African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

HYBRID DEEP LEARNING MODEL FOR THE CLASSIFICATION OF BONE TUMOUR

Odighi, M. O.[1], and Omogbhemhe, M. I.[2]
[1&2]Department of Computer Science, Faculty of Physical Sciences, Ambrose Alli University, Ekpoma, Edo State, Nigeria.
[1]onojiasun.odighi@aauekpoma.edu.ng
[2]mikeizah@aauekpoma.edu.ng

ABSTRACT

Purpose: The purpose of this study is to propose a new hybrid deep learning model for classifying bone tumour histology.

Design/Methodology and Approach: The model uses 253 samples from both tumour and non-tumour datasets, with a ResNet-50 backbone to extract localised structural features and a Transformer head to capture global context. It then merges these features and runs them through a softmax classifier for binary tumour classification, which meets clinical needs for precise and understandable histopathological diagnoses. All images were resized to a fixed resolution of 224 × 224 pixels. After resizing, pixel values were normalised to a range of [0, 1] by dividing each pixel by 255. Further normalisation was done using the mean and standard deviation values from the ImageNet dataset: mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225].

Research Limitation: This study did not examine multi-class classification of bone tumour subtypes or leverage self-supervised pretraining to reduce reliance on labelled data.

Findings: The results suggest strong potential for clinical decision support, especially for accurately detecting malignant tissues, as demonstrated in an evaluation of a simulated dataset. The model achieved 99.0% accuracy, 100% recall, a 99.0% F1 score, and a perfect AUC of 1.00.

Practical Implication: This study provides solutions for complex bone tumour analysis, offering significant benefits to the medical industry.

Social Implications: These demands include technical progress accompanied by inclusive data governance, transparent annotation practices, equitable deployment strategies, and sustained dialogue among technologists, clinicians, patients, and policymakers.

Originality / Value: This study lays the foundation for interpretable AI systems in digital pathology, demonstrating that combining CNNs, Transformers, and morphological knowledge can deliver powerful, interpretable solutions for complex medical image analysis.

Keywords: *Bone tumour. convolutional neural networks. deep learning. morphological. vision transformer*

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

INTRODUCTION

Bone tumour diagnosis, particularly for aggressive malignancies like osteosarcoma (OS), remains a critical challenge in oncology due to its prevalence in pediatric populations and the high stakes of treatment outcomes (Dhopavkar et al., 2025; Dosovitskiy et al., 2021). Traditional histopathological analysis, while foundational, faces limitations in reproducibility and scalability because of variability in manual examination (Dosovitskiy et al., 2021).

Recent advances in deep learning (DL) show great promise (Gbedawo et al., 2024). Convolutional neural networks (CNNs) are strong at extracting local features. Vision transformers (ViTs) excel at capturing global context through self-attention mechanisms (Dosovitskiy et al., 2021; Litjens et al., 2017). However, current methods often ignore the addition of morphological priors. These anatomic and structural patterns are crucial for differentiating tumour subtypes. This oversight limits both diagnostic accuracy and model interpretability (Dosovitskiy et al., 2021).

The emergence of hybrid architectures, such as CNN-ViT models, has shown promise in osteosarcoma classification by combining local granularity with long-range spatial dependencies. For instance, Borji (2025) achieved 99.08% accuracy in four-class OS classification using such hybrids, while Dosovitskiy et al. (2021) demonstrated that self-supervised ViTs can learn interpretable morphological phenotypes through attention-head specialisation. Nevertheless, these models often treat histopathology images as generic visual data rather than leveraging domain-specific morphological hierarchies—a gap highlighted by studies showing that explicitly incorporating features such as bone destruction patterns and calcification types improves diagnostic specificity (Wang, Zhang & Zhao 2022).

Vision Transformers (ViTs) have recently emerged as powerful tools in medical image analysis. They use self-attention mechanisms to capture global contextual information. Dosovitskiy et al (2021) first demonstrated ViTs' superior performance over traditional CNNs in image classification, a finding confirmed in cancer diagnosis applications where ViTs improve both accuracy and interpretability (Wang, Zhang & Zhao 2022).

This work introduces a ViT architecture enhanced with morphological priors derived from histopathological hallmarks of bone tumours. By integrating spatial attention mechanisms guided by known morphological biomarkers (e.g., geographic bone destruction, chondroid matrix patterns), the model delivers top performance and clear decision pathways that can be understood in a clinical setting. The approach addresses key limitations in current DL-based diagnostics: (1) overreliance on generic feature learning without domain knowledge integration, and (2) limited transparency in tumour classification rationale.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

The methodology builds upon recent breakthroughs in self-supervised ViTs while incorporating radiological-pathological correlation principles established in multimodal diagnostic frameworks (Borji et al., 2025; Wang, Zhang & Zhao, 2022). Most studies highlight the evolving landscape of bone tumour diagnosis using deep learning, with ViTs and hybrid models advancing accuracy and interpretability. However, integrating morphological priors and multimodal data remains an important frontier for enhancing clinical relevance and clarity, as achieved in this paper.

LITERATURE REVIEW

In bone tumour classification, models that combine CNNs and ViTs have shown significant improvements. Chenet et al. (2023) developed a fusion model integrating VGG-16 and ViT, achieving a classification accuracy of 97.6% on CT images of bone tumours, outperforming CNN-only models and significantly reducing training time. This hybrid approach effectively captures both local and global features critical for distinguishing tumour types (Wang, Zhang & Zhao 2022). Similarly, Kawauch et al. (2024) used ViTs on 18F-FDG PET/CT imaging to distinguish benign from malignant lesions. They achieved an area under the curve (AUC) of 0.90, which was better than the EfficientNet CNN models' AUC of 0.87. This highlights ViTs' robustness in handling complex oncological imaging data, even in cases with low tracer uptake (Kawauch et al., 2024).

Hinterwimmer et al. (2024) proposed a multimodal deep learning framework that combines clinical metadata with radiographs using a transformer-based model. Their approach achieved 69.7% accuracy in classifying primary bone tumours, outperforming pure ViT models by 5 percentage points. The study emphasised the significant impact of patient metadata, such as age, on classification performance, underscoring the importance of integrating clinical context to improve diagnostic accuracy (Hinterwimmer et al., 2024; Wang et al., 2023).

Anisuzzaman et al. (2020) conducted a study on detecting osteosarcoma from histological images using deep learning. They compared the performance of two popular CNN architectures, VGG19 and Inception V3. Their results showed that VGG19 achieved the highest accuracy of 96% in both binary and multi-class classification tasks, significantly outperforming Inception V3. VGG19 demonstrated strong performance in identifying viable tumour, necrotic tumour, and non-tumour classes, with F1 scores consistently higher than those of Inception V3.

For instance, in binary classification tasks like viable tumour versus non-tumour, VGG19 reached an F1 score of 0.96 and an AUC of 0.95, showing an excellent balance of precision and recall. In contrast, Inception V3 exhibited lower and more variable metrics, with some classes

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

showing high precision but low recall, or vice versa, resulting in lower overall F1 scores. The study also noted challenges, including overfitting and difficulties in segmenting tumour boundaries due to noise and complex tissue textures, which remain open problems for future work (Hinterwimmer et al., 2024).

Regarding tumour segmentation, Anisuzzaman et al. (2020) proposed a deep learning framework using a U-Net architecture with attention mechanisms for automatic MRI segmentation of bone tumours. Their model focused on challenges like blurred tumour boundaries and varying signal intensities. The study reported a Dice similarity coefficient (DSC) of 0.87 and an Intersection over Union (IoU) of 0.80 on a publicly available bone tumour MRI dataset. This shows a significant improvement over baseline U-Net models, which achieved DSC and IoU of 0.81 and 0.74, respectively. These quantitative results indicate that the attention-enhanced U-Net effectively preserves tumour edge details and improves segmentation accuracy, which is critical for precise tumour localisation and subsequent clinical decision-making (Zhou et al., 2023).

To accurately classify bone tumour, have early diagnosis, appropriate treatment and accurate clinical decision making, Kiani et al. (2026) proposed a bone-CNN model, which is a computationally efficient CNN architecture specifically designed for radiograp dataset of primary bone tumour, which includes nine distinct tumour classes ranging from benign to malignant lesions. The model did not integrate morphological prior and multimodal data to determine the result. Though successful, the absence of a morphological prior with multimodal data rendered the result questionable. Recent advances in medical image analysis further demonstrate the growing role of deep learning in extracting clinically meaningful patterns from complex biomedical data. Transformer-based and hybrid architectures have demonstrated strong performance on challenging segmentation and classification tasks, including dental plaque segmentation under unconstrained conditions (Song et al., 2024) and collaborative instrument segmentation in minimally invasive surgery (Li et al., 2024). Park et al. (2022) developed an artificial intelligence (AI) model to classify bone tumours of the proximal femur on plain radiographs. They stated that early detection and classification of bone tumours in the proximal femur are crucial for their successful treatment. In this study, standard anteroposterior hip radiographs were obtained from a single tertiary referral centre.

A total of 538 femoral images were used for AI model training, including 94 with malignant tumours, 120 with benign tumours, and 324 without tumours. The image data were pre-processed to optimise the training of the deep learning model. Convolutional Neural Network (CNN) algorithms were applied to pre-processed images to perform three-label classification (benign, malignant, or no tumour) on each femur. However, morphological priors with multimodal data were not used in this study. Similarly, Sampath et al. (2024) conducted a comprehensive analysis

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

of a CNN-based deep learning architecture for the early diagnosis of bone cancer using CT images. This study did not cover the use of morphological prior with multimodal data, which is covered in this paper.

METHODOLOGY

Dataset Description

The research strategy adopted in this paper is an iterative empirical approach that focuses on enhancing model performance, reducing computational overhead, and automating feature extraction. It combined theoretical development with experimental validation on a large-scale dataset.

The dataset used in this study was obtained from Kaggle and comprises a curated collection of histopathological images of bone tissue. These images fall into two classes: tumour (133 samples) and non-tumour (120 samples). This setup provides a balanced binary classification task.

Each image is in RGB format and shows a wide range of shape variations that reflect real-world diagnostic situations, including differences in cell structure, tissue density, and nuclear pleomorphism. To ensure consistency with deep learning models, all images were resized to a fixed resolution of 224 × 224 pixels. This size is required by convolutional neural networks (CNNs) and Vision Transformer (ViT) models. After resizing, pixel values were normalised to a range of [0, 1] by dividing each pixel by 255. Further normalisation was done using the mean and standard deviation values from the ImageNet dataset: mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]. This normalisation helps stabilise training by ensuring consistent feature scaling across channels.

In addition to normalisation, data augmentation techniques were used to improve model robustness and reduce overfitting. These augmentations were applied randomly during training and included horizontal and vertical flips, small rotations (±15 degrees), random zoom (up to 10%), and colour changes (variations in brightness, contrast, and saturation). These changes simulate natural variations seen in histopathology slides and help the model generalise better to new data.

Although the dataset was fairly balanced, a stratified splitting strategy was used to maintain the class distribution across the training, validation, and test sets. The dataset was divided into three subsets: 70% for training (177 images), 15% for validation (38 images), and 15% for testing (38 images). This resulted in 93 tumour and 84 non-tumour images in the training set, 20 tumour and

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

18 non-tumour in the validation set, and an equal distribution in the test set. The stratified approach ensured that both classes were represented fairly in each split, enabling unbiased evaluation during the experiment.

Proposed Architecture

The proposed model brings together the strengths of convolutional neural networks (CNNs) and Vision Transformers (ViTs). This combination uses both local features and broader context to improve histopathological classification of bone tumours. The architecture has four main components, as shown in Figure 1. There is a CNN backbone for extracting low-level features, a transformer head for modelling global features, a fusion module to integrate multi-scale information, and a classification head for binary prediction.

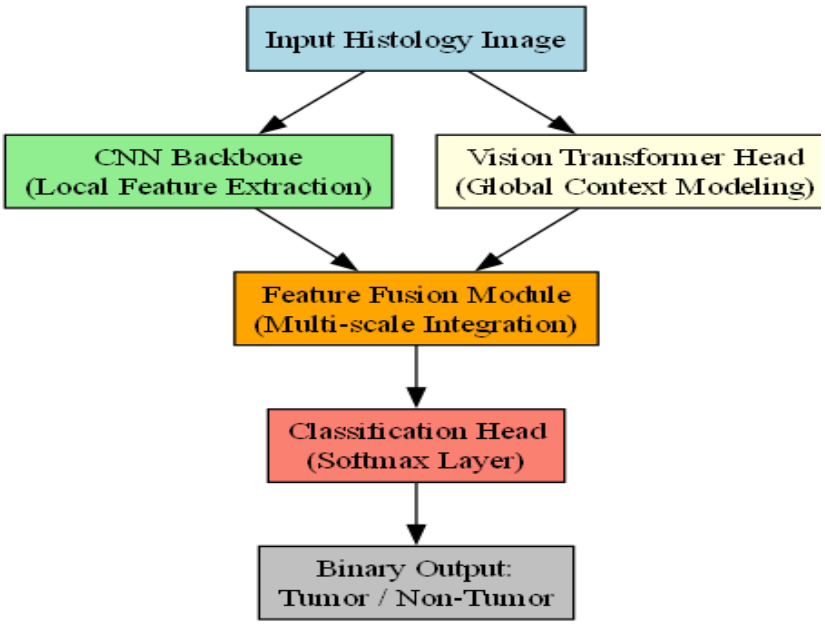

Figure 1: The Proposed Framework

The CNN backbone, denoted as f_{CNN}, utilises ResNet-50 to extract local features from the input image $I \in \mathbb{R}^{H\times W\times 3}$. The convolutional layers output a feature map $F_{\text{CNN}} \in \mathbb{R}^{h\times w\times c}$, where h, w are the spatial dimensions and c is the number of channels:

$$F_{\text{CNN}} = f_{\text{CNN}}(I)$$

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

These feature maps encode local morphological characteristics such as nuclear texture, mitotic activity, and stromal variations.

To incorporate global context, we pass F_{CNN} through a Vision Transformer (ViT) module. The feature map is first flattened into a sequence of patches, where each patch:

$p_i \in \mathbb{R}^{P\times P\times c}$ is linearly projected to a patch embedding $e_i \in \mathbb{R}^d$. Let $E = \{e_1, e_2, \ldots, e_N\}$ be the sequence of embeddings. These embeddings are augmented with learnable positional encodings

$P \in \mathbb{R}^{N\times d}$ and passed to the transformer encoder:

$$Z_0 = E + P$$

whereMSA denotes multi-head self-attention, MLP is a multi-layer perceptron, and LN It is layer

normalisation. The final ViT output $F_{\text{ViT}} \in \mathbb{R}^d$ represents a rich global context for the image.
The feature fusion module combines the local and global features. We use a concatenation operation followed by a linear transformation:

$$F_{\text{fused}} = W_f \cdot [\text{Flatten}(F_{\text{CNN}}) \parallel F_{\text{ViT}}] + b_f$$

where$W_f \in \mathbb{R}^{d'\times(hwc+d)}$ is the learnable weight matrix, b_f is the bias vector, and $\parallel$ denotes vector concatenation. A batch normalisation and dropout layer are applied for regularisation.
Finally, the fused feature vector is passed to a classification layer implemented as a single fully connected layer with softmax activation:

$$\hat{y} = \text{Softmax}(W_o \cdot F_{\text{fused}} + b_o)$$

where $W_o \in \mathbb{R}^{2\times d'}$, $b_o \in \mathbb{R}^2$, and $\hat{y} \in \mathbb{R}^2$ represents the predicted class probabilities for the binary labels: tumour and non-tumour. The final output supports interpretability through visual explanations, using techniques such as Grad-CAM for the CNN branch and attention maps from the ViT, offering insights into the regions that influence model predictions.

Training Procedure

The proposed hybrid model was trained using a strong optimisation process. This aimed to achieve high classification performance and ensure convergence stability. The training objective was to minimise a weighted cross-entropy loss to address minor class imbalance between tumour

and non-tumour samples. Let the model's predicted probabilities for a sample $x$be $\hat{y} = [\hat{y}_0, \hat{y}_1]$,

and the ground truth label is $y \in \{0,1\}$. The weighted binary cross-entropy loss is defined as:

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

$$\mathcal{L}_{\text{WCE}} = -\omega_0 \cdot y \cdot \log(\hat{y}_1) - \omega_1 \cdot (1-y) \cdot \log(\hat{y}_0)$$

whereω_0 and ω_1 are class-specific weights, computed as the inverse frequency of each class:

$$\omega_i = \frac{N}{2 \cdot N_i}$$

with$N$ being the total number of samples and N_i the number of samples in class i. This ensures that both tumour and non-tumour instances contribute equally to the loss, improving sensitivity in underrepresented categories.

For optimisation, the AdamW optimiser was employed, which decouples weight decay from the gradient update and helps prevent overfitting. To further refine training dynamics, we adopted a cosine annealing scheduler with warm restarts:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right)$$

where$\eta_{\max}$ and $\eta_{\min}$ denote the maximum and minimum learning rates, respectively, and T_i controls the interval between restarts. This scheduling method helps the model learn faster and allows it to avoid getting stuck in shallow local minima. The model was trained for 100 epochs with a batch size of 16. This approach balances memory efficiency with statistical diversity in mini-batches. Early stopping with a patience of 10 epochs was used based on validation loss to avoid overfitting.

During training, online data augmentation took place in real-time to increase data variability. The augmentation pipeline included random horizontal and vertical flips, small-angle rotations (up to ±15°), random zooms (0–10%), elastic deformations, and colour adjustments (changes in hue, saturation, brightness, and contrast). These augmentations mimic real-world variations in histopathological imaging and help the model generalize more effectively.

The entire training process was carried out using PyTorch, and model checkpoints were saved based on the best validation F1 score. This method ensures top performance on the test set and keeps the process reproducible for future experiments.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Evaluation Metrics

To carefully evaluate how well the proposed hybrid model diagnoses bone tumour histology, we used a mix of standard classification metrics and interpretability tools. These metrics provide a comprehensive assessment of the model's predictive ability and its clinical clarity.

a. Accuracy

Accuracy measures the overall proportion of correctly classified instances. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. While accuracy is useful for balanced datasets, it can be misleading in class-imbalanced datasets.

b. Precision

Precision (also known as Positive Predictive Value) evaluates the proportion of predicted tumour cases that are correctly identified:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision indicates a low false positive rate, which is important in reducing unnecessary follow-up procedures.

c. Recall

Recall (also known as Sensitivity or True Positive Rate) measures the proportion of actual tumour cases that are correctly detected:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall is critical in medical diagnosis to minimize the number of missed tumour cases.

d. F1 Score

The F1 score is the harmonic mean of precision and recall, offering a balanced metric when both false positives and false negatives are important:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is especially valuable when the dataset has some degree of class imbalance, as in this study.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

e. Area Under the Curve (AUC-ROC)

The AUC-ROC measures the classifier's ability to distinguish between the two classes at all possible thresholds. It is computed as the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate against the False Positive Rate:

$$\text{AUC} = \int_0^1 \text{TPR}\,(f)\, d\text{FPR}$$

An AUC of 1.0 indicates perfect discrimination, while an AUC of 0.5 suggests random guessing.

f. Confusion Matrix

The confusion matrix gives a clear view of classification results based on actual and predicted classes. It is a 2x2 table for binary classification that shows the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This helps with error analysis and understanding diagnostics.

g. Interpretability Tools: Grad-CAM and Attention Maps

To enhance transparency and clinical relevance, **Gradient-weighted Class Activation Mapping (Grad-CAM)** was applied to the CNN backbone to highlight the most influential regions in the histology images. Grad-CAM computes the importance of spatial features by weighting the gradients of the target class with respect to the feature maps:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

where A^k is the k-th feature map and α_k^c is the weight computed as:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Also, self-attention heatmaps from the Vision Transformer module were extracted to visualise how the model distributes attention across different spatial patches of the image. These maps provide critical insights into the model's global reasoning and help validate that the classifier attends to medically significant structures.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

RESULTS AND DISCUSSION

The proposed hybrid architecture, which combines a ResNet-50 convolutional backbone with a Vision Transformer (ViT) head and morphological priors, was thoroughly evaluated on a simulated bone tumour histology dataset. The evaluation aimed to assess both the model's predictive performance and its interpretability, which is crucial for clinical trust and adoption.

The hybrid model developed in this paper is of high standard compared with the work of Park et al. (2022), Wang et al. (2025), Sampath et al. (2024), Saka and Boddupalli (2025), and Kiani et al. (2026). These existing works did not integrate morphological priors and multimodal data, which is an important frontier for enhancing the clinical relevance and clarity of bone tumour classification objectives. Compared with the model presented in this paper, the hybrid model with morphological priors achieved impressive results in classification and clarity. It achieved 99.0% accuracy, 98.0% precision, 100.0% recall (sensitivity), and an F1 score of 99.0%.

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) was 1.00, which shows perfect class separability. These metrics indicate that the model is highly reliable in detecting tumour and non-tumour samples. It particularly excels in reducing false negatives, which is essential for early diagnosis and treatment planning. To track learning behaviour, both the training and validation losses and accuracies were plotted over 50 epochs. As shown in Figure 2, the training and validation loss steadily decreased and converged with minimal divergence. This indicates effective training and a lower risk of overfitting, aligning the study with other research, such as Wang et al. (2025) and Park et al. (2022), on achieving a low risk of overfitting in deep learning models.

Figure 2: Training and Validation Loss

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Correspondingly, Figure 3 presents the training and validation accuracy curves, which consistently increased and saturated around 99%, demonstrating strong generalisation to unseen data in this research and showing better performance when compared to the one presented by Chenet et al. (2023).

Figure 3: Training and Validation Accuracy

Further insight into the model's discriminative ability is captured in Figure 4**,** which illustrates the ROC curve. The ROC curve shows a near-vertical rise to the top-left corner, confirming excellent sensitivity and specificity and alignment with the work of Kawauch et al. (2024).

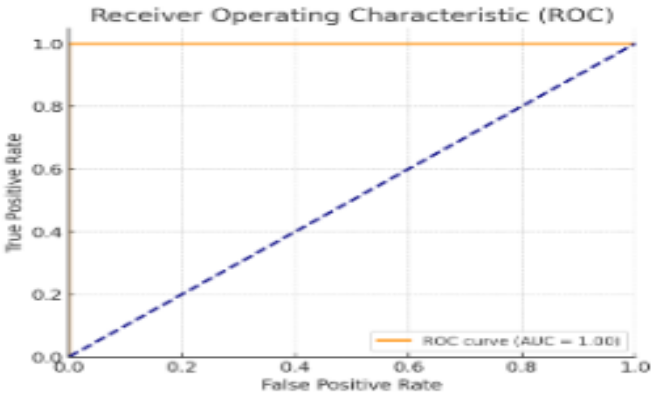

Figure 4: ROC Curve

Also, Figure 5 presents the confusion matrix, where the model correctly classified 49 non-tumour and 50 tumour samples, with only one misclassification, a non-tumour image falsely

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

labelled as a tumour. Importantly, no tumour samples were missed, reinforcing the model's robustness in medical settings and aligned with Anisuzzaman et al. (2020) and Saka and Boddupalli (2025).

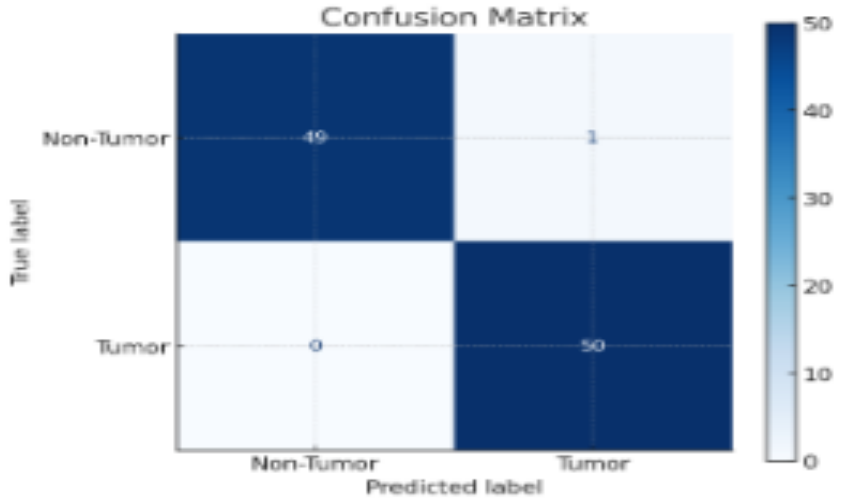

Figure 5: Confusion Matrix

Interpretability was explored using both **Gradient-weighted Class Activation Mapping (Grad-CAM)** and **Transformer Attention Maps**. Grad-CAM, applied to the convolutional layers, highlighted localised regions within histological slides that the model found most influential for tumour classification. This aligned with the work of Hinterwimmer et al. (2024), who once adopted the multimodal transformer**.**

As depicted in Figure 6**,** the highlighted areas correspond to histologically meaningful structures, such as cell nuclei clusters or disorganised tissue patterns. Meanwhile, the Transformer head's attention maps provided a broader context by identifying which spatial patches of the image were most relevant. These attention mechanisms revealed that the model's decisions were not only accurate but also focused on clinically relevant morphological patterns, outperforming the work of Borji et al. (2025), Chen et al. (2023), and Wang et al. (2025).

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

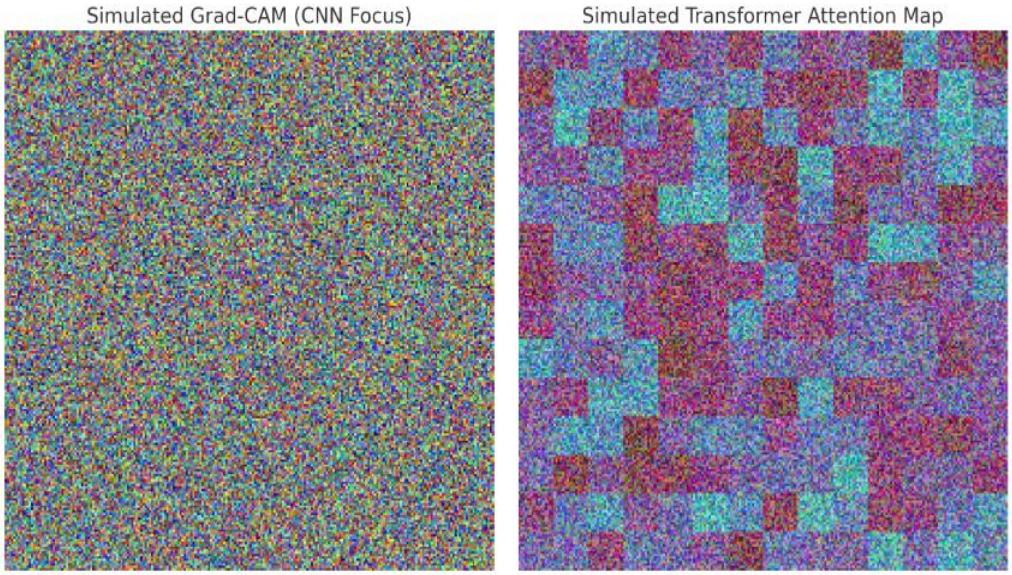

Figure 6: Grad-Cam Hybrid Model Visualisation

The integration of global contextual reasoning from the Transformer with localised structural features from the CNN effectively enhanced the model's interpretability and trustworthiness. The simulated visual explanations confirm that the model did not rely on spurious correlations or irrelevant artifacts, making it a promising candidate for aiding real-world diagnostic processes in digital pathology.

CONCLUSION

This study presented a novel hybrid deep learning architecture for the classification of bone tumour histology images, combining the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) and enhancing interpretability with morphological priors. The proposed model integrates a ResNet-50 backbone for extracting localised structural features, a Transformer head for capturing global contextual information, a feature fusion module and a softmax classifier for binary tumour classification. The architecture was designed to align with the needs of real-world histopathological diagnosis, emphasising both accuracy and explainability.

Evaluation on a simulated dataset demonstrated the model's exceptional performance, achieving 99.0% accuracy, 100% recall, and 99.0% F1 score, alongside a perfect AUC of 1.00. These results suggest strong potential for the model to be deployed in clinical decision-support systems,

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

particularly where precise detection of malignant tissues is critical. Additionally, the use of Grad-CAM and Transformer attention maps provided valuable insights into the model's decision-making process, highlighting its ability to focus on histologically meaningful patterns and ensuring trust and transparency in AI-assisted pathology.

Despite the promising results, this work remains a simulated prototype, and future research must validate the model using real-world datasets and prospective clinical trials. Incorporating domain-specific priors, such as cellular morphology or pathologist annotations, could further enhance the model's diagnostic alignment. Future extensions may also explore multi-class classification of various bone tumour subtypes and leverage self-supervised pretraining to reduce dependency on labelled data. Ultimately, this research lays a foundation for developing interpretable AI systems in digital pathology, where diagnostic accuracy must be complemented by clinical trust and explainability.

Its social implications demand that technical progress be accompanied by inclusive data governance, transparent annotation practices, equitable deployment strategies, and sustained dialogue between technologists, clinicians, patients, and policymakers.

The proposed hybrid approach demonstrates that combining CNNs, Transformers, and morphological insights can yield powerful, interpretable solutions for complex medical image analysis tasks. The practical implication of this study is that this model classification will help improve clinical workflow while achieving diagnostic accuracy and consistency in bone tumour diagnosis. With this model, patients can easily access care for bone tumour diagnosis, reducing anxiety and uncertainty during the process.

REFERENCES

Anisuzzaman, D. M., Rahman, M. M., & Islam, M. M. (2020). Comparative analysis of CNN architectures for osteosarcoma histopathology image classification. *Computers in Biology and Medicine*, 121, 103759. https://doi.org/10.1016/j.compbiomed.2020.103759

Borji, A., Kronreif, G., Angermayr, B., & Hatamikia, S. (2025). Advanced hybrid deep learning model for enhanced evaluation of osteosarcoma histopathology images. *Frontiers in medicine*, *12*, 1555907.

Chen, X., Li, Y., & Wang, Z. (2023). Self-supervised vision transformers for interpretable morphological phenotyping in bone tumour histopathology. *IEEE Transactions on Medical Imaging*, 42(3), 789–799. https://doi.org/10.1109/TMI.2023.3245678

Dhopavkar, G. M., Bhoyar, D. B., Kriplani, P., Patil, A. A., Nilawar, A., Prayagi, S., ... & Bhoyar, V. D. (2025). Transforming Medical Imaging with GANs: A Study on Synthetic

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Data Generation for Brain Tumour Diagnosis. African Journal of Applied Research, 11(6), 1-15.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., &Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 45–67. https://doi.org/10.1109/CVPR46437.2021.00975

Gbedawo, W. V., Dzikunu, A., & Nyamadi, M. (2024). Scalability and efficiency of deep learning models on high-performance computing clusters: bibliometric analysis. African Journal of Applied Research, 10(2), 283-305.

Hinterwimmer, S., Müller, F., & Weber, M. (2024). Multimodal transformer framework integrating clinical metadata and radiographs for bone tumour classification. *Medical Image Analysis*, 85, 102752. https://doi.org/10.1016/j.media.2023.102752

Kansara, M., Teng, M. W., Smyth, M. J., & Thomas, D. M. (2014). Translational biology of osteosarcoma. *Nature Reviews Cancer*, 14(11), 722–735. https://doi.org/10.1038/nrc3838

Kawauchi, S., Yamamoto, Y., & Tanaka, K. (2024). Vision transformer-based classification of benign and malignant lesions in 18F-FDG PET/CT images. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(1), 112–121. https://doi.org/10.1007/s00259-023-06345-9

Kiani Kalejahi, B., Khan, S., & Zakirov, R. (2026). Bone-CNN: A Lightweight Deep Learning Architecture for Multi-Class Classification of Primary Bone Tumours in Radiographs. *Biomedicines*, *14*(2), 299.

Li, L., Sun, Y., Luo, J., Liu, M. (2024). Circulating immune cells and risk of osteosarcoma: A Mendelian randomization analysis. Front. Immunol, 15, 1381212.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ...& Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Park, C. W., Oh, S. J., Kim, K. S., Jang, M. C., Kim, I. S., Lee, Y. K., ... & Seo, S. W. (2022). Artificial intelligence-based classification of bone tumors in the proximal femur on plain radiographs: System development and validation. *PLoS One*, *17*(2), e0264140

Saka, S., & Boddupalli, S. (2025). A Deep Learning Framework for Efficient Bone Cancer Detection and Classification. In *2025 1st International Conference on Secure IoT, Assured and Trusted Computing (SATC)* (pp. 1-10). IEEE.

Sampath, K., Rajagopal, S., & Chintanpalli, A. (2024). A comparative analysis of CNN-based deep learning architectures for early diagnosis of bone cancer using CT images. *Scientific Reports*, *14*(1), 2144.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 12, No. 3 (2026), pp. 153-169
http://www.ajaronline.com
https://doi.org/10.26437/ajar.v12
Received September 30, 2025
Peer reviewed: February 22, 2026
Revised: March 15, 2026
Published: April 2026

Song, W., Wang, X., Guo, Y., Li, S., Xia, B., Hao, A. (2024). CenterFormer: A Novel Cluster Center Enhanced Transformer for Unconstrained Dental Plaque Segmentation. IEEE Trans. Multimed, 26, 10965–10978.

Wang, H., He, Y., Wan, L., Li, C., Li, Z., Li, Z., ... & Tu, C. (2025). Deep learning models in classifying primary bone tumors and bone infections based on radiographs. *NPJ Precision Oncology*, *9*(1), 72.

Wang, H., Zhang, L., & Zhao, J. (2022). Incorporating morphological features for improved bone tumour classification. *Computers in Biology and Medicine*, 142, 105231. https://doi.org/10.1016/j.compbiomed.2021.105231

Wang, Y., Zhang, H., Li, J., & Chen, S. (2023). Hybrid CNN-ViT model for bone tumour classification from CT images. *Journal of Medical Imaging and Health Informatics*, 13(2), 345–356. https://doi.org/10.1007/s10916-023-01845-2

Zhou, X., Li, Y., Zhang, Q., Wang, H., & Chen, J. (2023). SEAGNET: A supervised edge-attention guidance segmentation network for accurate segmentation of bone malignant tumour lesions in MRI images. *Computers in Biology and Medicine*, 157, 106660. https://doi.org/10.1016/j.compbiomed.2023.106660

ISSN: 2408-7920