DOI: 10.14421/ijid.2024.4890 IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 13, No. 2, December 2024, Pp. 517-529

Improving Osteosarcoma Detection through SMOTE-Driven Machine Learning Approaches

Muhammad Ainul Fikri
Department of Information Engineering
Jember State Polytechnic
Jember, Indonesia
m.ainulfikri@polije.ac.id

Ajie Kusuma Wardhana
Faculty of Computer & Science
Amikom University of Yogyakarta
Yogyakarta, Indonesia
ajiekusuma@amikom.ac.id

Yudha Riwanto*
Faculty of Computer & Science
Amikom University of Yogyakarta
Yogyakarta, Indonesia
yudha.riwanto@amikom.ac.id

Inggrid Yanuar Risca Partiwi
Diploma Program in Information Technology
Malang State Polytechnic
Malang, Indonesia
inggrid.yanuar@polinema.ac.id

Fauzia Anis Sekar Ningrum
Faculty of Computer & Science
Amikom University of Yogyakarta
Yogyakarta, Indonesia
fauzianingrum@amikom.ac.id

Iqbal Kurniawan Asmar Putra
Department of Electronics Engineering
Padang State Polytechnic
Padang, Indonesia
iqbalkurniawan@pnp.ac.id

Article History
Received December 8th, 2024
Revised February 1st, 2025
Accepted February 3rd, 2025
Published February, 2025

***Abstract*— Osteosarcoma is an aggressive and highly malignant bone cancer primarily affecting adolescents and young adults, with males being more commonly affected. Although deep learning models such as YOLO (95.73% accuracy) and VGG19 (95.25% accuracy), have demonstrated effectiveness in osteosarcoma detection, their large model sizes and extensive computational requirements limit their feasibility in resource-constrained environments. This study proposes a lightweight AI approach that optimizes osteosarcoma detection while maintaining high diagnostic accuracy, leveraging machine learning models under 5MB, manually or semi-automatically extracted features, and SMOTE for data balancing. Experimental results show that Random Forest, SVM, and XGBoost achieve accuracies of 94.70%, 94.23%, and 94.39%, respectively, closely matching the performance of YOLO and VGG19 while maintaining computational efficiency. Furthermore, the inference time for SVM is under one second (0.97s), demonstrating the speed advantage of lightweight models. These findings highlight the potential of small-size (lightweight) machine learning models to deliver high diagnostic accuracy with minimal computational requirements, providing a scalable and practical solution for early osteosarcoma detection in resource-limited settings. By balancing simplicity, efficiency, and high performance, this study establishes a new benchmark for achieving state-of-the-art results with lightweight models and paving the way for improved healthcare accessibility in underserved regions.**

Keywords*— *Lightweight Machine Learning; medical diagnostic; Osteosarcoma detection; Random Forest; SMOTE

1 INTRODUCTION

Osteosarcoma is a highly aggressive and malignant form of bone cancer that primarily affects children and young adults, particularly those in their teenage years. It is considered the most common primary bone cancer in this age group, with more than 60% of cases diagnosed in individuals under the age of 25. The disease exhibits a bimodal age distribution, with a prominent peak during adolescence and a smaller peak in individuals over 60. This pattern suggests a strong correlation between osteosarcoma and periods of accelerated bone turnover. Epidemiologically, osteosarcoma exhibits a slight male predominance, with a male-to-female ratio of approximately 3:2. This disparity may be partly attributed to differences in growth spurts and bone remodeling rates between sexes. According to the World Health Organization (WHO), the overall incidence rate of osteosarcoma in the global population is approximately 4 to 5 cases per 1,000,000 individuals. However, this incidence nearly doubles to 8 to 11 cases per 1,000,000 individuals annually in the 15–19-year-old age group, emphasizing its predilection for adolescents.

In Indonesia, although osteosarcoma is relatively rare compared to other cancers, it remains a significant public health concern. Bone cancers, including osteosarcoma, account for approximately 1.6% of all cancers in the country. Notably, the incidence has shown a rising trend over the years, which could reflect improved diagnostic capabilities or a genuine increase in cases. A 13-year retrospective study at Cipto Mangunkusumo Hospital (RSCM) from 1995 to 2007 reported 219 cases of Osteosarcoma. This accounted for 70.59% of all bone malignancies treated at the institution during that period, with the highest prevalence observed in individuals in their second decade of life, reinforcing the link between osteosarcoma and adolescence [1], [2].

Osteosarcoma, a rare and aggressive form of bone cancer, is influenced by various risk factors that increase its likelihood of development. Age is a significant factor, with most cases occurring in individuals aged 10 to 30 years, particularly during adolescence when rapid bone growth occurs, and another smaller peak in those over 60, often associated with underlying bone conditions like Paget's disease. Studies have shown that children with osteosarcoma tend to be above average height, indicating a possible connection between rapid growth and the development of this disease. Gender differences are notable, with osteosarcoma being more common in males than females, potentially due to differences in bone growth and hormonal factors.

Osteosarcoma also exhibits racial disparities, with higher incidence rates observed in African American and Hispanic/Latino populations, although the factors contributing to these disparities are not yet fully understood. Exposure to high-dose radiation, particularly during childhood or adolescence, is a well-established risk factor, often linked to prior cancer treatments. Certain bone diseases, including Paget's disease, which alters normal bone remodeling processes, have been linked to an increased risk of osteosarcoma development.

Genetic predisposition is another critical aspect, as individuals with inherited conditions like Li-Fraumeni syndrome, Rothmund-Thomson syndrome, or hereditary retinoblastoma have a heightened risk of developing osteosarcoma. These syndromes are characterized by mutations in tumor suppressor genes, leading to impaired cell growth regulation and defective DNA repair mechanisms.

Beyond genetic factors, environmental exposures may also contribute to osteosarcoma risk. Studies have explored the potential role of heavy metals, pesticides, and industrial pollutants in disrupting normal bone cell function, though definitive conclusions are still lacking. Although no specific dietary causes have been identified, sufficient calcium and vitamin D intake is crucial for bone health, and deficiencies may indirectly influence bone disease susceptibility.

Treatment for osteosarcoma typically involves a combination of surgery, chemotherapy, and in some cases, radiation therapy. Advances in medical research have led to improved survival rates, particularly for patients whose cancer is detected early and has not metastasized. However, challenges remain, especially for individuals with tumors in difficult-to-operate locations or those who experience recurrence. Ongoing clinical trials and emerging therapies, such as targeted molecular treatments and immunotherapy, hold promise for further improving outcomes for osteosarcoma patients.

While osteosarcoma remains a complex disease with multiple influencing factors, increased awareness, early detection, and continued research are crucial in enhancing survival rates and improving the quality of life for affected individuals. Genetic predispositions also play a critical role, with inherited conditions like retinoblastoma, caused by mutations in the RB1 gene, and Li-Fraumeni syndrome, associated with mutations in the TP53 gene, significantly elevating the risk. Understanding these risk factors is essential for identifying high-risk populations and enabling earlier detection and intervention to improve outcomes [3], [4].

Symptoms of osteosarcoma are such as bone or joint pain that worsens over time, a lump in the arm or lower leg, uninjured fractures, back pain, or loss of bowel or bladder control [5], [6]. Diagnosis is made through physical examination, biopsy, imaging and laboratory tests. Treatment options depend on the stage and grade of the cancer, as well as the patient's overall condition, and may include surgery, chemotherapy and radiation therapy. Early detection and appropriate treatment can improve the patient's prognosis [7].

Osteosarcoma is the most common type of bone cancer in humans [8] and its aggressive nature and rapid progression make early detection critical for ensuring optimal treatment planning and improving patient outcomes. Delays in diagnosis can lead to cancer spreading, significantly reducing survival rates. Conventional diagnostic methods, relying heavily on expert analysis, are often hindered by delays and inaccuracies, especially in settings with limited resources. To address these challenges, artificial intelligence has emerged as a powerful tool to enhance the speed and accuracy of diagnostic processes, providing significant benefits in environments with limited access to specialized medical expertise. By leveraging AI-based approaches, such as automated image analysis and predictive modeling, subtle patterns in medical imaging—often overlooked by the human

eye—can be detected more accurately and quickly [9], [10], [11], [12], [13]. This capability accelerates diagnosis, enhances accuracy, and enables earlier intervention, ultimately leading to improved patient outcomes and prognoses.

A study by Anisuzzaman et al. demonstrated the efficacy of transfer learning models, such as VGG19 and Inception V3, in the classification of osteosarcoma histology images [14]. The evaluation showed that VGG19 achieved accuracy rates of up to 95.25% in binary and multi-class classification tasks, significantly outperforming Inception V3's accuracy of 84.20%. By leveraging pre-trained models on large public datasets, this approach not only enhances diagnostic accuracy but also significantly reduces the workload of pathologists. The use of transfer learning allows the models to generalize effectively from vast amounts of prior data, making them a valuable tool for improving diagnostic efficiency and consistency in clinical settings [14], [15].

Similar research by Gawade et al. focused on developing Convolutional Neural Network (CNN)-based deep learning models, including ResNet101, VGG19, and DenseNet201, tailored for bone radiography datasets. Among these, ResNet101 emerged as the most effective, achieving the highest accuracy of 90.36% in detecting bone tumors. These findings underscore the vast potential of deep learning models to revolutionize medical imaging diagnostics by dramatically enhancing precision and accuracy. However, the practical application of these models in real-world settings faces challenges, particularly in resource-constrained areas. The high computational power and large datasets required to train and deploy these models limit their feasibility in regions with limited infrastructure and technological resources, underscoring the need for more accessible and lightweight alternatives [16].

The utilization of cutting-edge segmentation methods, such as Multiple Supervised Fully Convolutional Networks (MSFCN), has yielded outstanding results in the automated detection and segmentation of osteosarcoma tumors in CT images. MSFCN leverages multilevel supervised output layers to enhance multi-scale learning, enabling the model to capture intricate details across varying scales. This approach has achieved a remarkable Dice similarity score of up to 87.8%, highlighting its accuracy in delineating tumor boundaries. However, despite its effectiveness, the technique faces challenges, including high computational demands and limited generalizability to diverse datasets. Although effective for CT images with distinct, complex textures, the current model's performance has limitations, highlighting the necessity for further refinement to enhance its versatility and applicability to diverse imaging modalities and datasets [17].

While these studies have shown promising results, they are constrained by significant limitations, including their dependence on sophisticated models, substantial computational resources, and large datasets, which restricts their feasibility in settings with limited resources. For instance, ResNet101 has a model size of 171 MB with 44.7 million parameters, VGG19 is 549 MB with 143.7 million parameters, and DenseNet201 is 80 MB with 20.2 million parameters [18]. The large model sizes and high parameter counts require substantial computational resources, posing significant deployment challenges in environments with constrained infrastructure.

In addition, the dataset available for osteosarcoma on the Cancer Imaging Archive (TCIA) is limited in size and imbalanced across the four available classes. This presents an additional challenge in the classification process of osteosarcoma bone cancer. The imbalance in the dataset tends to result in biased and inaccurate training and prediction outcomes [19].

To address these challenges, we aimed to enhance the approach to osteosarcoma detection. While previous studies have relied on large, resource-intensive deep learning models, this study introduces a lightweight AI approach that balances simplicity, computational efficiency, and diagnostic accuracy. In this context, "lightweight" refers to AI models that are small in size, consume minimal computational resources, and offer rapid inference times, making them highly practical for real-world applications, particularly in resource-limited environments. Unlike deep learning architectures such as VGG19 (549 MB), ResNet101 (171 MB), or DenseNet201 (80 MB), which require significant storage and computational power, our machine learning models remain under 5MB, allowing easy deployment on standard computing hardware, including edge devices.

Beyond storage efficiency, lightweight AI models also exhibit significantly lower CPU usage and faster processing speeds, making them well-suited for real-time applications where quick decision-making is crucial. Moreover, efficiency is boosted by leveraging manually or semi-automatically extracted features, which decreases dependence on the substantial computational resources usually needed for feature extraction using deep learning methods. To address data imbalance—a persistent challenge in medical imaging datasets—we incorporate the Synthetic Minority Over-sampling Technique (SMOTE), ensuring a more balanced representation of all classes. Despite their streamlined design, these lightweight models achieve diagnostic accuracy comparable to more complex deep learning architectures, demonstrating that high-performance osteosarcoma detection can be achieved without sacrificing computational efficiency or requiring high-end infrastructure.

2 METHOD

Various studies have employed artificial intelligence approaches to enhance the accuracy of osteosarcoma detection, primarily by leveraging pre-trained deep learning models. Some studies stand out for their promising results. For example, research such as [2]. Aziz et al. utilized pre-trained CNN models like ResNet and VGG. In these approaches, features from pre-trained models were further processed using multilayer perceptron (MLP) or Fast.ai algorithms, achieving high accuracy rates, with one study reporting up to 95.2% accuracy for multiclass classification. However, these approaches have a significant drawback: their reliance on large pre-trained models demands high computational power, making them unsuitable for low-specification devices.

To delve deeper into these challenges, this study leveraged data from the publicly accessible Cancer Imaging Archive (TCIA), focusing on the "Osteosarcoma Tumor Assessment" dataset. The dataset features a collection of annotated medical images, including CT and MRI scans, paired with pertinent clinical data, offering a robust resource for research and investigation. All ethical concerns were adequately addressed, as the dataset is fully anonymized and shared following stringent privacy protocols and regulatory guidelines. Despite its utility, the dataset presents challenges, including its limited size and an imbalance across its available classes, which complicates training and risks introducing bias in predictions.

This issue of dataset imbalance is widely acknowledged in the literature and has prompted various strategies to mitigate its impact. One common approach is data augmentation, as demonstrated in the study by Walid et al. [21]. By integrating pre-trained models like EfficientNet and NasNetMobile with a voting classifier, the researchers significantly improved classification performance, yielding an impressive Kappa score of 96.50%. Although this approach enhanced overall performance, it failed to completely mitigate the inherent bias toward the majority class, underscoring the ongoing difficulty of addressing class imbalance in osteosarcoma detection.

Apart from classification, some studies have focused more on osteosarcoma image segmentation. Examples of such research include the studies by Tang et al. [22] and Ouyang et al. [23], which investigated the accuracy of MRI image analysis. These approaches achieved Dice Similarity Coefficients (DSC) ranging from 0.921 to 0.949, demonstrating the advantages of attention mechanisms. Similar to classification methods, these techniques rely heavily on extensive datasets and intricate data processing, posing challenges for practical implementation. Researchers have also shifted focus toward developing lightweight models that strike a balance between computational efficiency and performance, offering a promising solution for resource-constrained applications. Studies such as [24], [25] emphasize enhancing image features like edge details and reducing noise. These approaches are computationally efficient alternatives to large pre-trained models but do not adequately address crucial challenges such as dataset imbalance or minority class classification.

Meanwhile, classical machine learning techniques have been explored as another potential pathway. An example is research by Liu et al. [26] which applied regression analysis to identify predictive biomarkers. While this method offers simplicity and computational efficiency, it similarly falls short of addressing the imbalance in datasets, a factor critical to improving model reliability.

The outlined methods indicate that combining classical machine learning algorithms with techniques like SMOTE offers a promising strategy to address existing challenges and limitations. SMOTE has the potential to tackle dataset imbalance more effectively without requiring excessive data augmentation, making it computationally more efficient. The subsequent section provides an in-depth examination of SMOTE, including its application and potential benefits in the context of osteosarcoma detection. For a clearer understanding of the framework of this study, it can be seen in the flowchart (Fig. 1) below.

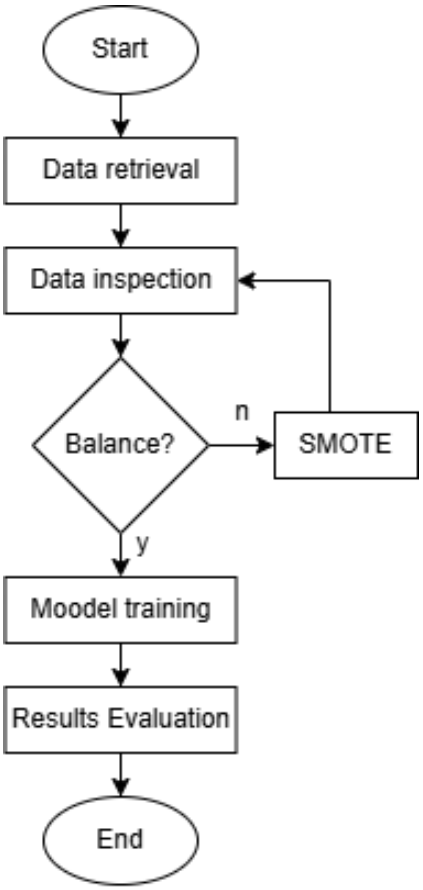

Figure 1. Research workflow

The first step after obtaining the dataset is to check the class balance within the dataset. This is important because medical data tends to have imbalanced classes in some available repositories. If the dataset is imbalanced, SMOTE is applied to achieve balance. The balanced dataset serves as the foundation for training a suite of machine-learning models, allowing for a thorough performance comparison. By comparing the performance of multiple models, this approach identifies the most accurate and reliable model for classifying osteosarcoma bone cancer.

2.1 Synthetic Minority Over-Sampling Technique

Given its effectiveness, the SMOTE has become a widely adopted method for addressing dataset imbalance. By generating synthetic samples for the minority class, SMOTE offers a robust solution that enhances model performance across various imbalanced datasets. Its versatility has made it a popular choice in numerous fields, enabling machine learning algorithms to achieve more reliable and balanced outcomes without relying on excessive data augmentation [27], [28], [29], [30]. In the following section, I will discuss SMOTE in more detail and explore its application to osteosarcoma detection.

The SMOTE is a method specifically designed to address the issue of imbalanced datasets, a common challenge in data analysis and machine learning. This technique was first introduced by Chawla et al. in their paper [31]. SMOTE generates synthetic data for the minority class through linear interpolation, thereby enhancing the representation of the minority class without duplicating existing data.

The approach works by calculating the distance between each minority class sample x_i and its k nearest neighbors using metrics such as Euclidean distance. Based on the desired level of oversampling, a certain number (m) of these neighbors are randomly selected. Synthetic samples are then generated using Eq. 1:

$$p_{ij} = x_i + rand(0,1)\left(x_{ij} - x_i\right) \quad (1)$$

Where rand $(0,1)$ is a uniformly distributed random number within the interval [0,1]. This process produces a new data point p_{ij}, which lies along the line segment connecting the original sample x_i and one of its nearest neighbors x_{ij}.

This ensures that the minority class distribution becomes more balanced, by reducing biases in machine learning models. In the formulation, the term $x_{ij} - x_i$ calculates the vector difference between a minority sample (x_i) and its nearest neighbor (x_{ij}), which dictates the direction of interpolation for generating synthetic samples. The multiplying factor rand $(0,1)$ controls the relative position of the synthetic sample along the line, ensuring that the new samples are randomly distributed along the segment. The result is an expansion of the decision space for the minority class, improving the model's ability to recognize patterns in the minority class that were previously underrepresented.

SMOTE has been widely applied across various fields that encounter data imbalance issues. For example, in breast cancer classification using datasets like the Breast-cancer-Wisconsin, SMOTE enhances the model's sensitivity to cancer cases [32]. In financial fraud detection, SMOTE improves accuracy by mitigating biases toward normal transactions, which typically far outnumber fraudulent transactions [33], [34], [35]. The technique has also been applied in medical image analysis, such as pixel classification to detect cancerous regions in mammograms [36].

2.2 Data Preparation

The data was directly retrieved from the TCIA repository, a trusted and widely utilized source for medical imaging research(https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52756935). The dataset, curated by clinical scientists at the University of Texas Southwestern Medical Center in Dallas, provides a valuable resource for research. Archival samples from 50 patients treated at Children's Medical Center, Dallas, between 1995 and 2015, were used to create the dataset. Pathologists selected four patients (out of the 50) based on the diversity of tumor specimens following surgical resection. The images were labeled as Non-Tumor, Viable Tumor, and Necrosis according to the dominant cancer type present in each image. Two medical experts performed the annotation, with the images split between the two pathologists. A single pathologist provided annotations for all images in the dataset. Based on information from the source, the dataset includes 1,144 images, each sized 1024 X 1024 at 10X resolution, with the following distribution: 536 (47%) non-tumor images, 263 (23%) necrotic tumor images, and 345 (30%) viable tumor tiles.

Upon downloading, you will find two main folders, TS1 and TS2. The TS1 folder has 11 subfolders (set1-set11), while the TS2 folder has 12 subfolders (set1-set12). After a series of data merging between TS1 and TS2, based on the information in the PathologistValidation.csv file for each subfolder set, we obtained the data count for each class, which has now been divided into four classes as shown in Fig. 2.

To prepare the data for model training, a comprehensive pre-processing pipeline is implemented to ensure the data is well-suited for the chosen machine learning algorithms. As shown in pseudocode Algorithm 1. The process begins with loading the dataset and separating the feature set (*X*) from the target labels (*y*). The target labels, which may be categorical, are encoded into numerical representations using a label encoder. This step formats the labels to ensure compatibility with machine learning models, enabling seamless processing and analysis.

Given the dataset contains a large number of features, dimensionality reduction is performed using Principal Component Analysis (PCA). PCA streamlines model training by reducing computational complexity while retaining essential data insights. The optimal number of PCA components (*k*) is determined by considering the explained variance or other domain-specific factors, striking a balance between dimensionality reduction and information retention.

Once the data is transformed, it is split into training and testing subsets with a 70:30 ratio. Stratified sampling is employed during the splitting process to ensure that the distribution of classes remains consistent across the training and testing sets. This is particularly important for datasets with class imbalance, as it prevents certain classes from being disproportionately represented in either the training or testing subsets.

The data is further standardized through feature scaling, utilizing a standard scaler to ensure consistent magnitude across all features. The standard scaler was selected as the pre-processing method to avoid a wide range of data values and ensure that all features have a mean of 0 and a standard deviation of 1. This helps many machine learning algorithms perform optimally by eliminating the influence of differing feature magnitudes. While alternatives like normalization, robust scaling, or log transformations could be considered, they are less suitable for this dataset. Normalization is not ideal with outliers, robust scaling is unnecessary given the absence of significant outliers, and log transformations would distort feature relationships. The scaler is fit on the training data and then applied to both the training and testing sets, ensuring consistent scaling across the dataset. Additionally, the same data preprocessing steps are applied to all models to ensure a fair comparison across their performance.

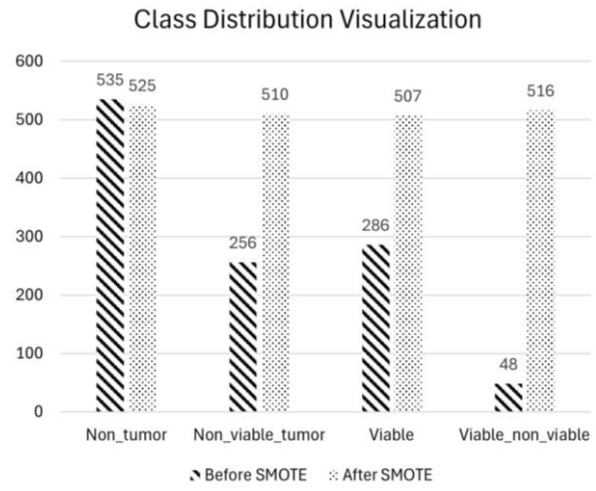

Figure 2. Class distribution

Algorithm 1: Function pre-process SMOTE

```
Input:
D = load_dataset(D)
X = D without column T
y = column T from D
Output: X_train, X_test, y_train, y_test
Function
      If y contains categorical values:
          Encoder = LabelEncoder()
          Y = encoder.fit_transform(y)

      If number of features in X is large:
          Pca = PCA(k)
          X = pca.fit_transform(X)

    X_train, X_test, y_train, y_test =
    split_data(X,y, test_size = 0.3,
    random_state=42, stratify=y)

        Scaler = StandardScaler()
        X_train =
scaler.fit_transform(X_train)
        X_test = scaler.transform(X_test)

        Return X_train, X_test, y_train,
y_test
```

By implementing this preprocessing pipeline, the dataset is optimized for training, minimizing noise and computational overhead while boosting model performance. This approach ensures data compatibility with machine learning models, while also guaranteeing reproducibility and robustness of the results.

3 RESULT AND DISCUSSION

Evaluating machine learning models for Osteosarcoma detection is crucial for identifying solutions that optimally balance accuracy, computational efficiency, and adaptability, particularly in resource-limited settings. Ten machine learning models, encompassing basic algorithms and advanced ensemble methods, were systematically evaluated to determine their efficacy on the dataset. To provide a thorough assessment, key performance metrics - including accuracy, precision, recall, and F1 score - were employed to highlight the strengths and weaknesses of each model. The results highlight significant differences in the ability of these models to generalize and handle the inherent complexities of the dataset, which are discussed in Table 1.

As shown in Table 1, ten models were evaluated based on metrics such as accuracy, precision, recall, and F1 score. Accuracy measures the overall correctness of a model's predictions, while precision indicates how well it identifies positive cases, and recall shows how many actual positives are correctly identified. High accuracy reflects general correctness but does not address the model's ability to distinguish positive and negative cases. High precision reduces false positives but may miss some positives, whereas high recall ensures most positives are identified, but may include false positives. A high F1 Score indicates a balanced performance in both precision and recall.

Overall, the results indicate that ensemble methods, particularly XGBoost, Random Forest (RF), and SVM, demonstrated superior generalization and accuracy compared to the other models. Both models excel in handling high-dimensional data and capturing complex features, which enabled them to deliver more reliable predictions on the testing data.

Table 1. Model Comparison Results

Models	**Results**			
	Accuracy	***Precision***	***Recall***	***F1 Score***
Logistic Regression	0.937695	0.938726	0.937695	0.937686
Decision Tree	0.811526	0.810877	0.811526	0.811118
Random Forest	**0.947040**	**0.948076**	**0.947040**	**0.947179**
Gradient Boosting	0.931464	0.931324	0.931464	0.931323
Support Vector Machine	**0.942368**	**0.944546**	**0.942368**	**0.941893**
K-Nearest Neighbors	0.813084	0.860779	0.813084	0.813475
Naive Bayes	0.665109	0.700222	0.665109	0.662887
AdaBoost	0.760125	0.776350	0.760125	0.760963
Bagging	0.884735	0.884476	0.884735	0.884540
XGBoost	**0.943925**	**0.943990**	**0.943925**	**0.943629**

Overall, the results indicate that ensemble methods, particularly XGBoost, RF and SVM, demonstrated superior generalization and accuracy compared to the other models. Both models excel in handling high-dimensional data and capturing complex features, which enabled them to deliver more reliable predictions on the testing data.

Although Gradient Boosting also performed well, it was slightly outperformed by XGBoost, SVM, and RF across most metrics. This suggests that while Gradient Boosting can effectively model patterns in the data, the computational efficiency and robustness of RF and XGBoost provided an additional edge in this context.

On the other hand, simpler models like Logistic Regression and Naive Bayes were less effective in capturing the complexity of the features. Although outperformed by more complex models, Logistic Regression delivered competitive accuracy. In contrast, Naive Bayes, which assumes feature independence, yielded the poorest results, ranking last in accuracy, precision, and recall among all evaluated models.

The Decision Tree model exhibited signs of overfitting, as evidenced by the disparity between its training and testing accuracies. This indicates that while the model learned well from the training data, it struggled to generalize to unseen data. Similarly, the K-Nearest Neighbors (KNN) model with 5 n-neighbors experienced a significant drop in performance when working with high-dimensional data, likely due to the limitations of the Euclidean distance metric in such spaces.

XGBoost, RF, and SVM demonstrated comparable accuracy to state-of-the-art deep learning models (ResNet101, MSFCN, and VGG19), yet demanded significantly fewer computational resources. This underscores their potential as practical solutions for deployment in resource-constrained environments without compromising diagnostic quality.

After evaluating the overall performance of the models through key metrics such as accuracy, precision, recall, and F1 score, it is important to delve deeper into their behavior across specific categories. This can be achieved by analyzing the confusion matrices, which provide detailed insights into how well each model classifies the four categories—non_tumor, non_viable_tumor, viable, and viable_non_viable. By examining these matrices, we can understand better the strengths and weaknesses of each model, particularly in handling misclassifications and distinguishing closely related categories.

The confusion matrices, presented in Figure 3, provide valuable insights into the performance of each model across the four categories: non_tumor (class 0), non_viable_tumor (class 1), viable (class 2), and viable_non_viable (class 3). Ensemble methods, particularly XGBoost and RF, showed outstanding performance in accurately classifying instances across all categories. Notably, they excelled in reducing misclassifications between closely related classes, such as non_viable_tumor and viable_non_viable. This indicates that these models are adept at handling subtle differences in the feature space, which are critical for distinguishing between these categories.

To delve further into the results, we can examine Figure 3a, which refers to the confusion matrix for the RF model. In this matrix, the prediction labels are as follows: label 0 represents the non-tumor category, label 1 represents the non-viable tumor category, label 2 represents the viable tumor category, and label 3 represents the viable/non-viable tumor category. The rows represent true labels, while the columns represent predicted labels, with the numbers in each cell indicating the number of samples that were correctly or incorrectly predicted. For prediction labels 0 and 1, each has 146 true positive samples; label 2 has 156 true positive samples; and label 3 has 157 true positive samples.

Focusing on the misclassifications, for the non-tumor category (label 0), 8 samples were misclassified as non-viable tumor, and 7 were misclassified as viable tumor. In the non-viable tumor category (label 1), 10 samples were misclassified as non-tumor, and 4 as viable tumor. These results suggest that true prediction rates are relatively high, indicating that the RF model performs well in classifying samples into the correct classes. Despite a low classification error rate, some mispredictions still occur, particularly with label 0 being misclassified as labels 1 or 2.

The viable tumor category (label 2) experienced 4 misclassifications as non-tumor and 1 as a non-viable tumor. The non-viable tumor category (label 1) also had 3 samples misclassified as viable tumors. Moving on, Figure 3b refers to the confusion matrix for the Support Vector Machine (SVM) model. The prediction labels are as follows: label 0 has 137 true positive samples, label 1 has 150 true positive samples, label 2 has 161 true positive samples, and label 3 has 157 true positive samples. For label 0, 12 testing samples were misclassified as non-viable tumor and 12 as viable tumor. For label 1, 6 testing samples were misclassified as non-tumor, and 4 as viable tumor.

Remarkably, for label 2, all testing samples were correctly predicted as viable tumor, showcasing the model's high accuracy for this class. For label 3, 3 testing samples were misclassified as viable tumor. These results demonstrate relatively high true prediction rates for the SVM model, with label 2 showing the best performance, as it was predicted correctly without error. Label 0 had the highest misclassification rate, with several samples mistakenly identified as non-viable tumor (label 1) or viable tumor (label 2).

Finally, Figure 3d refers to the confusion matrix for the XGBoost model. The prediction labels for this model are as follows: label 0 has 140 true positive samples, label 1 has 148 true positive samples, label 2 has 158 true positive samples, and label 3 has 156 true positive samples. For label 0, 14 testing samples were misclassified as non-viable tumor, 6 as viable tumor, and 1 as viable non-viable tumor. For label 1, 10 testing samples were misclassified as non-viable tumor, and 2 as viable tumor. For label 2, 1 testing sample was misclassified as non-tumor, and 2 as non-viable tumor. Testing samples from labels 3 and 4 were frequently misclassified as viable tumors (label 2).

Overall, true positives were high across all classes, indicating good performance of the XGBoost model. The largest misclassification error occurred with label 0, where

some testing samples were incorrectly predicted as label 1 and label 2.

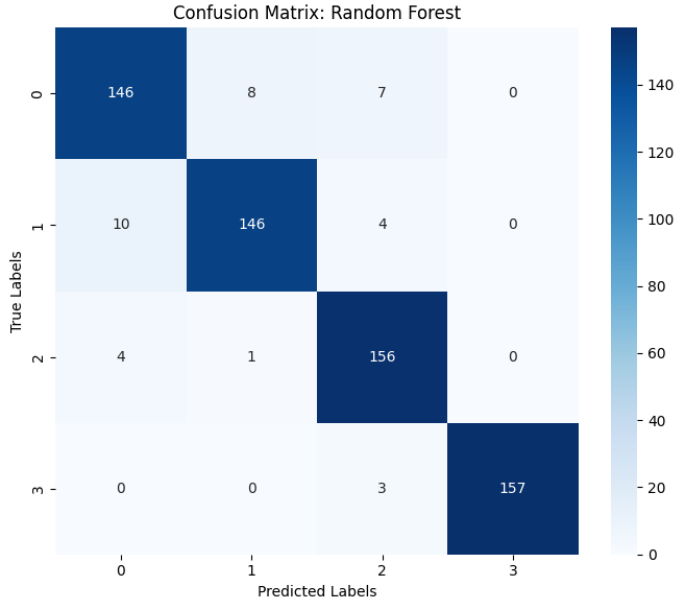

Figure 3a. Random Forest model

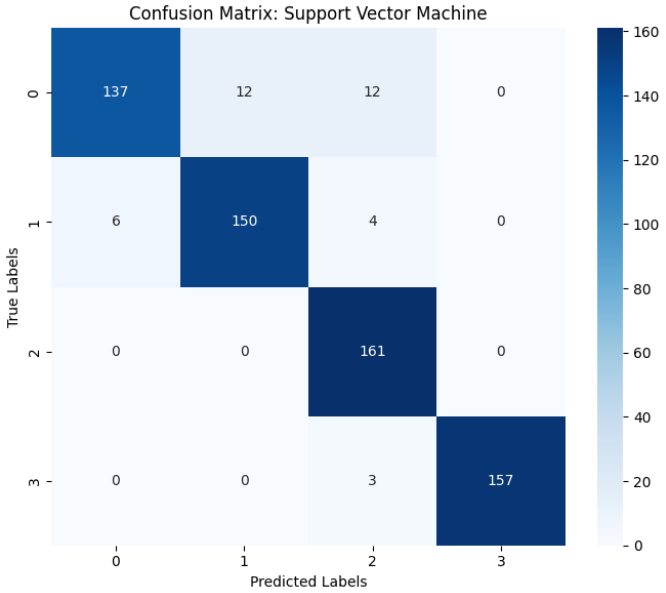

Figure 3b. Support Vector Machine model

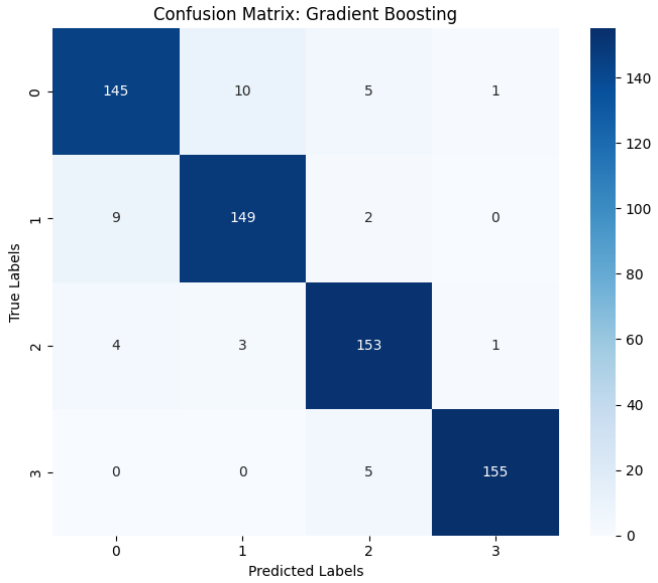

Figure 3c. Gradient Boosting model

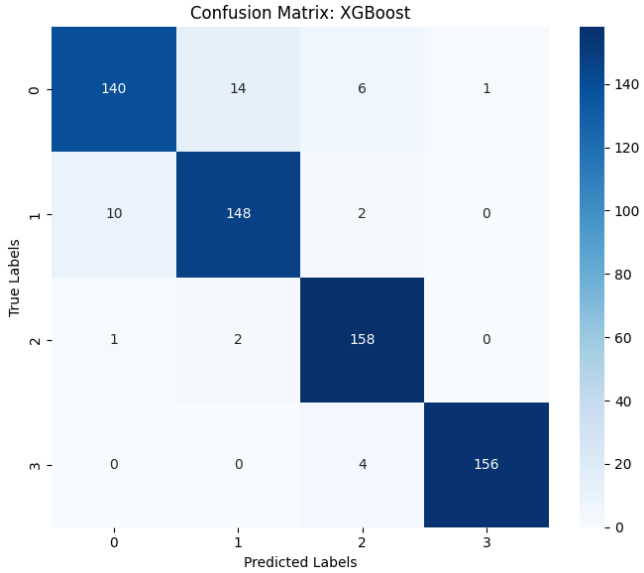

Figure 3d. XGBoost model

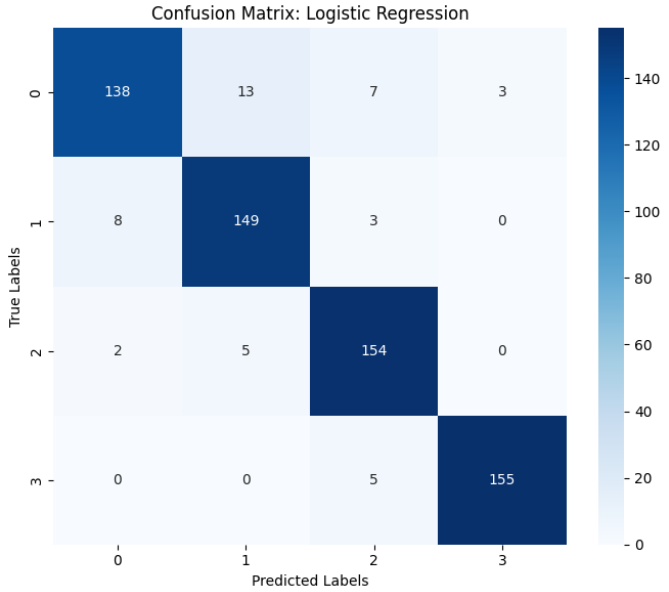

Figure 3e. Logistic Regression model

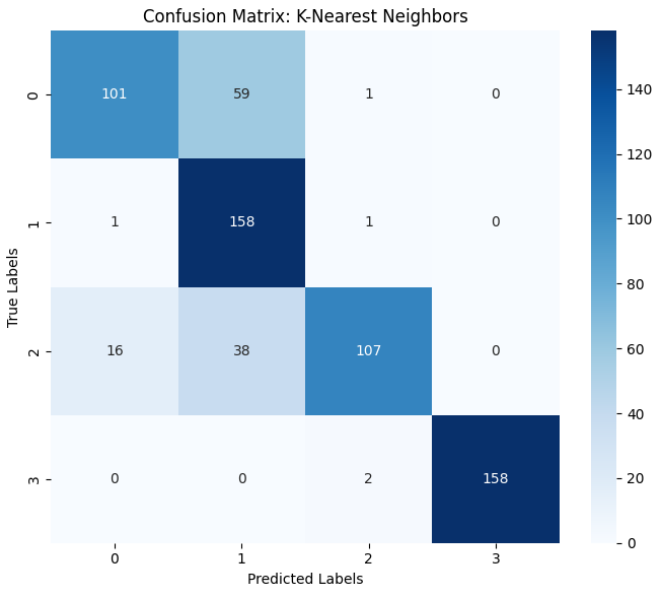

Figure 3f. K-Nearest Neighbors model

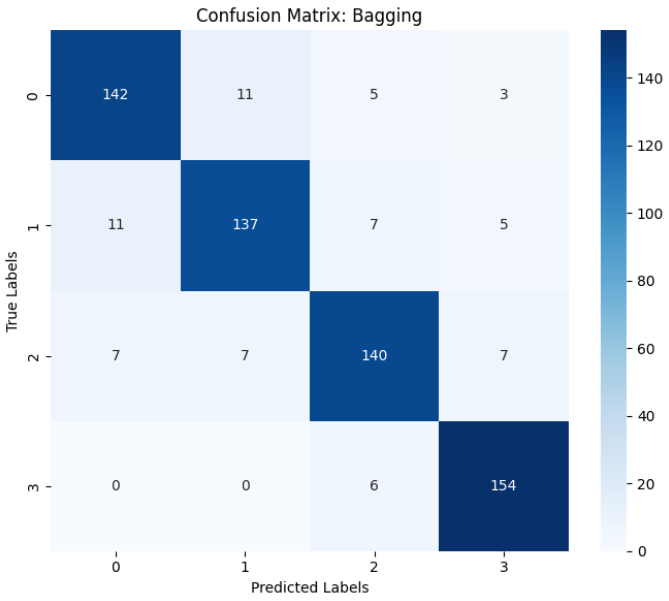

Figure 3g. Bagging model

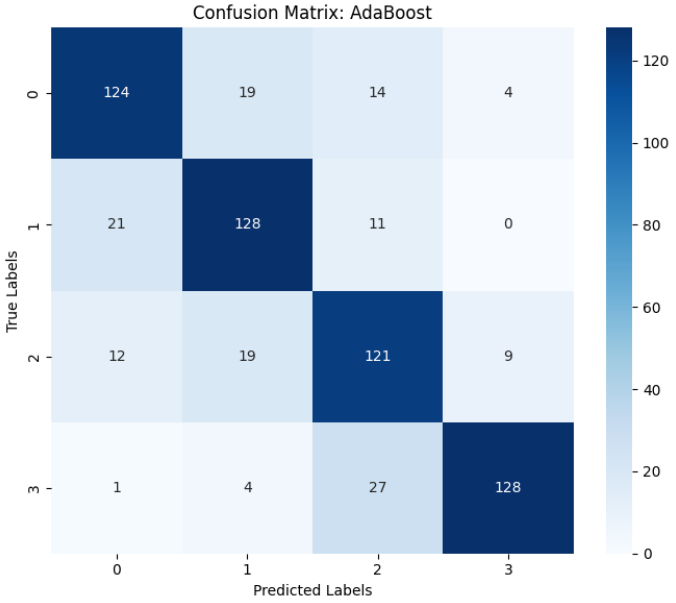

Figure 3h. AdaBoost model

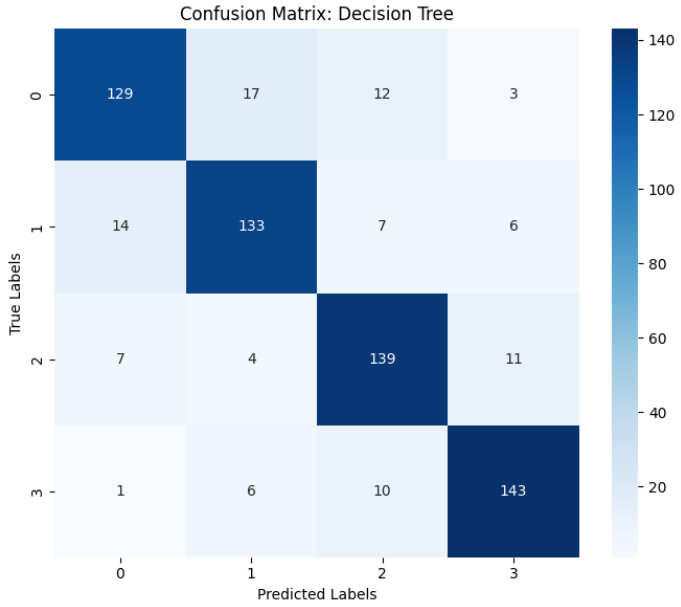

Figure 3i. Decision tree model

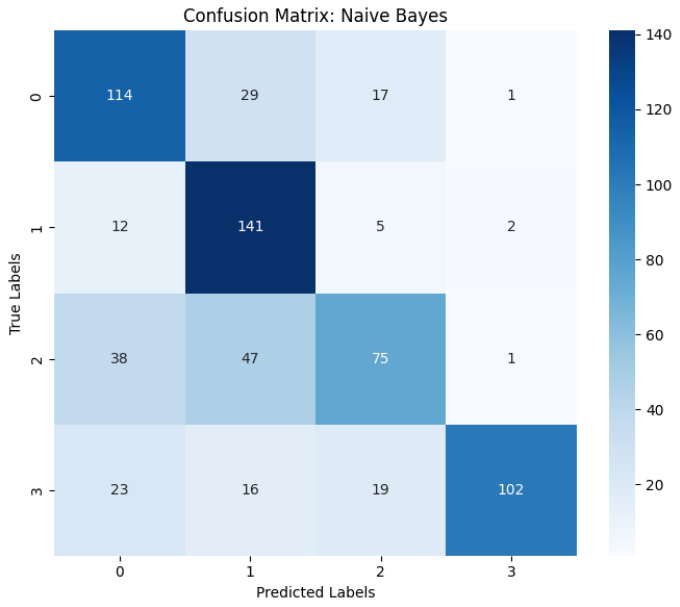

Figure 3j. Naïve Bayes model

Figure 3. Best model confusion matrix

The runtime analysis presented in Table 2 provides additional insights into the computational efficiency of the tested models. While accuracy and other performance metrics are crucial for evaluating a model's effectiveness, runtime is equally important, especially in real-world applications with limited computational resources. CPU usage was also observed to determine which models exhibited the highest and lowest processing demands during training. The saved model sizes were also examined to determine the most compact and largest models, providing insights for practical reuse in real-world applications. Among the tested models, RF demonstrated the lowest CPU consumption but had the largest saved model size. Given that SVM and XGBoost achieved slightly different evaluation scores with significantly smaller saved model sizes, they could serve as viable alternatives to RF.

Furthermore, the saved model sizes for each tested model were examined. The largest model identified was only 3.1 MB (RF)—just 0.56% of the size of VGG19 (approximately 549 MB). Despite this vast difference in size, the accuracy gap was minimal, at just 0.86%. This finding highlights a promising avenue for future research by demonstrating the potential of lightweight models to rival the performance of significantly larger deep learning architectures.

Finally, as a concluding part of the evaluation, we compared the testing accuracy of the proposed models with previous studies that employed larger and more complex deep learning architectures. The findings, summarized in Table 3, offer several noteworthy insights.

XGBoost, RF, and SVM achieved testing accuracies of 94.39%, 94.70%, and 94.23%, respectively. These results demonstrate that smaller, ensemble-based models can deliver performance comparable to, or even exceeding, some larger deep-learning models. For instance, ResNet101, a deep convolutional neural network with substantial computational demands, achieved an accuracy of 90.36%. Similarly, MSFCN, which specializes in segmentation tasks, reported an accuracy of 87.80%. These accuracies are significantly lower than those achieved by XGBoost and RF by emphasizing the efficiency of the lightweight models in this study.

Table 2. Time Comparison for Each Model

Model	Runtimes (second)	CPU	Saved Model Size
Logistic Regression	0.0742	47.35%	4 KB
Decision Tree	0.3694	74.70%	28 KB
Random Forest	**3.1412**	**50.55%**	**3.1 MB**
Gradient Boosting	38.1081	59.50%	704 KB
Support Vector Machine	**0.9753**	**51.95%**	**754 KB**
K-Nearest Neighbors	0.0011	50.50%	1.2 MB
Naive Bayes	0.0037	47.50%	7 KB
AdaBoost	1.8827	46.65%	36 KB
Bagging	1.3826	49.65%	204 KB
XGBoost	**2.8262**	**54.15%**	**511 KB**

Table 3. Model Comparison with Previous Results

Model		Accuracy
Current Research	XGBoost	94.39%
	Random Forest	94.70%
	SVM	94.23%
Previous Research	VGG19 [14]	95.25%
	Resnet101 [16]	90.36%
	MSFCN [37]	87.80%
	YOLO [17]	95.73%

The only model in the comparison that surpassed the proposed methods were YOLO and VGG19, with an accuracy of 95.73% and 95.25% respectively. However, YOLO is specifically designed for real-time object detection tasks and benefits from highly optimized architectures. Even so, the marginal difference between YOLO and Random Forest (1.03%) underscores the viability of ensemble models as competitive alternatives, particularly in scenarios where computational resources are limited.

The results highlight a critical advantage of the proposed models: their ability to deliver high accuracy without the need for extensive computational resources or large datasets. This makes them particularly well-suited for deployment in small devices or resource-constrained environments. Furthermore, the performance gap between XGBoost and Random Forest compared to ResNet101 and MSFCN underscores the efficiency of these models in handling complex data without requiring high-end hardware or prolonged training times.

Ultimately, the analysis of runtime, CPU usage, and model size highlights the delicate balance between model performance and computational efficiency. Models like Naive Bayes and K-Nearest Neighbours are suitable for scenarios where speed is paramount, but their lower accuracy makes them less ideal for complex tasks. Conversely, ensemble methods such as Random Forest, XGBoost, and SVM provide an excellent balance of runtime and accuracy, making them the most practical choices for real-world applications requiring reliability and efficiency. Gradient Boosting, while highly accurate, may be less feasible in scenarios where computational resources or time are constrained.

Building on these findings, this study opens new opportunities for further research utilizing diverse medical datasets and experimenting with other artificial intelligence models. Future research can build upon this foundation by exploring additional diseases and leveraging cutting-edge AI techniques, ultimately enhancing diagnostic precision, generalizability, and clinical applicability to improve healthcare accessibility and outcomes.

4 CONCLUSION

Based on our literature review, we recognize that osteosarcoma is a type of bone cancer that often appears suddenly, with symptoms that patients frequently overlook, leading to delayed treatment. Early detection is therefore essential to prevent further complications. With the assistance of AI, this has become feasible due to AI's ability to recognize patterns and retain data effectively. This makes AI an efficient solution for osteosarcoma detection.

This study showcases the effective detection of osteosarcoma using a straightforward approach that eliminates the need for pre-trained models or complex deep-learning infrastructure. The results show that several machine learning models successfully tackled challenges in medical data analysis, particularly in handling imbalanced datasets. By applying the SMOTE technique for data preparation, the performance achieved was comparable to that of larger models. One of the highest-performing models in our study was Random Forest, which achieved an accuracy of 94.70%, only 1.03% lower than YOLO. Notably, Random Forest demonstrated exceptional computational efficiency, characterized by a compact model size, rapid processing time, and moderate CPU usage, making it easy to deploy and maintain.

Our findings also indicate that by leveraging the SMOTE technique to address data imbalance and implementing lightweight machine learning algorithms like XGBoost and SVM, it is possible to achieve performance that surpasses larger deep learning models such as ResNet101 and MSFCN. Despite their simplicity, these lightweight models deliver high accuracy and computational efficiency, making them practical solutions for resource-constrained environments.

This study underscores the feasibility of deploying scalable and efficient diagnostic tools for osteosarcoma detection. By addressing the challenges of data imbalance and computational constraints, this approach lays a strong foundation for enhancing diagnostic accessibility and improving healthcare outcomes, particularly in underserved regions.

AUTHOR'S CONTRIBUTION

As the lead and corresponding author, Muhammad Ainul Fikri oversaw the research process, managed technical aspects, and drafted the manuscript, ensuring the study's effective progression and successful attainment of its objectives. Ajie Kusuma Wardhana contributed by developing the models and rigorously testing their performance, ensuring the reliability and robustness of the proposed methods. Yudha Riwanto significantly contributed to the documentation process and strengthened the analysis, yielding valuable insights that enhanced the research findings. Inggrid Yanuar Risca Partiwi's linguistic expertise was instrumental in refining the manuscript, elevating its clarity, coherence, and overall professional tone. Under the guidance of supervisors Fauzia Anis Sekar Ningrum and Iqbal Kurniawan Asmar Putra, the research was strategically directed and rigorously reviewed to ensure adherence to the highest academic standards. The collective contributions of the team members synergized to facilitate the successful completion of this research study.

COMPETING INTERESTS

By the publication ethics of this journal, Muhammad Ainul Fikri and the other authors of this article, confirm that there are no conflicts of interest (COI) or competing interests (CI) associated with this work.

ACKNOWLEDGMENT

Thank you to all individuals and institutions who have contributed to the development and completion of this project. We would like to express our sincere appreciation to the Department of Informatics Engineering, Jember State Polytechnic, Faculty of Computer & Science, Universitas Amikom Yogyakarta, Diploma Program in Information Technology, Malang State Polytechnic and Department of Electronics Engineering, Padang State Polytechnic. Their

invaluable support and assistance have been instrumental to the success of our efforts in completing this project.

REFERENCES

[1] F. Mahyudin, M. Edward, M. Hardian Basuki, Y. Abdul Bari, Y. Suwandani, and C. Author, "HOSPITAL SURABAYA 'A RETROSPECTIVE STUDY,'" no. 1, 2018, [Online]. Available: http://journal.unair.ac.id/ORTHO@journal-orthopaedi-and-traumatology-surabaya-media-104.html

[2] M. P. Link *et al.*, "The Effect of Adjuvant Chemotherapy on Relapse-Free Survival in Patients with Osteosarcoma of the Extremity," *New England Journal of Medicine*, vol. 314, no. 25, pp. 1600–1606, Jun. 1986, doi: 10.1056/NEJM198606193142502.

[3] "Key Statistics for Osteosarcoma." Accessed: Dec. 07, 2024. [Online]. Available: https://www.cancer.org/cancer/types/osteosarcoma/about/key-statistics.html

[4] M. A. Smith, S. F. Altekruse, P. C. Adamson, G. H. Reaman, and N. L. Seibel, "Declining childhood and adolescent cancer mortality," *Cancer*, vol. 120, no. 16, pp. 2497–2506, Aug. 2014, doi: 10.1002/cncr.28748.

[5] R. B. Guerra *et al.*, "COMPARATIVE ANALYSIS BETWEEN OSTEOSARCOMA AND EWING'S SARCOMA: EVALUATION OF THE TIME FROM ONSET OF SIGNS AND SYMPTOMS UNTIL DIAGNOSIS," *Clinics*, vol. 61, no. 2, pp. 99–106, Apr. 2006, doi: 10.1590/S1807-59322006000200003.

[6] K. L. Pan, W. H. Chan, and Y. Y. Chia, "Initial Symptoms and Delayed Diagnosis of Osteosarcoma around the Knee Joint," *Journal of Orthopaedic Surgery*, vol. 18, no. 1, pp. 55–57, Apr. 2010, doi: 10.1177/230949901001800112.

[7] A. F. Kamal, H. Widyawarman, K. Husodo, E. U. Hutagalung, W. Rajabto, and A. F. Kamal, "Clinical Outcome and Survival of Osteosarcoma Patients in Cipto Mangunkusumo Hospital: Limb Salvage Surgery versus Amputation."

[8] American Cancer Society, "What Is Osteosarcoma?" Accessed: Dec. 07, 2024. [Online]. Available: https://www.cancer.org/cancer/types/osteosarcoma/about/what-is-osteosarcoma.html

[9] L. Cai, J. Gao, and D. Zhao, "A review of the application of deep learning in medical image classification and segmentation," *Ann Transl Med*, vol. 8, no. 11, pp. 713–713, Jun. 2020, doi: 10.21037/atm.2020.02.44.

[10] W. Wang *et al.*, "Medical Image Classification Using Deep Learning," 2020, pp. 33–51. doi: 10.1007/978-3-030-32606-7_3.

[11] S. Tamuly, C. Jyotsna, and J. Amudha, "Deep Learning Model for Image Classification," 2020, pp. 312–320. doi: 10.1007/978-3-030-37218-7_36.

[12] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowl Based Syst*, vol. 215, p. 106771, Mar. 2021, doi: 10.1016/j.knosys.2021.106771.

[13] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep Learning for Hyperspectral Image Classification: An Overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019, doi: 10.1109/TGRS.2019.2907932.

[14] D. M. Anisuzzaman, H. Barzekar, L. Tong, J. Luo, and Z. Yu, "A deep learning study on osteosarcoma detection from histological images," *Biomed Signal Process Control*, vol. 69, Aug. 2021, doi: 10.1016/j.bspc.2021.102931.

[15] Patrick Leavey, Anita Sengupta, Dinesh Rakheja, Ovidiu Daescu, Harish Babu Arunachalam, and Rashika Mishra, "Osteosarcoma-Tumor-Assessment." Accessed: Dec. 07, 2024. [Online]. Available: https://www.cancerimagingarchive.net/collection/osteosarcoma-tumor-assessment/

[16] 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, Nov. 2023, doi: 10.1016/j.health.2023.100153.

[17] J. Li *et al.*, "Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model", doi: 10.1007/s00330-022-09289-y/Published.

[18] Keras.io, "Keras Applications." Accessed: Dec. 09, 2024. [Online]. Available: https://keras.io/api/applications/

[19] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif Intell Rev*, vol. 57, no. 6, p. 137, May 2024, doi: 10.1007/s10462-024-10759-6.

[20] M. T. Aziz *et al.*, "A Novel Hybrid Approach for Classifying Osteosarcoma Using Deep Feature Extraction and Multilayer Perceptron," *Diagnostics*, vol. 13, no. 12, Jun. 2023, doi: 10.3390/diagnostics13122106.

[21] M. A. A. Walid *et al.*, "Adapted Deep Ensemble Learning-Based Voting Classifier for Osteosarcoma Cancer Classification," *Diagnostics*, vol. 13, no. 19, Oct. 2023, doi: 10.3390/diagnostics13193155.

[22] H. Tang, H. Huang, J. Liu, J. Zhu, F. Gou, and J. Wu, "AI-Assisted Diagnosis and Decision-Making Method in Developing Countries for Osteosarcoma," *Healthcare (Switzerland)*, vol. 10, no. 11, Nov. 2022, doi: 10.3390/healthcare10112313.

[23] T. Ouyang, S. Yang, F. Gou, Z. Dai, and J. Wu, "Rethinking U-Net from an Attention Perspective with Transformers for Osteosarcoma MRI Image Segmentation," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/7973404.

[24] B. Lv, F. Liu, Y. Li, J. Nie, F. Gou, and J. Wu, "Artificial Intelligence-Aided Diagnosis Solution by Enhancing the Edge Features of Medical Images," *Diagnostics*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/diagnostics13061063.

[25] J. Wu, Y. Guo, F. Gou, and Z. Dai, "A medical assistant segmentation method for MRI images of osteosarcoma based on DecoupleSegNet," *International Journal of Intelligent Systems*, vol. 37, no. 11, pp. 8436–8461, Nov. 2022, doi: 10.1002/int.22949.

[26] F. Liu, L. Xing, X. Zhang, and X. Zhang, "A four-pseudogene classifier identified by machine learning serves as a novel prognostic marker for survival of osteosarcoma," *Genes (Basel)*, vol. 10, no. 6, Jun. 2019, doi: 10.3390/genes10060414.

[27] A. Fernandez, S. Garcia, 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, Apr. 2018, doi: 10.1613/jair.1.11192.

[28] M. Mukherjee and M. Khushi, "SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features," *Applied System Innovation*, vol. 4, no. 1, p. 18, Mar. 2021, doi: 10.3390/asi4010018.

[29] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[30] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, p. 106, Dec. 2013, doi: 10.1186/1471-2105-14-106.

[31] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[32] R. Oktafiani and E. Itje Sela, "Breast Cancer Classification with Principal Component Analysis and Smote using Random Forest Method and Support Vector Machine," 2024. [Online]. Available: https://archive.ics.uci.edu/

[33] Z. Zhao and T. Bai, "Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms," *Entropy*, vol. 24, no. 8, Aug. 2022, doi: 10.3390/e24081157.

[34] J. Wen, X. Tang, and J. Lu, "An imbalanced learning method based on graph tran-smote for fraud detection," *Sci Rep*, vol. 14, no. 1, p. 16560, Jul. 2024, doi: 10.1038/s41598-024-67550-4.

[35] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

[36] S. Hasan Abdulla, A. M. Sagheer, and H. Veisi, "Improving Breast Cancer Classification Using (Smote) Technique and Pectoral Muscle Removal in Mammographic Images," *MENDEL-Soft Computing Journal*, vol. 2, pp. 2571–3701, 2021, doi: 10.13164/mendel.2021..0.

[37] L. Huang, W. Xia, B. Zhang, B. Qiu, and X. Gao, "MSFCN-multiple supervised fully convolutional networks for the osteosarcoma segmentation of CT images," *Comput Methods Programs Biomed*, vol. 143, pp. 67–74, May 2017, doi: 10.1016/j.cmpb.2017.02.013.