Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 10, 2025

OSTEOSARCOMA BONE TUMOR DETECTION USING DEEP ENSEMBLE TRANSFER LEARNING ARCHITECTURE

Fouzia Nourin[1], Muhammad Shadab Alam Hashmi[*2]

[1]fouzianourin00@gmail.com, [*2]shadab.alam@kfueit.edu.pk

DOI: https://doi.org/10.5281/zenodo.18720941

Keywords
Osteosarcoma, Osteoblasts, CNN, Image Classification, Imbalanced Dataset, SMOTE.

Article History
Received: 19 August 2025
Accepted: 29 September 2025
Published: 10 October 2025

Corresponding Author: * Muhammad Shadab Alam Hashmi

Abstract
Being the most common bone cancer affecting children, osteosarcoma emerges during periods of rapid growth and development in bones which normally occurs during childhood and adolescence. Deriving from the under differentiated formation of osteoblasts, it leads to a haphazardly made, poorly mineralized bone matrix. Although it's the most common bone tumor type yet the exact mechanisms controlling the onset of osteosarcoma are unknown. Only a few proteomics analyses have been per- formed so far. Consequently, all available studies have mainly identified other cancer-related proteins rather than specifically osteosarcoma-associated ones. This paper aims to emphasize the growing scope of proteomics cutting-edge methodologies, especially their customized use to develop new diagnostic and therapeutic osteosarcoma targets. Moreover, it describes recent knowledge on signaling markers and pathways related to both the original and metastatic types of osteosarcomas from early- stage proteomic profiling studies. However, there is a major class imbalance problem in this dataset. To address this problem, we used the Synthetic Minority Over- sampling Technique (SMOTE). This technique helps distribute images across classes evenly to prevent class stability issues. We evaluate the model based on precision, which stands at 98%, sensitivity or recall at 96%, and F1 score at 97%. Thus, on analysis the final outcome, we found out that the proposed CNN model performs better than the other models on all the given evaluation metrics.

INTRODUCTION

Osteosarcoma, which is also known as the osteogenic sarcoma, is a rare but aggressive form of primary bone cancer that commonly affects children adolescent and young adults with a peak instance observed between the ages of 10 and 30. Specifically in the United States of America, almost 1,000 cases of Osteosarcoma are diagnosed each year [1], in which half of it are almost children and teens. Also, as the one of the most prevalent malignant bone tumors in this demographic, its account for approximately 56% of all the pediatric bone cancers, followed closely by Ewing sarcoma. Its occurrence is mostly commonly localized in the metaphysical regions of long bones with the distal femur, proximal tibia, and proximal humerus being frequent sites, where the humerus is involved in 10%, with 90% in the proximal humerus [2]. Despite advance in treatment modalities such as surgery and neoadjuvant chemotherapy, the prognosis for metastatic or recurrent osteosarcoma remains poor because the survival rate drop significantly below 20% after five years. At its early stages, somewhat heated, red, and painful bones in a specific area where the tumor is present are the classical indication of osteosarcoma [3]. Symptoms of Osteosarcoma may manifest as discomfort, edema, and stiffness in the damaged bone, as well as trouble limb movement. It's possible to see a bump or

tumor on the afflicted bone. Although the precise origin of osteosarcoma remains uncertain, a number of risk factors have been identified, such as radiation treatment history in the past, and the existence of certain genetic abnormalities including both a history of Paget's illness and Li-Fraumeni syndrome. A cancerous form of bone cancer that mostly affects the spine is called spinal osteosarcoma. Spinal osteosarcoma often affects older age groups [4], whereas the typical age of osteosarcoma of the extremities is 38. Its potential to develop quickly and spread (metastasize) to other bodily regions, especially the lungs, is what poses a threat. It may result in excruciating pain, neurological impairments, and even paralysis because of its proximity to vital nerves and the spinal cord. The mortality rate from osteosarcoma is notably higher than that from other malignancies. However, the overall survival rate of osteosarcoma may be raised by prompt diagnosis and close observations throughout the treatment cycle [5]. Thus, in order to spot the osteosarcoma with multifaceted method, we can utilize clinical evaluation, imaging systems such as X-rays, CT, and MRI, as well as histological inspection through biopsy. But all of these traditional approaches face challenges due to the overlap of radiographic features with other benign or malignant bone lesions, which make it necessary for a top level of expertise and knowledge for correct explanation and we can't only depend on humans in this field as of their subjective explanations. In order to lessen human dependency, Deep Learning (DL), a sub-set of Machine learning, is being used a lot by researchers. Transfer learning approaches, powered by Convolutional Neural Networks (CNN), these DL models can effectively and efficiently analyze high-dimensional histopathological images. These models are pre- trained on large datasets such as ImageNet and are fine-tuned to identify osteosarcoma-specific features with remarkable accuracy. Moreover, innovative data augmentation techniques like the Synthetic Minority Oversampling Technique (SMOTE) have been employed to mitigate class imbalance issues inherent in osteosarcoma datasets, which actually helped in improving model robustness and generalizability [6]. In this following research the major contribution we made in detecting the bone tumor are:

1) **Novel Architecture and Learning Strategy:** In this research we suggest a hybrid ensemble model which is based on a type of EfficientNet, which is specifically customized for histopathological osteosarcoma image classification, in such domains hybrid models are rarely used. This hybrid model is capable of obtaining detail pattern of tumor and in this way it increases diagnostic accuracy. This ensemble model gives more accuracy, robustness, scalability and better deployment, especially in those environments where resources are limited.

2) **Advanced Data Engineering and Class Imbalance Resolution:** We introduce a modified SMOTE strategy adapted for CNN-based image inputs, addressing severe class imbalance in medical datasets by synthesizing samples in latent space and reshaping them for compatibility with convolutional structures. Our label design departs from conventional binary schemes by using a four-class system that includes borderline and transitional tumor states (e.g., Viable Non-Viable), enabling more granular tumor classification and aiding clinical interpretability. The dataset preprocessing and reshaping pipeline is designed to preserve spatial integrity while aligning with ensemble-level input constraints, a nuance often overlooked in typical resizing strategies.

I. Literature Review

Osteosarcoma is an aggressive bone cancer that predominantly affects adolescents, and is also challenging to be diagnosed early to have timely treatment. Recent advancements in machine learning, particularly in deep learning have show potential in improving the accuracy of osteosarcoma detection through medical imaging. For Instance, Biopsy was the recommended method for a conclusive diagnosis, albeit it was a time consuming and challenging process that could benefit from automation. Recently, Transfer Learning (TL) techniques are used to get better result in less time [7]. A diagnostic technique for osteosarcoma aimed at reducing the burden of medical professionals involved in determining osteosarcoma through three essential aspects. Initially, the researchers developed a technique for improving

classification images incorporating resnet18 and Deep UPE. This method was designed to increase the clarity of the images and remove unnecessary ones, facilitating efficient professional oversight. Since we are talking about the Transfer Learning models, another research utilized it because they are pre-trained and we don't need to start everything from scratch, which reduces the need for complex datasets and improve efficiency. This specific research focuses on evaluating the performance of different TL models for bone metastasis detection. The dataset consists of H&E-stained images divided into four groups: potential tumors; Both the unexpected tumor metastasis and malignant tumor datasets resulted in an 80-20 split between training and testing datasets. Based on the data, four models previously trained on ImageNet were compared: EfficientNetB7, InceptionResNetV2, NasNetLarge and ResNet50. InceptionResNetV2 achieved the highest accuracy (93.29%), followed by EfficientNetB7 (62.77%), NasNetLarge (90.91%) and ResNet50 (89.83%). Furthermore, among the four models, InceptionResNetV2 achieved the best score and got Accuracy of (0.8658) [8]. Along with that, some researchers wanted to indicate the effectiveness of 12 previously trained DL models performed for classifying osteosarcoma which can highlight the significance of choosing models with smaller parameters values. For this, 30% of the dataset was put aside for testing and the remaining 70% was used for training. The PyTorch framework was used to fine-tune the pre-trained models and it was discovered which networks performed the best with appropriate image input sizes. Based on the macro-average F1 score, MobileNetV2 was shown to be the most effective model overall [9].

Another paper present a hybrid framework to increase the accuracy of classification of osteosarcoma tumors. The researchers actually targets three types of osteosarcoma tumors in it, starting from the necrosis tumor, then viable tumor and in the end the nontumor. This method entails combining different CNN-based architectures with the WSI dataset and an MLP algorithm. The five pre-trained CNN models were fine-tuned with several parameters' configurations after transfer learning across different preprocessing steps applied to WSI images to extract dominant features. Features extractors consisted of convolutional layers combined layers combined with pooling operations. Decision Tree-based Recursive Feature Elimination was used to select relevant features by recursively removing the least impactful ones to improve generalization and the quality of the model. The Decision Tree estimator has been applied for feature selection. Finally, a customized MLP classifier was used for two-class and multi-class osteosarcoma classification with five-fold cross validation as a robust test of the proposed model. The experimental results showed good performance in comparison with competing approaches. This suggests its use to support the diagnosis of osteoporosis in specialized hospital [10].

To conduct another study, researchers utilized a dataset comprising hematoxylin and eosin-stained images of bone cancer. The distribution is uneven. Mention worries about the effect on the reliability and generalizability of research. To solve this problem, CNN uses deep learning and classifies voters based on different learning methods. Bone cancer classification is known. This approach aims to reduce bias by increasing evenly distributed training data. And use additional data to improve technical skills. In the freezing phase and optimization phase, six pre-trained CNN models were implemented and evaluated: MobileNetV1, MobileNetV2, ResNetV250, InceptionV2, EfficientNetV2B0, and NasNetMobile. Furthermore, for the purpose of classifying osteosarcomas, both a modified ensemble learning-based vote classifier and a novel CNN model were introduced. These were constructed using CNN model, refined NasNetMobile model, and refined EfficientNetV2B0 model. The study's conclusions have applications in the fields of telemedicine, mobile healthcare, and medical professionals' support tools [11]. The model's approaches that are used in osteosarcoma bone tumor detection are shown in Table I. Latest Research utilizes the Group Teaching Optimization Algorithm with Deep Learning-Driven Osteosarcoma Detection on Histopathological Images (GTOADL-ODHI) [12]. It actually, introduced a multi-stage framework that begins with Gaussian filtering (GF) for noise reduction in HIs, followed by the feature extraction using a capsule network (CapsNet), which is fine-tuned using Group Teaching Optimization

Algorithm (GTOA) to ensure optimal performance. Eventually, a self- attention bidirectional long short-term memory (SA-BiLSTM) model is employed for the recognition and classification of osteosarcoma. Thus, this model performed exceptionally well in contrast to the regularized CNNs models like AlexNet, DBN, XGBoost and NB because it outperformed them with an accuracy of 99.13%, precision of 98.45%, recall of 98.28%, and F1 score of 98.36%. Subclassifying osteosarcoma tumors based on molecular characteristics is a challenging yet crucial task due to the complexity of cellular origins and intra-class variations but this study introduces a novel approach that integrates GhostNet with an upgraded ResNet for enhanced feature extraction and classification [13]. It leverages a augmented Faster Region Convolutional Neural Network (FRCNN) and the Sooty Tern Optimization Algorithm (STOA) for parameter fine-tuning by which the proposed model achieves impressive accuracy rates of 97% for binary and 96.83% for secondary class classification. Image augmentation techniques further mitigate data imbalance, which helped to increase the model robustness and generalizability.

The model's approaches that are used in osteosarcoma bone tumor detection are shown in Table I. Meanwhile these recent studies have successfully applied transfer learning and CNN architectures like ResNet, MobileNet, and InceptionResNetV2 for osteosarcoma detection, most focus on binary or limited multi-class classification, often neglecting fine-grained tumor categories crucial for clinical relevance. Many rely on off- the-shelf models with minimal architectural adaptation, and class imbalance is typically addressed with basic augmentation rather than structured solutions. Though hybrid models and optimization algorithms show improved performance, they often come at the cost of complexity and reduced scalability. These limitations highlight the need for efficient, ensemble- based frameworks tailored to diverse tumor types and designed with deployment feasibility in mind , which is an area our proposed model directly addresses.

II. METHODOLOGY

A. Modeling of Proposed EfficientNet-CNN Ensemble Model

This research introduce a deep learning (DL) framework using an EfficientNet-CNN ensemble model for the detection of the Osteosarcoma bone tumors. The purposed model utilize the Efficient architecture, which is known for its balance be- tween computational efficiency and high accuracy by optimizing network depth, breath, and precision through hybrid scaling methods. This optimization ensures exceptional performance in computer vision tasks, such as image classification (which we actually going to use) and object detection, while being able to adapt to resource constraints [14]. Some elaborative key feature for the EfficentNet-CNN model is following:

1) **Ensemble Approach:** Combines multiple CNN architectures to capture diverse features and patterns from complex datasets to train over model on diversity. Uses integration strategies, for instance weighted predictions, to strengthen the overall model performance and mitigate individual weak-nesses.

2) **Image Preprocessing:**

- Reshaping: Converts input images into a standard for- mat (150x150x3) for compatibility with the network.
- **Rescaling:** Normalizes pixel values to a [0,1] range to ensure consistent input and reduce computational complexity.
- **Label Assignment:** Categorizes images into four classes—Non-Tumor, Viable Tumor, Non-Viable Tu- mor, and Viable Non-Viable—for effective model training.

3) **Synthetic Minority Oversampling Technique (SMOTE):** Addresses data imbalance by generating synthetic samples, reshaping data into a 1D vector for processing and then back to 2D arrays for compatibility with the CNN.

4) **Data Splitting:** Divides datasets into training, validation, and testing sets to ensure robust model performance and prevent overfitting.

5) **Model Compilation:** Specifies optimization parameters, loss functions, and performance metrics to prepare the model for training.

6) **Training and Evaluation:** Supervised learning is used with user-defined parameters like

batch size and iteration count to optimize model learning. Performance is assessed through metrics like accuracy, ensuring the model generalizes well to unseen data.

Figure 1 shows Work flows of proposed EfficientNet-CNN classification for Osteosarcoma bone tumors.

B. Addressing the Class Imbalance Problem

The class imbalance problem occurs in medical datasets when one class (e.g. patients with a certain disease) is under- represented compared to another class (e.g. healthy patients). This can happen for multiple motives like a lack of funding for research on a particular disease, or a higher prevalence of one disease over another. The problem with class imbalance is that it could result in a biased model, as the model is more probable to forecast the class majority, even when the minority class is present. Thus, to mitigate this issue some modern techniques like oversampling the minority class, under sampling the majority class, and advance methods such as cost-sensitive learning, ensemble approaches, and meta-learning could be utilized [15].

1) SMOTE: Cutting to the chase, the technique we utilized to mitigate the concern of the class imbalance was Syn- thetic Minority Over-sampling Technique (SMOTE), which functions by generating synthetic samples from the minority class instead of merely duplicating existing ones [17]. It achieves this by identifying two or more similar instances (nearest neighbors) from the minority class and fabricating new instances along the line segments connecting these points in the feature space. This process aids in re-balancing the class distribution, thereby enhancing classifier performance, particularly in cases of severe class imbalance [18]. Along with that, data augmentation techniques are also used which expands the dataset diversity by applying transformations like rotations, flips, or noise, particularly in image-based tasks, which are later integrated into training dataset that help to improve the model generalization and resilience [19] [20]. Figure 2 illustrate these methods' application to enhance model robustness against class imbalance.

TABLE I: Comparison of different methods on various metrics

Approach	Method	ACC	AUC	Recall	F1-Score	Precision
[7]	VGG16 + VGG19 + DenseNet201 + ResNet101	90.36%	0.946	89.59 %	89.35%	89.51%
[10]	VGG16 + VGG19 + ResNet50 + InceptionV3 + DenseNet201 + NASNet-Large	93.9%	-	94%	94%	94%
[8]	EfficientNet + Inception-ResNetV2 + NasNet-Large + ResNet50	93.29%	0.9195	86.58 %	-	86.58%
[9]	MobileNetV2	91%	-	95%	-	-
[12]	GTOADL-ODHI	99.13%	-	98.28 %	98.36%	98.45%
[13]	GhostNet + ResNeXt + STOA	97% (Binary), 96.83% (Secondary)	-	89.68 %	-	-

C. Dataset Composition and Overview

The dataset curated by Arunachalam [21] contain histology images stained with hematoxylin and eosin (H&E), which is a standard method in cancer studies. H&E staining highlights tissues with pink hues and nuclei with blue which aids in identifying cellular structures. Specifically for Osteosarcoma, while both normal and tumor cells appear blue, their shapes vary:

normal cells round and regular, whereas tumor cells are irregular and diverse, which poses difficulties for conventional analysis methods. Anyhow, the dataset we used in our research was originated from achieved samples of 50 patients treated at the Children's Medical Center, Dallas, between 1995 and 2015. Clinical investigators from the University of Texas Southwest- ern Medical Center selected four representative patients based on tumor variety after surgical excision. Pathologists annotated the images collaboratively, with each image receiving a single annotation. The dataset contains 1144 numbers of images with a resolution of 10x at 1024 x 1024 pixels of magnifications. The categories for that images were as follow:

TABLE II: Classification of Tumor and Non-Tumor Tissues

Categories	Description
Non-Viable Tumor	Tumors rendered inactive, often due to successful treatment or cell death.
Viable Tumor	Actively growing tumors which can be lethal.
Viable Non-Tumor	Living, non-cancerous tissue.
Non-Tumor	Healthy tissue without any tumor characteristics.

Here in more elaborate manner, a" Non-viable Tumor" denotes as a tumor that is inactive or no longer capable of surviving. In medical terms," non-viable" suggests that the tissue or organism cannot function or grow. In the context of tumors, a non-viable tumor could indicate effective treatment, inactivation, or necrosis (cell death) induced by therapies like chemotherapy or radiation. Essentially, it signifies a tumor that is no longer able to grow or cause harm. "Non-tumor" refers to tissue or cells that do not display characteristics of tumor growth or malignancy. In medical terminology, this term describes healthy tissue or cells that lack signs of abnormal growth associated with tumors. Within datasets or medical imaging, "non-tumor" categorizes regions devoid of tumors or cancerous cells. Similarly, a "viable tumor" refers to a tumor that is alive and actively growing. In medical terminology, "viable" indicates the ability to survive, function, or proliferate. Consequently, a viable tumor is one that remains active and has the potential to expand or cause harm within the body.

D. Description of Pretrained Models

MobileNet-V2 is an efficiency-oriented neural network proposed to reduce computationally expensive computations without sacrificing too much performance for devices such as smart phones and any other kind of embedded system. It employs reverse residual blocks together with the Linear Link feature; this reduces a lot of lost data during performance, and noise-distributed convolution, where conventional convolutions in the network architecture are broken up into point to point and noised, results in the major improvement of fast processing. These perks give the MobileNet-V2 model an edge for real-time applications in low-resource environments [22]. Next comes our VGG-19, which is a convolutional neural network architecture that was developed by researchers at the University of Oxford. It is famous for its simplicity and robustness. It has 19 layers, with 16 convolutional layers and 3 fully connected layers. VGG-19 uses 3×3 filters to effectively capture image features. Despite the fact that it requires a lot of computational resources for training and deployment, it performs very well on standard datasets such as ImageNet, making it highly suitable for image classification tasks. Its strength to interpret objects and scenes accurately ensure a lasting influence on subsequent deep learning architectures [23].

EfficientNet-B0 is a cornerstone of the proposed EfficientNet-CNN, because it is known for striking an optimal balance between computational efficiency and accuracy, which we actually yearn for due to the issue of class imbalance. Anyway, it has innovative design principles that improve its performance in detection of objects, image classification, among others. The EfficientNet-B0 model's architecture is scalable and resource-efficient for complex computer vision applications. Its elegant and lightweight nature makes it an effective tool in a wide array of machine learning applications, hence providing a good

alternative for the environments that both speed and accuracy are needed.

More alike VGG-19, the VGG-16 is another influential convolutional neural network that is known for its simplicity

and robust performance. It has 16 layers, which include 13 convolutional and 3 fully connected layers. Its deep generalization capability makes it suitable for diverse computer vision applications, including object recognition and image segmentation. With complementary CNN feedback algorithms, VGG-16 significantly adds to diversity and performance, proving to be versatile in dealing with complex machine learning challenges.

DenseNet121 is crucial in improving the efficiency and robustness of CNN architectures mainly through its densely connected design. The architecture will effectively cause feature propagation and minimize redundant computation thus making DenseNet121 highly efficient for applications such as object detection and image classification. The proposed EfficientNet- CNN benefits from a comprehensive set of features achieved by integrating DenseNet121 with other CNN frameworks. Its innovative design and proven effectiveness assure its relevance across a broad spectrum of machine learning applications [24]. In the end we got, ResNet-50, which is a central deep learning model that resolves the vanishing gradient problem in deep networks with residual learning architecture. ResNet-50 is known to handle complex data efficiently, and it reduces degradation issues in data and enhances model accuracy. Within the proposed EfficientNet-CNN, ResNet-50 brings its unique capabilities to the table which let the model to perform a wide variety of computer vision tasks, including object recognition and image classification [25].

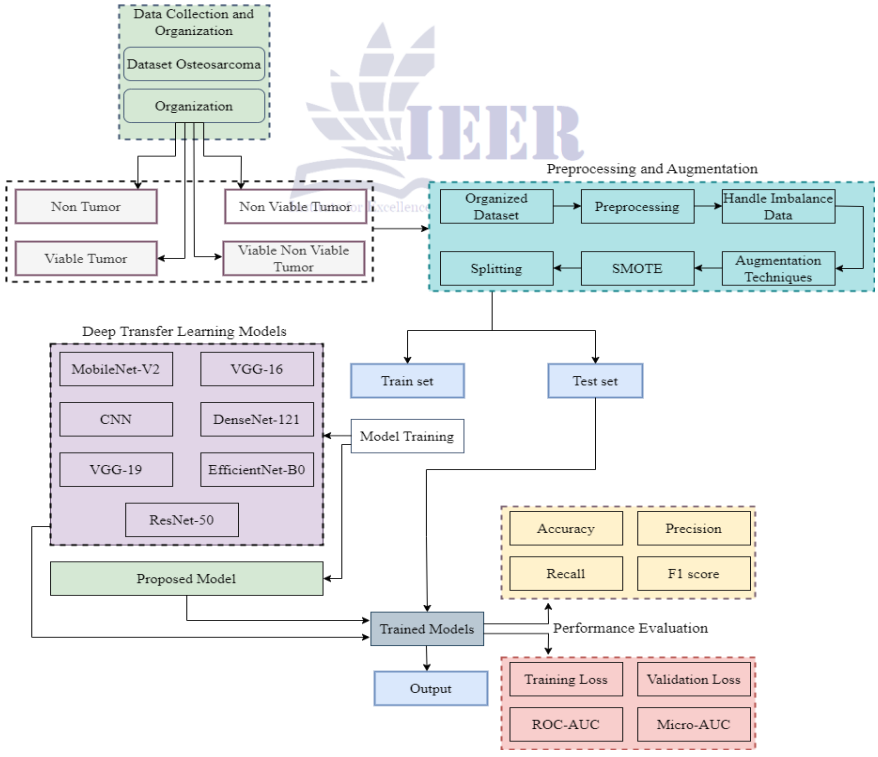

Fig. 1: Work flows of proposed Efficient Net- CNN classification for Osteosis- coma bone tumors

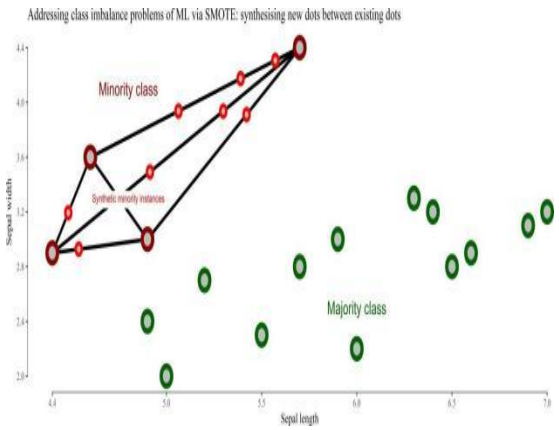

Fig. 2: Showing the oversampling to the unstable dataset in the minor class [16].

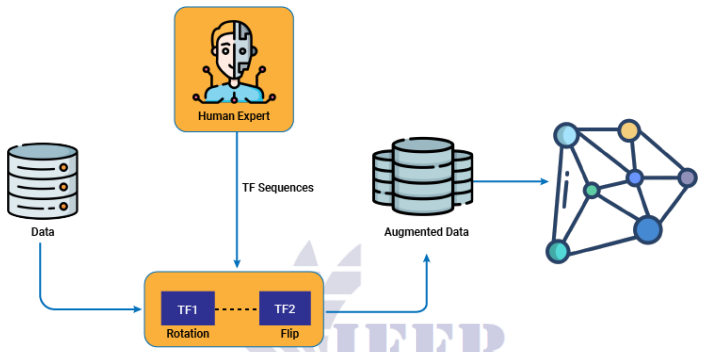

Fig. 3: Data augmentations guided by human experts and applied using heuristic transformation functions.

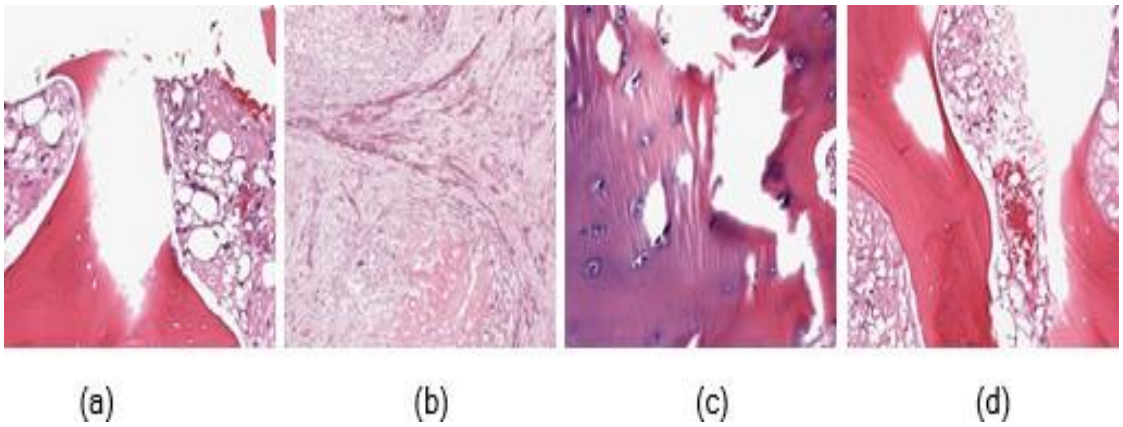

Fig. 4: Types of Osteosarcoma Images- (a) Viable (b) Non- viable Tumor, (c) non-tumor, (d) Viable Non-viable Tumor

E. Architecture of Proposed EfficientNet-CNN

The proposed EfficientNet-CNN architecture integrates convolutional layer with the pre-trained EfficientNetB0 models using the Keras which is an open-source python library that provides interface for Artificial Neural Networks. It was being used to leverage the transfer learning from ImageNet. This architecture adapts to specific tasks by removing

the top layer of the pre-trained model and appending specialized layers. The input dimensions are set to (150,150,3), where the layers were constructed using the sequential API in Keras. The model incorporates a Conv2D layer with 32 filters and ReLU activation which is followed by a MaxPooling2D layer for spatial reduction and a GlobalAveragePooling2D layer to create a distinct feature map for each filter. In order to prevent the overfitting, dropout layer was being used for regularization, while the final output layer employs softmax activation for multi-class classification. The ensemble model synergizes EfficientNetB0's feature extraction capabilities with tailored layers for enhanced classification performance. Figure 5 showcase the complete development of an ensemble model in which the Convolutional blocks serves as the foundation of the architecture, compromising 2D Convolution, 2D Average Pooling and ReLU activation. The kernel weights are initialized using the GlorotUniform-V2 kernel initializer. ReLU activation actually eradicate the vanishing gradient problem and expedites network training. Dropout layers further enhance training by deactivating nodes during each iteration based on a probability distribution, which helps in preventing overfitting.

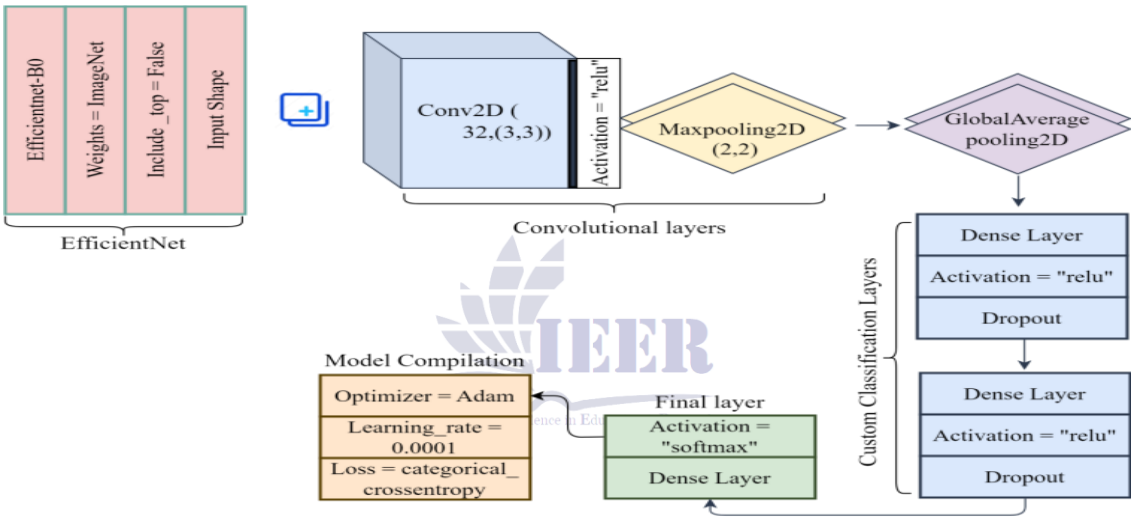

Fig. 5: Detailed architecture of proposed Efficient Net-CNN

Dense blocks which combine activation functions and a series of layers, also contribute to the architecture's performance. The dense layers, which is also known as the fully connected layers, are imperative for image classification because they map input vectors to output classes through back-propagation across interconnected layers. A softmax activation function in the output layer facilitates multi-class classification by assigning probabilities to each class. The flatten layer bridges convolutional and dense layers by converting feather maps into a 1D vector format making them compatible with dense layer. Convolutional layers extract local features such as edges and curves which enable the model to learn essential patterns effectively. Likewise, the model evaluation was done using the Osteosarcoma bone tumor dataset on Kaggle with a GPU P100. Metrics like accuracy, precision, recall, F1-score, and ROC- AUC were employed to measure the performance. Accuracy was calculated as the ratio of true predictions total predictions, which precision quantified the correctly predicted positives. Recall measured the ration of true positives to actual positives and the F1-score represented the harmonic mean of precision and recall. The ROC curve illustrated the trade-off between sensitivity and specificity across thresholds. Additionally, a categorical cross-entropy loss function quantified the disparity between predicted and actual values which ensures the robust model training and evaluation.

F. KFOLD Cross-Validation

For our result, we have explicitly implemented K-Fold, which is a robust technique used to assess a model's performance by splitting the dataset into multiple subsets known as folds. The data is divided into K equal parts, where the model is trained on (K-1) folds and tested on the remaining fold. This process is repeated K times, with each fold serving as the validation set once. The final performance metric is obtained by averaging the results across all iterations.

Here in this research, we implement a deep learning-based image classification model using EfficientNetB0 as the base model and then combine it with K-Fold Cross-Validation for robust evaluation. The dataset was composed of synthetic im- ages and their corresponding labels, which are first reshaped to match the expected dimensions of the model. The labels, which were originally in string format, are encoded into numeric values and then transformed into a categorical format for multi- class classification. We utilized 5-fold cross-validation, which means the dataset is split into five subsets, and the model is trained and evaluated five times, each time using a different subset as the validation set while training on the remaining data. The EfficientNetB0 model, pre-trained on ImageNet, serves as the backbone for feature extraction, followed by custom layers including convolution, pooling, dropout, and fully connected layers to refine the classification process

G. Experimental Setup

This work uses a research method using systematic planning and carefully selected collection tools to ensure the reliability and accuracy of the surgical model that diagnoses bone diseases. In terms of hardware, we use a Windows system running at 2.40 GHz. The development process is performed on Google Collab Notebook platform with 12 GB RAM Google Collab or Google Collaboratory. It is a cloud-based platform in which users can write and execute python programs in single browser. Scientist is a data scientist and machine learning experts highly respect the ability to code in powerful tools without the need for coding or programming Users may access 12GB of random-access memory (RAM) on the platform for temporary storage of data and code instructions. Users can effectively handle extensive datasets, do intricate calculations, and enhance application performance.

III. RESULTS

A. Results of Pretrained Models

Table III shows that many pretrained models effectively identified osteosarcoma bone tumors from a multiclass dataset. MobileNet-V2 stands out from other models because of its exceptional recall, accuracy, and F1 score, particularly for class 0 (non-tumorous data). It shows an inadequate ability to distinguish viable tumor groupings from non-viable ones. VGG-16 demonstrates outstanding accuracy and F1 score for class 0, as well as increased precision for active tumors. It achieves a high degree of recall for non-viable tumors and a good level of accuracy for both viable and non-viable tumors. DenseNet-121 shows different degrees of precision in identifying cancers, with notably high accuracy in the viable and non-tumor categories. Yet, its effectiveness notably declines when dealing with non-viable tumors and the combined group of viable and non-viable tumors. VGG-19 shows high accuracy when used on non-cancerous tumors but poorer precision when used on non-viable tumors compared to other models. ResNet- 50 outperforms other models and emerges as the top performer in all categories. The model demonstrates its strength and efficiency in detecting cancers by achieving outstanding accuracy, recall, and F1 scores for all tumor types. EfficientNet-B0 shows outstanding performance in categorizing non- tumorous occurrences with a high F1 score and amazing accuracy in classifying both viable and non-viable tumors. The findings show the varying performance of each pretrained model in identifying osteosarcoma bone cancers. Some models excel in some tumor categories but may perform poorly in others. MobileNet-V2 excels in identifying non-tumor cases but has challenges in classifying cancers. VGG-16 and ResNet-50 models are more dependable options for tumor detection tasks because of their fair distribution of performance across all classes. It is crucial to choose the most appropriate pretrained model based on the unique needs and features of the dataset. They highlight the potential of sophisticated convolutional neural

network designs such as ResNet-50 and EfficientNet-B0 to achieve better results in medical image processing. The MobileNet-V2 model achieved better precision, recall and f1 score for class 0 (non-tumor), but poor results for viable non-viable tumor. VGG-16 achieved 96% precision, 93% f1 score for non-tumor, 88% precision for viable tumor,86% recall for non-viable tumor, and 92% precision for viable non- viable tumor. DenseNet-121 achieved 92%, 86%, 62%, and 78% precision for non-tumor, viable tumor, non-viable tumor, and viable non-viable tumor, respectively. VGG- 19 achieved the highest results for non-tumors and the lowest for non- viable tumors. ResNet-50 achieved better results for all classes as compared to other models. EfficientNet-B0 achieved 99% precision for non-tumor and a 95% f1 score for viable non- tumor.

TABLE III: Performance Metrics of Different Models

Model	Class	Precision	Recall	F1 Score
MobileNet-V2	0	51	67	58
	1	50	44	46
	2	34	60	43
	3	99	03	01
VGG-16	0	96	91	93
	1	88	84	86
	2	78	86	82
	3	92	90	91
DenseNet-121	0	92	87	90
	1	86	76	80
	2	62	69	65
	3	78	82	80
VGG-19	0	96	95	95
	1	87	83	85
	2	80	85	83
	3	89	89	89
ResNet-50	0	96	96	96
	1	86	92	89
	2	91	75	82
	3	87	95	91
EfficientNet-B0	0	99	91	95
	1	89	96	92
	2	92	86	89
	3	91	95	95

B. Results of Proposed Model

Table IV includes findings from two popular optimizers, Adam and RMSProp, along with data demonstrating the efficacy of the proposed model. The model performs better on a variety of evaluation metrics, including accuracy, F1 score, and recall. The proposed model, using the Adam optimizer, outperforms RMSProp in all classes. Its consistent attainment of higher accuracy, recall, and F1 scores in several courses is evidence of this. Additionally, RMSProp performs well, particularly in the non-tumor and possible non-tumor classes, although the proposed model clearly beats the others. In every class, it regularly beats other models, even ones with different optimizers. The proposed model's design or training approach may improve the Adam optimizer, resulting in outstanding performance for different classes in the dataset.

TABLE IV: Results of the Proposed Model using Adam and RMSProp

Model	Class	Precision	F1 Score
Proposed using Adam optimizer	0	98	97
	1	92	94
	2	97	94
	3	96	98
Proposed using RMSprop optimizer	0	98	97
	1	94	94
	2	93	94
	3	97	97

1) **Accuracy and Loss:** Deep Learning (DL) models acquire knowledge from data in the training phase, which involves input and labeled data. Over time, the model's prediction accuracy increases as an algorithm identifies the relationship between inputs and outcomes. Validation occurs after training, where a separate dataset is used to assess the model's performance and generalize to new data. Accuracy is one of commonly used metric for evaluating the execution of model, especially in classification work. It is the ratio of correct predictions to the total number of fore- casts generated by the model. Loss quantifies the inaccuracies in the model's predictions, enabling the optimization method to alter the model parameters to minimize errors.

TABLE V: Training and Validation Accuracy of Different Models

Models	Training Accuracy	Validation Accuracy
MobileNet-V2	84.70	42.88
VGG-16	97.75	87.97
DenseNet-121	78.01	78.50
VGG-19	91.44	87.97
ResNet-50	92.48	89.61
EfficientNet-B0	97.77	92.61
Proposed	99.14	95.75

ROC-AUC Scores: The score used (ROC) or Area under the Curve (AUC) is an important statistic in machine learning (ML) or DL to measure the efficiency of a classification algorithm. ROC is a graphical representation of the sample rating system, showing two categories TPR measures the proportion of true parameters that are correctly classified by the model, while FPR is the proportion of true parameters that are incorrectly predicted. The ROC curve shows these bars on the Y and X axis AUC to measure discrimination. A continuous classification has an AUC of 1, with 0.5 indicating no discrimination. A ratio less than 0.5 indicate that performance below the bias usually indicates inadequate sampling. The ROC-AUC is unbiased, measuring the ability of the model to distinguish groups without classification, which increases its reliability. Figure 6 (a) displays the ROC and AUC values achieved with the Adam.

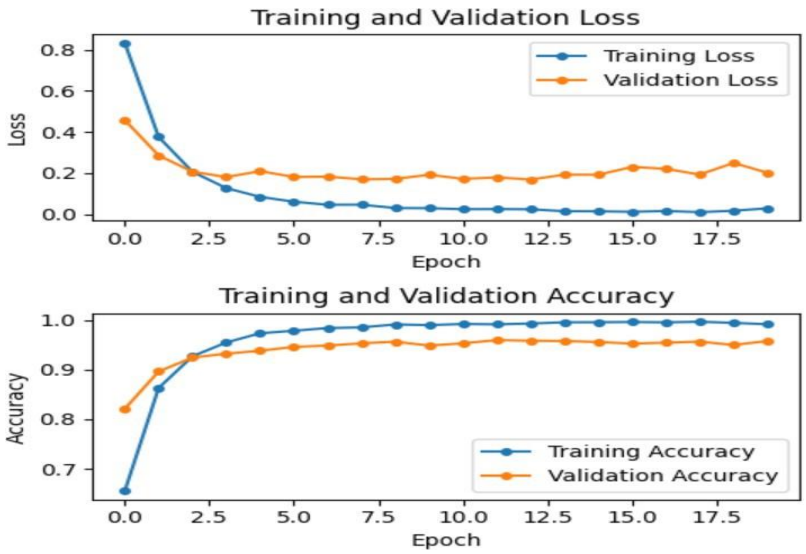

Fig. 6: Accuracy and loss values using Adam optimizer

2) **Confusion Matrix Result:** Confusion matrices are an important tool in the evaluation of classification models, especially in ML and statistics. It displays actual and predicted category tables for a given data set. An array is divided into rows and columns. Since the cells in the matrix contain correctly or incorrectly classified cases, the performance of the model can be better evaluated by classifying the strengths and weaknesses of the model. The confusion matrix provides a clear dissection of true-positive, true-negative, false-positive, and false-negative predictions, allowing a comprehensive assessment of model accuracy and robustness. Figure 9 (a) shows the values of the confusion matrix generated by ADAM optimization and Figure 9 (b) shows the values obtained by RMSPROP. In Figure 9 (a), the values of the confusion matrix produced with the Adam optimizer are shown, while in Figure 9 (b), the values obtained using RMSProp are shown.

Result of KFOLD: The model performed high and consistent validation accuracy in all the five folds, reflecting strong classification performance for images. Accuracy readings for every fold were a tiny bit distinct with respect to different training and validation data splits but were more than 95% in every fold, reflecting consistency in performance for the model. The overall accuracy of the model across all five folds was 96.45%, reflecting just 1.23% variation in the average. This low standard deviation emphasizes the reliability and generalization ability of the model. These results demonstrate that EfficientNetB0, along with other custom layers and K-Fold Cross-Validation, was able to effectively extract and process valid features, minimizing overfitting and maximizing performance.

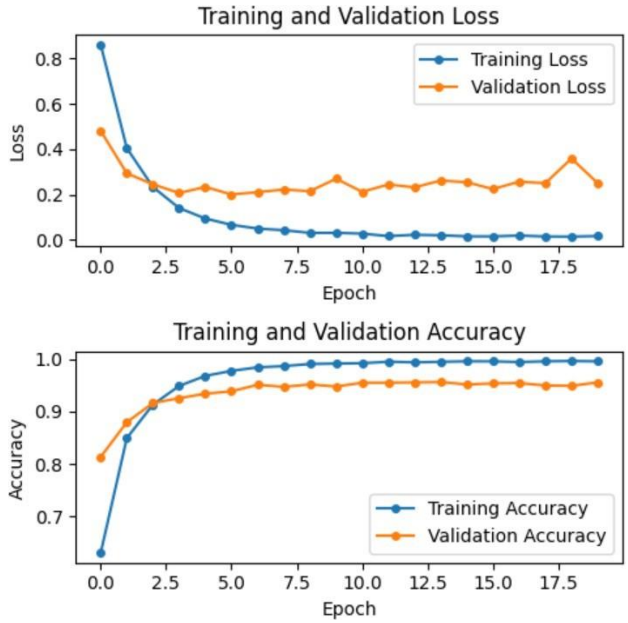

Fig. 7: Accuracy and loss values using RMSProp optimizer

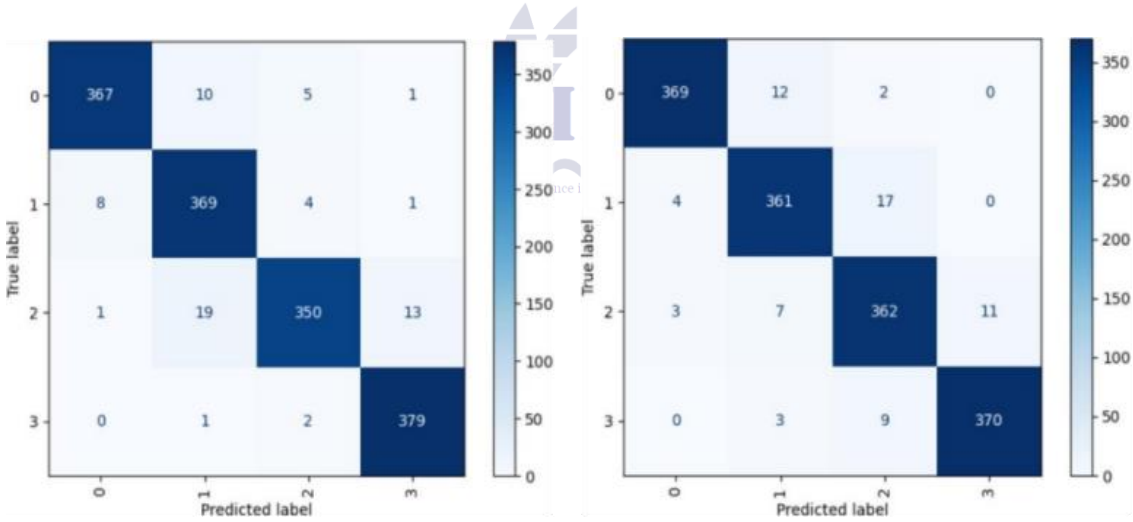

Fig. 8: Confusion matrix acquired with Adam optimizer (left) and RMSProp optimizer(right)

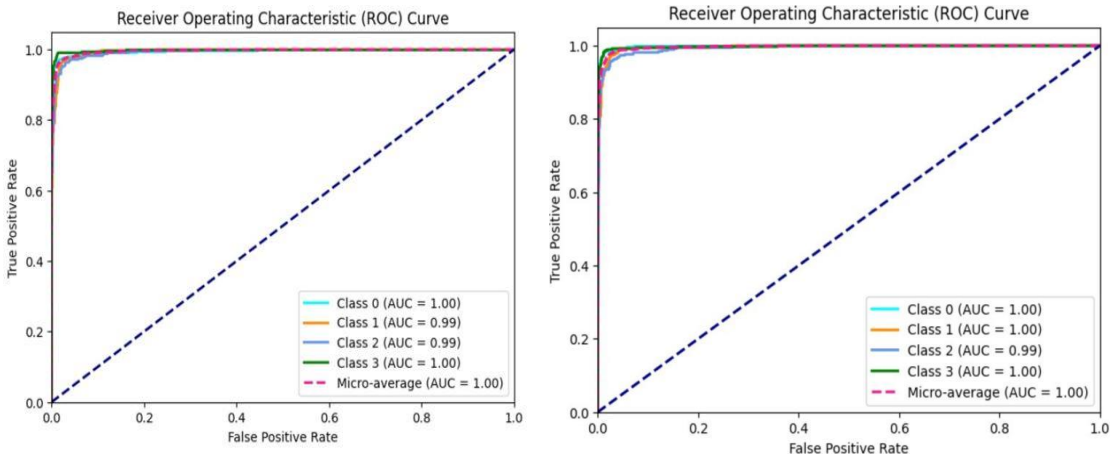

Fig. 9: ROC values obtained using the (a) Adam Optimizer(left) and (b) RMSProp Optimizer

IV. CONCLUSIONS

Ensemble CNN models have the potential to revolutionize deep image analysis in medical imaging, image processing, and assessment, by capitalizing on emergent workflows and technological advances. However, in medical datasets, class imbalances are significant in most cases and hinder the efficacy of DL models. Advanced data re-sampling techniques must therefore be applied to handle class imbalance in medical datasets to overcome the problem and to improve the detection of positive cases. The proposed research develops a carefully designed ensemble CNN model with minimal parameters for early diagnosis and image classification of Osteosarcoma bone tumors. The architecture involves a number of meticulously crafted layers within each convolutional block for early-stage detection and categorization of the tumor. Augment- mentation and SMOTE techniques are applied to the imbalanced dataset in order to produce synthetic data instances for each class. Moreover, the operational principles of Ensemble CNN layers are explained for better understanding. Our proposed ensemble CNN model obtains outstanding performance metrics, includ- ing accuracy, precision, recall, AUC, F1 score, loss, and ROC. Looking forward, we will include advanced architectures and state-of-the-art technologies to further improve model accuracy and performance.

TABLE VI: Validation Accuracy for Each Fold

Fold	Validation Accuracy (%)
1	95.8
2	97.2
3	96.1
4	95.3
5	97.6

REFERENCES

J. C. W. et al., "Osteosarcoma: a multidisciplinary approach to diagnosis and treatment," *American Family Physician*, vol. 65, no. 6, pp. 1123– 1133, 2002.

G. Ottaviani and N. Jaffe, *The epidemiology of osteosarcoma*, 2010, pp. 3–13.

D. Anisuzzaman, H. Barzekar, L. Tong, J. Luo, and Z. Yu, "A deep learning study on osteosarcoma detection from histological images," *Biomedical Signal Processing and Control*, vol. 69, p. 102931, 2021.

T. O. et al., "Osteosarcoma of the spine: experience of the cooperative osteosarcoma study group," *Cancer*, vol. 94, no. 4, pp. 1069–1077, 2002.

R. M. et al., "Osteosarcoma segmentation in mri using dynamic harmony search based clustering," in *2010 International Conference of Soft Computing and Pattern Recognition*, 2010, pp. 423–429.

M. Haider, M. S. A. Hashmi, A. Raza, M. Ibrahim, N. L. Fitriyani, M. Syafrudin, and S. W. Lee, "Novel ensemble learning algorithm for early detection of lower back pain using spinal anomalies," *Mathematics (MDPI)*, vol. 12, no. 13, p. 1955, 2024.

S. Gawade, A. Bhansali, K. Patil, and D. Shaikh, "Application of the convolutional neural networks and supervised deep-learning methods for osteosarcoma bone cancer detection," *Healthcare Analytics*, vol. 3,p. 100153, 2023.

R. F. Meem and K. T. Hasan, "Osteosarcoma tumor detection using different transfer learning models," 2023.

I. A. Vezakis, G. I. Lambrou, and G. K. Matsopoulos, "Deep learning approaches to osteosarcoma diagnosis and classification: A comparative methodological approach," *Cancers*, vol. 15, no. 8, p. 2290, 2023.

M. T. A. et al., "A novel hybrid approach for classifying osteosarcoma using deep feature extraction and multilayer perceptron," *Diagnostics*, vol. 13, no. 12, p. 2106, 2023.

M. A. A. W. et al., "Adapted deep ensemble learning-based voting classifier for osteosarcoma cancer classification," *Diagnostics*, vol. 13, no. 19, p. 3155, 2023.

M. A. et al., "Group teaching optimization with deep learning-driven osteosarcoma detection using histopathological images," *IEEE Access*, vol. 12, pp. 34 089–34 098, 2024.

S. Stephe, B. Manjunatha, and V. Revathi, "Osteosarcoma cancer detection using ghost-faster rcnn model from histopathological images," *Iran Journal of Computer Science*, 2024. [Online]. Available: https://doi.org/10.1007/s42044-024-00217-5

M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114. [Online]. Available: https://arxiv.org/abs/1905.11946

Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Frontiers in Digital Health*, vol. 6, 2024. [Online]. Available: https://doi.org/10.3389/fdgth.2024.1430245

K. P. Mahesh, S. A. Afrouz, and A. S. Areeckal, "Detection of fraud- ulent credit card transactions: A comparative analysis of data sampling and classification techniques," *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012072, 2022, 1st International Conference on Artificial Intelligence, Computational Electronics and Communication System (AICECS 2021), 28–30 October 2021, Manipal, India.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

A. Ferna´ndez, S. Garc´ıa, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmen- tation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

Z. Feng and H. Lu, "Data augmentation techniques for deep learning- based medical image analyses," *Journal of Imaging Science and Technology*, vol. 67, no. 1, p. 010402, 2023.

H. B. A. et al., "Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models," *PloS One*, vol. 14, no. 4, p. e0210706, 2019.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.