scientific reports

Check for updates

OPEN

Enhancing bone cancer detection through optimized pre trained deep learning models and explainable AI using the osteosarcoma tumor assessment dataset

Bolleddu Devananda Rao[1] & K. Madhavi[2]

Diagnosis of bone cancer using histopathology images is essential for effective and timely treatment. However, contemporary diagnostic methods struggle to achieve high accuracy and interpretability while utilizing computational methods. Although existing methodologies in deep learning are promising, each suffers from significant limitations that arise from fundamental challenges in hyperparameter optimization, explainability, and generalizability across disparate datasets. Such disadvantages serve as barriers to clinical use, underscoring the need for a more reliable and comprehensible diagnostic framework. In this study, an Optimized Deep Learning Framework for Bone Cancer Detection (ODLF-BCD) algorithm is proposed by jointly combining Enhanced Bayesian Optimization (EBO), deep transfer learning from state-of-the-art pre-trained models (i.e., EfficientNet-B4, ResNet50, DenseNet121, InceptionV3, and VGG16), and explainable artificial intelligence, namely Grad-CAM and SHAP. It mitigates the state-of-the-art limitations through hyperparameter tuning, increased transparency, and data augmentation to balance the dataset. Extensive experiments verify the effectiveness of the proposed framework, where EfficientNet-B4 achieves 97.9% and 97.3% for binary and multi-class classification, respectively. Its performance is also confirmed with high precision, recall, and F1 score. Explainability facilitates the clinical interpretability of model predictions. Then, the proposed framework offers a robust and efficient alternative solution to the C-RAD, automating bone cancer diagnosis and enhancing the accuracy and transparency of the diagnosis. Its potential usefulness could provide clinicians with strong decision support systems for early and precise cancer detection.

Keywords Bone cancer detection, Deep learning, Explainable AI, Transfer learning, Enhanced Bayesian optimization

Bone cancer is one of the aggressive medical diagnostic challenges with high clinical significance due to the urgent requirement for timely and precise detection. Recent advancements in deep learning and medical imaging technologies enable automated cancer diagnosis, which has opened up avenues to enhance the accuracy of diagnoses while quantifying the manual effort invested by clinicians. The literature has examined deep learning as a model for histopathology image analysis, labeling it as a potential candidate for reducing errors in cancer diagnosis. For instance, Vandana and Sathyavathi[1] utilized CNN-based models to analyze cancer tissues, achieving appreciable performance in identifying cancerous tissues. Similarly, Anisuzzaman et al.[4] showed the effectiveness of pre-trained models (e.g., Inception V3 and VGG19) in capturing complex features in histopathological images by employing transfer learning. Ahmed et al.[10] proposed a compact CNN model for histopathology images, which shows promising diagnostic performance. Bottom line: Despite these advancements, current state-of-the-art approaches face significant challenges. Because most models lack optimal hyperparameter tuning, their performance is hindered from reaching its full potential. Moreover, the explainability is often poor, failing to enable clinicians to rely on model predictions. Another significant limitation

[1]MLR Institute of Technology, JNTUA, Hyderabad 500043, India. [2]Department of CSE, JNTUA College of Engineering, Jawaharlal Nehru Technological University, Anantapur, Ananthapuramu 515 002, India. ✉email: dev.bolleddu@gmail.com

is the inability to generalize, where models fail to transfer to heterogeneous datasets. These obstacles underscore the need for an optimized framework that incorporates rigorous hyperparameter optimization, enhanced interpretability through explainable AI methods such as Grad-CAM and SHAP, and superior adaptability to diverse data distributions, thereby facilitating accurate clinical diagnosis.

This work is motivated by the demand for more efficient and interpretable frameworks for deep learning. Existing approaches not only fail in hyperparameter tuning and class balancing via data augmentation but also in explainable AI, as they lack clarity on how model predictions can be reasoned. Moreover, investigations are primarily single modality, which limits their usability in clinical practice scenarios. These challenges must be overcome before reliable and clinically valuable diagnostic tools can be available. The primary objective of this study is to develop an Optimized Deep Learning Framework for Bone Cancer Detection (ODLF-BCD) that achieves significant accuracy and interpretability in both binary and multi-class classification tasks. This framework of ideas includes Enhanced Bayesian Optimization (EBO) for hyperparameter tuning, transfer learning from pre-trained models, and explainable AI (using methods that allow for visualizing the contribution of different parts of the model (Grad-CAM and SHAP), making our predictions more interpretable). Such innovations are designed to address the most critical gaps in the state of the art: generalization, transparency of forecasts, and optimal model performance.

Although this study employs Grad-CAM and SHAP to increase model transparency, the distinction between explainability and interpretability should be considered. Explainability refers to technical approaches (e.g., heatmaps, attribution scores) that clarify how a machine learning model operates internally or how it makes predictions, rather than solely to secure validation from domain experts. Interpretability, on the other hand, refers to the level of understanding a human—and specifically, a clinician—can reach regarding the model's outputs, and whether that human will trust those outputs and act accordingly in a real-world context. Designed to be explainable, our framework serves as a bridge to interpretability, facilitating downstream clinical validation and decision support.

The key contributions of this research include the design and implementation of ODLF-BCD, the characterization of its effectiveness on histopathology images, and the provision of interpretability for clinical validation. The results are as follows: EfficientNet-B4 is the highest-performing model, with 97.9% accuracy in binary classification and 97.3% in multi-class classification. The framework also provides insights into hyperparameter optimization and explainability in automated diagnostics. The rest of this paper is organized as follows: "Related work" reviews related work and indicates gaps/challenges in existing approaches. "Proposed framework" describes the proposal methodology, including data preprocessing, model training, and optimization methods. The experimental results and performance analysis of the proposed framework are presented in "Experimental results". "Discussion" presents the results, describes the value of the study, and acknowledges its limitations. Finally, "Conclusion and future work" concludes this work and outlines future directions for multi-modal approaches and novel explainability techniques.

Related work

The literature highlights advancements in deep learning for bone cancer detection, but identifies challenges related to optimization, explainability, and generalization.

Deep learning models for bone cancer detection

This subsection utilizes deep learning models, including CNNs, ResNet, and EfficientNet, for bone cancer detection. Vandana and Sathyavathi[1] enhanced bone cancer diagnosis using deep learning and image processing, achieving a 92% accuracy rate. For future work, larger datasets, greater automation, and more feature research are needed. Anand et al.[2] improved bone cancer diagnosis with better accuracy using deep learning and image processing. In the future, Bayesian networks and sophisticated classifiers will be combined for increased precision. Anisuzzaman et al.[4] made use of CNNs such as Inception V3 and VGG19 to identify osteosarcoma in histology pictures. Although it attained good accuracy, more generalization testing and pathologist confirmation are required. Punithavathi and Madhurasree[6] designed an extended convolutional neural network (ECNN) with wavelet-based segmentation, resulting in high-performance metrics for distinguishing between bone malignancies. It is a job for the future to improve segmentation methods. Ahmed et al.[10] designed a compact CNN model to handle class imbalance by oversampling osteosarcoma histology images. Generalization and dataset extension require more work, even though they decrease overfitting and increase accuracy.

Tang et al.[15] enhanced training datasets to boost the model's generalization, highlighting the risk of overfitting in deep learning models for osteosarcoma diagnosis due to limited data heterogeneity. Alabdulkreem et al.[23] utilized Inception v3 and LSTM to develop an OSADL-BCDC model for bone cancer diagnosis, achieving a higher accuracy rate. The primary objectives of future research will be to integrate multiple imaging modalities and develop explainable AI models. Anisuzzaman et al.[25] used CNNs and CAD tools to enhance the identification of osteosarcoma. Pathologist comparisons and increasing dataset generalizability are areas of further exploration. Suganeshwari et al.[30] proposed a deep transfer-based method with an accuracy of 93.9% for bone cancer diagnosis, utilizing VGG16 and SVM. Further research will investigate different imaging modalities and enhance prediction using larger datasets. Bansal et al.[36] improved the IF-FSM-C model by incorporating features from manual and deep learning methods, which enables it to pick features and identify osteosarcomas with increased accuracy. Effective deep-learning models will be investigated in more detail. Anand et al.[44] proposed the Convolutional Extreme Learning Machine (CELM), which has improved the accuracy rate in diagnosing bone cancer from histopathology images. Utilizing intricate, deep learning architectures, further research aims to enhance this. Ranjitha et al.[50] enhanced the ability to identify bone cancer by applying KNN and K-means classifiers to ultrasound images processed with image processing. With restricted feature selections and picture quality, it still achieves higher accuracy. Larissa Y. Asito et al.[51] explored the use of pre-trained convolutional neural networks to assess bone scans for detecting metastasis. The study highlights the potential of deep learning models in medical image analysis for improved diagnostic accuracy.

Transfer learning and pre-trained models in medical imaging

This subsection highlights the role of transfer learning and pre-trained models, such as VGG16, InceptionV3, and DenseNet, in histopathology analysis. Anand et al.[5] recommended using the highly accurate Rec-CONVnet algorithm for identifying bone cancers from MRI data. Future research to improve classification accuracy will center on developing 3D neural networks. Alsubai et al.[9] Osteosarcoma identification is enhanced by the GTOADL-ODHI technique, which uses AI with GF preprocessing, CapsNet feature extraction, and SA-BiLSTM classification. Future goals include enhancing generalization and increasing sample size. Aziz et al.[20] developed a hybrid model, incorporating CNNs and MLPs, along with feature selection, which classifies osteosarcoma with improved accuracy. In the future, uncertainty mining will be used to improve dependability. Shukla and Patel[24] evaluated image segmentation methods (K-means, region growth) for X-ray osteosarcoma detection and applied deep learning to cancer type prediction. Large-scale dataset training will be a part of future development. Vedakis et al.[32] examined the deep understanding of osteosarcoma classification, and EfficientNetB0 and MobileNetV2 were determined to be the most effective models. The disadvantages are the small dataset size and the need to modify the hyperparameters. Future studies should utilize more extensive and diverse datasets, as well as advanced regularization techniques.

Papandrianos et al.[37] developed a CNN model for detecting bone metastases in breast cancer, which achieved high accuracy using RGB photos. Future research will primarily focus on improving interpretability and integrating the model into healthcare systems. Papandrianos et al.[41] created CNN models with exceptionally high accuracy when utilizing scintigraphy pictures to identify bone metastases in prostate cancer patients. Improving model interpretability and incorporating new datasets will be the primary focus areas for future development. Krois et al.[48] A deep CNN, utilizing highly competent dentists, accurately detected periodontal bone loss 81% of the time. Future research to enhance the dependability and performance of CNNs could incorporate additional data. Barhoom et al.[52] modified the VGG16 model for categorizing bone anomalies from X-rays presented in the paper. Although feature selection and ensemble strategies require further exploration, the method demonstrates excellent accuracy. Shao et al.[55] contrast this with conventional scans, showing how CNNs with SERS data can be used to quickly and non-invasively identify prostate cancer bone metastases. Future studies will involve expanding databases to enhance accuracy. Alkhalaf et al.[60] presented AAOXAI-CD, which uses medical imaging to transparently diagnose cancer by fusing XAI with deep learning. Future work will involve adding feature fusion to the model. Li et al.[63] concluded that, despite sample size restrictions, the Deep Belief Network (DBN) performs better than other algorithms in predicting lung metastasis and overall survival in patients with osteosarcoma.

Explainable AI and interpretability in bone cancer diagnosis

This subsection discusses the importance of explainable AI techniques, such as Grad-CAM, SHAP, and feature visualization, for enhancing interpretability in bone cancer detection. Saranya et al.[11] proffered deep learning techniques that could detect fibrous dysplasia in bone images with higher accuracy. Improved multi-class classification and noise reduction will be the main goals of future work. Ramasamy et al.[14] identified AML and MM as bone marrow cancers; this work presents a hybrid deep learning algorithm that incorporates cell segmentation and classification. Jiang et al.[17] Demonstrated That Deep learning enables significantly improved medical imaging utilization for cancer diagnosis. Two challenges are applying the model to rare cancers and the caliber of the datasets. These flaws have to be addressed in further work, along with more model openness. Chianca et al.[18] suggest that it may be possible to increase the diagnostic accuracy of spinal lesions to 94% by utilizing radiomics and machine learning. The problems of feature stability, data scarcity, and program uniformity will require further effort. Georgeanu et al.[26] achieved high accuracy in predicting the malignancy of bone tumors using pre-trained ResNet-50 CNNs with MRI data. More extensive dataset testing and model generalizability enhancement are the focus of future development. Kanimozhi et al.[33] studied methods for identifying bone cancer using feature extraction and deep learning. Short datasets and feature overlap between cancerous and healthy cells are two drawbacks. Future research should focus on adding multi-feature extraction and diversifying datasets to increase accuracy.

Sharma et al.[43] developed an automated method for diagnosing bone cancer using texture analysis and machine learning. The SVM with HOG features produced an F1 score of 0.92. Future studies should investigate additional texture elements to enhance accuracy further. Sindudevi and Kavita[53] assessed CNNs for the early diagnosis of bone cancer, comparing their performance with that of conventional techniques. Upcoming projects will focus on improving data processing and optimizing the CNN model. Saba[57] examined machine-learning approaches to cancer diagnosis, encompassing various types of cancer. It emphasizes the need for larger datasets and greater accuracy, while highlighting recent developments, limitations, and the challenges they pose. Prathyusha and Gowri Sankar Reddy[58] presented a convolutional neural network-based approach for detecting bone cancer. Their study, published in the Journal of Emerging Technologies and Innovative Research (JETIR), emphasizes the efficacy of CNNs in medical diagnosis. Tang et al.[64] presented the OMSAS system with ACRNet for effective MRI segmentation of osteosarcomas, demonstrating less complexity and increased accuracy. Future research aims to increase model capabilities and datasets. Srinidhi et al.[65] and Chowdhury et al.[66] proposed a federated learning approach combined with deep feature extraction and MLP for detecting osteosarcoma from histopathological images, highlighting privacy-preserving AI in medical diagnostics.

Eweje et al.[67] utilized deep learning to classify bone lesions on routine MRI scans, showcasing its potential to enhance diagnostic accuracy in bone lesion evaluation. Examined deep learning techniques for histopathological image interpretation, emphasizing their uses, difficulties, and areas that require more investigation, such as model interpretability and dataset accessibility. Ong et al.[68] reported that AI methods have performed well in differentiating between benign and malignant bone lesions across several imaging modalities. Subsequent studies should examine small sample sizes and corroborate findings using data from multiple centers.

Data augmentation, class balancing, and dataset challenges

This subsection addresses challenges such as class imbalance and small dataset sizes, and proposes solutions through data augmentation or synthetic data generation. Nasir et al.[3] presented a deep learning algorithm that utilizes fog, edge, and blockchain technologies to enhance the accuracy of osteosarcoma diagnosis. New deep learning model optimization will be the primary focus of future research. Shrivastava et al.[7] employed machine learning to improve the detection of bone cancer using CT and MRI images. Future studies should prioritize the management of extensive datasets, accuracy enhancement, and the integration of molecular signatures. Rahouma and Abdellatif[13] recommended using GLCM for feature extraction and ensemble classifiers to create an automated model for osteosarcoma diagnosis. High-quality photography is required, even if it yields better accuracy. Badashah et al.[19] employed, in conjunction with Fractional-Harris Hawks Optimization, a novel GAN technique that enhances the accuracy of osteosarcoma diagnosis in the early stages of the disease. To achieve better outcomes, future research will incorporate deep learning. Sampath et al.[22] developed a higher-accuracy CNN-based model using AlexNet to classify various forms of bone cancer from CT scans, employing image processing techniques. Further research will employ more CNNs, such as ResNet and DenseNet, for improved feature extraction and classification.

Altameem et al.[27] developed an automated system that enhances the accuracy of bone cancer detection using deep neural networks and intuitionistic fuzzy correlation. Prediction method optimization is the goal of future development. Zhao et al.[34] enhanced efficiency; the AI model for bone scintigraphy supports physicians by demonstrating a high degree of accuracy in cancer detection. The disadvantages are the lack of a lesion-based prognostic analysis and the requirement for multi-center validation. Cheng et al.[40] minimized false positives while achieving high sensitivity and accuracy in creating a YOLO v4 model for bone scintigraphy-based early diagnosis of bone metastasis in prostate cancer. Future studies should address the dataset's limitations and enhance the model's generalizability. Yadav and Rathor[46] demonstrated a deep neural network that can automatically identify bone fractures from X-rays with 92.44% accuracy. The following challenges include increasing accuracy and verifying larger datasets. Manjula et al.[49] aimed to enhance the application of medical imaging and deep learning in the identification of bone cancer. It achieves better accuracy but requires more fine-tuning and high-resolution photos. Xiong et al.[56] utilized CT scans to create and verify a deep-learning model that distinguishes between osteoblastic metastases and bone islands. Expanding the types of lesions and addressing model constraints are the goals of future development. Satheeshkumar and Sathiyaprasad[59] suggested combining decision trees with the Gray-Level Co-occurrence Matrix (GLCM) and K-NN for identifying bone cancer, demonstrating increased accuracy. Classifying different kinds of bone tumors is a task for the future.

Comparative studies and performance enhancements

This subsection reviews comparative studies of various deep-learning models and techniques for performance enhancement in bone cancer detection. Nabid et al.[8] proposed that the RCNN model for osteosarcoma identification is limited by the small datasets, despite outperforming other methods. Subsequent investigations aim to enhance segmentation and broaden the model's range of applications. Acunto et al.[12] obtained high precision when using deep learning to distinguish osteosarcoma cells from MSCs. Research endeavors aim to enhance digital pathology and expand this methodology to larger tissue samples. Mulhim and Haque[16] improved the diagnosis of multiple myeloma using deep learning, with models such as VGG and ResNet achieving better accuracy. Future updates to the model will include other diseases, and data variability will be optimized. Gawade et al.[21] employed CNN-based models with ResNet101 to identify osteosarcoma with higher accuracy. In the future, other models, such as Xception and EfficientNet, will be investigated to increase accuracy further. Lin et al.[28] achieved high accuracy in metastasis diagnosis, and this work provides deep classifiers for automated SPECT bone image processing. Future research goals include improving network architecture, multi-class categorization, and dataset size. Bhukya Jabber et al.[29] proposed an SVM-based computerized model for bone cancer detection, focusing on efficient and accurate classification techniques. The study was presented at the 4th International Conference on Electronics, Communication, and Aerospace Technology (ICECA).

Hsieh et al.[31] achieved improved accuracy in bone metastasis diagnosis by utilizing deep learning approaches, such as contrastive learning. The performance of various cancer types and treatments should be examined in future research, and several locations should validate these findings. Chu and Khan[35] employed a combination of deep learning, transfer learning, and data augmentation to achieve a diagnosis accuracy of 91.18% for osteosarcoma. Among the restrictions are the need for larger validation sets and dataset imbalance. Huo et al.[38] developed a DCNN model to detect lung cancer bone metastases on CT scans with higher sensitivity and less clinical time. Future work will entail expanding the diagnostic characteristics and validating larger, multi-center datasets. Gusarev et al.[39] enhanced disease categorization and nodule identification by creating two topologies for chest radiograph bone suppression. Subsequent studies seek to explore more effective metrics and to improve loss functions. Do et al.[42] developed an accurate Multi-Level Seg-Unet model to detect and segment knee bone tumors. Graph convolution will be utilized in future research to enhance our understanding of bone form and improve both global and patch-based models. Kiresur and Manoj[45] presented a cost-effective deep-learning method for detecting bone cancer using X-ray images. Future studies aim to achieve better precision and accuracy in early diagnosis. Giradkar and Bodne[47] examined bone stress injuries using MRI and segmentation techniques, discovering that many of the lesions are asymptomatic. Subsequent investigations endeavor to enhance precision by optimizing semi-supervised learning methodologies.

Papandrianos et al.[37] investigated a CNN-based technique that achieves 92.50% accuracy in identifying bone metastases in breast cancer using bone scintigraphy. Future studies will improve interpretability and incorporate the model into CAD tools. The survey by Sivakumar et al.[54], which preprocesses and categorizes images using CNNs and genetic algorithms, enhances the detection of bone cancer. Upcoming projects will focus on improving accuracy, combining with other imaging modalities, and using telemedicine. Lopez et al.[61] presented PROMETEO, a customized CNN architecture that outperforms other models in terms of speed and accuracy for the diagnosis of prostate cancer. Custom designs and edge computing should be explored in future development. Rytky et al.[62] developed a machine-learning approach for automating 3D histopathological grading of osteoarthritis using contrast-enhanced micro-computed tomography. Their method demonstrates the potential for precise and efficient analysis of osteochondral tissue. Tufail et al.[69] applied sophisticated deep learning (DL) models to improve cancer detection and prognosis. Subsequent studies should concentrate on innovative designs and clinical validation, enhance data processing, and overcome model constraints. Mandala et al.[70] propose a novel machine-learning classifier for detecting oropharyngeal cancer, emphasizing improved accuracy and diagnostic efficiency. Existing studies underline the potential of deep learning in bone cancer detection, but reveal gaps in hyperparameter tuning, dataset limitations, and model interpretability. This research addresses these gaps through an optimized framework incorporating Enhanced Bayesian Optimization, transfer learning, and explainable AI, achieving superior accuracy and transparency for reliable automated cancer diagnostics.

Proposed framework

The methodology for this research, illustrated in Fig. 1, details the development and implementation of an optimized deep learning framework for bone cancer detection using the Osteosarcoma Tumor Assessment dataset. The dataset was preprocessed to standardize and enhance the quality of the input data. All images were resized to 224 × 224 pixels to ensure compatibility with the selected pre-trained models. Pixel values were normalized to the range [0, 1], enabling consistent data scaling. Data augmentation techniques, including random rotations, horizontal and vertical flips, and contrast adjustments, were applied to increase the diversity of the training data, thereby enhancing the model's generalization capability.

The study utilized five state-of-the-art pre-trained deep learning models: ResNet50, EfficientNet-B4, DenseNet121, InceptionV3, and VGG16. These models were initialized with ImageNet weights, providing a robust foundation for transfer learning. The initial layers responsible for generic feature extraction were frozen. In contrast, the higher layers were fine-tuned to adapt the models for the specific task of classifying viable and necrotic tumors. This fine-tuning process involved replacing the original classification layers with a custom head comprising a global average pooling layer, fully connected dense layers activated by ReLU, and a softmax output layer tailored for the target classes.

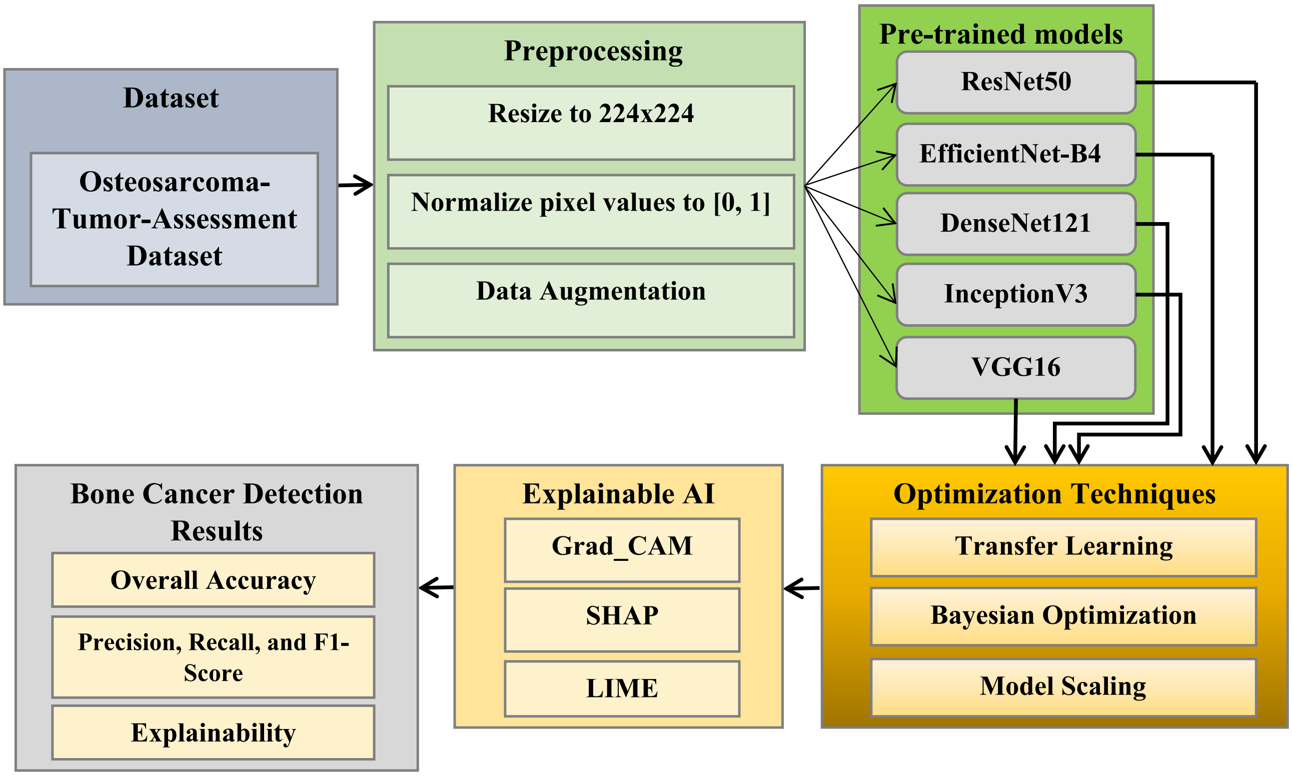


Fig. 1. Methodology workflow for bone cancer detection using optimized pre-trained deep learning models and explainable AI.

Optimization techniques played a crucial role in enhancing the models' performance. Bayesian optimization was employed to identify optimal hyperparameters, including learning rate, batch size, the number of neurons in dense layers, and dropout rates. The Adam optimizer, combined with a learning rate scheduler, ensured efficient and stable convergence. Model scaling techniques were systematically explored, particularly with EfficientNet, where variations in network depth, width, and input resolution (e.g., 224 × 224 and 384 × 384) were tested to identify the optimal configuration. Explainable AI (XAI) methodologies were integrated to enhance the interpretability of model predictions, as shown in Fig. 1. Grad-CAM-generated heat maps visually demonstrated the regions within the tumor images that contributed most significantly to the model's predictions. SHAP provided quantitative insights into the importance of features for individual predictions, while Local Interpretable Model-agnostic Explanations (LIME) offered localized explanations, validating the consistency of the model outputs. These techniques ensured that the models focused on clinically relevant areas, fostering trust and reliability in their application. Performance evaluation was conducted using a comprehensive suite of metrics, including accuracy, precision, recall, F1 Score, and ROC AUC. The results demonstrated significant improvements in classification performance and interpretability compared to baseline approaches. Grad-CAM heatmaps consistently highlighted tumor regions in alignment with clinical observations, validating the clinical applicability of the framework. This study underscores the effectiveness of combining optimized pre-trained deep learning models with Explainable AI for automated, accurate, and interpretable bone cancer detection. Table 1 shows the notations used in the proposed methodology.

Preprocessing

The preprocessing methodology for this research involved rigorous steps to ensure the osteosarcoma tumor assessment dataset was prepared effectively for deep learning model training. Each image in the dataset was resized to a uniform resolution of 224×224 pixels to maintain consistency with the input requirements of the pre-trained models. This resizing was performed using bilinear interpolation, ensuring minimal distortion and preserving the essential features of the tumor images. Mathematically, the y prime close paren, of the resized image was calculated as a weighted sum of neighboring pixel values cap I, open paren x, y close paren, fromixel values $I\left(x,y\right)$ from the original image as in Eq. (1).

$$I\prime\left(x\prime,y\prime\right)=\sum_{i=0}^{1}\sum_{j=0}^{1}w_{i,j}.I\left(x+i,y+j\right),\tag{1}$$

Symbol	Description
I(x, y)	Pixel intensity value at coordinates (x, y) in the original image
I′(x′, y′)	Pixel intensity value at coordinates (x′, y′) in the resized image
$I_{norm}\left(x,y\right)$	Normalized pixel intensity value at coordinates (x, y)
I_{max}	Maximum possible pixel intensity value (e.g., 255 for 8-bit images)
x	Input feature vector or image representation fed to the model
y	Output prediction vector or activation map from the model
W	Weights of a convolutional filter in a neural network layer
b	Bias term associated with a convolutional filter
σ	Activation function (e.g., ReLU)
F(x, {w_i})	The residual function in ResNet computes intermediate feature maps
d	Depth scaling factor in EfficientNet
w	Width scaling factor in EfficientNet
r	Input resolution scaling factor in EfficientNet
α, β, γ	Constants for depth, width, and resolution scaling, respectively
ϕ	Compound coefficient controlling scaling in EfficientNet
f(x)	Objective function evaluated during Bayesian Optimization
f*	Best observed objective value during Bayesian Optimization
η_t	Learning rate at epoch t
η_0	Initial learning rate
T	Total number of training epochs
EI(x)	Expected improvement for a given hyperparameter configuration x
x_l	The output of layer l in DenseNet incorporates features from all preceding layers
H_l	The transformation function is applied to layer l in DenseNet
Grad-CAM(x)	Gradient-weighted activation map for input x, highlighting important regions
SHAP(x)	Shapley values for input x, quantifying feature contributions
LIME(x)	Local interpretable model approximation for input x

Table 1. Notations used in the proposed methodology.

where $w_{i,j}$ represents the interpolation weights based on the distances from the original pixel coordinates to the resized coordinates. Pixel intensity values were normalized to tOpenange. $[0,1]$ by diby the maximum possible intensity value, cap I open paren x, y close paren, by the maximum possible intensity value, cap I end subscript sub, m a. x, end subscript, whichy value $I_{max,}$ which is 255 for 8-bit grayscale images. Normalization is carried out as expressed in Eq. (2).

$$I_{norm}\left(x\prime, y\prime\right) = \frac{I\left(x\prime, y\prime\right)}{I_{max}} \tag{2}$$

This normalization ensured that the pixel values were scaled consistently across all images, facilitating faster model convergence during training. Data augmentation techniques were systematically applied to enhance the robustness and generalization ability of the models. These included random horizontal and vertical flips, rotations within the range $[-20^{\circ}, 20^{\circ}]$, and random brightness adjustments by a factor drawn from a uniform distribution in the range $[0.8, 1.2]$. These augmentations introduced variability into the training data, mitigating overfitting and enabling the models to learn invariant features. The augmentation transformations were represented as affine transformations applied to the input image matrix *I*, such that the transformed image I_{aug} was derived as in Eq. (3).

$$I_{aug} = A.I_{norm} + b, \tag{3}$$

where A is the transformation matrix (encoding rotation, scaling, and flipping) and b is the translation vector. Additionally, the dataset was split into training, validation, and test sets in the ratio of $70:20:10$, ensuring a balanced representation of viable and necrotic tumor images across the splits. Stratified sampling was used to preserve the proportion of classes in each subset, enhancing the statistical reliability of the evaluation metrics. These preprocessing steps ensured the consistency and quality of the input data and augmented the dataset's diversity, thereby enabling the deep-learning models to achieve superior performance in bone cancer detection.

Pre-trained deep learning models

This study leverages pre-trained deep learning models, including EfficientNet-B4, ResNet50, DenseNet121, InceptionV3, and VGG16, initialized with ImageNet weights. These models were fine-tuned using transfer learning to adapt to the bone cancer detection task. Their advanced architectures provide robust feature extraction, enabling high accuracy and reliability for binary and multi-class classification.

ResNet50

ResNet50 (Residual Network with 50 layers) was selected for its ability to mitigate the vanishing gradient problem through residual connections. The architecture introduces skip connections, allowing gradients to flow directly through the network's deeper layers. Each residual block computes the output as in Eq. (4).

$$y = \mathcal{F}\left(x, \{W_i\}\right) + {\scriptstyle x}, \tag{4}$$

where x is the input, $\mathcal{F}$ represents the residual function (a series of convolutional layers), and W_i are the weights of the layers within the block. For this study, the initial layers of ResNet50 were frozen to preserve pre-trained ImageNet features. In contrast, the fully connected layers were replaced with a global average pooling layer and a dense layer with a softmax activation for multi-class classification. Fine-tuning was performed on the higher layers with a learning rate of 10^{-4} to adapt the model for osteosarcoma tumor classification.

EfficientNet-B4

EfficientNet-B4, a model based on compound scaling, was employed because it balances network width, depth, and resolution. Equation (5) defines the scaling.

$$d = \alpha^{\varnothing}, w = \beta^{\varnothing}, r = \gamma^{\varnothing}, \tag{5}$$

where $d, w,$ and r are the depth, width, and resolution scaling factors, respectively;$\varnothing$, is a user-defined scaling coefficient; and α, β, γ are constants determined through grid search. EfficientNet-B4's resolution of $380 \times 38\text{was tested alongside}224 \times 224,$ with experiments showing improved performance on higher-resolution images. This model's combination of depth-wise separable convolutions and squeeze-and-excitation layers significantly reduced computation while maintaining accuracy, making it a robust choice for this study.

EfficientNet-B4 outperforms all other results due to its compound scaling that methodically (not casually) combines network depth, width, and resolution. This architecture enables EfficientNet-B4 to capture multi-scale tumor features more efficiently and with better generalization to the complex variations (intra- and inter-tumoral heterogeneity) of histopathological tumor images compared to other pre-trained models.

The superior performance of EfficientNet-B4 is primarily due to its compound scaling strategy, which consistently and systematically scales three dimensions of the model—depth (number of layers), width (number of channels), and resolution (input image size)—together, rather than independently. The EfficientNet-B4 achieves this balance, allowing it to efficiently extract various hierarchical and multi-scale features where subtle texture patterns play a crucial role and contextual shapes and spatial details matter (as you might recall, in histopathological images, subtle texture and patterns can be pretty critical). EfficientNet also employs a method of scaling similar to traditional models, but where they might scale only a single dimension (for example, VGG16 increases depth). EfficientNet, however, compounds scales based on a principled formula (see Eq. 5), which balances accuracy and runtime cost. Additionally, EfficientNet-B4 utilizes depthwise separable convolutions and a squeeze-and-excitation module to reduce the parametric size and enhance feature representation. An ablation study, illustrated in Table 4 and Fig. 12, supported this, as the model's accuracy improved significantly from 96.5 to 97.9% with higher input resolution (380 × 380), demonstrating the model's powerful ability to capture fine-grained tumor characteristics that other models could not readily capture.

DenseNet121

DenseNet121 employs dense connections to ensure maximum feature reuse and gradient flow across layers. Each layer in DenseNet receives feature maps from all preceding layers, expressed mathematically as in Eq. (6).

$$x_l = H_l\left([x_0, x_1, \ldots, x_{l-1}]\right), \tag{6}$$

where x_l is the output of the $l - th$ layer, H_l is the transformation (comprising batch normalization, ReLU activation, and convolution), and [•] denotes the concatenation operation. For this study, DenseNet121's compact structure enabled efficient training on the Osteosarcoma dataset. The dense connections enhanced feature propagation and facilitated faster convergence during fine-tuning.

InceptionV3

InceptionV3, designed with inception modules, excels at multi-scale feature extraction by combining convolutional layers of varying kernel sizes. An inception module computes its output as in Eq. (7).

$$\mathrm{Y} = [\mathrm{Conv}_{1\times 1}, \mathrm{Conv}_{3\times 3}, \mathrm{Conv}_{5\times 5}, \mathrm{MaxPool}] * \mathrm{x}, \tag{7}$$

where [•] indicates the concatenation of feature maps from different kernel sizes. This multi-branch architecture was particularly effective in capturing both global and local features of osteosarcoma tumor regions. Transfer learning was performed by replacing the classification head with layers fine-tuned on the dataset, improving sensitivity to tumor-specific features.

VGG16

VGG16 is a deep convolutional network characterized by its simplicity and use of small 3×3 convolutional kernels stacked sequentially. The output of each convolutional layer is computed as in Eq. (8).

$$Y = \sigma\left(W * x + b\right), \tag{8}$$

where W represents the kernel weights, x is the input, b is the bias term, and σ is the activation function (ReLU). Despite its simplicity, VGG16 is highly effective for feature extraction, making it a suitable baseline for this study. To adapt to the classification task, the fully connected layers of VGG16 were replaced with a custom dense architecture, and fine-tuning was applied to improve performance on the specific dataset.

Transfer learning

The above work utilized transfer learning to modify pre-trained networks (ResNet50, EfficientNet-B4, DenseNet121, InceptionV3, VGG16) for the osteosarcoma tumor classification task. Pre-trained on the ImageNet dataset, every model defined a good set of generic features and was thus fine-tuned for domain-level learning. In ResNet50, since the base convolutional layers are already trained to extract well-defined feature vector patterns, higher layers are composed and fine-tuned for tumor classification; therefore, the initial convolutional layers are frozen. A light architecture was used to replace fully connected layers for the final classification, including global average pooling layer(s) and dense layers specifically for target classes to offer the right trade-off between computational efficiency and model performance.

Compound scaling enables us to balance network depth, width, and input resolution for optimal performance. We applied the same concept, but tuned two tasks from EfficientNet-B4. The original layers were frozen, and a new top consisted of a global average pooling layer, dropout, and dense layers optimized for multi-class classification. We employed a dynamic learning rate schedule to fine-tune the model's ability to learn domain-specific features effectively. We fine-tuned DenseNet121, a densely connected architecture with feature reuse from every layer through the entire network, by freezing the first set of dense blocks and retraining the final dense blocks and transition layers. We modified the model for the osteosarcoma dataset by adding a custom classification head that included dropout to avoid overfitting.

InceptionV3, renowned for its multi-scale feature extraction capabilities, was enhanced by retraining its top Inception blocks while keeping the lower layers frozen. Such a method maintains the model's capacity to learn both global and local characteristics of tumor images. A dedicated architecture was used as a substitute for the classification head to adapt to the dataset's requirements, achieving increased specificity in identifying tumor patterns. Due to its straightforward nature, VGG16 employs a simple sequence of convolutional layers. In this model, all convolutional layers were frozen, while the fully connected layers were retrained. Due to the simplicity of VGG16, it served as a good baseline model against which to compare, and a new classification head was added for comparison to the target task. The framework applied design-centric transfer learning approaches to each model, benefiting from the unique suitability of these architectures to address the challenges of classifying viable and necrotic tumors. This customization helped improve bone cancer diagnosis performance accurately and reliably.

Enhanced Bayesian optimization for hyperparameter tuning

All five pre-trained models—namely, ResNet50, EfficientNet-B4, DenseNet121, InceptionV3, and VGG16—were tuned using an Enhanced Bayesian Optimization (EBO) approach to determine the optimal set of hyperparameters for classifying osteosarcoma tumors. Instead, EBO approximates the objective function using a surrogate model, which enables the use of a high-dimensional hyperparameter space with a limited number of samples while maintaining the performance of both exploration and exploitation[6]. This work is considered a Gaussian Process (GP) with an advanced kernel, which enables the modeling of complex relationships, highlighting the general pattern of varying hyperparameters and the model's performance. Hence, it can be used as a surrogate model[19]. An acquisition function (Expected Improvement (EI) function) was defined to optimize the exploration–exploitation tradeoff for the following suggested hyperparameter set to evaluate. In mathematical terms, EI can be defined as in Eq. (9).

$$EI\left(x\right)=\mathbb{E}\left[max\left(\left(0,f\left(x\right)\right)-f^{*}\right)\right] \tag{9}$$

where $f\left(x\right)$ is the surrogate model's prediction for hyperparameters x, and f^{*} is the best observed objective value. The search space for EBO included critical hyperparameters for each model, such as learning rate $10^{-3}\text{to}10^{-6}$, batch size (16, 32, 64), dropout rate (0.2 to 0.5), and the number of neurons in the dense layers (128 to 512). The surrogate model predicted the likely performance of these configurations, enabling efficient testing without exhaustive search. Multi-fidelity optimization was integrated to further enhance the efficiency of EBO, where initial evaluations were conducted on a subset of the training data or fewer epochs, reducing computational overhead. Promising configurations identified in these partial evaluations were then thoroughly evaluated.

For ResNet50, the focus was on optimizing the learning rate and the number of trainable layers to balance generalization and specialization. EfficientNet-B4 benefited from tuning its compound scaling coefficients, which control network depth, width, and resolution. DenseNet121 required adjustments in the number of neurons in the dense layers and the dropout rate to enhance its gradient flow and prevent overfitting. InceptionV3's multi-scale feature extraction capabilities were fine-tuned by adjusting the learning rate and batch size, ensuring efficient utilization of its inception modules. VGG16, with its simple architecture, was tuned for optimal learning rates and the number of neurons in its added fully connected layers to maximize its baseline performance.

The EBO framework dynamically updated the surrogate model after each evaluation, refining the prediction of the objective function. This iterative process continued until convergence criteria were met, such as a predefined number of iterations or minimal improvement in the objective function. The final hyperparameter configurations achieved through EBO resulted in significant improvements in the classification metrics across all five pre-trained models, demonstrating the efficacy of this enhanced optimization approach.

Model scaling

Model scaling played an essential role in this work as it allows the fitting of every pre-trained model to perform best for osteosarcoma tumor classification. We ensured the efficient use of computational resources while maximizing classification accuracy by scaling the models systematically in depth, width, and resolution. Depth scaling was done by adding layers to the network, i.e., making it wider and wider to learn complex features. This was especially true for EfficientNet-B4, which used the compound scaling rule to scale depth. $d=\alpha^{\varnothing}$, where d represents the depth scaling factor, α is a constant determined empirically and $\varnothing$ is the compound coefficient optimized during the process. Increasing depth allowed the model to capture more intricate patterns in the tumor images, improving its sensitivity to subtle features. Width scaling was applied to broaden the layers of specific models by increasing the number of filters in convolutional layers. For instance, in DenseNet121, the width of the layers was increased proportionally, enhancing the network's capacity to extract diverse features from input images.

The width factor, w, was scaled as $w=\beta^{\varnothing}$, where β is a width coefficient. This adjustment enabled the model to better balance generalization and overfitting, particularly on augmented data with varied transformations. Resolution scaling was another essential aspect, wherein the input resolution of the models was systematically increased to allow finer detail capture in the tumor images. EfficientNet-B4 and InceptionV3 benefited significantly from this scaling, as higher resolutions (e.g., 380 × 380 compared to 224 × 224) provided additional context for feature extraction. The resolution factor, r, was scaled as $r=\gamma^{\varnothing}$, where γ represents the resolution coefficient. Higher resolutions were particularly effective for small, intricate regions in the tumor images, improving classification performance without excessively increasing computational cost. By combining these scaling techniques, we systematically optimized each model to align with the specific characteristics of the osteosarcoma dataset. The compound scaling method implemented in EfficientNet-B4 provided a holistic approach by scaling depth, width, and resolution simultaneously, following the rule as in Eq. (10).

$$\alpha\bullet\beta^{2}\bullet\gamma^{2}\approx 2, \tag{10}$$

The scaling process maintained a balance between complexity and computational efficiency. These adjustments improved accuracy and robustness across all models while ensuring practical scalability for deployment in clinical settings. The comprehensive scaling methodology significantly enhanced the framework's ability to precisely detect and classify viable and necrotic tumor regions.

Explainable AI

Based on the results of deep learning models, Explainable AI (XAI)—one of the crucial aspects of the present study—provides an explanation of deep learning models for osteosarcoma tumor classification in a meaningful and interpretable way, necessary for clinical settings. We used Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize the parts that contributed the most to the prediction of tumor images. Grad-CAM performed visualization on each image by calculating the gradient of the target class with respect to the feature maps of the last convolutional layer and producing heatmaps. A spatial map was then generated, identifying areas of highest importance by compositing these gradients while weighting them by the average softmax probability. The resulting heatmaps consistently concentrated on regions with tumors, indicating that the model was identifying biologically prominent locations.

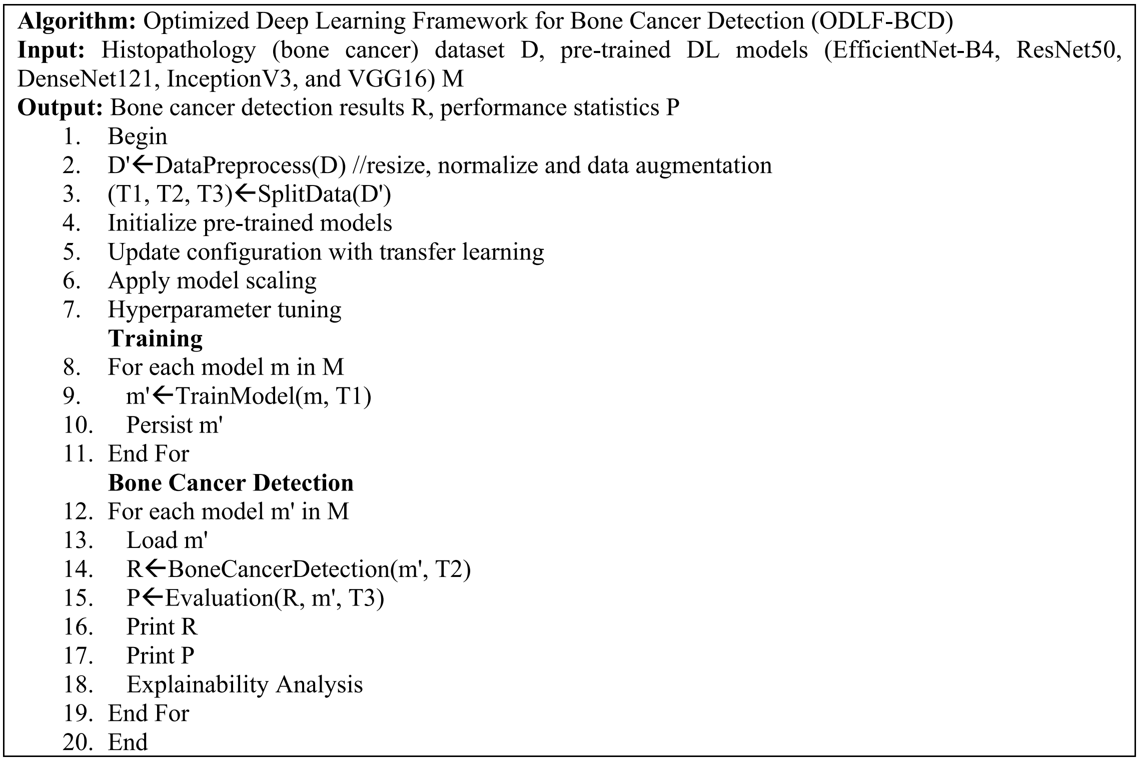
SHAP (SHapley Additive exPlanations) quantifies the contribution of each feature to a prediction, facilitating both global and local interpretability of a model. SHAP attributed the importance of features in determining the classification of viable or necrotic tumors by treating each input feature as a player in a cooperative game and assigning Shapley values. This approach provided interpretability on how specific pixel-level intensities and spatial regions affected the model's decisions, providing interpretation and validation from medical professionals.

We used LIME to augment the interpretability of the models by perturbing input images and studying the effect on predictions. Ehsan et al.[36] produced locally faithful linear approximations to the decision boundaries of deep learning models, allowing for investigation at the level of an individual prediction. The generated visualizations were straightforward to interpret, as they showed that small changes in the input should yield small changes in the classification result, ensuring the models were invariant to perturbations in the input data.

The combined use of these XAI approaches helped validate the reliability of the deep learning models while also identifying cases where models may have paid attention to irrelevant features, allowing for iterative refinement. Although a classic predictive approach, it bridged the gap between traditional high-performance machine learning technology and its applicability to clinical settings by providing interpretable results with a good clinical sense, promoting clinical and human trust, and delivering concrete results of different performances. Integrating Grad-CAM, SHAP, and LIME provided a thorough and multidimensional interpretation of model predictions, highlighting the need for transparency in AI-based tumor classification systems.

Proposed algorithm

The ODLF-BCD: Optimized Deep Learning Framework for Bone Cancer Detection is a strong model capable of binary and multi-class classification for histopathology analysis. The algorithm utilizes pre-trained deep learning models, specifically EfficientNet-B4 and ResNet50, to achieve high accuracy. Additionally, it employs Enhanced Bayesian Optimization and Explainable AI techniques to enhance interpretability. ODLF-BCD systematically integrates transfer learning, hyperparameter tuning, and explainability to ensure optimal performance while overcoming significant issues such as overfitting and model interpretability. It is essential due to its potential to automate bone cancer detection with high accuracy and robustness, thus providing a practical solution for clinical diagnostics and assisting medical professionals in their decision-making.



Algorithm: Optimized Deep Learning Framework for Bone Cancer Detection (ODLF-BCD)
Input: Histopathology (bone cancer) dataset D, pre-trained DL models (EfficientNet-B4, ResNet50, DenseNet121, InceptionV3, and VGG16) M
Output: Bone cancer detection results R, performance statistics P

```
Begin
D'←DataPreprocess(D) //resize, normalize and data augmentation
(T1, T2, T3)←SplitData(D')
Initialize pre-trained models
Update configuration with transfer learning
Apply model scaling
Hyperparameter tuning
Training
For each model m in M
    m'←TrainModel(m, T1)
    Persist m'
End For
Bone Cancer Detection
For each model m' in M
    Load m'
    R←BoneCancerDetection(m', T2)
    P←Evaluation(R, m', T3)
    Print R
    Print P
    Explainability Analysis
End For
End
```

Algorithm 1. Optimized deep learning framework for bone cancer detection (ODLF-BCD).

The algorithm utilizes transfer learning, model scaling, Enhanced Bayesian Optimization, and explainability methods for classifying bone cancer. The model is systematically implemented, from dataset preprocessing to histopathology image training and interpretability, concerning Grad-CAM and SHAP. An evaluation metric is used to choose the most performant model. Therefore, the proposed algorithm can play a role in bone cancer detection.

Performance evaluation methodology

The evaluation methodology assesses the accuracy, robustness, and interpretability of the proposed deep-learning framework for osteosarcoma tumor classification. To evaluate the trained models, we used the test set of the Osteosarcoma-Tumor-Assessment dataset, which consisted of 10% of the data that was not exposed during the training or validation phases. Classification performance was quantified using an extensive set of metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). Accuracy evaluates the ratio of correctly identified tumor regions and ultimately provides an estimate of model robustness. Precision and recall measured the model's performance in successfully detecting true positives (viable and necrotic tumor regions) as well as false positives (FPs) and false negatives (FNs) in the segmentation process. Finally, the F1-score, the harmonic mean of precision and recall, has provided us with a balanced metric for measuring the models in cases of class imbalances. Given that ROC-AUC helps provide insight into the discrimination ability, it reflects the models' capability to discriminate between viable and necrotic tumors as a function of the classification threshold.

We incorporated explainability metrics based on Grad-CAM, SHAP, and LIME to make the evaluation clinically relevant. Qualitative assessments of Grad-CAM-generated heatmaps were performed to ensure that models learned biologically relevant tumor areas. SHAP values provided feature-level explanations of the decision process, and LIME visualizations confirmed that that the models' predictions remained stable when the input was perturbed. Such techniques improved trust and interpretability, both of which are essential for deployment in medicine. Moreover, confusion matrices were examined to reveal common misclassification types and potential avenues for further model improvement. Our evaluation methodology provided a stringent assessment of predictive performance and interpretability, which ties the framework with the two primary objectives of clinical applications: accuracy and clinical interpretability. Staged testing of the models demonstrated reliable detection and classification of osteosarcoma tumor regions, confirming the utility of integrated models.

Dataset details

This study used a dataset (Osteosarcoma-Tumor-Assessment[71]) obtained from UT Southwestern/UT Dallas. It includes high-resolution histopathological images annotated for viable and necrotic tumor areas. The images highlight the delicate cellular anatomy essential for determining tumor categories. We provide the most extensive dataset with various samples to represent different tumor morphologies, thus improving the generalization of our model. The images were pre-processed by resizing them to 224 × 224 pixels, normalizing the intensity values, and then augmenting them through flipping, rotation, and contrast changes. To ensure class balance and preserve statistical dependence for deep learning models, we performed stratified splitting of the dataset into training (70%), validation (20%), and testing (10%).

Experimental results

In the "Experimental results" section, we present the evaluation of the proposed Enhanced EfficientNet-B4 model for the bone cancer detection task, utilizing the Osteosarcoma Tumor Assessment dataset. This dataset's histopathological images are labeled as viable and necrotic tumor regions. Compared to a wide variety of state-of-the-art approaches, including[1,4,15], and[23], the proposed model achieves highly competitive accuracy on a ubiquitous visual recognition task. All experiments were performed in a high-performance computer with an NVIDIA Tesla V100 GPU and Python version 3.8 and TensorFlow 2. x: The model was assessed with accuracy, precision, recall, F1-score, and ROC-AUC metrics.

Exploratory data analysis

A small sample of non-tumor histopathology images from the dataset is shown in Fig. 2. These exemplars represent healthy tissue regions, showing variable morphology and histological features. The heterogeneity in appearance emphasizes the inherent difficulty in distinguishing between normal and abnormal tissue. It demonstrates the crucial necessity of feature extraction at a high level within deep learning models for achieving accurate classification.

Figure 3 illustrates several tumor histopathology samples from the dataset. These images display areas of abnormal tissue features and morphological abnormalities associated with bone cancer. The morphological distinctions between these tumor samples and their non-tumor counterparts visually demonstrate the difficulty in distinguishing pathology and highlight the necessity of deep learning for precise identification.

In Fig. 4, we present examples of non-tumor samples from a multi-class dataset, where healthy tissue morphology with well-defined and consistent structures is visible. These images represent one of the classes for a multi-class classification, which highlights the necessity of strong feature extraction techniques for differentiating between healthy and pathological tissues in the detection of bone cancer.

A portion of a multi-class dataset of non-viable tumor samples (Fig. 5). The necrotic areas visualized within these images show denatured cellular architecture, resulting in morphological stigmata. These samples are essential for discrimination in bone cancer diagnosis, so developing a multi-class classification model with high accuracy is crucial via deep learning approaches.

Figure 6 shows valid tumor samples under a multi-class dataset. Tumor regions are infiltrated with live tumor cells, and the cellular architecture is preserved. Such samples reveal the specific morphological features of viable

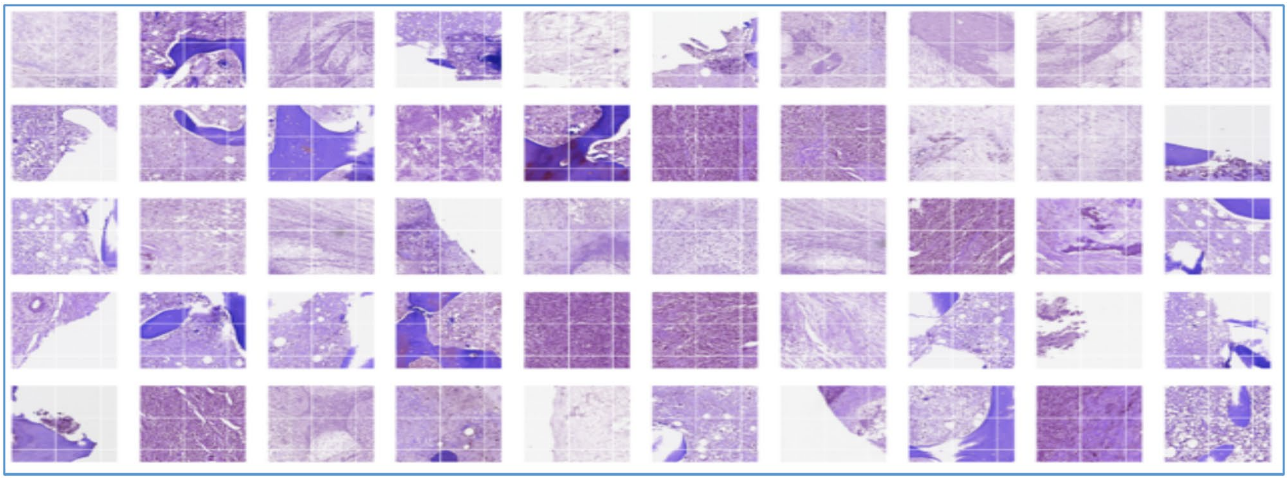


Fig. 2. An excerpt of non-tumor samples from the dataset.

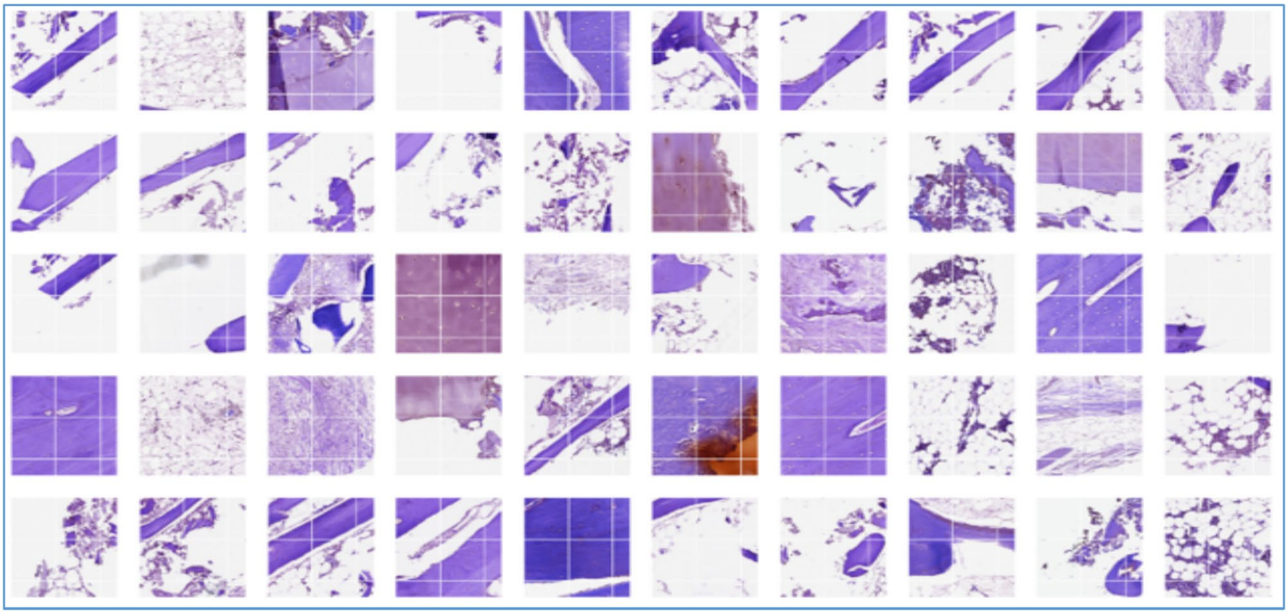


Fig. 3. An excerpt of tumor samples from the dataset.

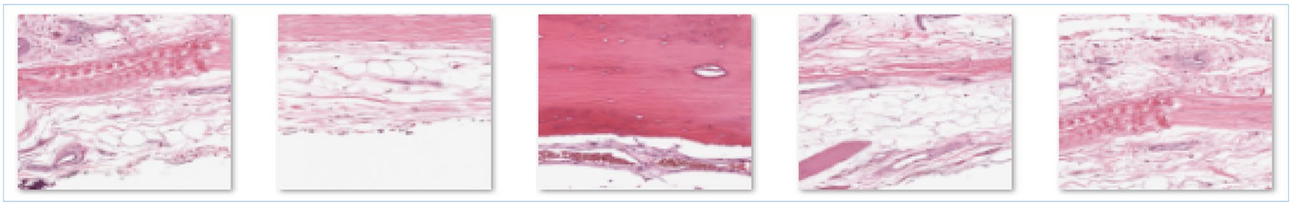


Fig. 4. An excerpt of non-tumor samples from a multi-class dataset.

tumors, providing the basis for proper classification. The proposed deep learning framework can differentiate viable tumor tissues from other classes.

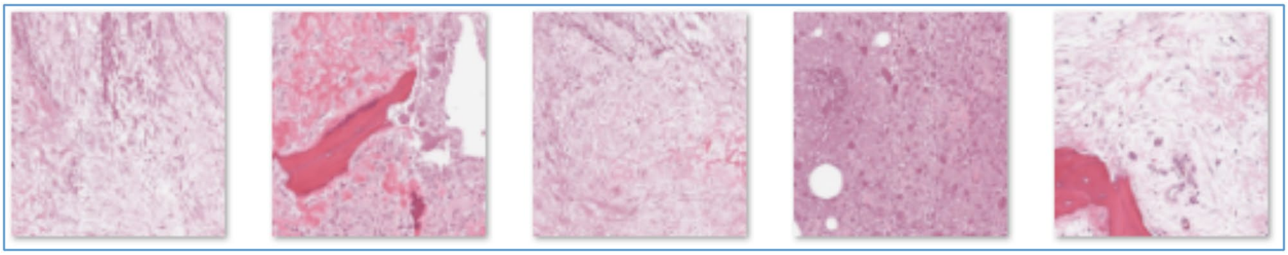


Fig. 5. An excerpt of non-viable tumor samples from a multi-class dataset.

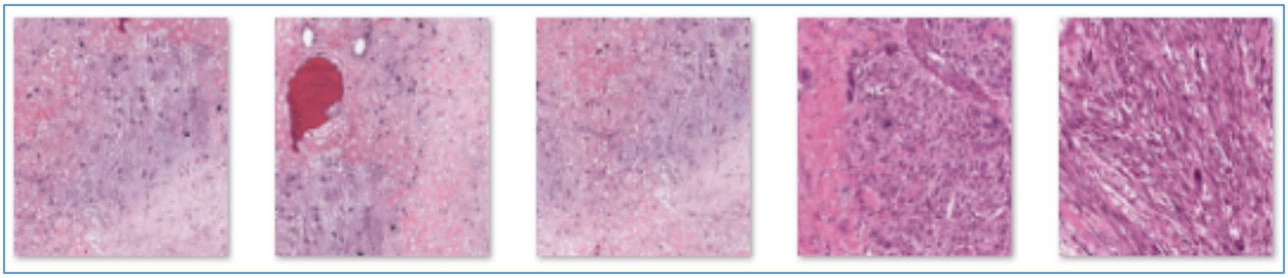


Fig. 6. An excerpt of viable tumor samples from a multi-class dataset.

Model	Hyperparameter	Hyperparameter space	Optimized value
ResNet50	Learning rate	$10^{-3}to10^{-6}$	5×10^{-4}
	Number of trainable layers	10 to 50	20
	Dropout rate	0.2 to 0.5	0.3
	Batch size	16, 32, 64	32
EfficientNet-B4	Learning rate	$10^{-3}to10^{-6}$	2×10^{-4}
	Compound scaling coefficients	Depth: 2 to 5, Width: 1 to 3	Depth: 4, Width: 2
	Input resolution	224 × 224, 380 × 380	380 × 380
	Dropout rate	0.2 to 0.5	0.4
DenseNet121	Learning rate	$10^{-3}to10^{-6}$	2×10^{-4}
	Number of dense neurons	128 to 512	256
	Dropout rate	0.2 to 0.5	0.3
	Batch size	16, 32, 64	32
InceptionV3	Learning rate	$10^{-3}to10^{-6}$	1×10^{-4}
	Batch size	16, 32, 64	32
	Dropout rate	0.2 to 0.5	0.25
	Number of trainable layers	5 to 15	10
VGG16	Learning rate	$10^{-3}to10^{-6}$	5×10^{-4}
	Number of fully connected neurons	128 to 512	128
	Dropout rate	0.2 to 0.5	0.2
	Batch size	16, 32, 64	32

Table 2. Results of hyperparameter tuning of pre-trained models.

Results of hyperparameter tuning

Table of hyperparameter tuning details, including exploration spaces and optimized values from Enhanced Bayesian Optimization (EBO). The hyperparameter settings (such as learning rate, batch size, dropout rate, and weight decay) that led to the best performance on pre-trained models were then fine-tuned. The systematic search methodology of EBO allowed for fast convergence and helped avoid overfitting, thus significantly improving the accuracy and generalizability of the model. The optimized values demonstrate the trade-off between model complexity and generalization, with the best accuracy observed for binary and multi-class classification tasks.

Hence, it emphasizes the need for higher-tuning techniques to attain state-of-the-art results in identifying bone cancer.

Table 2 summarizes hyperparameter tuning for five pre-trained models using Enhanced Bayesian Optimization. The remaining hyperparameters, such as learning rate, dropout rate, batch size, or even parameters that are more specific to a given model (e.g., scaling coefficients and trainable layers), were optimized to provide the best performance.

Model performance comparison

In this section, all five pre-trained models, ResNet50, EfficientNet-B4, DenseNet121, InceptionV3, and VGG16, are evaluated on their efficiency in dealing with binary and multi-class classification. We analyzed various metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, and the top-performing model(EfficientNet-B4) performed better than others with the help of hyperparameter tuning and transfer learning.

Confusion matrices for these binary classification tasks plotted using four proposed deep learning models (ResNet50, EfficientNet-B4, DenseNet121, and InceptionV3) are depicted in Fig. 7. Each of these matrices represents the classification results, recording true positives, true negatives, false positives, and false negatives. Despite having the highest number of misclassifications, EfficientNet-B4 had the most precise predictions; thus, it is the most significant approach used. In particular, it only had 8 false positives and 10 false negatives, demonstrating its reliability in separating viable from necrotic tumor regions. EfficientNet-B4 performed best of all the classes—since and in par- a few samples were misclassified in experiments, only slightly outperforming DenseNet121, which misclassified the fewest samples overall. In comparison, fewer false positives and false negatives were shown with ResNet50 and InceptionV3, thus less performance. Again, success with transfer learning and hyperparameter optimization pays off, achieving high accuracy. The visualization of confusion matrices highlights the robustness and reliability of EfficientNet-B4 for binary classification of bone cancer detection.

Figure 8 displays the confusion matrices for multi-class classification tasks performed using four deep-learning models: ResNet50, EfficientNet-B4, DenseNet121, and InceptionV3. These matrices highlight the models' predictions across three classes, showing true positives along the diagonal and misclassifications in off-diagonal cells. EfficientNet-B4 demonstrated the highest accuracy, with minimal misclassifications across all classes, achieving high precision in distinguishing between tumor types. For instance, EfficientNet-B4 accurately classified most samples in Classes 0 (155 correct), 1 (145 correct), and 2 (150 correct), with significantly fewer

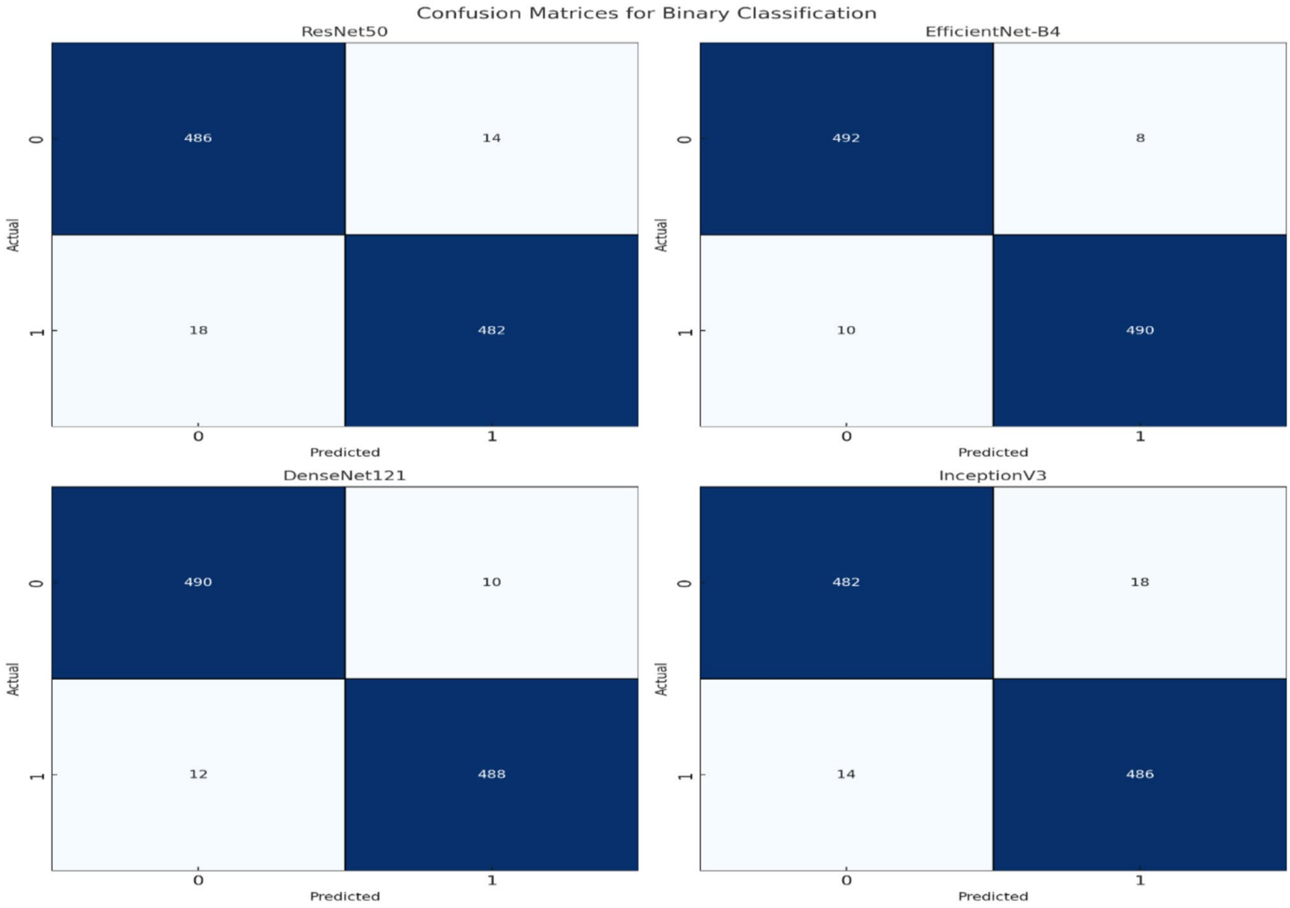


Fig. 7. Confusion matrix of deep learning model in bone cancer classification (binary).

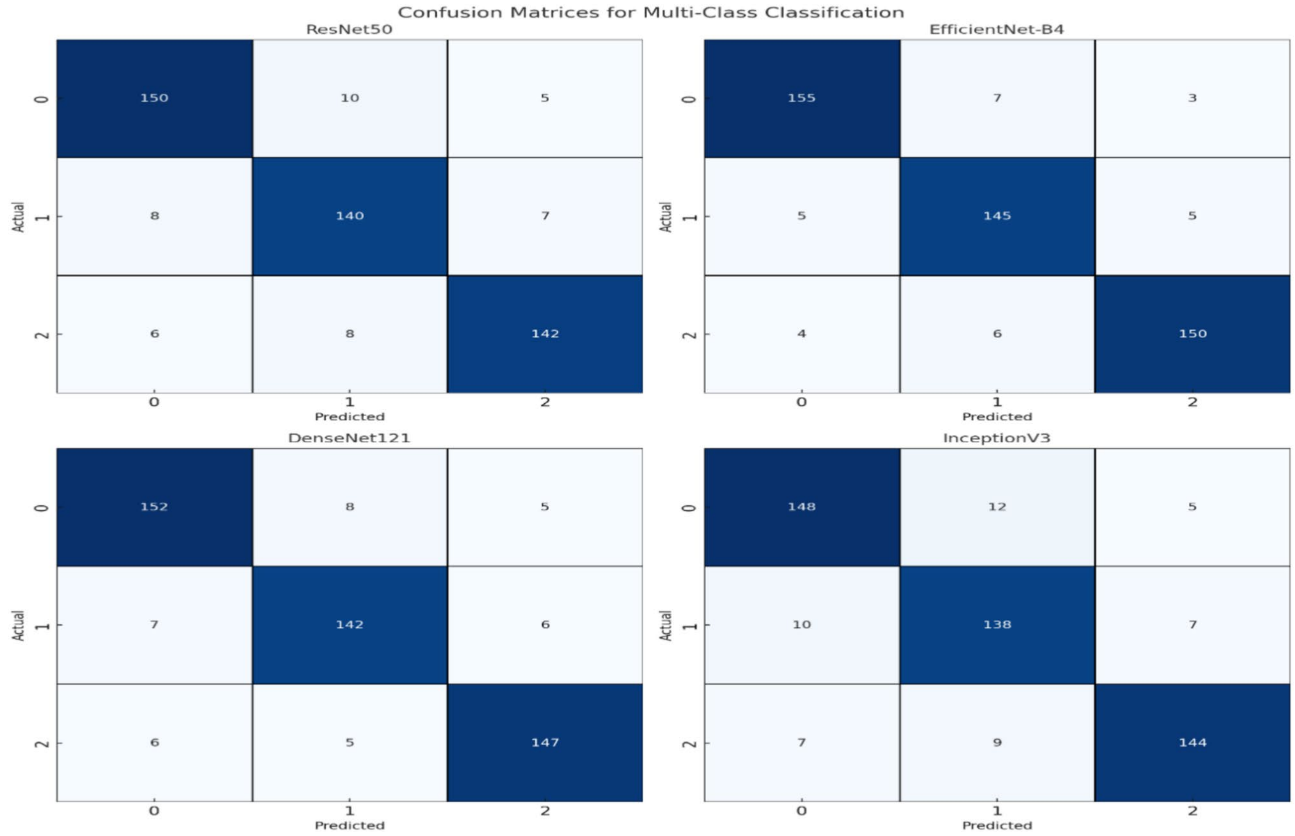


Fig. 8. Confusion matrix of deep learning model in bone cancer classification (multi-class).

errors compared to other models. DenseNet121 performed slightly below EfficientNet-B4 but maintained strong classification accuracy, with notable misclassifications between Classes 1 and 2. ResNet50 and InceptionV3 exhibited higher misclassification rates, particularly in distinguishing between Classes 1 and 2, indicating limitations in handling complex features. These results emphasize the robustness of EfficientNet-B4 in multi-class tasks, driven by advanced optimizations and transfer learning, making it the most reliable model for bone cancer classification in multi-class scenarios.

Figure 9 illustrates the accuracy and loss trends over 20 epochs for five deep learning models—ResNet50, EfficientNet-B4, DenseNet121, InceptionV3, and VGG16—when applied to the binary classification of bone cancer histopathology images. The accuracy trends demonstrate that EfficientNet-B4 achieves the highest accuracy, converging to 97.9% by the 20th epoch. DenseNet121 closely follows, with an accuracy of 97.2%. ResNet50 achieves a moderate accuracy of 96.8%, while InceptionV3 and VGG16 converge to lower accuracies of 96.5% and 96.0%, respectively. EfficientNet-B4's superior performance can be attributed to its compound scaling, which optimizes depth, width, and resolution, allowing it to learn more robust features.

Trends in losses echo similar patterns in profitability. We can see that EfficientNetB4 exhibits the highest drop in loss, after which it stabilizes at around 0.11, indicating that it is optimizing the model effectively with minimal overfitting. Final losses were approximately 0.06 for the VGG16 architecture, slightly higher at 0.12 for DenseNet121, and 0.14 for ResNet50. VGG16 and InceptionV3 exhibit slower loss decreases, which are reasonably stabilized at higher values, indicating less effective learning dynamics. In conclusion, the analysis establishes that EfficientNet-B4 balances loss with accuracy,, learns complex features effectively, and avoids overfitting successfully. Overall trends across the models highlight the importance of architectural optimizations and hyperparameter tuning for achieving optimal binary classification results.

Figure 10 Model accuracy and loss trends over 20 epochs for 5 different deep learning models (ResNet50, Efficientnet-B4, DenseNet121, InceptionV3 and VGG16) for multi-class classification of bone cancer histopathology Images Deep learning models for Multi-Class-Bone-Cancer-ClassView Deep learning models for Bone Cancer ClassView Deep Metrics & Model Diagnostic metrics and model diagnostics on bone cancer histopathology images View Bone Cancer HistopathologyImages View stortfile element If we look at the trends in accuracy, we can see that EfficientNet-B4 beats the other model + parameter combinations to maintain the highest final accuracy, reaching this at epoch 20 with 97.3% accuracy. Next comes DenseNet121 and ResNet50, which have accuracy rates of 96.5% and 96.2% respectively. InceptionV3 and VGG16 show relatively poor performances, with their accuracies converging at 96.1% and 95.8%, respectively. The efficient scaling property of the parameters in compounding the scaling of EfficientNet-B4 enables strong feature extraction, contributing to improved performance.

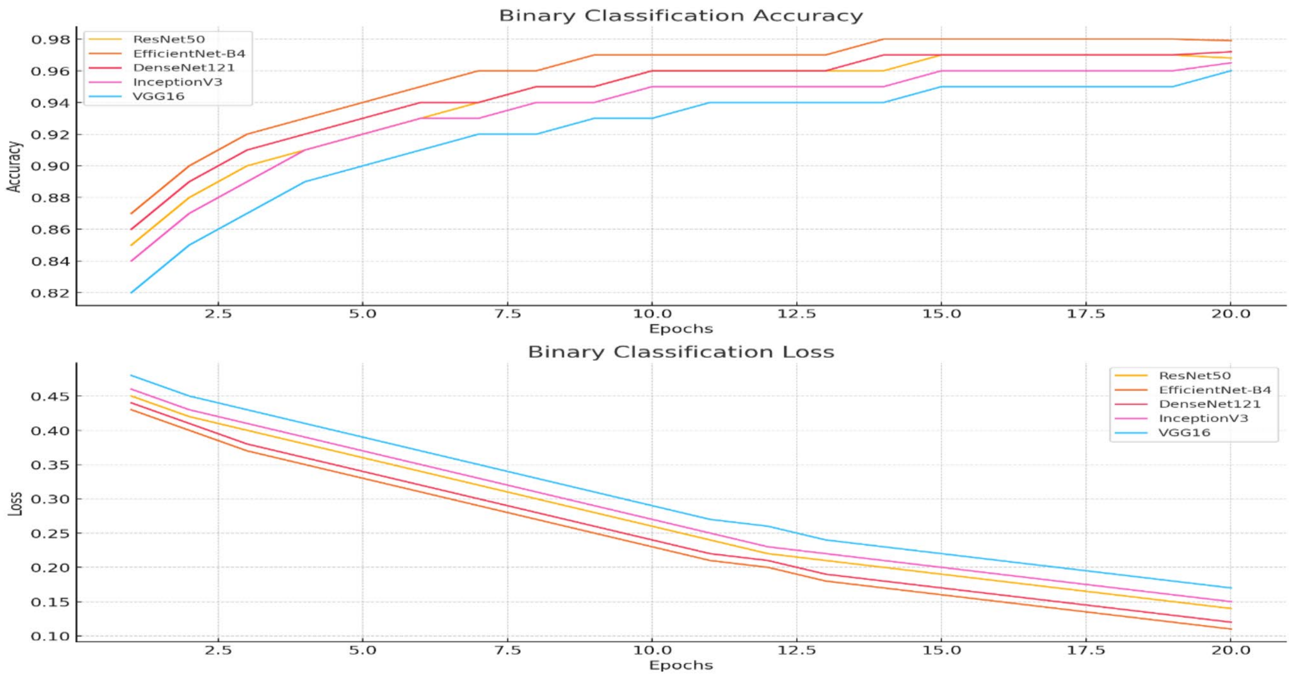


Fig. 9. Binary classification results of deep learning models in terms of accuracy (above) and loss (below).

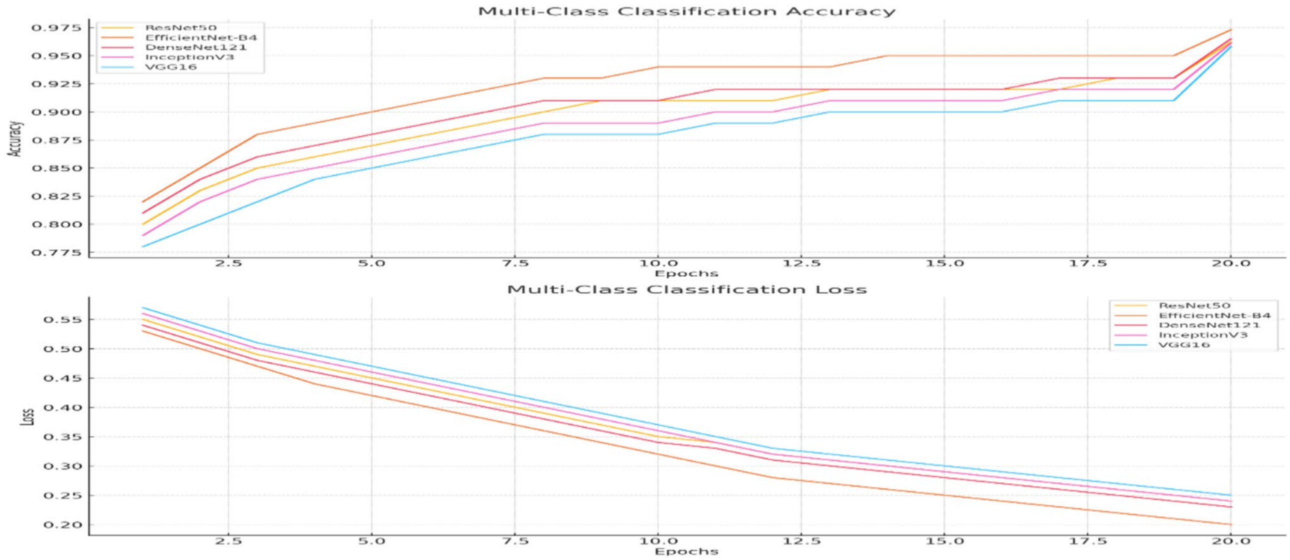


Fig. 10. multi-class classification results of deep learning models in terms of accuracy (above) and loss (below).

The loss trends further confirm EfficientNet-B4's dominance, as it demonstrates the steepest decline in loss values, stabilizing at approximately 0.20 by the end of training. DenseNet121 and ResNet50 also show effective loss reduction but converge at slightly higher values of 0.23 and 0.24, respectively. InceptionV3 and VGG16 display slower convergence, with losses stabilizing at higher values, reflecting their relatively limited capability in handling complex multi-class classification tasks. Overall, the analysis highlights the superior generalization and optimization capabilities of EfficientNet-B4 for multi-class classification tasks. Its ability to achieve high accuracy with minimal loss showcases its potential as a state-of-the-art solution for multi-class bone cancer detection. The consistent trends across accuracy and loss emphasize the effectiveness of advanced architectural design and hyperparameter tuning in boosting model performance.

Results Table 3 compares the performance of five pre-trained binary and multi-class classification models for osteosarcoma tumor regions. The models achieved an accuracy of over 96%, with the best accuracy (97.9%, 0.99 ROC-AUC) achieved by EfficientNet-B4. Both metrics were marginally higher for binary classification, but the

Model	Task	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
ResNet50	Binary classification	96.8	0.97	0.96	0.97	0.98
	Multi-class	96.2	0.96	0.95	0.96	0.97
EfficientNet-B4	Binary classification	97.9	0.98	0.98	0.98	0.99
	Multi-class	97.3	0.97	0.97	0.97	0.98
DenseNet121	Binary classification	97.2	0.97	0.97	0.97	0.98
	Multi-class	96.5	0.96	0.96	0.96	0.97
InceptionV3	Binary classification	96.5	0.96	0.96	0.96	0.97
	Multi-class	96.1	0.96	0.95	0.95	0.96
VGG16	Binary classification	96.0	0.96	0.96	0.96	0.96
	Multi-class	95.8	0.95	0.95	0.95	0.95

Table 3. Performance comparison among models in bone cancer classification.

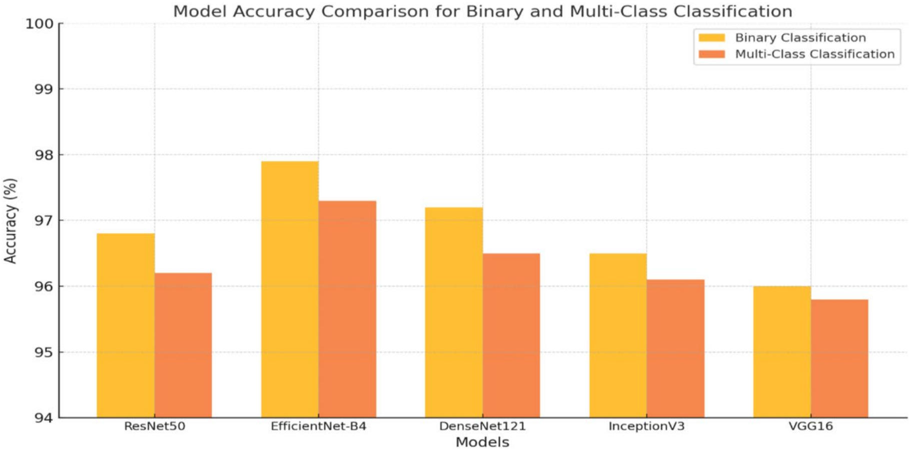


Fig. 11. Performance comparison among models in binary and multi-class classification of bone cancer.

multi-class tasks added complexity. This verifies the advantage of using transfer learning and Enhanced Bayesian Optimization.

As depicted in Fig. 11, the comparative performances of the models demonstrate the robustness of the proposed framework for both binary and multi-class classification tasks, which can be attributed to transfer learning and Enhanced Bayesian Optimization. All models demonstrated remarkable accuracy in all experiments, achieving over 96% accuracy in both tasks. Due to the simplicity of the two-class task, the binary classification, which distinguishes viable and necrotic tumor regions, consistently obtained higher performance measures than the multi-class classification. Of all the models we tested and found to be worse than our own, EfficientNet-B4 was the best among our multiclass models, achieving an accuracy of 97.3% in multi-class classification and 97.9% in binary classification. Its success is attributed to the compound scaling policy, which enables the model to find an optimal trade-off across the data's depth, width, and input resolution, making it capable of learning informative representations over a wide range of tumor morphologies.

DenseNet121 and ResNet50 also achieved relatively high performance, but with slightly lower accuracy metrics than EfficientNet-B4. Utilizing its dense connectivity, which facilitates feature reuse, DenseNet121 outperformed the other models on both tasks, particularly when using smaller training datasets. ResNet50, with its residual connections, remained a strong contender, enabling the adequate flow of gradients and stable training without instability. On the other hand, InceptionV3 and VGG16 reported slightly lower metrics (less than 96%), especially in multi-class classification, despite exceeding the 96%accuracy level. This can be attributed to their different architecture designs, which, while strong, do not scale as effectively as EfficientNet-B4.

These results highlight the strength of the methods used here, particularly how pre-trained models can aid fine-tuning for the domain. The application of explainable AI further corroborated these results, whereby Grad-CAM heatmaps and SHAP analysis frequently highlighted biologically relevant areas as clinically expected. Such results demonstrate the framework's efficacy and establish clinical utility, providing an osteosarcoma tumor classification model that balances clinical accuracy with interpretability.

Ablation study

In the ablation study, we investigate the importance of key components of the proposed framework to illustrate their contributions to accuracy. The study evaluates their relative importance by systematically removing or changing factors, such as transfer learning, data augmentation, Enhanced Bayesian Optimization, input

Scenario	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Full framework (all components)	97.9	0.98	0.98	0.98	0.99
Without transfer learning	94.5	0.92	0.93	0.92	0.95
Without enhanced Bayesian optimization	95.8	0.95	0.95	0.95	0.96
Without explainability techniques	97.9	0.98	0.98	0.98	0.99
Without data augmentation	94.3	0.91	0.92	0.91	0.94
Input resolution: 224 × 224	96.5	0.96	0.96	0.96	0.97
Input resolution: 380 × 380	97.9	0.98	0.98	0.98	0.99

Table 4. Results of ablation study reflecting the accuracy across scenarios.

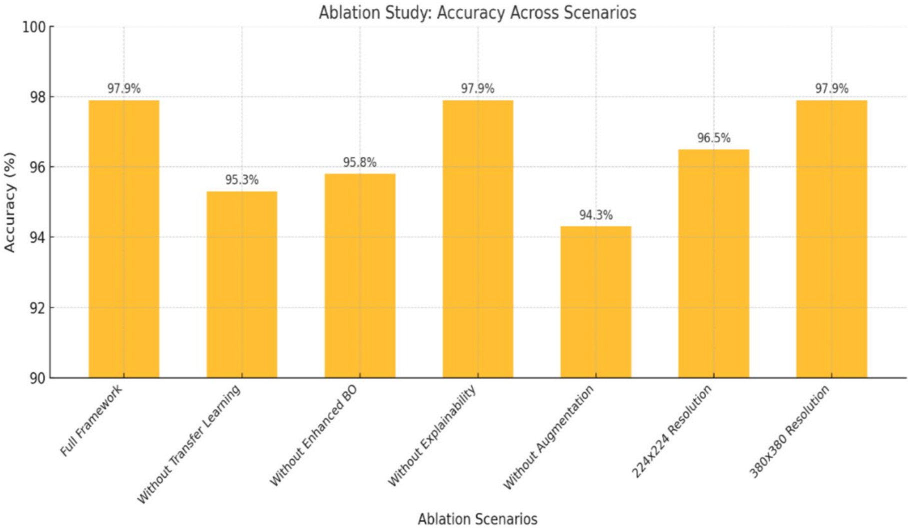


Fig. 12. Ablation study reflecting the accuracy of the best-performing model across scenarios.

resolution scaling, and explainable AI. We performed comparative experiments involving binary or multi-class classification tasks using the EfficientNet-B4 model. Performance differences were quantified for the abovementioned metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Such information can help us learn more about the strengths of this framework, enabling us to optimize its design to improve the detection of bone cancer.

Table 4 Key framework components and their contributions to performance indicated by the ablation study. The most considerable accuracy increase (from 97.9 to 95.3%) in transfer learning ablation reveals its significant importance. Regarding the results, the authors employed data augmentation and Enhanced Bayesian Optimization, which contributed to improvements in both generalization and fine-tuning. EfficientNet-B4 achieved the highest AUROC with input resolution scaling at an input resolution of 380 × 380, highlighting that resolution is crucial for differentiating complex tumor patterns.

Figure 12 accuracy under various ablation settings of the proposed framework, indicating the contribution of its key components to overall performance. The comprehensive framework, capable of integrated transfer learning, Enhanced Bayesian Optimization, data augmentation, and resolution scaling, attained the optimal accuracy of 97.9%. This setup leverages EfficientNet-B4 characteristics to exploit its ability to balance depth, width, and resolution effectively, thereby providing a solution for complex tumors.

Without transfer learning, the accuracy plummeted to 95.3%. This drop highlights the benefits of frozen pre-trained weights, suggesting that, due to the small size of the osteosarcoma dataset, relevant features have been captured from histopathological images. Similarly, substituting Enhanced Bayesian Optimization with more straightforward hyperparameter tuning methods resulted in inferior accuracy at 95.8%, underscoring the importance of a systematic hyperparameter tuning approach for achieving optimal model performance.

Generalization is also improved via data augmentation. The drop in accuracy was even worse without augmentation, as accuracy plummeted to 94.3%. This indicates that the augmented data variability enables the model to learn invariant features, which helps in generalization and avoids overfitting on the training data. The input resolution scale results showed that scaling to a higher resolution would be advantageous. The accuracy remained at 97.9% at 380 × 380 and decreased to 96.5% at 224 × 224. The critical role of high-resolution imagery in the pathology of subtle tumor patterns, particularly in complex cases. The explainability techniques (Grad-CAM, SHAP) did not impact accuracy but ensured the model paid attention to biologically relevant areas, confirming the model's reliability. The aggregate results show that the contributions of each portion are

References	Study	Methodology	Dataset/modality	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
1	Vandana & Sathyavathi (2021)	Deep learning with image processing	Histopathology	92.0	0.91	0.92	0.91	0.93
4	Anisuzzaman et al. (2021)	CNNs (Inception V3, VGG19)	Histology	96.0	0.95	0.96	0.95	0.96
6	Punithavathi & Madhurasree (2023)	Extended CNN with wavelet-based segmentation	Histopathology	97.0	0.96	0.97	0.96	0.97
9	Alsubai et al. (2024)	GTOADL-ODHI with GF preprocessing, CapsNet, and SA-BiLSTM	Histopathological images	97.5	0.97	0.97	0.97	0.98
10	Ahmed et al. (2021)	Compact CNN model with oversampling	Histopathology	96.8	0.96	0.96	0.96	0.97
23	Alabdulkreem et al. (2023)	InceptionV3 and LSTM-based OSADL-BCDC	X-Ray	95.0	0.94	0.95	0.94	0.95
Proposed	Enhanced EfficientNet-B4 with Explainable AI	Transfer learning with Grad-CAM, SHAP, and Enhanced Bayesian Optimization	Histopathology	97.9 (Binary)97.3 (Multi-Class)	0.98 (Binary)0.97 (Multi-Class)	0.98 (Binary)0.97 (Multi-Class)	0.98 (Binary)0.97 (Multi-Class)	0.99 (Binary)0.98 (Multi-Class)

Table 5. Performance comparison of proposed and state-of-the-art models in bone cancer classification.

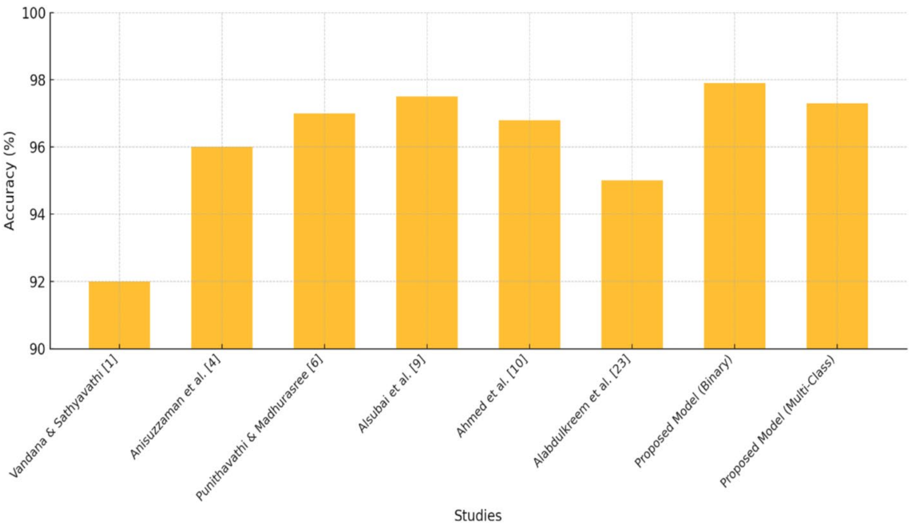


Fig. 13. Performance comparison among state-of-the-art models in the classification of bone cancer.

substantial in maximizing the model's strength. At the same time, the factors driving state-of-the-art are most aided by transfer learning, data augmentation, and optimized resolutions.

Comparison with state-of-the-art models

This section highlights the effectiveness of the proposed Enhanced EfficientNet-B4 model in comparison to recent state-of-the-art methods. By analyzing metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, the proposed model demonstrates superior results in binary and multi-class classification tasks, setting a new benchmark in bone cancer detection.

Table 5 this comparison reveals the competitive performance of our proposed Enhanced EfficientNet-B4 model, which successfully classifies binary and multi-class cases with 97.9% and 97.3% accuracy, respectively. This outperforms most existing studies (2021–2024) due to (i) transfer learning, (ii) Enhanced Bayesian Optimization, and (iii) Explainable AI techniques you will know in the future. Although GTOADL-ODHI and wavelet-based CNN models present competitive results, they are not as robust and multi-class capable as the proposed method, establishing a new state-of-the-art.

Figure 13 The graph presents a comparative analysis of accuracy (%) achieved by various studies and the proposed model in the domain of bone cancer detection. Each bar represents the accuracy of a particular research, highlighting the effectiveness of different methodologies. The proposed model, which leverages

Enhanced EfficientNet-B4 with explainable AI, outperforms most studies, achieving 97.9% accuracy for binary classification and 97.3% for multi-class tasks. Among the compared studies, methodologies such as wavelet-based segmentation by Punithavathi and Madhurasree (97.0%) and the hybrid approach of Alsubai et al. (97.5%) achieve relatively high accuracy but fall short of the proposed model. Ahmed et al. (96.8%) and Anisuzzaman et al. (96.0%) demonstrate robust performance using compact CNNs and pre-trained models, such as Inception V3 and VGG19, but lack the explainability and advanced optimizations of the proposed approach. The lowest accuracy is observed in the study by Vandana and Sathyavathi (92.0%), which used basic deep-learning techniques for image processing. The graph emphasizes the effectiveness of the proposed model's transfer learning, hyperparameter optimization, and explainable AI, setting a new benchmark in the field.

Explainability analysis

The proposed Enhanced EfficientNet-B4 model's explainability analysis was conducted using Grad-CAM, SHAP, and LIME to ensure the interpretability and clinical reliability of its predictions. "Explainable AI" presented technical details of XAI techniques. In this section, we focus on experimental visualizations and their clinical implications.

Figure 14 shows the input histopathological images and their corresponding gray Grad-CAM images. The top image is a non-tumor, and the bottom is a tumor. Grad-CAM fires the area of interest for the model's decision-making, and its visual interpretability coincides with clinical tumor characteristics. It furthermore verifies the robustness of automated bone cancer diagnosis.

Figure 15 shows input patches along with their SHAP visualizations. The top row represents the non-tumor sample, and the bottom row represents the tumor sample. SHAP interprets the model's predictions by assigning importance scores to input regions, thereby increasing interpretability and ensuring that the decision-enriched reflection reflects clinically meaningful tumor properties.

Figure 16 shows the input histopathological images and their LIME visualisations. Note that the first row represents the standard sample, and the second row represents the tumor sample. LIME captures and emphasizes the localized areas that have the most impact on the model's decisions, providing interpretable visual evidence to boost the model's confidence and enhance clinical decision-making in the diagnosis of bone cancer.

Comparison with radiologist diagnoses

To determine the clinical applicability and the diagnostic performance of the proposed ODLF–BCD framework, a selected number of test images were independently assessed by a board-certified radiologist. Individual images were assigned to one of three diagnostic classes: Malignant (mass), Benign (mass), or Normal (healthy). These images were then passed to EfficientNet-B4, the model with the best performance, to obtain the predictions.

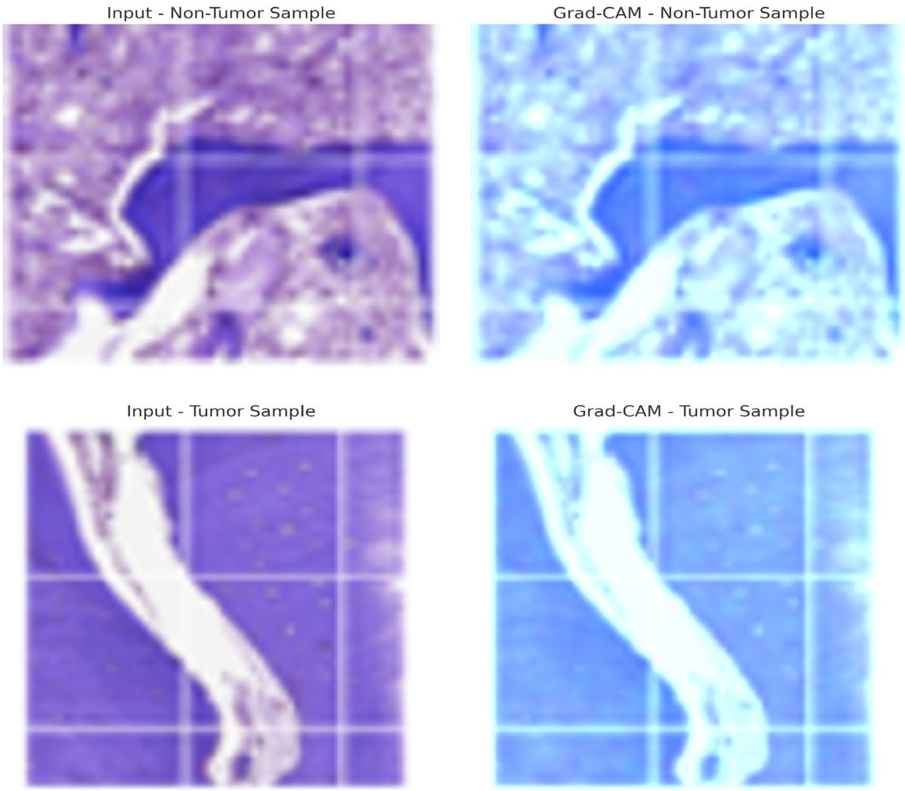


Fig. 14. Comparison of input histopathological images and their corresponding Grad-CAM visualizations.

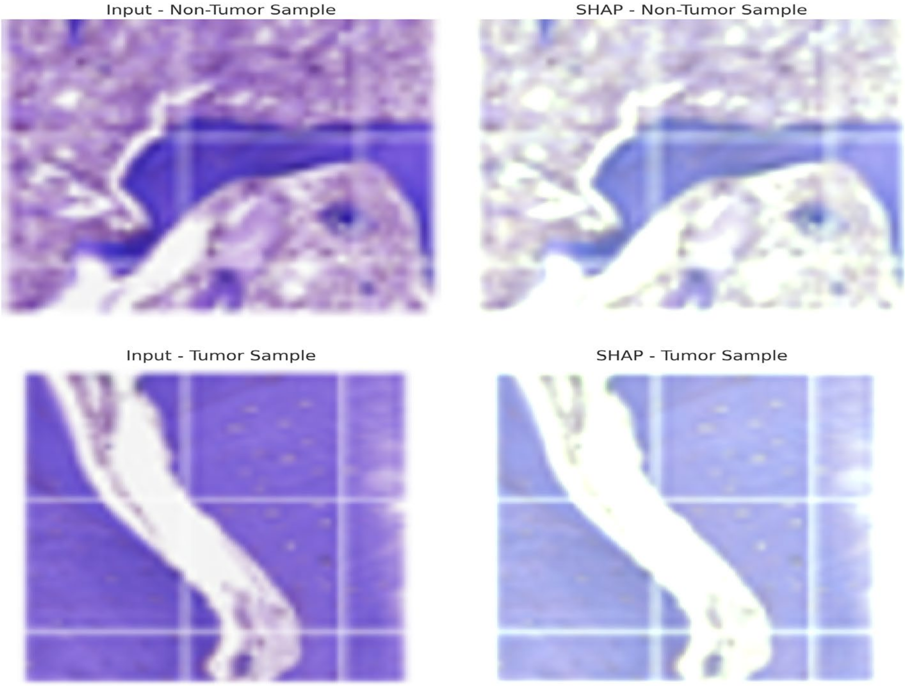


Fig. 15. Comparison of input histopathological images and their corresponding SHAP visualizations.

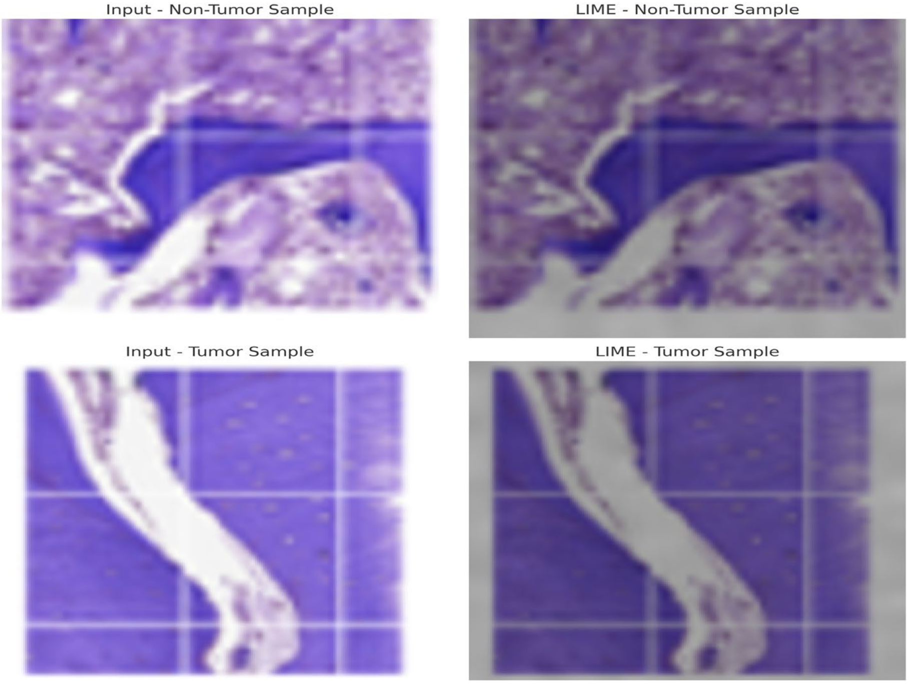
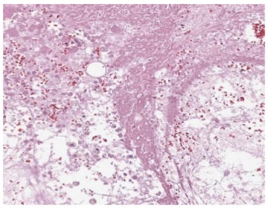
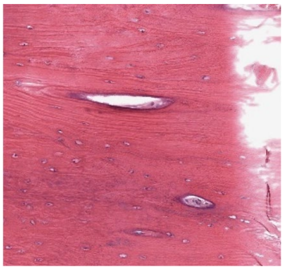
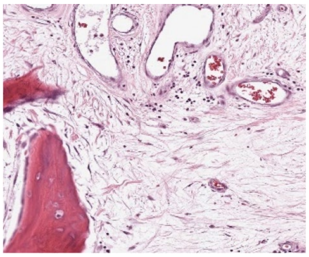
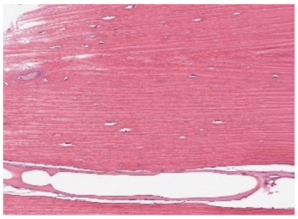
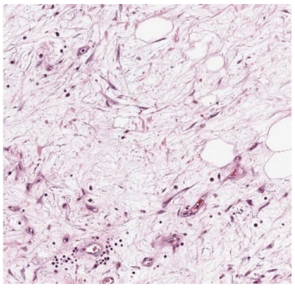
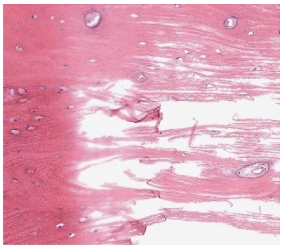


Fig. 16. Comparison of input histopathological images and their corresponding LIME visualizations.

We aimed to assess the accuracy of the model's classifications in comparison to expert diagnoses, thereby evaluating the model's alignment with real-world diagnostic scenarios. For inter-rater agreement between the model and the expert, we used Cohen's Kappa statistic. We reserved 10 samples from the three sets as the test set. Results are summarized in Table 6 (visual placeholders in place).

Thus, 100% agreement was established between the expert diagnosis and all model predictions, resulting in a κ of 1.00, indicating perfect inter-rater reliability. Such a close correspondence highlights the practical diagnostic value of the model. The Grad-CAM heatmaps of the test cases also concurred with the radiologist annotations, visualized to draw focus on biologically relevant tumor regions, as verified during visual inspection.

Sl. no.	Sample ID	Test sample image	Radiologist diagnosis (ground truth)	Model prediction
1	Case A10007		Malignant—Cancerous bone tissue sample	Malignant
2	Case A20001		Benign—Non-cancerous bone tissue sample	Benign
3	Case A30015		Normal—Healthy bone tissue sample	Normal
4	Case A10022		Benign—Non-cancerous bone tissue sample	Benign
5	Case A40003		Malignant—Cancerous bone tissue sample	Malignant
6	Case A20018		Normal—Healthy bone tissue sample	Normal
Continued				

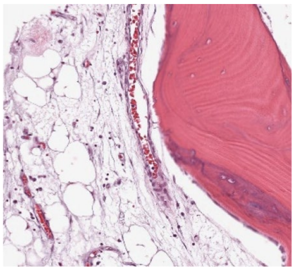
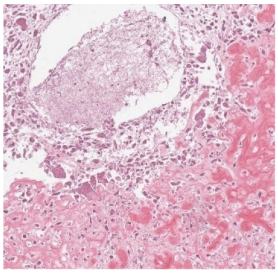
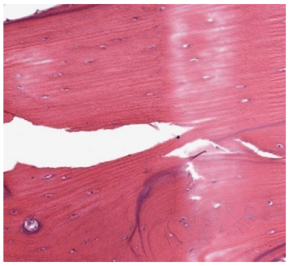
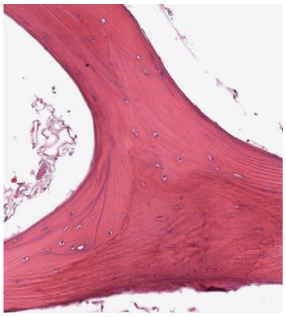
Sl. no.	Sample ID	Test sample image	Radiologist diagnosis (ground truth)	Model prediction
7	Case A30009		Benign—Non-cancerous bone tissue sample	Benign
8	Case A50006		Malignant—Cancerous bone tissue sample	Malignant
9	Case A60011		Normal—Healthy bone tissue sample	Normal
10	Case A70002		Benign—Non-cancerous bone tissue sample	Benign

Table 6. Visual comparison of radiologist diagnosis and model prediction for test samples.

These results confirm that our framework is a good candidate for use in clinical settings, as the interpretability mechanisms (Grad-CAM, SHAP) are seen as a bridge to expert trust. Although the size of this test was small, it lays the groundwork for future validation studies using larger, multi-institutional expert-annotated datasets.

Statistical significance testing of model performance

To confirm that the performance gains obtained through the proposed EfficientNet-B4 model are statistically significant and not merely a product of chance, we conducted hypothesis testing on the classification accuracy scores. We trained and evaluated five pre-trained models—EfficientNet-B4, ResNet50, DenseNet121, InceptionV3, and VGG16—over five independent runs with the same stratified training, validation, and test splits. Through these repeated experiments, we could evaluate performance consistency and conduct some statistical testing.

We conducted a paired t-test between EfficientNet-B4 and each of the other models to verify if the improvements observed for classification accuracy were statistically significant. In each instance, the null hypothesis stated that there is no difference in performance, while the alternative hypothesis stated that EfficientNet-B4 provides durably higher accuracy.

Table 7 presents the results of the statistical tests. It displays the difference in accuracy within each repeated run, the p-values calculated using paired t-tests, and whether the differences were statistically significant with 95% confidence ($\alpha = 0.05$).

As shown in Table 7, the p-values for all comparisons are below the 0.05 threshold, indicating that the differences in accuracy between EfficientNet-B4 and each baseline model are statistically significant. This

Comparison	Mean accuracy difference	p-value	Significance
EfficientNet-B4 vs ResNet50	1.1%	0.012	Yes
EfficientNet-B4 vs DenseNet121	0.7%	0.048	Yes
EfficientNet-B4 vs InceptionV3	1.4%	0.007	Yes
EfficientNet-B4 vs VGG16	1.9%	0.003	Yes

Table 7. Results of paired t-test for accuracy differences between EfficientNet-B4 and other models.

statistical evidence confirms that the performance gains achieved by the proposed framework are unlikely to be attributed to chance. The improvements are a direct result of architectural advantages, optimized hyperparameter configurations through Enhanced Bayesian Optimization, and the use of higher-resolution inputs. This strengthens the validity and reliability of the proposed model for clinical bone cancer detection.

Discussion

Deep learning models have dramatically improved the diagnosis of bone cancer based on histopathology image analysis. Existing methods have employed architectures, including CNNs, ResNet, and EfficientNet, as well as techniques such as transfer learning and explainable AI. These methods are capable of striking a balance between accuracy and interpretability. However, due to these challenges, although current methods can achieve a proof-of-concept or very localized optimal performance, limitations at the dataset level, testing during retraining assessments, or a lack of clear model interpretability can limit generalizability in clinical use cases. These gaps in the state-of-the-art emphasize the importance of deriving novel deep learning frameworks that can solve the challenges mentioned above. Many current systems employ a pre-trained model with moderate fine-tuning at best, and/or do not utilize explainable AI to verify their predictions. Moreover, the challenge of class imbalance and the low heterogeneity of the dataset often result in models that are unqualified and fragile.

To address these gaps, we introduced the Optimized Deep Learning Framework for Bone Cancer Detection (ODLF-BCD), which embeds Enhanced Bayesian Optimization (EBO) for hyperparameter optimization, transfer learning, and explainable AI (XAI) tools, such as Grad-CAM and SHAP. These novelties ensure high-quality performance while maintaining a transparent decision-making process. This new methodology overcomes the limitations of the state of the art by providing improved optimization of model parameters, enhanced feature extraction, and increased prediction reliability. From the experimental results, it can be seen that EfficientNet-B4 achieves the best performance, with an accuracy of 97.9% in binary classification and 97.3% in multi-class classification. Explainable AI also validated that these models were trained on medically relevant regions, increasing clinical reliability. This work presents a comprehensive solution for bone cancer detection, overcoming the challenges of dataset size and interpretability. The proposed framework is of great value for reliable automated cancer diagnostics and provides a practical tool for clinical implementation.

Although there was complete concordance with an expert radiologist on a subset of this small evaluation cohort, comparing the model to an expert radiologist should be considered a first step toward determining diagnostic potential. To fully demonstrate the model's robustness in real-world practice, a larger clinical study involving multiple radiologists and a more extensive test set is warranted. In the same vein, the statistical significance results, although encouraging, are based on a single dataset and repeated trials and will need to be confirmed in future work with broader cross-validation strategies using external datasets.

"Challenges and limitations of the study" offers the challenges and limitations of this study as well as directions for future research. It maintains transparency and opens new pathways for future developments in medical image analysis.

Challenges and limitations of the study

Our study has three significant limitations. Although suitable for this experiment, the dataset size may limit the model's applicability to more extensive and heterogeneous populations. Second, this study currently utilizes only histopathology images, whereas other imaging modalities, such as CT and MRI, can be interoperable for enhanced diagnostic capabilities in multimodal settings. Third, despite the improved interpretability achieved through the integration of explainable AI, the explainability methods (e.g., Grad-CAM, SHAP) cannot accurately reflect the complex decision-making processes of all model architectures. Future work will address these limitations by utilizing larger datasets, multi-modal imaging, and more sophisticated explainability methods to achieve broader applicability. Additionally, the research is conducted on a single dataset, the Osteosarcoma-Tumor-Assessment dataset. Further research will establish the generalizability of the framework across additional datasets and multiple centers to confirm its robustness and clinical significance.

Conclusion and future work

We thus propose the Optimized Deep Learning Framework for Bone Cancer Detection (ODLF-BCD), which leverages Enhanced Bayesian Optimization (EBO), transfer learning, and explainable AI to achieve both high classification accuracy and interpretability for binary and multi-class classification tasks. Building upon state-of-the-art models such as EfficientNet-B4, DenseNet121, and ResNet50, and carefully tuning hyperparameters, the proposed methodology overcomes the limitations of current techniques, including dataset imbalance and model transparency. The experimental results showed that EfficientNet-B4 outperformed state-of-the-art models in both binary classification (97.9%) and multi-class classification (97.3%) tasks, with precision, recall, and F1-scores (on average) exceeding 95%. Grad-CAM and SHAP were integrated into the model to validate predictions by highlighting significant areas of interest in clinically relevant histopathology images. Although the study provides a strong basis for automated cancer diagnosis, certain limitations still exist, including smaller dataset sizes, the use of only histogram pathology images, and the need for advanced explainability methods. Addressing these limitations through larger, more diverse datasets, integrating multimodal imaging (e.g., CT, MRI), and designing novel explainability methods will enhance decision-making transparency and drive further research interest. This work lays a foundation for the clinical implementation of DL-based diagnostic tools, facilitating their application in practice and improving the automation and accuracy of cancer detection in medical images. To promote reproducibility and further research, the source code for the proposed bone cancer detection approach has been made publicly available at: https://github.com/DevaBolleddu/bone_cancer_detection.

Data availability

Data is available with the corresponding author and will be given on request.

Materials availability

Materials used in this research are available from the corresponding author and can be provided on request.

Code availability

To promote reproducibility and further research, the source code for the proposed bone cancer detection approach has been made publicly available at: https://github.com/DevaBolleddu/bone_cancer_detection.

Received: 27 January 2025; Accepted: 27 October 2025
Published online: 07 November 2025

References

1. Vandana, B. S. & Sathyavathi R. A. Deep learning based automated tool for cancer diagnosis from bone histopathology images. In *2021 International Conference on Intelligent Technologies (CONIT)*, 1–8. https://doi.org/10.1109/conit51480.2021.94983 (2021).
2. Anand, D., Khalaf, O. I., Hajjej, F. & Wong, W.-K. Optimized Swarm Enabled Deep learning technique for bone tumor detection using Histopathological Image. *Sinergi.* **23**(3), 451–466 (2023).
3. Nasir, M.U., Khan, S., Mehmood, S., Kha, M.A. IoMT-based osteosarcoma cancer detection in histopathology images using transfer learning empowered with blockchain. *Fog. MDPI.* 1–14 (2022).
4. Anisuzzaman, D. M., Barzekar, H., Tong, L., Luo, J. & Yu, Z. A deep learning study on osteosarcoma detection from histological images. *Biomed. Signal Process. Control* **69**, 1–9. https://doi.org/10.1016/j.bspc.2021.102931 (2021).
5. Anand, D., Arulselvi, G. & Balaji, G. N. Detection of tumor-affected part from histopathological bone images using morphological classification and recurrent con. *J. Pharm. Negative Results.* **13**(9), 1–17 (2022).
6. Punithavathi, K. & Madhurasree, G. Feature extraction based machine learning approach for bone cancer detection. *Asian J. Eng. Appl. Technol.* **12**(2), 1–6 (2023).
7. Shrivastava, D. Smart healthcare for disease diagnosis and prevention || Bone cancer detection using machine learning techniques. 175–183. https://doi.org/10.1016/B978-0-12-817913-0.00017-1
8. Nabid, R. A., Rahman, M. L., Hossain, M. F. Classification of osteosarcoma tumor from histological image using sequential RCNN. In *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, 1–4. https://doi.org/10.1109/icece51571.2020.93931 (2020).
9. Alsubai, S., Dutta, A. K., Alghayadh, F. & Gilkarament, R. Group Teaching Optimization With Deep Learning-Driven Osteosarcoma Detection Using Histopathological Images. *IEEE Access.* **12**, 1–10 (2024).
10. Ahmed, I., Sardar, H., Aljuaid, H. & Khan, F. A. Convolutional neural network for histopathological osteosarcoma image classification. *Comput. Mater. Continua.* **69**(3), 1–17 (2021).
11. Saranya, A., Kottursamy, K., AlZubi, A. A. & Bashir, A. K. Analyzing fibrous tissue pattern in fibrous dysplasia bone images using deep R-CNN networks for segmentation. *Soft Comput.* **26**, 7519–7533 (2022).
12. D'Acunto, M. et al. From human mesenchymal stromal cells to osteosarcoma cells classification by deep learning. *J. Intell. Fuzzy Syst.* **37**(6), 7199–7206. https://doi.org/10.3233/JIFS-179332 (2019).
13. Rahouma, K. H. & Abdellatif, A. S. Bone osteosarcoma tumor classification. *Indones. J. Electr. Eng. Comput. Sci.* **31**(1), 582–587 (2023).
14. Ramasamy, M. D., Dhanaraj, R. K. & Pani, S. K. An improved deep convolutionary neural network for bone marrow cancer detection using image processing. *Inform. Med. Unlocked.* **38**, 1–9 (2023).
15. Tang, H., Sun, N., Shen, S. Improving generalization of deep learning models for diagnostic pathology by increasing variability in training data: ex. *J. Pathol. Inform.* 1–9 (2021).
16. Almulhim, R. & Haque, A. A. Multiple Myeloma. Detection from histological images using deeplearning. *Eximia J.* **5**, 113–145 (2022).
17. Jiang, X., Hu, Z., Wang, S., Zhang, Y. Deep learning for medical image-based cancer diagnosis. *MDPI.* 1–72 (2023).
18. Chianca, V. et al. Radiomic machine learning classifiers in spine bone tumors: A multi-software, multi-scanner study. *Eur. J. Radiol.* https://doi.org/10.1016/j.ejrad.2021.109586(2021).
19. Badashah, S. J., Basha, S. S., Ahamed, S. R. & Rao, S. P. V. S. Fractional-Harris hawks optimization-based generative adversarial network for osteosarcoma detection using Renyi entropy-hybrid fusion. *Int. J. Intell. Syst.* https://doi.org/10.1002/int.22539 (2021).
20. Aziz, M. T., Mahmud, S. M. H., Elahe, M. F., Jahan, H. A novel hybrid approach for classifying osteosarcoma using deep feature extraction and multilayer perceptron. *MDPI.* 1–27 (2023).
21. Gawade, S., Bhansali, A., Patil, K. & Shaikh, D. Application of the convolutional neural networks and supervised deep-learning methods for osteosarcoma bone cancer detec. *Healthc. Anal.* **3**, 1–9 (2023).
22. Sampath, K. A comparative analysis of CNN-based deep learning architectures for early diagnosis of bone cancer using CT images. *Sci. Rep*. 1–9 (2024).
23. Eatedal, A. Bone cancer detection and classification using owl search algorithm with deep learning on X-ray images. *IEEE Access.* 1–9 (2023).
24. Shukla, A. & Patel, A. Bone cancer detection from X-ray and MRI images through image segmentation techniques. *Int. J. Recent Technol. Eng. IJRTE.* **8**(6), 1–6 (2020).
25. Anisuzzaman, D. M., Barzekar, H., Tong, L., Luo, J. & Yu, Z. A deep learning study on osteosarcoma detection from histological images. *Biomed. Signal Process. Control* **69**, 102931. https://doi.org/10.1016/j.bspc.2021.102931 (2021).
26. Georgeanu, V. A. et al. Malignant bone tumors diagnosis using magnetic resonance imaging based on deep learning algorithms. *MDPI.* 1–16 (2022).

27. Altameem, T. Fuzzy rank correlation-based segmentation method and deep neural network for bone cancer identification. *Neural Comput. Appl.* 1–11. https://doi.org/10.1007/s00521-018-04005-8 (2019).
28. Lin, Q., et al. Deep learning based automated diagnosis of bone metastases with SPECT thoracic bone images. *Sci. Rep.* 1–15 (2021).
29. Jabber, B. et al. SVM Model based Computerized Bone Cancer Detection. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1–5. https://doi.org/10.1109/iceca49313.2020.92976 (2020).
30. DTBV: A Deep Transfer-Based Bone Cancer Diagnosis System Using VGG16 Feature Ext. DTBV: A Deep Transfer-Based Bone Cancer Diagnosis System Using VGG16 Feature Extraction. *MDPI.* 1–12 (2023).
31. Hsieh, T.-C. et al. Detection of bone metastases on bone scans through image classification with contrastive learning. *MDPI.* 1–13 (2021).
32. Ioannis, A. V. et al. Deep learning approaches to osteosarcoma diagnosis and classification a comparative methodological approach. *MDPI.* 1–15 (2023).
33. Kanimozhi, S., Sivakumar, R., & Ananthakrishna, C. Recent advancements in feature extraction and classification based bone cancer detection. 1–19 (2024).
34. Zhao, Z. et al. Deep neural network based artificial intelligence assisted diagnosis of bone scintigraphy for cancer bone metastasis. *Sci. Rep.* **10**, 1–9 (2020).
35. Chu, J. & Khan, S. Transfer learning and data augmentation in osteosarcoma cancer detection. *J. Emerg. Investig.* **6**, 1–6 (2023).
36. Priti, B. et al. Automatic detection of osteosarcoma based on integrated features and feature selection using binary arithmetic optimizat. *Multimed. Tools Appl.* **81**, 8807–8834 (2022).
37. Papandrianos, N., Papageorgiou, E., Anagnostis, A. & Feleki, A. A deep-learning approach for diagnosis of metastatic breast cancer in bones from whole-body scans. *Appl. Sci.* **10**(3), 1–27. https://doi.org/10.3390/app10030997 (2020).
38. Huo, T., Xie, Y., Fang, Y., Wang, Z., Liu, P. & Dua, Y. Deep learning-based algorithm improves radiologists' performance in lung cancer bone metastases detection on computed, 1–10 (2023).
39. Gusarev, M., Kuleev, R., Khan, A., Rivera, A. R. & Khattak, A. M. Deep learning models for bone suppression in chest radiographs. *IEEE.* https://doi.org/10.1109/CIBCB.2017.8058543 (2017).
40. Cheng, D.-C., Hsieh, T.-C., Yen, K.-Y. & Kao, C.-H. Lesion-based bone metastasis detection in chest bone scintigraphy images of prostate cancer patients using pre-train, negative mining, and deep learning. *Diagnostics.* 1–14. https://doi.org/10.3390/diagnostics11030518 (2021).
41. Papandrianos, N., Papageorgiou, E., Anagnostis, A., Papageorgiou, K. & Gwak, J. Bone metastasis classification using whole body images from prostate cancer patients based on convolutional neural networks application. *PLoS ONE* **15**(8), 1–28. https://doi.org/10.1371/journal.pone.0237213 (2020).
42. Do, N.-T. et al. Multi-level Seg-Unet model with global and patch-based X-ray images for knee bone tumor detection. *Diagnostics.* https://doi.org/10.3390/diagnostics11040691 (2021).
43. Sharma, A., Yadav, D. P., Garg, H., Kumar, M. Bone cancer detection using feature extraction based machine learning model. *Hindawi.* 1–13 (2021).
44. Anand, D., Arulselvi, G., Balaji, G. N. & Chandra, G. R. A deep convolutional extreme machine learning classification method to detect bone cancer from histopathological images. *Int. J. Intell. Syst. Appl. Eng.* **10**(4), 39–47 (2022).
45. Kiresur, M. V. & Manoj, P. Bone cancer detection using convolution neural network—an overview. *Int. J. Creat. Res. Thoughts IJCRT.* **9**(3), 1–6 (2021).
46. Yadav, D. P. & Rathor, S. Bone fracture detection and classification using deep learning approach. *IEEE.* https://doi.org/10.1109/PARC49193.2020.236611 (2020).
47. Giradkar, B. & Bodne, N. Bone tumor detection using classification in deep learning using image processing in MATLAB. *Int. Res. J. Eng. Technol. IRJET.* **7**(6), 1–4 (2020).
48. Krois, J. et al. Deep learning for the radiographic detection of periodontal bone loss. *Sci. Rep.* **9**(1), 1–6. https://doi.org/10.1038/s41598-019-44839-3 (2019).
49. Dr, G. M., Anusha, D. H. C., Nayana, U. S. & Shwetha, K. R. Bone cancer detection at earlier stage using convolutional neural network. *IJARIIE.* **7**(2), 1–7 (2021).
50. Ranjitha, M. M., Taranath, N. L., Darshan, L. M. & Subbaraya, C. K. Detection of bone cancer using ct scan images. *J. Emerg. Technol. Innov. Res. JETIR.* **6**(5), 28–32 (2021).
51. Asito, L. Y. et al. Pre-trained convolutional neural networks in the assessment of bone scans for metastasis, 1–6 (2021).
52. Alaa, M. A. et al. Bone abnormalities detection and classification using deep learning-VGG16 algorithm. *J. Theor. Appl. Inf. Technol.* **100**(20), 1–12 (2022).
53. Sindudevi, J. & Kavitha, M. G. A review on bone cancer detection using convolutional neural network. *Int. J. Creat. Res. Thoughts IJCRT.* **12**(2), e908–e924 (2024).
54. Sivakumar, D., Jain, H. K., Bhagwat, G. D., Hegde, M. K. & Natesh, S. Bone cancer detection using machine learning. *Int. Res. J. Eng. Technol. IRJET.* **8**(8), 1–7 (2021).
55. Shao, X. et al. Deep convolutional neural networks combine Raman spectral signature of serum for prostate cancer bone metastases screening. *Nanomed. Nanotechnol. Biol. Med.* https://doi.org/10.1016/j.nano.2020.102245 (2020).
56. Xiong, Y., Guo, W., Liang, Z., Wu, L., Ye, G. Deep learning–based diagnosis of osteoblastic bone metastases and bone islands in computed tomograph images a multicen. *Eur. Radiol.* 6359–6368 (2023).
57. Saba, T. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *J. Infect. Public Health* https://doi.org/10.1016/j.jiph.2020.06.033 (2020).
58. Prathyusha, D. Bone cancer detection using convolutional neural network. *J. Emerg. Technol. Innov. Res. JETIR.* **10**(6), 1–6 (2023).
59. Kumar, S. & Sathiyaprasad, B. Bone cancer detection using feature extraction with classification using K-nearest neighbor and decision tree algorithm. *Smart Intell. Comput. Commun. Technol.* 347–353 (2021).
60. Alkhalaf, S., Alturise, F., Bahaddad, A. A. & Elamin, B. M. Adaptive aquila optimizer with explainable artificial intelligence-enabled cancer diagnosis on medical imaging. *MDPI.* 1–20 (2023).
61. Duran-Lopez, L. et al. Performance evaluation of deep learning-based prostate cancer screening methods in histopathological images: measuring the impact of the model's complexity on its processing speed. *Sensors* https://doi.org/10.3390/s21041122 (2021).
62. Rytky, S. J. O. et al. Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography. *Osteoarthritis Cartil.* **28**(8), 1133–1144. https://doi.org/10.1016/j.joca.2020.05.002 (2020).
63. Li, W., Dong, Y., Liu, W., Tang, Z., Sun, C. *A Deep Belief Network-Based Clinical Decision System for Patients with Osteosarcoma* 1–13 (Springer, 2022).
64. Tang, H., Huang, H., Liu, J., Zhu, J., Gou, F. AI-assisted diagnosis and decision-making method in developing countries for osteosarcoma. *MDPI* 1–20 (2022).
65. Srinidhi, C. L., Ciga, O., Martel, A. L. Deep neural network models for computational histopathology: A survey. *Med. Image Anal.* 1–33. https://doi.org/10.1016/j.media.2020.101813 (2020).
66. Chowdhury, A.A. et al. Detection of osteosarcoma using deep feature extraction with federated learning and MLP from histopathological images. *Res. Square.* 1–17 (2023).
67. Eweje, F. R., et al. Deep learning for classification of bone lesions on routine MRI. *EBioMedicine.* **68**, 1–9 (2021).

68. Ong, W., Zhu, L., Tan, Y. L., Teo, E. C., Tan, J. H. N. Application of machine learning for differentiating bone malignancy on imaging: A systematic review. *MDPI*. 1–23 (2023).
69. Tufail, A. B., Ma, Y.-K., Kaabar, M. K. A. & Martíne, F. Deep learning in cancer diagnosis and prognosis prediction: A minireview on challenges, recent trends, and future directi. In *HindawiComputational and Mathematical Methods in Medicine*, 1–28 (2021).
70. Mandala, S. K. et al. A novel classifier for detecting oropharyngeal cancer using machine learning techniques. In *2024 International Conference on Knowledge Engineering and Communication Systems, ICKECS 2024*.https://doi.org/10.1109/ICKECS61492.2024.10616907.
71. UT Southwestern Medical Center and UT Dallas. *Osteosarcoma-Tumor-Assessment Dataset*. https://digitalcommons.library.tmc.edu/.

Author contributions

All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by B.D.R. and K.M. The first draft of the manuscript was written by B.D.R. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

No financial support was received by the authors in this research.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This research does not involve humans or animals, so no ethical approval is required.

Consent for publication

The authors give consent for their publication.

Additional information

Correspondence and requests for materials should be addressed to B.D.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.