


Article

Deep Learning Approaches to Osteosarcoma Diagnosis and Classification: A Comparative Methodological Approach

Ioannis A. Vezakis ¹, George I. Lambrou ^{1,2,3} and George K. Matsopoulos ^{1,*}

- ¹ Biomedical Engineering Laboratory, School of Electrical & Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou St., 15780 Athens, Greece; ivezakis@biomed.ntua.gr (I.A.V.); glamprou@med.uoa.gr (G.I.L.)
- ² Choremeio Research Laboratory, First Department of Pediatrics, National and Kapodistrian University of Athens, Thivon & Levadeias 8, 11527 Athens, Greece
- ³ University Research Institute of Maternal and Child Health & Precision Medicine, National and Kapodistrian University of Athens, Thivon & Levadeias 8, 11527 Athens, Greece
- * Correspondence: gmatsopoulos@biomed.ntua.gr

Simple Summary: Osteosarcoma is a rare form of bone cancer that primarily affects children and adolescents during their growth years. Known to be one of the most aggressive tumors, its 5-year survival rate ranges from 27% to 65% across all age groups. Despite the availability of treatment options such as surgery, chemotherapy, and limb-salvage surgery, the risk of recurrence and metastasis remains high even after remission. To improve disease prognosis, it is crucial to explore new diagnostic and treatment methods. Machine learning and artificial intelligence hold promise in this regard. In this study, we adopted a comparative methodological approach to evaluate various deep learning networks for disease diagnosis and classification, aiming to contribute to the advancement of these promising technologies in the field of osteosarcoma research.

Abstract: Background: Osteosarcoma is the most common primary malignancy of the bone, being most prevalent in childhood and adolescence. Despite recent progress in diagnostic methods, histopathology remains the gold standard for disease staging and therapy decisions. Machine learning and deep learning methods have shown potential for evaluating and classifying histopathological cross-sections. Methods: This study used publicly available images of osteosarcoma cross-sections to analyze and compare the performance of state-of-the-art deep neural networks for histopathological evaluation of osteosarcomas. Results: The classification performance did not necessarily improve when using larger networks on our dataset. In fact, the smallest network combined with the smallest image input size achieved the best overall performance. When trained using 5-fold cross-validation, the MobileNetV2 network achieved 91% overall accuracy. Conclusions: The present study highlights the importance of careful selection of network and input image size. Our results indicate that a larger number of parameters is not always better, and the best results can be achieved on smaller and more efficient networks. The identification of an optimal network and training configuration could greatly improve the accuracy of osteosarcoma diagnoses and ultimately lead to better disease outcomes for patients.

Keywords: osteosarcoma; neural networks; machine learning; deep learning



Citation: Vezakis, I.A.; Lambrou, G.I.; Matsopoulos, G.K. Deep Learning Approaches to Osteosarcoma Diagnosis and Classification: A Comparative Methodological Approach. *Cancers* **2023**, *15*, 2290. <https://doi.org/10.3390/cancers15082290>

Academic Editor: Shinji Miwa

Received: 1 February 2023

Revised: 5 April 2023

Accepted: 11 April 2023

Published: 13 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Osteosarcoma (OS), or osteogenic sarcoma, is a malignancy of the bone derived from cells of mesenchymatic origin that exhibit osteoblastic differentiation [1,2]. It is a very aggressive type of tumor and is known to be the most common primary bone cancer. In OS, mesenchymal cells produce osteoid and immature bone tissue, primarily at the loci of bone growth. The immature bone tissue is known to be linked to osteoblast proliferation, and thus it is probable that cells acquire genomic aberrations as well as epigenetic changes.

OS is a disease with a high degree of heterogeneity, which partly explains the difficulty in understanding the molecular machinery underlying its pathogenesis. Several studies have investigated the molecular pathogenetic mechanisms of OS.

Proposed key players in OS pathology include the genes (or proteins) *Rb* [3], *TP53* [4], *Grim-19* [5,6], *P21/RAS* [7,8], and *NF-κB* [9]. *Rb*, a tumor suppressor gene, is aberrantly expressed in several tumors. This gene is known for its role in cell cycle progression and regulation of G1-to-S phase transition. The Rb protein is able to inhibit excessive cell growth, and upon phosphorylation, the pRb inhibits cell proliferation. One of the main findings concerning OS is that both the Rb and p53 (*TP53*) proteins are dysfunctional [10]. *TP53*, or tumor protein 53 is a “transcription factor that regulates critical genes in DNA damage response, cell cycle progression, and apoptosis pathways” [4]. *TP53* acts as a tumor suppressor in all tumor types. In normal cells, the *TP53* levels are low [11], while in tumors, *TP53* is either mutated or down-regulated [11]. The major regulator of *TP53* is MDM2, which acts as a negative regulator and can trigger the degradation of the p53-ubiquitin complex [12]. *TP53* germline mutations are linked to early childhood OS. Another tumor suppressor gene known for its role in OS is *GRIM-19*. Its main function is to mediate apoptosis. Recent studies have indicated that *GRIM-19* is downregulated in OS, while radiation-induced apoptosis in OS is tightly linked to *TP53* upregulation, whose down-regulation infers resistance to radiation-induced apoptosis [5].

Another interesting key molecule in the biology of OS is the transcription factor *NF-κB* [13]. This is one of the well-studied transcription factors for its role in inflammation [14], tumor progression [15], chemotherapy (and radiation) resistance [16], and in particular for its role in OS radiation-related resistance [9,17,18]. Recent studies have shown that *NF-κB* activation is equivalent to tumor resistance in both chemotherapy as well as radiation therapy [17,18]. The mechanism of resistance to chemo- and radiation therapy in OS is still largely unknown, yet it has been reported that *NF-κB* functions through the Akt/*NF-κB* pathway [19,20]. In a very recent report, an explanation was given for this effect, which included evidence that osteosarcoma resistance comes about due to BMI-1 overactivation.

OS is the most prevalent primary skeletal malignancy of childhood and adolescence. It primarily occurs during the adolescent growth spurt between the ages of 10 and 14, and it accounts for 2% of all childhood neoplasms [21]. OS is considered to be a devastating disease. Although in recent years therapeutic advances and options have greatly improved patient survival, metastasis remains the main obstacle in patient prognosis [1]. Patient five-year survival has reached 60–65% in recent years, but the overall prognosis remains poor [1]. Unfortunately, in metastatic cases of the disease, overall survival is as low as 27% [22,23].

Although it is considered a rare form of cancer, approximately 1000 children are newly diagnosed with OS each year in the US alone. In the past, amputations used to be the first line of treatment, aiming to remove the tumor completely. However, advances in imaging techniques, as well as neoadjuvant (preoperative) chemotherapy (NPC) and adjuvant (postoperative) chemotherapy (APC), have facilitated the shift to limb-salvage surgery and increased the five-year survival rate from <20% in the 1970s to 65% in 2015 [24,25]. Following NPC, the achievement of high tumor necrosis rates (>90%) is associated with a significantly higher survival rate and better prognosis [25,26]. Despite recent progress in diagnostic methods, both molecular and imaging histopathological assessment remains the current gold standard for treatment evaluation.

Histopathology includes the evaluation of tissue samples using microscopic examination. In the case of OS, microscopic examination facilitates the estimation of tumor differentiation, invasion, and the presence of necrosis. Still, microscopy is a lengthy, tedious process that is prone to observer bias [27,28]. OS’s high degree of heterogeneity further complicates this process. As such, automating the histopathological evaluation of OS could result in more accurate, fast, and cost-effective examinations [28].

Machine learning approaches are the current state-of-the-art (SotA) method for image classification. Conventional machine learning algorithms such as Support Vector Machines

(SVMs) [29,30] and Random Forests (RF) [31–34] have been widely used in the past for image classification tasks. They rely on a set of features extracted from the images, and their performance is tightly coupled to feature quality. The feature extraction step itself is no easy task; it usually includes hand-crafted methods that fall short on large datasets with a high degree of variability. On the other hand, deep learning architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have achieved impressive results, comparable to human performance in many tasks. Numerous recent reports have highlighted the importance and advantages of deep learning over conventional machine learning approaches in microscopy [35–37]. Unlike SVMs and RFs, these methods do not depend on a feature extraction step; instead, they learn to perform the feature extraction on their own. Their success is highly dependent on the quality and size of the dataset used for training, as well as the overall design of the network architecture. For this reason, new network architectures emerge every year.

The use of machine learning in histopathology has a relatively short history, with the earliest reported studies dating back to the 1990s. These early studies primarily focused on the use of simple image analysis techniques, such as thresholding and edge detection, to identify specific structures within the images [38]. As technology progressed, researchers started exploring more sophisticated methods, such as texture analysis and pattern recognition, to improve the classification of Whole Slide Images (WSIs). WSIs are high-resolution digital images of stained tissue samples captured by a digital slide scanner. The scanner-produced images are large, typically several gigabytes in size and containing millions of pixels. Analyzing an image of this size poses a significant challenge, as current hardware capabilities are insufficient to process potentially thousands of images of this size.

In the present study, we trained several state-of-the-art networks using the same dataset and compared their results to determine which architecture, depth, and input image size was most effective in detecting viable and necrotic tumors, as well as healthy tissue.

2. Methodology

2.1. Methodological Description of Deep Learning Methodologies

The current state-of-the-art CNNs have been designed to work with images from ImageNet or similar databases, with an average resolution of 469×387 pixels [39,40]. In practice, these images are cropped or resized to 224×224 or 256×256 pixels to conserve memory and improve computational efficiency [41]. However, these images are much smaller than WSIs, which can contain up to $100,000 \times 100,000$ pixels [42]. To overcome current technology limitations, researchers typically analyze local mini-patches cropped from the WSIs, and each patch is classified independently. With increasing computational power and memory, larger patch sizes are becoming possible. In theory, a larger patch size should produce more accurate results, as it incorporates a much larger image context and more data points (pixels) [36]. Yet, in the present case, processing large images posed several challenges due to limitations in memory and processing power capacities.

2.2. Dataset

The dataset used in this study was the publicly available Osteosarcoma data from UT Southwestern/UT Dallas for Viable and Necrotic Tumor Assessment (<https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52756935>, accessed 31 January 2023 [43]) [44–47]. The patients included in the dataset were treated at the Children's Medical Center in Dallas between 1995 and 2015, where they underwent surgical resection. The resected bone was then cut into pieces, de-calcified, treated with an H&E stain, and converted to slides [45]. The slides were scanned into digital WSIs, 40 of which were manually selected by two pathologists based on tumor heterogeneity and response characteristics. From each of these WSIs, 30 tiles of size 1024×1024 pixels were randomly selected, resulting in 1200 tiles. After filtering out non-tissue, ink-mark regions, and blurry images, 1144 tiles were selected for analysis. Each of these tiles was manually classified by the pathologists as Non-Tumor (NT), Viable Tumor (VT), or Necrosis (NC), with the

following distributions: 536 (47%) NT, 263 (23%) NC, and 345 (30%) VT. It should be noted that 53 out of the 263 (20%) NC images also contained segments of VT. An example of the images is shown in Figure 1.

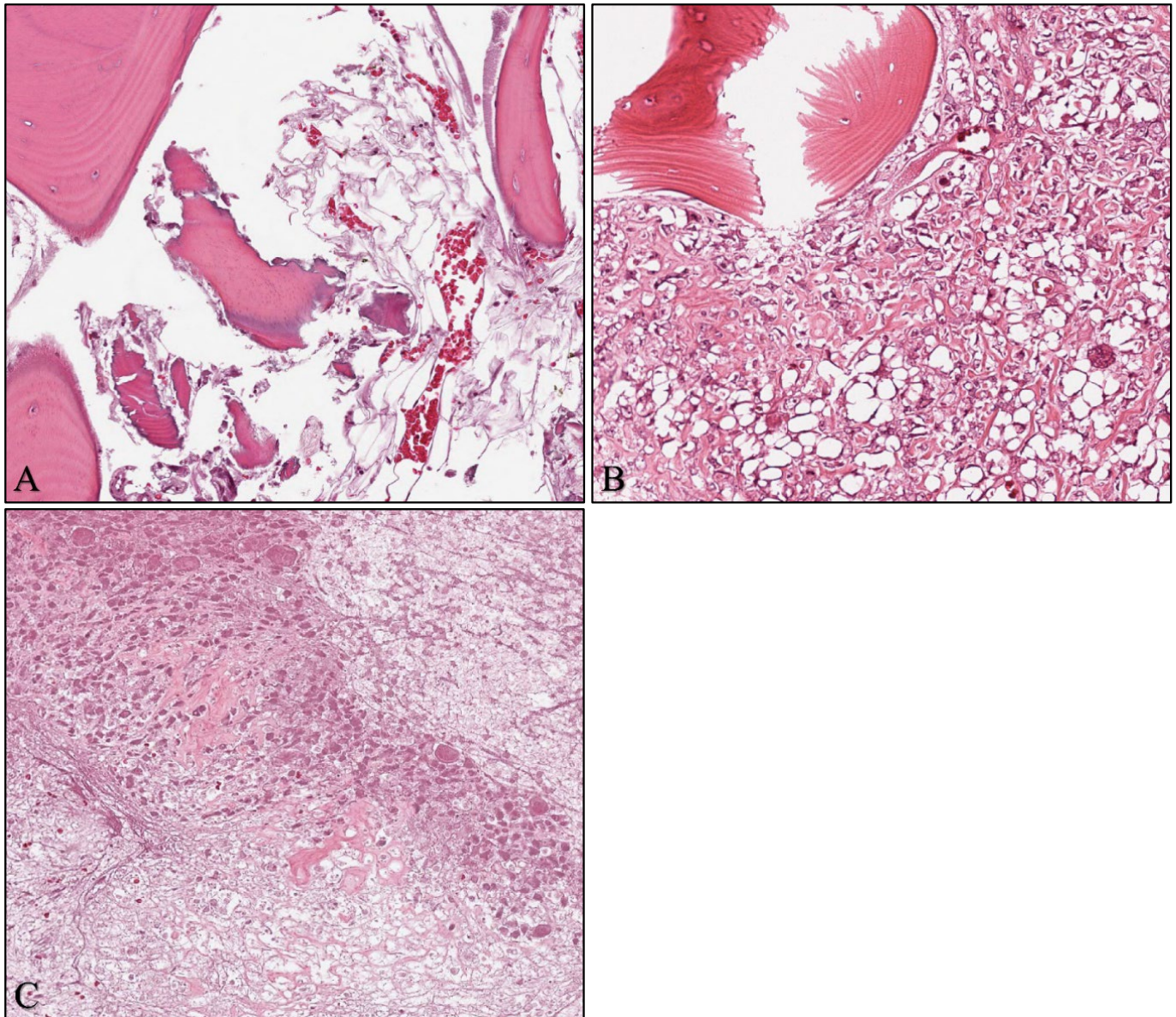


Figure 1. Example images from the Osteosarcoma dataset (magnification $\times 10$). (A) Non-tumor; (B) viable tumor; (C) necrosis.

2.3. Experimental Setup

We compared the performance of various state-of-the-art deep learning architectures on the Osteosarcoma dataset. To train our models, we utilized transfer learning by fine-tuning networks that had been pre-trained on the ImageNet dataset. This allowed us to leverage the knowledge learned by the pre-trained model and apply it, with adjustments, to the much smaller Osteosarcoma dataset.

All computations were performed with the PyTorch framework. The networks were trained on a single NVIDIA Titan Xp GPU with 12GB of memory. The source code is available on Zenodo (<https://doi.org/10.5281/zenodo.7765031>, accessed 23 March 2023).

The chosen deep learning architectures included the Visual Geometry Group network (VGG) [48], the Residual Network (ResNet) [49], the MobileNetV2 [50], the EfficientNet family of networks [51], and the Vision Transformer [52]. We chose these architectures because they are well-established with proven success in image classification, as they have all previously achieved state-of-the-art results on ImageNet.

A particular network architecture can usually be scaled up or down in terms of size (i.e., number of parameters) depending on the requirements of a specific use case. While larger networks have greater learning capacities, they are also more prone to overfitting, particularly on small datasets. Therefore, in our study, we selected several variants of each network architecture, ranging from small (~2.2M parameters) to large (~86M parameters) models. Table 1 shows the number of parameters for each chosen network variant, which may differ from those reported in their original publications (ResNet [49], VGG [48], MobileNet [50], EfficientNet [51]) due to modifications made to their last layers. For example, a VGG16 normally has 138M parameters. However, in this study, we followed the work of Anisuzzaman et al. (2021) [53], where they substituted the last fully connected layer of size 4096 neurons (which makes up a classifier containing ~120M parameters), with two fully connected layers of sizes 512 and 1024 neurons, containing only ~13M parameters. This modification resulted in a much smaller network without significant changes to its architecture. Modifying the last layer of the networks was necessary because they were originally designed to classify images among 1000 categories, whereas in our use case, we only required classification among three categories.

Table 1. The number of parameters for each network variant.

Model	Number of Parameters
EfficientNetB0	4.0 M
EfficientNetB1	6.5 M
EfficientNetB3	11 M
EfficientNetB5	28 M
EfficientNetB7	64 M
MobileNetV2	2.2 M
ResNet18	11 M
ResNet34	21 M
ResNet50	24 M
VGG16	28 M
VGG19	33 M
ViT-B/16	86 M

To compare the performance of the different deep learning architectures, we split the dataset into a training and a test set, with a 70/30 split. We ensured that each network was trained and evaluated on the same set of images by using the same split each time, thus providing a fair comparison.

For all experiments, we used the Adam optimizer [54] with a decoupled weight decay [55] and a learning rate (LR) of 0.0003. We used a cosine annealing learning rate to reduce the LR from 3×10^{-4} to 1×10^{-5} over the 100 epochs of training. Although Adam may not have been the best optimizer for all tasks, it is a widely used default choice in the literature [46,53,54,56,57]. We choose to use it because our main goal was to compare the performance of different architectures rather than to optimize hyperparameters.

The batch size varied depending on the network architecture and the size of the input images. For the largest image size of 1024×1024 pixels, we set the batch size to 2, as larger batch sizes could not fit in the GPU memory for most networks. For networks where even a batch size of 2 did not fit into the GPU memory, we trained them on a smaller image size of 896×896 pixels instead. A notable exception was EfficientNetB7, which was too large to fit in memory even with the reduced image size.

In addition to training on the full-resolution images, we also trained the networks on down-sampled versions of the images, with resolutions of 512×512 and 256×256 pixels. We performed down-sampling using bilinear interpolation and doubled the batch size when the image size was halved. Down-sampling is frequently performed when the goal is image classification as opposed to segmentation, to reduce the computational cost and memory requirements while still achieving good results. Although down-sampling is a destructive process that removes information, it is not a major problem for image classification as the network is only interested in the class of the image, not the exact location of an object. Moreover, down-sampling can even be beneficial as it effectively enlarges the receptive field of the CNN's convolutional layers [16], allowing the network to learn more global features. On the other hand, there is a trade-off between the benefits of the enlarged receptive field and the loss of information due to down-sampling. Therefore, it is not always clear which input image size is optimal.

Further to image resizing, we normalized each RGB channel of the input image independently by subtracting the mean and dividing by the standard deviation. To match the input images with those the networks were pre-trained with, we used the means and standard deviations of the ImageNet dataset, rather than the OS. The mean values used were 0.485, 0.456, and 0.406, and the standard deviation values were 0.229, 0.224, and 0.225 for the R, G, and B channels, respectively.

In addition, data augmentation techniques were applied during training to increase the diversity of the training set. Specifically, we used random horizontal and vertical flipping, as well as random rotation within 20 degrees.

2.4. Network Evaluation

We trained the following network architectures on the Osteosarcoma dataset: EfficientNetB0, EfficientNetB1, EfficientNetB3, EfficientNetB5, EfficientNetB7, MobileNetV2, ResNet18, ResNet34, ResNet50, VGG16, VGG19, and ViT-B/16. The ViT's architecture was designed to treat the input image as a sequence of 14 patches with a size of 16×16 pixels, thus limiting the image size to exactly 224×224 pixels. As a result, we could not experiment with different image sizes for this model due to the need to load pre-trained weights.

We evaluated the performance of each trained network on the test set and computed the F1 score using the One-vs-Rest (OvR) multiclass strategy. More specifically:

$$F1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (3)$$

where TP , TN , FP , and FN are the true positives, true negatives, false positives, and false negatives, respectively. According to the OvR strategy, each class is treated as a binary classification problem, with the positive class being the class of interest and the negative class comprising all other classes. Therefore, the terms TP and FP refer to the number of images that were classified as belonging to the class of interest, with TP being the number of correct classifications and FP being the number of incorrect classifications. Similarly, TN and FN refer to the number of images that were classified as not belonging to the class of interest, with TN being the number of correct classifications and FN being the number of incorrect classifications.

2.5. Follow-Up Experiment

Following network evaluation and result analysis, we proceeded to select the best-performing combination of network and image-input size, and retrained it using five-fold cross-validation. This was done in order to provide a more accurate estimation of the model's performance, which was not tied to a specific training-validation split. Specifically, we split the dataset into five parts (folds), using four of them to train the network and the remaining one to validate it. We repeated the process five times, with each fold serving as the validation set once. The final performance was then computed as the average over the five folds. The details of the retraining process remained the same as before, i.e., we used the same number of epochs, optimizer, learning rate, and augmentation strategy.

In this experiment, we computed additional performance metrics in order to conduct a more thorough investigation into the network's performance and interpret its usefulness in a clinical setting. To this end, we computed the mean and standard deviation of the F1 Score, Accuracy, Specificity, Recall, and Precision across the five folds, as well as the combination of the Confusion Matrices through summation, and the Receiver Operating Characteristic (ROC) curve.

Expanding on the previous equations, the accuracy (Equation (4)) and specificity (Equation (5)) of the classifier were calculated as:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

$$specificity = \frac{TN}{TN + FP} \quad (5)$$

The confusion matrix is a table that summarizes the performance of a classifier by comparing its predicted labels with the true labels of a test dataset. It provides a breakdown of the number of *TP*, *TN*, *FP*, and *FN* for each class, allowing for a more detailed evaluation of a model's performance.

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classifier as its discrimination threshold is varied. It plots the true positive rate (recall) against the false positive rate for different threshold values. The Area under the ROC curve (AUC) is a commonly used metric to evaluate the overall performance of the classifier, with a higher AUC indicating better performance.

3. Results

3.1. Network Comparison

We compared the performance of the different architectures using the *F1* score, which is a robust measure of classification performance considering both precision and recall. Precision measures the proportion of true positive predictions out of all positive predictions made by the model (Equation (2)), while recall measures the proportion of true positive predictions out of all actual positive samples (Equation (3)). The *F1* score is calculated as the harmonic mean of the two values (Equation (1)), requiring both to contribute evenly in order to get a high *F1* score value. Table 2 shows the *F1* scores of the networks, providing a comparative analysis of their performance in each of the three classes (NT, VT, and NC). In order to compare the overall efficacy of the applied methodologies, we estimated the macro-average *F1* score, which is presented in Table S1.

Table 2. F1 scores for each network and image size using pre-trained weight initialization.

Network	Image Size	F1 Score		
		Non-Tumor	Viable Tumor	Necrosis
EfficientNetB0	1024 × 1024	0.93	0.89	0.84
EfficientNetB0	512 × 512	0.93	0.88	0.83
EfficientNetB0	256 × 256	0.95	0.87	0.85
EfficientNetB1	1024 × 1024	0.93	0.85	0.82
EfficientNetB1	512 × 512	0.94	0.88	0.82
EfficientNetB1	256 × 256	0.95	0.86	0.84
EfficientNetB3	1024 × 1024	0.94	0.89	0.84
EfficientNetB3	512 × 512	0.93	0.86	0.81
EfficientNetB3	256 × 256	0.93	0.87	0.81
EfficientNetB5	896 × 896	0.92	0.89	0.81
EfficientNetB5	512 × 512	0.93	0.87	0.82
EfficientNetB5	256 × 256	0.94	0.84	0.80
EfficientNetB7	512 × 512	0.94	0.88	0.84
EfficientNetB7	256 × 256	0.95	0.87	0.83
MobileNetV2	1024 × 1024	0.82	0.84	0.66
MobileNetV2	512 × 512	0.92	0.85	0.81
MobileNetV2	256 × 256	0.94	0.89	0.85
ResNet18	1024 × 1024	0.83	0.86	0.72
ResNet18	512 × 512	0.92	0.85	0.78
ResNet18	256 × 256	0.92	0.88	0.81
ResNet34	1024 × 1024	0.82	0.87	0.70
ResNet34	512 × 512	0.93	0.92	0.82
ResNet34	256 × 256	0.92	0.92	0.82
ResNet50	896 × 896	0.90	0.89	0.77
ResNet50	512 × 512	0.92	0.88	0.82
ResNet50	256 × 256	0.94	0.89	0.82
VGG16	1024 × 1024	0.63	-	-
VGG16	512 × 512	0.63	-	-
VGG16	256 × 256	0.93	0.89	0.81
VGG19	896 × 896	0.63	-	-
VGG19	512 × 512	0.63	-	-
VGG19	256 × 256	0.63	-	-
ViT-B/16	224 × 224	0.88	0.83	0.72

We found that deeper architectures do not necessarily perform better than shallower ones (Table 2, Table S1). MobileNetV2, by far the smallest network in our experiments, yielded the highest macro-averaged F1 score. EfficientNetB0 performed equally or better than its variants B1, B3, B5, and B7. Furthermore, ResNet34 outperformed its bigger counterpart, ResNet50. This behavior is not surprising, given that we observed overfitting with almost every architecture where the accuracy on the training set approached 100% (Figure S1). Notable exceptions were the VGG networks, which failed to learn altogether except for VGG16, which achieved good results only on image input size of 256 × 256 pixels. This is due to the initial learning rate of 3×10^{-4} for the Adam optimizer being too large, causing large updates to the networks' weights and converging on a suboptimal solution where every sample was classified as NT. Later experiments with the learning rate set to 1×10^{-5} provided results comparable to similarly sized networks. We chose not to include them in this study because (a) they did not impact our conclusions, and (b) we did not perform hyper-parameter optimization for other networks.

Increasing the image size did not appear to provide significant benefits, as it introduced new data points that could cause the model to overfit. This trend was observed in networks (EfficientNetB0, EfficientNetB1, ResNet18, ResNet34, ResNet50, and MobileNetV2). Some networks did achieve slightly higher results when trained on a larger input image size, including EfficientNetB3, EfficientNetB5, and EfficientNetB7. This small increase in performance could be attributed to EfficientNets utilizing compound scaling, a technique

which uniformly scales the network's width, depth, and input resolution between variants. Consequently, bigger EfficientNets were pre-trained on larger input image sizes, which could give them a slight advantage when fine-tuning on larger images.

3.2. Follow-Up Experiment

Based on the macro-averaged *F1* score, we selected the MobileNetV2 with an input image size of 256×256 pixels as the best-performing configuration. The network was re-trained (as described in Section 2.5), and the mean \pm standard deviation of the *F1* score, accuracy, specificity, recall, and precision are summarized in Table 3. To provide a comprehensive overview of the classification performance, we aggregated the confusion matrices obtained from all folds by summation and presented the resulting confusion matrix containing all 1144 samples in Table 4.

Table 3. Means and standard deviations of the performance metrics of MobileNetV2 over 5 folds.

Metrics	Non-Tumor	Viable Tumor	Necrosis
<i>F1</i> Score	0.95 ± 0.02	0.90 ± 0.04	0.85 ± 0.03
Accuracy	0.95 ± 0.02	0.95 ± 0.02	0.92 ± 0.02
Specificity	0.96 ± 0.03	0.96 ± 0.02	0.96 ± 0.02
Recall	0.95 ± 0.03	0.93 ± 0.05	0.83 ± 0.05
Precision	0.95 ± 0.03	0.88 ± 0.05	0.88 ± 0.05

Table 4. Confusion matrix of MobileNetV2 with aggregated results over 5 folds.

		Predicted		
		Non-Tumor	Viable Tumor	Necrosis
Actual	Non-Tumor	510	7	19
	Viable Tumor	3	272	17
	Necrosis	24	30	262

The MobileNetV2 architecture was re-trained using 5-fold cross-validation to confirm the consistency of the results. The obtained mean *F1* scores were 0.95 and 0.90 for NT and VT, respectively, indicating that the model performed well in identifying these categories. However, the *F1* score for NC was lower, with a value of 0.85.

The precision values for NC and VT were similar, indicating that the model was equally confident when predicting either class. However, the lower recall value observed for NC suggests that the model had difficulty distinguishing this category from the other classes. Upon examination of the confusion matrix presented in Table 4, we found that out of 316 NC images, 24 were classified as NT and 30 were classified as VT. We also observed that when NT and VT were misclassified, they were seldom mistaken for each other but rather labeled as NC. This suggests that the observed lower accuracy for NC was not due to class imbalance within the dataset, as the network would have been biased towards classifying more samples as NT if that were the case. Instead, these findings suggest that NC shares visual features with both NT and VT, leading to misclassifications by the network. This behavior was observed throughout all of our experiments (Table 2). Indeed, as explained in Section 2.2, approximately 20% of the NC images also depicted VT. Re-training the network without these ambiguous images significantly increased the network's accuracy (Table 5).

Table 5. Means and standard deviations of the performance metrics of MobileNetV2 over 5-folds, excluding ambiguous images from the dataset under investigation (ambiguous images were considered those that simultaneously included both NC and VT tissue).

Metrics	Non-Tumor	Viable Tumor	Necrosis
F1 Score	0.96 ± 0.03	0.97 ± 0.02	0.93 ± 0.03
Accuracy	0.96 ± 0.03	0.99 ± 0.01	0.97 ± 0.02
Specificity	0.97 ± 0.02	0.99 ± 0.01	0.97 ± 0.04
Recall	0.95 ± 0.06	0.98 ± 0.05	0.93 ± 0.09
Precision	0.97 ± 0.02	0.97 ± 0.03	0.93 ± 0.08

In order to provide an overview of MobileNetV2's performance on each fold, we used ROC analysis and plotted the results in Figure 2. The different curves computed on each fold were superimposed in order to assist in performance comparison. The performance across all folds was consistent, with very high AUC values ranging from 0.98 to 1 for NT (Figure 2A), 0.97 to 0.99 for VT (Figure 2B), and 0.95 to 0.97 for NC (Figure 2C). The observed similarities across all AUCs for all classes indicates that the selected network is likely to achieve similar classification performance on different but comparable datasets.

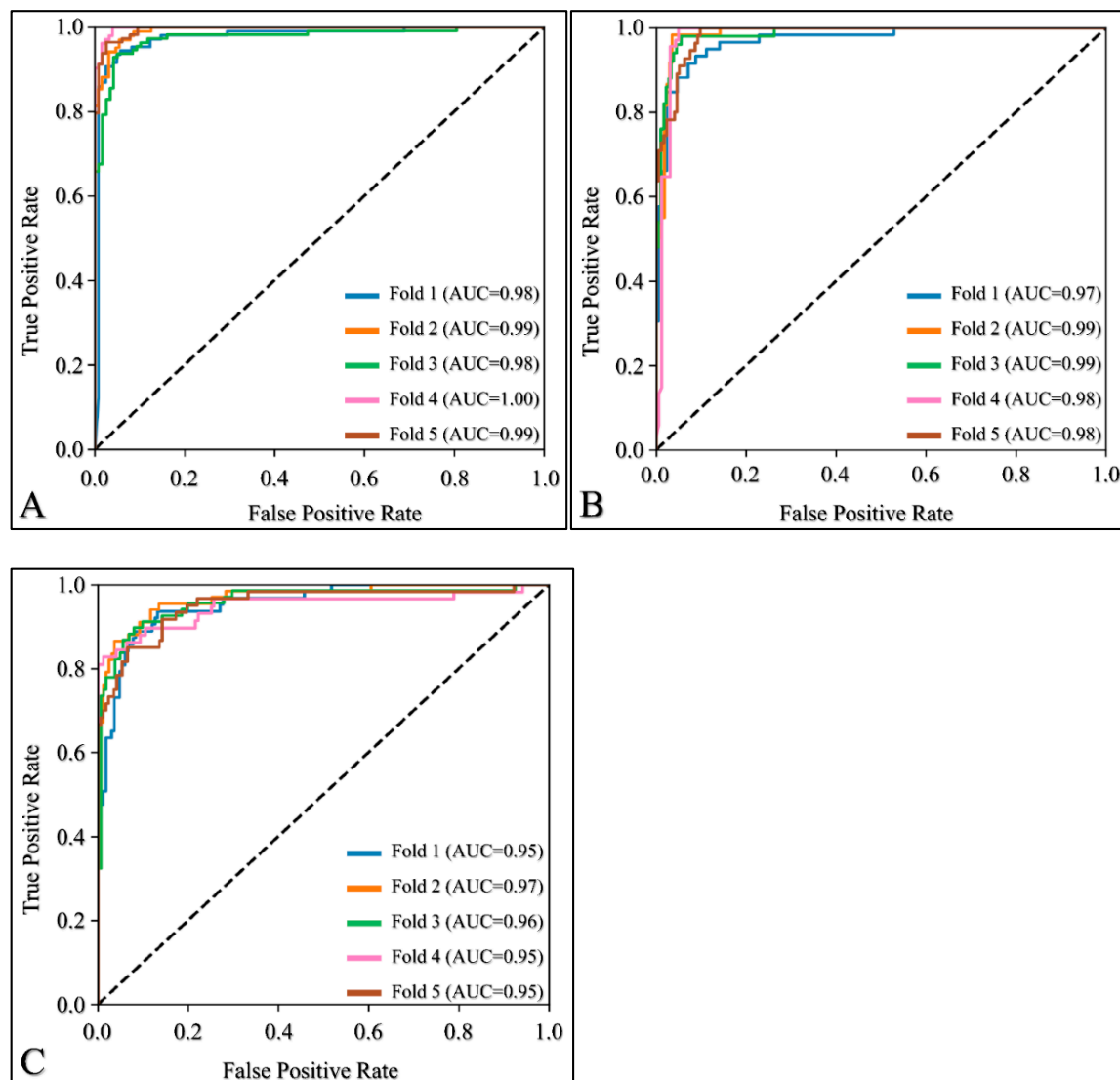


Figure 2. Receiver Operating Characteristic (ROC) curves for (A) non-Tumor, (B) viable tumor, and (C) necrosis.

4. Discussion

In the present study, we investigated various network architectures in order to examine their use in OS microscopy image classification.

4.1. Comparing Neural Networks

In order to obtain a better understanding of network efficiencies, we compared the results of MobileNetV2's re-training with other studies that used the same osteosarcoma dataset (Table 6).

To ensure a fair comparison, we excluded studies that further modified the images using methods such as cropping patches to artificially enlarge the dataset. Such modifications have been shown before to improve performance [35] compared to uncropped tiles, but can also lead to data leakage if not done correctly. For example, two patches from the same tile located next to each other could end up in the training and test set. Consequently, the network would be evaluated on images that are extremely similar to the training set, resulting in accuracies that are too optimistic [58].

Table 6. Comparison of overall accuracy with related work in the same dataset.

Study	Method	Validation Strategy	Overall Accuracy
Arunachalam et al. (2019) [35]	Custom CNN	Holdout	0.910
Anisuzzaman et al. (2021) [53]	VGG19	Holdout	0.940
Bansal et al. (2022) [59]	Combination of HC and DL features	Holdout	0.995
Present study	MobileNetV2	Cross-Validation	0.910

Arunachalam et al. (2019) [35] developed a custom CNN architecture to classify the same osteosarcoma dataset. They reported their results on a limited subset of 230 images, as they performed an 80-20 split between the training and test set (holdout validation strategy). They achieved an overall accuracy of 0.91 and a recall of 89.5, 92.6, and 91.5 for NT, VT, and NC, respectively. Although the authors did not report specific implementation details of the re-training procedure, such as the image input size when using full-size image tiles, a key difference seems to be that their classification approach consisted of two stages. The first stage involved classifying the image as either a tumor or non-tumor. The second stage was executed conditionally: if the image was classified as a tumor, it was then further classified as VT or NC. By using this approach, the authors achieved results similar to ours, despite using a much simpler network with just three convolutional layers that was trained from scratch (as opposed to our transfer-learning approach). This suggests that their hierarchical approach could improve our results as well.

Anisuzzaman et al. (2021) [53] trained several CNNs and reported the best result with VGG19, achieving an overall accuracy of 0.94 using the Adam optimizer and a learning rate of 0.01. Similar to the work of Arunachalam et al. (2019) [35], they followed a holdout strategy and reported their results on a small subset of 230 images. However, our results are markedly different, as we found that VGG19 did not converge on a solution due to the learning rate of 3×10^{-4} for the Adam optimizer being too high. When we re-trained VGG19 with a learning rate of 0.01, we still found that it did not converge (data not shown). In addition, the authors reported a very low average *F1* score with ResNet50, whereas in our experiments, ResNet50 performed well in all cases. Although we tried to follow the authors' implementation as closely as possible by adding two Fully Connected (FC) layers containing 512 and 1024 neurons to the end of the network, there were still some differences that could have influenced the outcome. Firstly, the authors used Keras applications for importing the VGG19 model, whereas we used the Torchvision implementation contained in PyTorch. Secondly, we used a batch size of 4 for the 512×512 input image sizes and 8 for the 256×256 image sizes, which differed from the authors' batch size of 80. Thirdly, the authors downsampled all images to dimensions of 375×375 for training and evaluation,

while our implementation did not. Lastly, we used a slightly different implementation of the Adam optimizer called AdamW, which corrects the way weight decay is implemented [55]).

Bansal et al. (2022) [59] used a combination of handcrafted (HC) features and Deep Learning (DL) features extracted from the Xception Network to train a Singular Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel. They reported an extremely high overall accuracy of 0.995 when tested on a small subset of 219 images, demonstrating a clear advantage when combining features from different methods. In the same work, an overall accuracy of 0.968 with EfficientNetB0 was reported, which is higher compared to our findings. On one hand, this deviation is expected when validating the networks on different sets of samples. We observed folds where our MobileNetV2 achieved an overall accuracy of more than 0.96, but this was not a realistic estimate, as indicated by our cross-validation results. On the other hand, the increased reported accuracy might also be due to a difference in the authors' approach. They modified the network after training by removing its last FC layer to extract a set of features, which were then filtered using a Binary Arithmetic Optimization Algorithm (BAOA) and classified using an RBF-SVM. Further evaluation using cross-validation with both approaches is required to assess whether this technique can further improve network performance.

4.2. Limitations and Future Perspectives

While the present study has yielded valuable insights into the use of deep learning networks for the classification of OS tissue samples, there are several limitations that need to be acknowledged.

Firstly, our study compared the performance of different network architectures using a default set of hyperparameters. While the results showed that smaller networks can outperform larger ones and that MobileNetV2 and EfficientNetB0 were the most effective models, it is possible that the performance of other networks could be improved by optimizing their hyperparameters. Therefore, future research could explore the impact of hyperparameter optimization on the performance of the evaluated networks.

Secondly, the OS dataset contained only a limited number of images that depicted multiple tissue categories, while the majority of images exclusively contained NT, VT, or NC. We observed that after removing ambiguous images depicting both VT and NC, the network's performance was improved. Thus, we expect the accuracy of the network to be reduced in a scenario where images contain any combination of these tissue categories. Increasing the dataset to contain more ambiguous samples or expanding the annotations to include pixel-level classification for segmentation approaches could result in more reliable results.

Lastly, our study was limited by the size of the OS dataset, which contained images from just four patients that were selected by pathologists based on the diversity of tumor specimens. Different tumor types, stages, or even demographics, may result in unique imaging characteristics. These factors would be likely to impact the generalizability of our findings due to the dataset not being representative of the broader population of OS patients. Thus, larger and more diverse OS datasets are required to confirm the effectiveness of the identified networks and to ensure their generalizability to new datasets and populations. The inclusion of data from multiple centers could help further increase dataset diversity.

5. Conclusions

The present study evaluated various deep-learning networks for the classification of osteosarcoma tissue samples. Our results suggested that commonly used deep networks exhibited overfitting, and that smaller networks could outperform larger ones on the present dataset. Specifically, the MobileNetV2 and EfficientNetB0 models led to the most effective overall classification of non-tumors, viable tumors, and necrosis when the original images were downsampled from 1024×1024 pixels to 256×256 . Re-training MobileNetV2 using five-fold cross-validation showed consistent results across all folds, achieving an overall accuracy of 0.91 and mean recalls of 0.95, 0.93, and 0.83 for non-tumors, viable

tumors, and necrosis, respectively. Removing images containing both viable tumors and necrosis further improved our results to mean recalls of 0.95, 0.98, and 0.93, and an overall accuracy of 0.96.

These findings suggest that smaller and more efficient networks can be used to improve results on the osteosarcoma dataset without resorting to increasingly bigger and more complex models. Therefore, we recommend that future research focus on evaluating the results of more aggressive regularization techniques, such as pre-training the models on similar but larger datasets, using more creative augmentation techniques, reducing input dimensionality and batch size, and adding dropout [60]. These techniques could achieve greater results in osteosarcoma datasets and serve as an invaluable tool that, when used in conjunction with the expertise and experience of pathologists, could ultimately lead to improved disease outcomes for patients.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers15082290/s1>, Table S1: F1 scores for each network and image size. Figure S1: The overall training accuracies, plotted over 100 epochs, were superimposed for all networks with the following input image sizes: (A) 256×256 , (B) 512×512 , and (C) 1024×1024 . In all cases, the training accuracy tended to approach a perfect score. VGG networks have been excluded as they failed to learn in our experiments.

Author Contributions: Conceptualization, I.A.V. and G.K.M.; methodology, I.A.V.; software, I.A.V.; validation, I.A.V. and G.K.M.; formal analysis, I.A.V.; investigation, I.A.V.; resources, G.K.M.; data curation, I.A.V.; writing—original draft preparation, I.A.V. and G.I.L.; writing—review and editing, I.A.V. and G.I.L.; visualization, I.A.V.; supervision, G.K.M.; project administration, G.K.M.; funding acquisition, G.I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luetke, A.; Meyers, P.A.; Lewis, I.; Juergens, H. Osteosarcoma treatment—Where do we stand? A state of the art review. *Cancer Treat. Rev.* **2014**, *40*, 523–532. [CrossRef] [PubMed]
2. Koutsomplia, G.; Lambrou, G.I. Resistance mechanisms in the radiation therapy of osteosarcoma: A brief review. *J. Res. Pract. Musculoskelet. Syst.* **2020**, *4*, 15–19. [CrossRef]
3. Thomas, N.A.; Abraham, R.G.; Dedi, B.; Krucher, N.A. Targeting retinoblastoma protein phosphorylation in combination with egfr inhibition in pancreatic cancer cells. *Int. J. Oncol.* **2019**, *54*, 527–536.
4. Matlashewski, G.; Lamb, P.; Pim, D.; Peacock, J.; Crawford, L.; Benchimol, S. Isolation and characterization of a human p53 cDNA clone: Expression of the human p53 gene. *EMBO J.* **1984**, *3*, 3257–3262. [CrossRef] [PubMed]
5. Chen, W.; Liu, Q.; Fu, B.; Liu, K.; Jiang, W. Overexpression of grim-19 accelerates radiation-induced osteosarcoma cells apoptosis by p53 stabilization. *Life Sci.* **2018**, *208*, 232–238. [CrossRef] [PubMed]
6. Lambrou, G.I.; Vlahopoulos, S.; Papathanasiou, C.; Papanikolaou, M.; Karpusas, M.; Zoumakis, E.; Tzortzatou-Stathopoulou, F. Prednisolone exerts late mitogenic and biphasic effects on resistant acute lymphoblastic leukemia cells: Relation to early gene expression. *Leuk. Res.* **2009**, *33*, 1684–1695. [CrossRef]
7. Miller, A.C.; Kariko, K.; Myers, C.E.; Clark, E.P.; Samid, D. Increased radioresistance of ejras-transformed human osteosarcoma cells and its modulation by lovastatin, an inhibitor of p21ras isoprenylation. *Int. J. Cancer* **1993**, *53*, 302–307. [CrossRef]
8. Miller, A.C.; Gafner, J.; Clark, E.P.; Samid, D. Differences in radiation-induced micronuclei yields of human cells: Influence of ras gene expression and protein localization. *Int. J. Radiat. Biol.* **1993**, *64*, 547–554. [CrossRef]
9. Campbell, K.J.; Chapman, N.R.; Perkins, N.D. Uv stimulation induces nuclear factor kappa b (nf-kappa b) DNA-binding activity but not transcriptional activation. *Biochem. Soc. Trans.* **2001**, *29*, 688–691. [CrossRef]
10. Chaussade, A.; Millot, G.; Wells, C.; Brisse, H.; Lae, M.; Savignoni, A.; Desjardins, L.; Dendale, R.; Doz, F.; Aerts, I.; et al. Correlation between rb1 germline mutations and second primary malignancies in hereditary retinoblastoma patients treated with external beam radiotherapy. *Eur. J. Med. Genet.* **2019**, *62*, 217–223. [CrossRef]

11. Surget, S.; Khoury, M.P.; Bourdon, J.C. Uncovering the role of p53 splice variants in human malignancy: A clinical perspective. *OncoTargets Ther.* **2013**, *7*, 57–68.
12. Park, D.E.; Cheng, J.; Berrios, C.; Montero, J.; Cortes-Cros, M.; Ferretti, S.; Arora, R.; Tillgren, M.L.; Gokhale, P.C.; DeCaprio, J.A. Dual inhibition of mdm2 and mdm4 in virus-positive merkel cell carcinoma enhances the p53 response. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 1027–1032. [\[CrossRef\]](#)
13. Gilmore, T.D. Introduction to nf-kappab: Players, pathways, perspectives. *Oncogene* **2006**, *25*, 6680–6684. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Taniguchi, K.; Karin, M. Nf-kappab, inflammation, immunity and cancer: Coming of age. *Nat. Rev. Immunol.* **2018**, *18*, 309–324. [\[CrossRef\]](#)
15. Vlahopoulos, S.A.; Cen, O.; Hengen, N.; Agan, J.; Moschovi, M.; Critselis, E.; Adamaki, M.; Bacopoulou, F.; Copland, J.A.; Boldogh, I.; et al. Dynamic aberrant nf-kappab spurs tumorigenesis: A new model encompassing the microenvironment. *Cytokine Growth Factor Rev.* **2015**, *26*, 389–403. [\[CrossRef\]](#)
16. Nouri, M.; Massah, S.; Caradec, J.; Lubik, A.A.; Li, N.; Truong, S.; Lee, A.R.; Fazli, L.; Ramnarine, V.R.; Lovnicki, J.M.; et al. Transient sox9 expression facilitates resistance to androgen-targeted therapy in prostate cancer. *Clin Cancer Res* **2020**, *26*, 1678–1689. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Zhang, B.; Shi, Z.L.; Liu, B.; Yan, X.B.; Feng, J.; Tao, H.M. Enhanced anticancer effect of gemcitabine by genistein in osteosarcoma: The role of akt and nuclear factor-kappab. *Anti-Cancer Drugs* **2010**, *21*, 288–296. [\[CrossRef\]](#)
18. Tsagaraki, I.; Tsilibary, E.C.; Tzinia, A.K. Timp-1 interaction with alphavbeta3 integrin confers resistance to human osteosarcoma cell line mg-63 against tnfr-alpha-induced apoptosis. *Cell Tissue Res.* **2010**, *342*, 87–96. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Li, K.; Li, X.; Tian, J.; Wang, H.; Pan, J.; Li, J. Downregulation of DNA-pkcs suppresses p-gp expression via inhibition of the akt/nf-kappab pathway in cd133-positive osteosarcoma mg-63 cells. *Oncol. Rep.* **2016**, *36*, 1973–1980. [\[CrossRef\]](#)
20. Yan, M.; Ni, J.; Song, D.; Ding, M.; Huang, J. Activation of unfolded protein response protects osteosarcoma cells from cisplatin-induced apoptosis through nf-kappab pathway. *Int. J. Clin. Exp. Pathol.* **2015**, *8*, 10204–10215.
21. Taran, S.J.; Taran, R.; Malipatil, N.B. Pediatric osteosarcoma: An updated review. *Indian J. Med. Paediatr. Oncol.* **2017**, *38*, 33–43. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Siegel, R.L.; Miller, K.D.; Fuchs, H.E.; Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **2022**, *72*, 7–33. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Smeland, S.; Bielack, S.S.; Whelan, J.; Bernstein, M.; Hogendoorn, P.; Krailo, M.D.; Gorlick, R.; Janeway, K.A.; Ingleby, F.C.; Anninga, J.; et al. Survival and prognosis with osteosarcoma: Outcomes in more than 2000 patients in the euramos-1 (european and american osteosarcoma study) cohort. *Eur. J. Cancer* **2019**, *109*, 36–50. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Friebele, J.C.; Peck, J.; Pan, X.; Abdel-Rasoul, M.; Mayerson, J.L. Osteosarcoma: A meta-analysis and review of the literature. *Am. J. Orthop.* **2015**, *44*, 547–553.
25. Jiang, Y.; Wang, X.; Cheng, Y.; Peng, J.; Xiao, J.; Tang, D.; Yi, Y. Associations between inflammatory gene polymorphisms (tnf-alpha 308g/a, tnfr-alpha 238g/a, tnfr-beta 252a/g, tgfr-beta1 29t/c, il-6 174g/c and il-10 1082a/g) and susceptibility to osteosarcoma: A meta-analysis and literature review. *Oncotarget* **2017**, *8*, 97571–97583. [\[CrossRef\]](#)
26. Bajpai, J.; Kumar, R.; Sreenivas, V.; Sharma, M.C.; Khan, S.A.; Rastogi, S.; Malhotra, A.; Gamnagatti, S.; Kumar, R.; Safaya, R.; et al. Prediction of chemotherapy response by pet-ct in osteosarcoma: Correlation with histologic necrosis. *J. Pediatr. Hematol. Oncol.* **2011**, *33*, e271–e278. [\[CrossRef\]](#)
27. Emerson, J.L. Observer bias in histopathological examinations. In *Carcinogenicity: The Design, Analysis, and Interpretation of Long-Term Animal Studies*; Grice, H.C., Ciminera, J.L., Eds.; Springer: Berlin/Heidelberg, Germany, 1988; pp. 137–147.
28. Asilian Bidgoli, A.; Rahnamayan, S.; Dehkharghanian, T.; Grami, A.; Tizhoosh, H.R. Bias reduction in representation of histopathology images using deep feature selection. *Sci. Rep.* **2022**, *12*, 19994. [\[CrossRef\]](#)
29. Jeong, S.Y.; Kim, W.; Byun, B.H.; Kong, C.B.; Song, W.S.; Lim, I.; Lim, S.M.; Woo, S.K. Prediction of chemotherapy response of osteosarcoma using baseline (18)f-fdg textural features machine learning approaches with pca. *Contrast Media Mol. Imaging* **2019**, *2019*, 3515080. [\[CrossRef\]](#)
30. Zhang, L.; Ge, Y.; Gao, Q.; Zhao, F.; Cheng, T.; Li, H.; Xia, Y. Machine learning-based radiomics nomogram with dynamic contrast-enhanced mri of the osteosarcoma for evaluation of efficacy of neoadjuvant chemotherapy. *Front. Oncol.* **2021**, *11*, 758921. [\[CrossRef\]](#)
31. Buhnemann, C.; Li, S.; Yu, H.; Branford White, H.; Schafer, K.L.; Llombart-Bosch, A.; Machado, I.; Picci, P.; Hogendoorn, P.C.; Athanasou, N.A.; et al. Quantification of the heterogeneity of prognostic cellular biomarkers in ewing sarcoma using automated image and random survival forest analysis. *PLoS ONE* **2014**, *9*, e107105. [\[CrossRef\]](#)
32. Essa, E.; Xie, X.; Errington, R.J.; White, N. A multi-stage random forest classifier for phase contrast cell segmentation. In Proceedings of the In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25–29 August 2015; pp. 3865–3868.
33. Li, W.; Dong, Y.; Liu, W.; Tang, Z.; Sun, C.; Lowe, S.; Chen, S.; Bentley, R.; Zhou, Q.; Xu, C.; et al. A deep belief network-based clinical decision system for patients with osteosarcoma. *Front. Immunol.* **2022**, *13*, 1003347. [\[CrossRef\]](#)
34. Shen, R.; Li, Z.; Zhang, L.; Hua, Y.; Mao, M.; Li, Z.; Cai, Z.; Qiu, Y.; Gryak, J.; Najarian, K. Osteosarcoma patients classification using plain x-rays and metabolomic data. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 690–693.

35. Arunachalam, H.B.; Mishra, R.; Daescu, O.; Cederberg, K.; Rakheja, D.; Sengupta, A.; Leonard, D.; Hallac, R.; Leavey, P. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PLoS ONE* **2019**, *14*, e0210706. [CrossRef] [PubMed]
36. Komura, D.; Ishikawa, S. Machine learning methods for histopathological image analysis. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 34–42. [CrossRef]
37. Melanthota, S.K.; Gopal, D.; Chakrabarti, S.; Kashyap, A.A.; Radhakrishnan, R.; Mazumder, N. Deep learning-based image processing in optical microscopy. *Biophys. Rev.* **2022**, *14*, 463–481. [CrossRef]
38. Ong, S.H.; Jin, X.C.; Jayasooriah; Sinniah, R. Image analysis of tissue sections. *Comput. Biol. Med.* **1996**, *26*, 269–279. [CrossRef] [PubMed]
39. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285–1298. [CrossRef]
40. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Kai, L.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
41. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
42. Wang, F.; Oh, T.W.; Vergara-Niedermayr, C.; Kurc, T.; Saltz, J. Managing and querying whole slide images. *Proc. SPIE—Int. Soc. Opt. Eng.* **2012**, *8319*, 137–148.
43. Leavey, P.; Sengupta, A.; Rakheja, D.; Daescu, O.; Arunachalam, H.B.; Mishra, R. Osteosarcoma data from ut southwestern/ut dallas for viable and necrotic tumor assessment [data set]. *Cancer Imaging Arch.* **2019**, *14*. [CrossRef]
44. Mishra, R.; Daescu, O.; Leavey, P.; Rakheja, D.; Sengupta, A. *Histopathological Diagnosis for Viable and Non-Viable Tumor Prediction for Osteosarcoma Using Convolutional Neural Network*; Springer International Publishing: Cham, Switzerland, 2017; pp. 12–23.
45. Arunachalam, H.B.; Mishra, R.; Armaselu, B.; Daescu, O.; Martinez, M.; Leavey, P.; Rakheja, D.; Cederberg, K.; Sengupta, A.; Ni'suilleabhain, M. Computer aided image segmentation and classification for viable and non-viable tumor identification in osteosarcoma. *Pac. Symp. Biocomput.* **2017**, *22*, 195–206.
46. Mishra, R.; Daescu, O.; Leavey, P.; Rakheja, D.; Sengupta, A. Convolutional neural network for histopathological analysis of osteosarcoma. *J. Comput. Biol.* **2018**, *25*, 313–325. [CrossRef]
47. Leavey, P.; Arunachalam, H.B.; Armaselu, B.; Sengupta, A.; Rakheja, D.; Skapek, S.; Cederberg, K.; Bach, J.-P.; Glick, S.; Ni'Suilleabhain, M. *Implementation of Computer-Based Image Pattern Recognition Algorithms to Interpret Tumor Necrosis; A First Step in Development of a Novel Biomarker in Osteosarcoma*; Pediatric Blood & Cancer; Wiley: Hoboken, NJ, USA, 2017; p. S52.
48. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**, arXiv:1512.03385.
50. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
51. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Chaudhuri, K., Salakhutdinov, R., Eds.; Volume 97, pp. 6105–6114.
52. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
53. Anisuzzaman, D.M.; Barzekar, H.; Tong, L.; Luo, J.K.; Yu, Z.Y. A deep learning study on osteosarcoma detection from histological images. *Biomed. Signal Process. Control* **2021**, *69*, 102931. [CrossRef]
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
55. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
56. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neur.* **2017**, *30*, 1–11.
57. Stanford. Cs231n Convolutional Neural Networks for Visual Recognition. Available online: <https://cs231n.github.io/neural-networks-3/> (accessed on 23 March 2023).
58. Tampu, I.E.; Eklund, A.; Haj-Hosseini, N. Inflation of test accuracy due to data leakage in deep learning-based classification of oct images. *Sci. Data* **2022**, *9*, 580. [CrossRef]
59. Bansal, P.; Gehlot, K.; Singhal, A.; Gupta, A. Automatic detection of osteosarcoma based on integrated features and feature selection using binary arithmetic optimization algorithm. *Multimed. Tools Appl.* **2022**, *81*, 8807–8834. [CrossRef] [PubMed]
60. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.