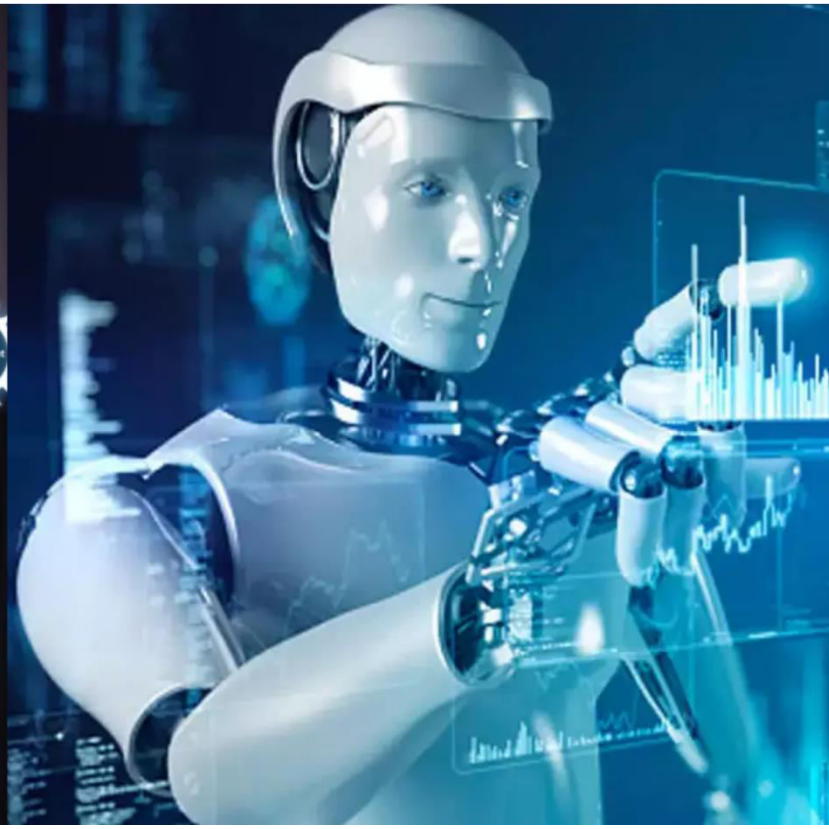
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503223|

Secure and Reliable AI Framework for Orthopedic Disease Detection

S .SenthilKumar[1], V. Aashika[2], S. Janani[3], A. Abinayasri[4], M. Dhinisha[5]

Assistant Professor, Department of Computer Science Engineering, Mahendra Engineering College for Women, Namakkal, Tamilnadu, India[1]

UG Students, Department of Computer Science Engineering, Mahendra Engineering College for Women, Namakkal, Tamilnadu, India[2345]

ABSTRACT: This study introduces a novel explainable orthopaedic disease detection framework that is robust to both white-box and black-box adversarial attack examples based on hybrid fuzzy Multi-Criteria Decision-Making (MCDM). It addresses the lack of a resilient framework against adversarial attacks with complex evaluation matrices and the absence of a reliable method for selecting optimal orthopedic disease detection models using four-phase methodology. The dataset was augmented with edge-based techniques, improving generalizability and adversarial robustness. A comprehensive evaluation of nine Machine Learning (ML) models was conducted under Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Label Poisoning Method (LPM), and Model Extraction Method (MEM) adversarial attacks to increase the robustness of the developed models. The selection of the most robust developed model is established via a novel decision matrix framework and the optimized Weight Fuzzy Judgment Method (WFJM) and Z-Score Multi-Attribute-Based Aggregation (Z-MABAC) method. According to the WFJM, the fooling rate and prevalence obtained the highest average importance weights of 0.090 and 0.094, respectively, across the four attack scenarios. In contrast, specificity, precision, and Cohen's kappa received the lowest weights, with values of 0.057, 0.058, and 0.069, respectively. The benchmarking results revealed the following top-performing models across adversarial scenarios: the LPM-Decision Tree, BIM-Support Vector Machin, FGSM-Gradient Boosting, and MEM-Random Forest with scores of 0.122, 0.180, 0.122 and 0.170, respectively. The proposed framework validated through sensitivity-correlation analysis and explainability analysis via Local Interpretable Model-agnostic Explanations (LIME). The results underscore the importance of adversarial robustness in AI-based orthopaedic detection systems, contributing to the broader field of secure and ethical AI adoption in healthcare..

KEYWORDS: Adversarial machine learning, multi-criteria decision-making (MCDM), orthopaedic disease detection, robustness evaluation, trustworthy artificial intelligence.

I. INTRODUCTION

Orthopaedic diseases, which affect millions of people globally, present significant diagnostic challenges, often leading to long-term disability and chronic pain. There is an ongoing debate across the literature regarding the trustworthiness of artificial intelligence (AI) in detecting orthopaedic diseases. This systematic review aims to provide a comprehensive taxonomy of AI applications in orthopaedic disease detection. A thorough literature search was conducted across five major databases (Science Direct, Scopus, IEEE Xplore, PubMed, and Web of Science) covering publications from January 2019 to 2024. Following rigorous screening on the basis of predefined inclusion criteria, 85 relevant studies were identified and critically evaluated. For the first time, this review classifies AI contributions into six key categories of orthopaedic conditions on the basis of medical perspective: arthritis, tumours, deformities, fractures, osteoporosis, and general bone abnormalities. In addition to analyzing motivations, challenges, and recommendations for future research, this review highlights the various AI techniques employed, including deep learning (DL), machine learning (ML), explainable AI (XAI), fuzzy logic, and multicriteria decision-making (MCDM), as well as the datasets utilized. Furthermore, the trustworthiness of AI models is evaluated on the basis of seven AI trustworthiness components, aligned with European Union guidelines, within each category. These findings underscore the need for high-quality research to ensure that AI computational systems in orthopaedic disease detection are reliable, safe, and ethical. Future research should focus on optimizing AI algorithms, improving dataset diversity, and addressing ethical and regulatory challenges to ensure successful integration into clinical practice.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503223|

Orthopaedic diseases have a great impact on the musculoskeletal system, such as bones, cartilage, ligaments, tendons, and connective tissues, leading to various problems, including chronic pain, reduced mobility, and chronic disability [1]. According to the Global Burden of Disease Study, 1.71 billion people globally have musculoskeletal disorders such as osteoarthritis, rheumatoid arthritis, and low back pain, highlighting the need for better diagnosis and treatment [2]. Various major types of diseases exist under this line of thought, including fracture, bone tomour, arthritis (e.g., osteoarthritis and rheumatoid arthritis), and degenerative diseases such as osteoporosis and musculoskeletal injuries [3]. Recent advancements in artificial intelligence (AI) have had a tremendous impact on the enhancement of orthopaedic diseases diagnosis and paved the way for the implementation of intelligent systems in diagnostics [4]. Machine learning (ML), deep learning (DL) and multicriteria decision-making (MCDM) techniques have been applied in different branches of orthopaedics to assist specialists in handling complex radiological images (including magnetic resonance imaging (MRI), X-ray, and computed tomography (CT) [5]. To illustrate, ML/DL computational models have been effectively deployed to enable the detection of pathologies such as complex fractures, mild forms of osteoarthritis, and even early-stage bone tumours in a quick and efficient manner [6, 7]. Furthermore, clinicians use MCDM and fuzzy logic methods in orthopaedics to support sensible decision-making and assess several medical standards, promoting accuracy in measurements as well as tailored treatment recommendations [8,9,10]. These results indicate that the evolution and ongoing advancement of modern mechanisms will be vital for the operative management of contemporary orthopaedic disorders.

While computational artificial intelligence (AI) systems have transformed orthopaedic disease detection, they also pose considerable challenges regarding their trustworthiness and interpretability. The dark side is the complexity of the detection models, specifically the ML/DL algorithms in intelligent detection systems. This commonly results in a loss of transparency regarding the decision-making process, which may lead to reluctance among clinicians to fully trust such systems [11]. This is exactly where trustworthy AI comes into play. Trustworthy AI systems are designed with fundamental principles such as fairness, transparency, privacy, and accountability, thereby enabling their safe inclusion in clinical practice [12, 13]. Several disciplines have explored the area of trustworthy AI, for example, disaster management, where AI plays a pivotal role in early warning and decision-making regarding real-time interventions, and healthcare where AI is utilized for diagnosis and treatment planning [14, 15], 16. In orthopaedic disease detection, trustworthy AI must not only provide highly accurate diagnoses but also offer explainable and interpretable outcomes so that clinicians can understand and validate the AI decision-making process [17, 18]. The AI trustworthiness guidelines of the European Union emphasize legality, ethics, and sustainability to promote robust and fair AI solutions for orthopaedic disease detection and stakeholder trust on the basis of seven key components [19]. These components include ensuring human agency and oversight by upholding human rights and involving humans in critical decisions [20]. Additionally, technical robustness and safety are emphasized through security measures, backup systems, and ensuring the accuracy, reproducibility, and reliability of the AI-based computational detection system. Additionally, privacy and data governance prioritize maintaining data quality, protecting privacy, and ensuring data availability. Transparency involves clearly explaining decision-making processes to stakeholders. Diversity, nondiscrimination, and fairness aim to eliminate bias, ensure accessibility, adopt a universal design, and advocate for all users, including those with disabilities [21]. Environmental and societal well-being consider environmental, sustainability, and social impacts and promote democracy. Finally, accountability minimizes and reports negative impacts, manages trade-offs and provides redress when necessary [22].

II. LITERATURE SERVEY

In the realm of medical diagnostics, the utilization of deep learning techniques, notably in the context of radiology images, has emerged as a transformative force. The significance of artificial intelligence (AI), specifically machine learning (ML) and deep learning (DL), lies in their capacity to rapidly and accurately diagnose diseases from radiology images. This capability has been particularly vital during the COVID-19 pandemic, where rapid and precise diagnosis played a pivotal role in managing the spread of the virus. DL models, trained on vast datasets of radiology images, have showcased remarkable proficiency in distinguishing between normal and COVID-19-affected cases, offering a ray of hope amidst the crisis. However, as with any technological advancement, vulnerabilities emerge. Deep learning-based diagnostic models, although proficient, are not immune to adversarial attacks. These attacks, characterized by carefully crafted perturbations to input data, can potentially disrupt the models' decision-making processes. In the medical context, such vulnerabilities could have dire consequences, leading to misdiagnoses and compromised patient care. To address this, we propose a two-phase defenseframework that combines advanced adversarial learning and adversarial

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503223|

image filtering techniques. We use a modified adversarial learning algorithm to enhance the model's resilience against adversarial examples during the training phase. During the inference phase, we apply JPEG compression to mitigate perturbations that cause misclassification. We evaluate our approach on three models based on ResNet-50, VGG-16, and Inception-V3. These models perform exceptionally in classifying radiology images (X-ray and CT) of lung regions into normal, pneumonia, and COVID-19 pneumonia categories. We then assess the vulnerability of these models to three targeted adversarial attacks: fast gradient sign method (FGSM), projected gradient descent (PGD), and basic iterative method (BIM). The results show a significant drop in model performance after the attacks. However, our defense framework greatly improves the models' resistance to adversarial attacks, maintaining high accuracy on adversarial examples. Importantly, our framework ensures the reliability of the models in diagnosing COVID-19 from clean images.

III. IMPLEMENTATION

EXISTING SYSTEM

- Traditional orthopedic diagnosis relies on:
- Manual medical image inspection
- Standard CNN-based disease detection models
- Limitations
- Sensitive to noise and adversarial perturbations
- Lack of model explainability
- No security validation mechanism
- Reduced reliability in clinical environments

3.1. DISADVANTAGES OF EXISTING SYSTEM

- Vulnerability to Adversarial Attacks
- Lack of Model Reliability
- Absence of Security Mechanisms
- Overfitting Due to Limited Medical Data
- Low Interpretability
- High Misclassification Risk
- Poor Generalization Ability
- No Trust Evaluation Framework

3.2. PROPOSED SYSTEM

- The proposed framework integrates secure deep learning, adversarial defense, and trust evaluation mechanisms.
- Framework Components
- Medical Image Acquisition (X-ray/MRI Dataset)
- Image Preprocessing& Enhancement
- Orthopedic Disease Classification Model (CNN/ResNet)
- Adversarial Attack Detection Module
- Defense Mechanism (Adversarial Training)
- Trust Score Evaluation System
- Final Reliable Diagnosis Output

3.2.1.ADVANTAGES

✓High diagnostic accuracy
✓ Resistant to adversarial attacks
✓Secure medical AI framework
✓ Improved reliability and trust
✓ Suitable for clinical decision support

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503223|

IV. METODOLOGY

This study introduces a novel explainable orthopaedic disease detection framework that is robust to both white-box and black-box adversarial attack examples based on hybrid fuzzy Multi-Criteria Decision-Making (MCDM). It addresses the lack of a resilient framework against adversarial attacks with complex evaluation matrices and the absence of a reliable method for selecting optimal orthopedic disease detection models using four-phase methodology. The dataset was augmented with edge-based techniques, improving generalizability and adversarial robustness. A comprehensive evaluation of nine Machine Learning (ML) models was conducted under Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Label Poisoning Method (LPM), and Model Extraction Method (MEM) adversarial attacks to increase the robustness of the developed models. The selection of the most robust developed model is established via a novel decision matrix framework and the optimized Weight Fuzzy Judgment Method (WFJM) and Z-Score Multi-Attribute-Based Aggregation (Z-MABAC) method. According to the WFJM, the fooling rate and prevalence obtained the highest average importance weights of 0.090 and 0.094, respectively, across the four attack scenarios. In contrast, specificity, precision, and Cohen's kappa received the lowest weights, with values of 0.057, 0.058, and 0.069, respectively. The benchmarking results revealed the following top-performing models across adversarial scenarios: the LPM-Decision Tree, BIM-Support Vector Machin, FGSM-Gradient Boosting, and MEM-Random Forest with scores of 0.122, 0.180, 0.122 and 0.170, respectively. The proposed framework validated through sensitivity-correlation analysis and explainability analysis via Local Interpretable Model-agnostic Explanations (LIME). The results underscore the importance of adversarial robustness in AI-based orthopaedic detection systems, contributing to the broader field of secure and ethical AI adoption in healthcare.

Medical imaging has a profound impact on healthcare by enabling the visualization of organs, cells, and pathological specimens, revolutionizing the diagnosis of diseases. It encompasses a broad range of tasks, including image processing, storage, analysis, retrieval, and interpretation. This technology plays a crucial role in advancing our understanding of various medical conditions, leading to more accurate diagnoses and improved treatment options.
Deep Neural Networks (DNNs), recognized as highly effective artificial intelligence techniques, gained significant prominence in medical imaging. They are essential in enhancing diagnostic accuracy and accelerating decision-making for healthcare professionals. DNNs achieved remarkable success in processing large volumes of data, extracting valuable insights from vast datasets across various fields. Examples of DNN applications in medical imaging include the early diagnosis of skin cancer through photographic images [1], classification of diabetic retinopathy using Optical Coherence Tomography Angiography (OCTA) images [2], pneumonia detection from chest X-rays [3], and nodule segmentation from CT images [4]. Various studies demonstrated that DNNs in medical imaging achieve performance levels comparable to human experts, with diagnostic results often matching those of professional medical staff.

Learning systems are proven highly cost-effective, delivering remarkable results with accuracy comparable to standard clinical practices. They received the United States Food and Drug Administration (FDA) approval. However, ensuring the security and reliability of medical DNNs remains a critical concern for the scientific community. Recent studies addressing classification and segmentation tasks in medical imaging revealed that even state-of-the-art DNNs are significantly susceptible to adversarial attacks. Medical imaging DNNs are more vulnerable to adversarial attacks than DNNs processing natural images as their input [5]. These vulnerabilities enable slight, carefully crafted manipulations in image samples, often invisible to the human eye, to significantly impact the DNN's performance (see Figure 1). Addressing these malicious adversaries is one of the most crucial challenges in medical Deep Learning (DL) systems.
Various methods were proposed to generate adversarial attacks, including the Fast Gradient Sign Method (FGSM) [6] and its more potent variants like Projected Gradient Descent (PGD) Madry2018, as well as the "Carlini& Wagner (C&W)" method [7]. A study by Ref. [8] explained different incentives for attacking medical learning systems primarily from financial motivations. This is exacerbated by the rise in healthcare costs. Yet, medical DNNs cannot serve as replacements for physicians and medical specialists. Given the unique characteristics of medical images and the limited availability of annotated data, even with the guidance of human experts such as engineers, physicians, and radiologists, these neural networks were proven to be susceptible to rapidly evolving adversarial examples, highlighting the challenges of relying solely on automated systems for medical decision-making.

To counter adversarial attacks, researchers proposed various mitigation and detection techniques. One widely used approach involves adversarial training, incorporating adversarial samples into the training dataset to enhance the neural networks' robustness against adversarial attacks. Despite the effectiveness of these defense methods against adversarial

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503223|

attacks, they consistently show superior performance on CNNs with natural image datasets, such as CIFAR-10, compared to medical CNNs. This disparity in performance can be attributed to the insufficient availability of high-quality images and labeled data, proving the importance of addressing this data scarcity to strengthen the defenses of medical DNNs.

This study introduces a novel approach to adversarial medical machine-learning training focusing on robustness and efficiency. Unlike previous methods that augment adversarial samples into the training dataset, our approach emphasizes advanced feature transformation. Traditionally, augmenting adversarial samples demands significant computational resources and does not work well with the specific challenges posed by medical CNNs.

V. CONCLUSION

In the evolving field of healthcare, centralized cloud-based medical image retrieval faces challenges related to security, availability, and adversarial threats. Existing deep learning-based solutions improve retrieval but remain vulnerable to adversarial attacks and quantum threats, necessitating a shift to more secure distributed cloud solutions. This article proposes SFMedIR, a secure and fault tolerant medical image retrieval framework that contains an adversarial attack-resistant federated learning for hashcode generation, utilizing a ConvNeXt-based model to improve accuracy and generalizability. The framework integrates quantum-chaos-based encryption for security, dynamic threshold-based shadow storage for fault tolerance, and a distributed cloud architecture to mitigate single points of failure. Unlike conventional methods, this approach significantly improves security and availability in cloud-based medical image retrieval systems, providing a resilient and efficient solution for healthcare applications. The framework is validated on Brain MRI and Kidney CT datasets, achieving a 60-70% improvement in retrieval accuracy for adversarial queries and an overall 90% retrieval accuracy, outperforming existing models by 5-10%. The results demonstrate superior performance in terms of both security and retrieval efficiency, making this framework a valuable contribution to the future of secure medical image management.

REFERENCES

[1] P. Gupta, E. M. Marigi, and J. Sanchez-Sotelo, ‘‘Research on artificial intelligence in shoulder and elbow surgery is increasing,’’ JSES Int., vol. 7, no. 1, pp. 158–161, Jan. 2023, doi: 10.1016/j.jseint. 2022.10.004.
[2] H. Baid, J. Holland, and F. Pirro, ‘‘Environmentally sustainable orthopaedics and trauma: Systems and behaviour change,’’ Orthopaedics Trauma, vol. 36, no. 5, pp. 256–264, Oct. 2022, doi: 10.1016/j.mporth.2022.07.002.
[3] X. Yang, D. Zhao, F. Yu, A. A. Heidari, Y. Bano, A. Ibrohimov, Y. Liu, Z. Cai, H. Chen, and X. Chen, ‘‘Boosted machine learning model for predicting intradialytic hypotension using serum biomarkers of nutrition,’’ Comput. Biol. Med., vol. 147, Aug. 2022, Art. no. 105752, doi: 10.1016/j.compbiomed.2022.105752.
[4] D. S. Overstreet, L. J. Strath, M. Jordan, I. A. Jordan, J. M. Hobson, M. A. Owens, A. C. Williams, R. R. Edwards, and S. M. Meints, ‘‘A brief overview: Sex differences in prevalent chronic musculoskeletal conditions,’’ Int. J. Environ. Res. Public Health, vol. 20, no. 5, p. 4521, Mar. 2023, doi: 10.3390/ijerph20054521.
[5] Y. Sato, Y. Takegami, T. Asamoto, Y. Ono, T. Hidetoshi, R. Goto, A. Kitamura, and S. Honda, ‘‘Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: A multicenter study,’’ BMC Musculoskeletal Disorders, vol. 22, no. 1, p. 407, May 2021, doi: 10.1186/s12891-021-04260-2.
[6] Z. He, S. Wang, and D. Liu, ‘‘A degradation modeling method based on artificial neural network supported tweedie exponential dispersion process,’’ Adv. Eng. Informat., vol. 65, May 2025, Art. no. 103376, doi: 10.1016/j.aei.2025.103376.
[7] V. Bousson, N. Benoist, P. Guetat, G. Attané, C. Salvat, and L. Perronne, ‘‘Application of artificial intelligence to imaging interpretations in the musculoskeletal area: Where are we? Where are we going?’’ Joint Bone Spine, vol. 90, no. 1, Jan. 2023, Art. no. 105493, doi: 10.1016/j.jbspin.2022.105493.
[8] G. Hajianfar, M. Sabouri, S. Bagheri, Y. Salimi, M. Oveisi, I. Shiri, and H. Zaidi, ‘‘Dual input scintigraphy image-based fused deep neural networks for bone abnormalities detection and differentiation,’’ in Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC), Oct. 2021, pp. 1–3, doi: 10.1109/NSS/MIC44867.2021.9875765.
[9] M. Sufiyan, A. Haleem, S. Khan, and M. I. Khan, ‘‘Evaluating food supply chain performance using hybrid fuzzy MCDM technique,’’ Sustain. Prod. Consumption, vol. 20, pp. 40–57, Oct. 2019, doi: 10.1016/j.spc.2019.03.004.
[10] I. M. Wani and S. Arora, ‘‘Computer-aided diagnosis systems for osteoporosis detection: A comprehensive survey,’’ Med. Biol. Eng. Comput., vol. 58, no. 9, pp. 1873–1917, Sep. 2020, doi: 10.1007/s11517-020- 02171-3.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503223|

[11] K. Ren, T. Zheng, Z. Qin, and X. Liu, ''Adversarial attacks and defenses in deep learning,'' Engineering, vol. 6, no. 3, pp. 346–360, Mar. 2020, doi: 10.1016/j.eng.2019.12.012.
[12] J. Huang, H. Shen, J. Wu, X. Hu, Z. Zhu, X. Lv, Y. Liu, and Y. Wang, ''Spine explorer: A deep learning based fully automated program for efficient and reliable quantifications of the vertebrae and discs on sagittal lumbar spine MR images,'' Spine J., vol. 20, no. 4, pp. 590–599, Apr. 2020, doi: 10.1016/j.spinee.2019.11.010.
[13] A. P. Kurmis and J. R. Ianunzio, ''Artificial intelligence in orthopedic surgery: Evolution, current state and future directions,'' Arthroplasty, vol. 4, no. 1, p. 9, Mar. 2022, doi: 10.1186/ s42836-022-00112-z.
[14] B. U. H. Sheikh, ''Mitigating adversarial threats in deep CT image diagnosis models via a dual-stage inference-time defense,'' Appl. Soft Comput., vol. 163, Sep. 2024, Art. no. 111909, doi: 10.1016/j.asoc.2024.111909.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details