualization, Data curation,

Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by Beijing Jishuitan Hospital Research Funding (NO.ZR202315) and 2020 SKY Image Research Funding(NO. Z-2014-07-2003-16).

References

[1] C. Errani, J. Kreshak, P. Ruggieri, M. Alberghini, P. Picci, D. Vanel, Imaging of bone tumors for the musculoskeletal oncologic surgeon, Eur. J. Radiol. 82 (12) (2013) 2083–2091, https://doi.org/10.1016/j.ejrad.2011.11.034.
[2] S.S. Biswas, Role of chat GPT in public health, Ann. Biomed. Eng. 51 (5) (2023) 868–869, https://doi.org/10.1007/s10439-023-03172-7.
[3] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, H. Aerts, Artificial intelligence in radiology, Nat. Rev. Cancer 18 (8) (2018) 500–510, https://doi.org/10.1038/s41568-018-0016-5.
[4] M. Sallam, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, Healthcare (Basel) 11(6) (2023) 887, https://doi.org/10.3390/healthcare11060887.
[5] R.K. Garg, V.L. Urs, A.A. Agrawal, S.K. Chaudhary, V. Paliwal, S.K. Kar, Exploring the Role of Chat GPT in patient care (diagnosis and Treatment) and medical research: A Systematic Review, medRxiv (2023) 23291311, https://doi.org/10.1101/2023.06.13.23291311.
[6] E. Faiella, D. Santucci, A. Calabrese, F. Russo, G. Vadalà, B.B. Zobel, P. Soda, G. Iannello, C. de Felice, V. Denaro, Artificial intelligence in bone metastases: an mri and ct imaging review, Int. J. Environ. Res. Public Health 19 (3) (2022) 1880, https://doi.org/10.3390/ijerph19031880.
[7] P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (12) (2017) 749–762, https://doi.org/10.1038/nrclinonc.2017.141.
[8] B. Norgeot, B.S. Glicksberg, A.J. Butte, A call for deep-learning healthcare, Nat. Med. 25 (1) (2019) 14–15, https://doi.org/10.1038/s41591-018-0320-3.
[9] B. Vasey, M. Nagendran, B. Campbell, D.A. Clifton, G.S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, X. Liu, B.A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D.S.W. Ting, P. Watkinson, W. Weber, P. Wheatstone, P. McCulloch, Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, Nat. Med. 28 (5) (2022) 924–933, https://doi.org/10.1038/s41591-022-01772-9.
[10] L. Floridi, M. Chiriatti, GPT-3: its nature, scope, limits, and consequences, Mind. Mach. 30 (2020) 681–694, https://doi.org/10.1007/s11023-020-09548-1.
[11] SuvarnaKadam, VinayVaidya, SuvarnaKadam, VinayVaidya, Review and Analysis of Zero, One and Few Shot Learning Approaches, Springer, Cham (2020) 100–112, https://doi.org/10.1007/978-3-030-16657-1_10.
[12] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56, https://doi.org/10.1038/s41591-018-0300-7.
[13] D.S. Geller, R. Gorlick, Osteosarcoma: a review of diagnosis, management, and treatment strategies, Clin. Adv. Hematol. Oncol. 8 (10) (2010) 705–718.
[14] E.O. Nsoesie, Evaluating artificial intelligence applications in clinical settings, JAMA Netw. Open 1 (5) (2018) e182658.
[15] M.D. Murphey, M.R. Robbin, G.A. McRae, D.J. Flemming, H.T. Temple, M. J. Kransdorf, The many faces of osteosarcoma, Radiographics 17 (5) (1997) 1205–1231, https://doi.org/10.1148/radiographics.17.5.9308111.
[16] J. Ritter, S.S. Bielack, Osteosarcoma, Ann. Oncol. 21 Suppl 7 (2010) vii320-325, https://doi.org/10.1093/annonc/mdq276.
[17] A. Hirschmann, J. Cyriac, B. Stieltjes, T. Kober, J. Richiardi, P. Omoumi, Artificial intelligence in musculoskeletal imaging: review of current literature, challenges, and trends, Semin Musculoskelet Radiol 23 (3) (2019) 304–311, https://doi.org/10.1055/s-0039-1684024.
[18] E.K. Oikonomou, M. Siddique, C. Antoniades, Artificial intelligence in medical imaging: a radiomic guide to precision phenotyping of cardiovascular disease, Cardiovasc. Res. 116 (13) (2020) 2040–2054, https://doi.org/10.1093/cvr/cvaa021.
[19] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (1) (2019) 195, https://doi.org/10.1186/s12916-019-1426-2.
[20] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.
[21] M. Nagendran, Y. Chen, C.A. Lovejoy, A.C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J.P.A. Ioannidis, G.S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, BMJ 368 (2020) m689, https://doi.org/10.1136/bmj.m689.
[22] E. Baylor, E. Beede, F. Hersch, A. Iurchenko, L. Wilcox, A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy, ACM CHI (2020) 1–12, https://doi.org/10.1145/3313831.3376718.
[23] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nat. Med. 25 (1) (2019) 24–29, https://doi.org/10.1038/s41591-018-0316-z.
[24] D.S.W. Ting, L.R. Pasquale, L. Peng, J.P. Campbell, A.Y. Lee, R. Raman, G.S.W. Tan, L. Schmetterer, P.A. Keane, T.Y. Wong, Artificial intelligence and deep learning in ophthalmology, Br. J. Ophthalmol. 103 (2) (2019) 167–175, https://doi.org/10.1136/bjophthalmol-2018-313173.
[25] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke Vasc Neurol 2 (4) (2017) 230–243, https://doi.org/10.1136/svn-2017-000101.
[26] M. Boudiaf, Z.I. Masud, J. Rony, J. Dolz, P. Piantanida, I.B, Ayed, Transductive Information Maximization For Few-Shot Learning, 2020.
[27] P. Pilavaki, A. Gahanbani Ardakani, P. Gikas, A. Constantinidou, Osteosarcoma: current concepts and evolutions in management principles, J. Clin. Med. 12 (8) (2023) 2785, https://doi.org/10.3390/jcm12082785.
[28] N. Tajbakhsh, Y. Hu, J. Cao, X. Yan, X. Ding, Surrogate supervision for medical image analysis: effective deep learning from limited quantities of labeled data, IEEE (2019), https://doi.org/10.1109/ISBI.2019.8759553.
[29] G. Nguyen, M.T. StefanBobak, AlvaroHeredia Garcia, PeterHluchy IgnacioMalik, Ladislav, Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey, Artif. Intell. Rev.: an Int. Sci. Eng. J. 52 (1) (2019), https://doi.org/10.1007/s10462-018-09679-z.
[30] R. Manne, S.C. Kantheti, Application of artificial intelligence in healthcare: chances and challenges, Current J. Appl. Sci. Technol. 40 (6) (2021) 78–89, https://doi.org/10.9734/CJAST/2021/V40I631320.
[31] J. Crim, L.J. Layfield, Bone and soft tissue tumors at the borderlands of malignancy, Skeletal Radiol. 52 (3) (2023) 379–392, https://doi.org/10.1007/s00256-022-04099-1.
[32] O. Savvidou, O. Papakonstantinou, E. Lakiotaki, D. Melissaridou, P. Korkolopoulou, P.J. Papagelopoulos, Post-traumatic myositis ossificans: a benign lesion that simulates malignant bone and soft tissue tumours, EFORT Open Reviews 6 (7) (2021) 572–583, https://doi.org/10.1302/2058-5241.6.210002.
[33] A. Al Khader, E. Habaibeh, H.Y. Tawalbeh, B.M. Mansour, S.M. Mansour, A.N. Haj Ahmad, T.T. Owais, H. Odeh, Myositis ossificans of the chest wall in an 8-year-old boy: a case report of a diagnostic pitfall, Indian J. Thoracic Cardiovascular Surgery 39 (2) (2023) 186–189, https://doi.org/10.1007/s12055-022-01463-7.
[34] Al-Chalabi, M. M. M., Jamil, I., & Wan Sulaiman, W. A.. Unusual Location of Bone Tumor Easily Misdiagnosed: Distal Radius Osteosarcoma Treated as Osteomyelitis. *Cureus 13*(11) (2021) e19905,https://doi.org/10.7759/cureus.19905.
[35] Salman, R., Mcgraw, M., & Naffaa, L.. Chronic Osteomyelitis of Long Bones: Imaging Pearls and Pitfalls in Pediatrics. *Seminars in ultrasound, CT, and MR 43*(1) (2022) 88–96, https://doi.org/10.1053/j.sult.2021.05.009 doi:10.1053/j.sult.2021.05.009.
[36] D. Tafti, N.D. Cecava, Fibrous Dysplasia, StatPearls, StatPearls Publishing Copyright © 2023, StatPearls Publishing LLC., Treasure Island (FL) ineligible companies. Disclosure: Nathan Cecava declares no relevant financial relationships with ineligible companies., 2023.
[37] Momodu, II, V. Savaliya, Osteomyelitis, StatPearls, StatPearls Publishing Copyright © 2023, StatPearls Publishing LLC., Treasure Island (FL) ineligible companies. Disclosure: Vipul Savaliya declares no relevant financial relationships with ineligible companies., 2023.
[38] C. Pineda, R. Espinosa, A. Pena, Radiographic imaging in osteomyelitis: the role of plain radiography, computed tomography, ultrasonography, magnetic resonance imaging, and scintigraphy, Semin. Plast. Surg. 23 (2) (2009) 80–89, https://doi.org/10.1055/s-0029-1214160.
[39] Z. Obermeyer, E.J. Emanuel, Predicting the future - big data, machine learning, and clinical medicine, N. Engl. J. Med. 375 (13) (2016) 1216–1219, https://doi.org/10.1056/NEJMp1606181.
[40] E.H. Shortliffe, M.J. Sepúlveda, Clinical decision support in the era of artificial intelligence, J. Am. Med. Assoc. 320 (21) (2018) 2199–2200, https://doi.org/10.1001/jama.2018.17163.