Applied Data Science and Analysis
Vol.2024, **pp.** 52–68
DOI: https://doi.org/10.58496/ADSA/2024/006; ISSN: 3005-317X
https://mc04.manuscriptcentral.com/adsa

Research Article

Advanced Ensemble Classifier Techniques for Predicting Tumor Viability in Osteosarcoma Histological Slide Images

Tahsien Al-Quraishi [1], *, , Chee Keong NG [1], , Osama A. Mahdi [2], Amoakoh Gyasi [2], , Naseer Al-Quraishi [3],

[1] *Victorian Institute of Technology, School of IT, Melbourne, Victoria, Australia*
[2] *Melbourne Institute of Technology, School of IT, Melbourne, Victoria, Australia*
[3] *Alayen Iraqi University, College of Computer Science, Computer Science Department, Nasiriyah, Iraq*

ARTICLEINFO

Article History

Received 01 Mar 2024
Revised 13 Apr 2024
Accepted 08 May 2024
Published 29 May 2024

Keywords

Osteosarcoma Cancer Histological Slide Images Machine Learning

Feature Selection Ensemble Classifier

Voting Classifier

Multi-Layer Perceptron Random Forest

Logistic Regression AdaBoost Classifier

GB Classifier

ABSTRACT

Osteosarcoma is considered as the primary malignant tumor of the bone, emanating from primitive mesenchymal cells that form osteoid or immature bone. Accurate diagnosis and classification play a key role in management planning to achieve improved patient outcomes. Machine learning techniques may be used to augment and surpass existing conventional methods towards an analysis of medical data. In the present study, the combination of feature selection techniques and classification methods was used in the development of predictive models of osteosarcoma cases. The techniques include L1 Regularization (Lasso), Recursive Feature Elimination (RFE), SelectKBest, Tree-based Feature Importance, while the following classification methods were applied: Voting Classifier, Decision Tree, Naive Bayes, Multi-Layer Perceptron, Random Forest, Logistic Regression, AdaBoost, and Gradient Boosting. Some model assessment was done by combining metrics such as accuracy, precision, recall, F1 score, AUC, and V score. The combination of the Tree-Based Feature Importance for feature selection and Voting Classifier with Decision Tree Classifier proved to be giving a higher performance compared to all other combinations, where such combinations helped in correct classification of positive instances and wonderful minimization of false positives. Other combinations also gave significant performances but slightly less effective, for example, L1 Regularization with the Voting Classifier, RFE with the Voting Classifier. This work presents strong evidence that advanced machine learning with ensemble classifiers and robust feature selection can result in overall improvement of the diagnostic accuracy and robustness for the classification of osteosarcoma. Research on class imbalance and computational efficiency will be its future research priority.

1. INTRODUCTION

Osteosarcoma is the most common bone cancer, characterized by the development of osteoid or immature bone by malignant mesenchymal cells [1]. In 2023, it is expected that around 1,000 Americans of all ages will be told they have osteosarcoma. About half of these cases will be in children and teens. Osteosarcomas comprise about 2% of all cancers in children under age 14 and 3% of all cancers in children between the ages of 15 to 19 years. It's diagnosed more commonly between the ages of 10 to 30, with most teenagers diagnosed in this age category. However, it can be identified at any age, including elderly individuals. This form of sarcoma represents only ten percent of all osteosarcoma diagnoses and tends to occur in those over 60 years old [2]. Presently, in all cases, a detailed history and physical inquest are required for diagnosis. Any bone is susceptible to osteosarcoma, though the knee is commonly affected because of the involvement of the metaphysis of the proximal tibia, distal femur, or proximal humerus. Diagnosis utilizes X-rays, CT scans, and MRI. [3]. As illustrated in Figure 1, the five-year survival rates for osteosarcoma cases from 2012 to 2018 indicate significant trends

**Corresponding author. Email: tahsien.a@vit.edu.au*

in patient outcomes. Additionally, Figure 2 provides a comparative analysis of osteosarcoma diagnosis through imaging, showcasing (A) a CT scan and (B) an MR coronal image of an osteosarcoma case in the right distal femur. These figures highlight the importance of accurate diagnostic tools in assessing tumor progression and treatment efficacy

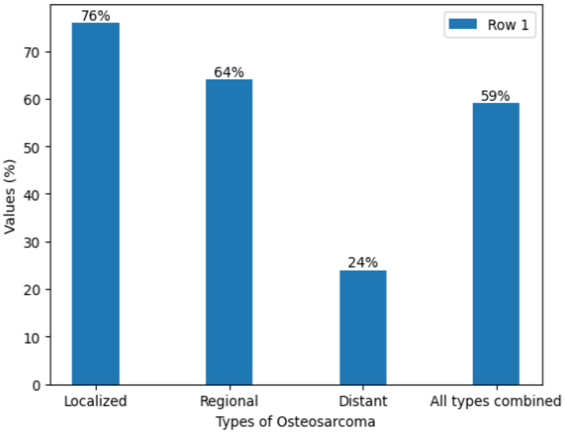


Fig. 1. Five-year Survival Rates for Osteosarcoma 2012 and 2018

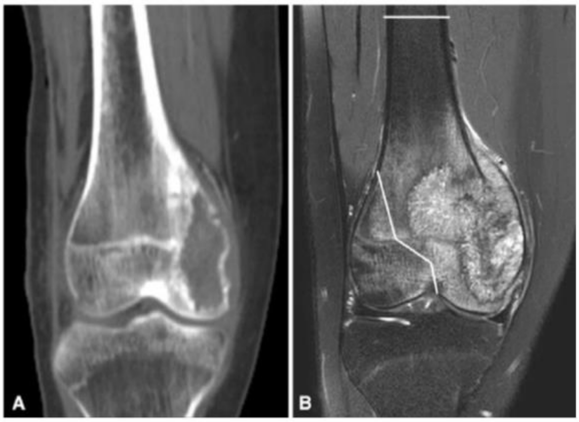


Fig. 2. A–B (A) CT and (B) MR coronal images show osteosarcoma in the right distal femur of our patient.

X-rays with high spatial resolution can quite effectively show the size and location of the tumor, as well as the degree of bone degradation that has taken place [4]. They bring out even tiny periosteal reactions and the Codman triangles [5].

CT illustrates the extent of infiltrates inside and outside, and the micro-mineralized bone-like development of tumors, which are not observed on the X-rays Pathological fractures can be diagnosed with the guidance of CT. Conventional MRI remains the mainstay for preoperative evaluation in surgical planning and optimally defines the intramedullary extent of the tumor, the size of the soft tissue mass or swelling, and the surrounding structures [6]. Biopsy is essential to achieve a definite diagnosis of osteosarcoma. Tissue biopsy is still considered to be the gold standard for diagnosing osteosarcoma [7]. Although cancer detection based on biopsy has proven effective, histologically guided biopsies and MRI scans have rather poor sensitivity for detecting cancer. The preparation of histology samples is rather time-consuming. For example, in the case of the reliable diagnosis of the malignancy of osteosarcoma, at least 50 histology slides are required to simulate the plane of a large three-dimensional tumor [8].

1.1 Problem

The diagnosis and treatment of cancer are becoming increasingly complex due to the growing number of types of cancer and the diversity of therapies tailored to individual patients [9]. Traditionally, pathologists use a microscope for the interpretation of digital whole slide images (WSI) in their process of manual evaluation. The digitization of histological slides of a tumor into a digital image database and the use of machine learning to analyze patterns in MRI and histology digital images may potentially enhance pathological processes [10].

Such predictive abilities produced by the model would be of immense assistance to healthcare providers for making guided decisions regarding treatment modalities, thereby improving patients' quality of life [11]. Among these, ML and AI methods find significant importance in the health sector. This kind of model can be used for estimation, prediction, and developing a design process toward a good model that can be improved through experience [12] .As technology continues to evolve rapidly, new processing systems become more powerful instantly. Consequently, the medical field has developed automated diagnostic methods that predict whether a tumor is likely to be benign or malignant [13].

1.2 Contribution and Novelty

Applying feature selection techniques at the pre-processing level of learning algorithms enhances the accuracy of learning, and the results of the feature selection domain provide better performance [14]. For this reason, inclusion of a feature selection process is crucial in all phases of image processing. Performance of one's classifier heavily depends on selected features. The study discusses a few effective machine learning approaches, including feature selection algorithms like L1 regularization and the Recursive Feature Elimination (RFE) method, coupled with the Tree-Based Feature Importance and SelectKBest. More recently, with the growth of the ensemble approach, it has been used to be quite an effective way to combine many learning algorithms to improve prediction capabilities overall [15].

Our model proposed here will show the successful development of automated learning techniques for classifying osteosarcoma WSIs into non-tumor, viable tumor, and non-viable tumor correctly stating the regions of tumor and necrosis. The following are the points that make our study very much new:

- Finding optimal hyperparameters for decision tree classifiers using the GridsearchCV method.
- The concept of concatenation doing data augmentation that keeps increasing the dataset's sample count
- Provide a comparative evaluation of the different techniques with an emphasis on KPIs such as accuracy, precision, and recall, some combined metrics: F-score, V-score, and AUC.
- To prove the superiority of the Decision Tree Voting Classifier with the Tree-Based Feature Importance framework over the existing method [16].

The rest of the paper is organized as follows: Section II reviews various ML algorithms. The proposed framework appears in Section III, and in Section IV, there is a discussion on the plurality classifiers with different feature selection techniques and finally describing the proposed ensemble technique. Section V elaborates the experiment, findings, and results. The paper concludes in the conclusion section, which is Section VI.

2. RELATED WORK

The application of machine learning for disease classification is gaining traction [17]. able 1 offers a concise overview of recent studies related to predicting osteosarcoma development using machine learning methods. Firstly, The authors in [18] tackled the challenge of assessing osteosarcoma treatment response. They proposed a digital image analysis method to automate this process, improving accuracy and efficiency. Their method combines pixel-based and object-based techniques to segment tumor and non-tumor areas in high-resolution Whole Slide Images (WSIs) of osteosarcoma. This approach analyzes tumor properties such as nuclei clustering, density, and circularity to differentiate between viable and non-viable regions. Initially, K-Means clustering with color normalization was used to isolate tumors, followed by a Flood-Fill algorithm to group similar pixels into cellular objects for further analysis. Their results showed approximately 90% accuracy.

Similarly, in [19], a convolutional neural network (CNN) was used on the osteosarcoma WSIs dataset to enhance the classification accuracy of osteosarcoma tumors into viable tumor, necrosis, and non-tumor categories. Due to training limitations, the original 1024 x 1024 images were cropped to 128 x 128 patches. Using CNNs, the classification accuracy for distinguishing between tumor classes (viable tumor and necrosis) and non-tumor regions increased significantly to 92%.

Vezakis et al. [20] explored the use of CNNs for predicting osteosarcoma malignancy. They utilized transfer learning with VGG19 and InceptionV3 models on a dataset of 40 WSIs of osteosarcoma tumors, achieving a 2% accuracy improvement over previous studies. The VGG19 model achieved 93.91% accuracy, while the InceptionV3 model reached 78.26% in multiclass classification tasks, highlighting the potential of CNNs in improving osteosarcoma diagnostic accuracy. Zhou et al. [21] reviewed the application of deep learning (DL) in bone tumor diagnosis, emphasizing its potential in enhancing detection, classification, segmentation, and volume calculation of primary and metastatic bone tumors from radiological and pathological images. One DL model they highlighted achieved 91.2% accuracy in classifying viable and necrotic tumor regions in osteosarcoma. Despite these promising results, challenges such as the need for extensive labeled data, model generalization, and integrating clinical parameters persist. Future research should address these issues to improve DL model implementation in clinical settings for better bone tumor diagnosis and prognosis.

Gawade et al. [16] dealt with the use of convolutional neural network algorithms and supervised deep learning methods for detecting osteosarcoma. They tested models such as VGG16, VGG19, DenseNet201, and ResNet101 on an osteosarcoma image dataset. It was found from the results that the ResNet101 model outperforms other models by achieving a maximum accuracy of 90.36%, while the VGG19 model achieves 89.44% accuracy. Additionally, a similar research project pointed out that these models could attain the level of automation required for the enhancement of overall accuracy in the diagnosis of osteosarcomas—a disease that is traditionally diagnosed through rather time-consuming and difficult-to-carry-out biopsy procedures.

TABLE. I : SUMMARY OF COMPARISON WITH OTHER STATE-OF-THE METHODS FOR THE PREDICTION OF OSTEOSARCOMA CANCER

Ref	Dataset	Architecture	Advantage	Disadvantage
[20]	WSIs dataset	K-Means	90% accuracy highlights its effectiveness in differentiating viable and non-viable regions	The K-Means & Flood-Fill algorithm negatively impact accuracy
[21]	Cancer Imaging Archive (TCIA)	CNN	Improve the average classification accuracy to 92%	labelled data, generalization, and preprocessing.
[22]	Cancer Imaging Archive (TCIA)	CNN	93.91% accuracy improving the classification	labelled data, generalization, and complexity
[23]	Medical Center of Children in Dallas	Deep learning model	91.2% accuracy indicates the efficacy of the Deep learning model	Need for substantial labeled data, model generalisation, and Integration of clinical parameters
[18]	University of Texas Dallas	ResNet101	90.36% indicating its effectiveness in detection of bone cancer	Sensitivity specificity and precision are not provided

3. PROPOSED FRAMEWORK

The methodology suggested in this research critically examines the dataset of osteosarcoma and includes both the process of description of data and data cleaning. Large datasets and large predictive models need powerful algorithms to work with a lot of computational time [22]. Figure 3 shows that the mechanism of selecting features keeps the significant variables of the osteosarcoma dataset for enhanced performance of the model and lower complexity. We further explain that the classification approach trains machine learning models with the osteosarcoma dataset for classifying patients with osteosarcoma into several groups: non-tumor, non-viable tumor, viable, and viable non-viable. Figure 3 presents the architectural blueprint of the proposed osteosarcoma classification model, outlining the key steps involved in feature selection, data preprocessing, and classification techniques used to enhance diagnostic accuracy

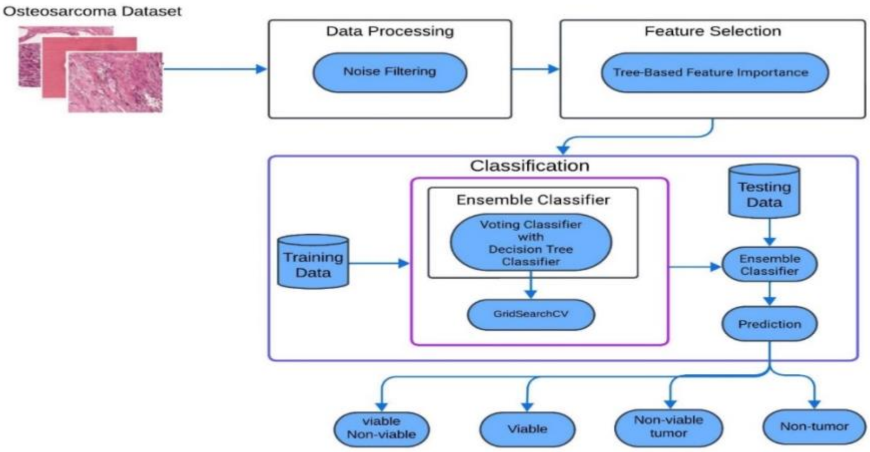


Fig. 3. Architectural Blueprint of the Proposed Osteosarcoma Model

3.1 Dataset Description

This study used the osteosarcoma dataset from Leavey et al. [23]. Fig. 4 classifies cases into non-tumor, non-viable tumor, viable tumor, and viable non-viable tumor groups. The dataset contains 1,144 samples with 69 features across these categories. There was a class imbalance, with 'non-tumor' being the most common, potentially hindering weight updates for minority classes. However, only one minority class, 'viable: non-viable,' was identified, which doesn't significantly affect model performance, as noted by Dal Pozzolo et al. [24]. The issue of class imbalance has been widely studied, as He and Garcia [25] highlighted.

The dataset's variables are divided into three categories: essential for predicting class labels, unnecessary with no impact on predictions, and redundant ones that can be replaced by others [26]. These variables should correlate with the target variable but not with each other. During data cleaning, two variables, image.name and unnamed, were removed, leaving 67 features. Among these, 66 are numeric (either floats or integers), and one is nominal, representing the class label. Additionally, the attributes 'area' and 'circularity' each had a missing value. Table 2 shows the basic statistics of the first five samples.

3.2 Data Processing

Data preprocessing is crucial to ensure the dataset's quality and consistency. This involves correcting erroneous values, often referred to as noise, which can come from human errors or faulty data collection tools. Noise includes duplicates, missing values, inaccurate entries (both inliers and outliers), and mislabeled data points. Our analysis found no duplicates, NaN (Not a Number) values, or categorical features, but we did find missing values in the 'circularity' and 'area' variables. How we handled these missing values depended on whether there were outliers and their proportion in the dataset.

Outliers are data points that significantly deviate from other values in the dataset, potentially distorting patterns and compromising model accuracy, leading to misleading predictions [27]. Box plot inspections revealed outliers in both variables. Since these outliers represented less than 3% (34 out of 1,144) of the dataset, we decided to exclude them and replace the missing values with the mean values of their respective variables. Alternatively, we could have omitted the samples with missing values, given their minimal count of two. Figure 4 provides a statistical analysis of the osteosarcoma dataset, classifying cases into different tumor categories and highlighting key dataset characteristics relevant to the machine learning model's performance.

To identify outliers, we used a distance-based, unsupervised outlier detection approach. Specifically, a data point, denoted as xk, is considered an outlier if it meets the following condition:

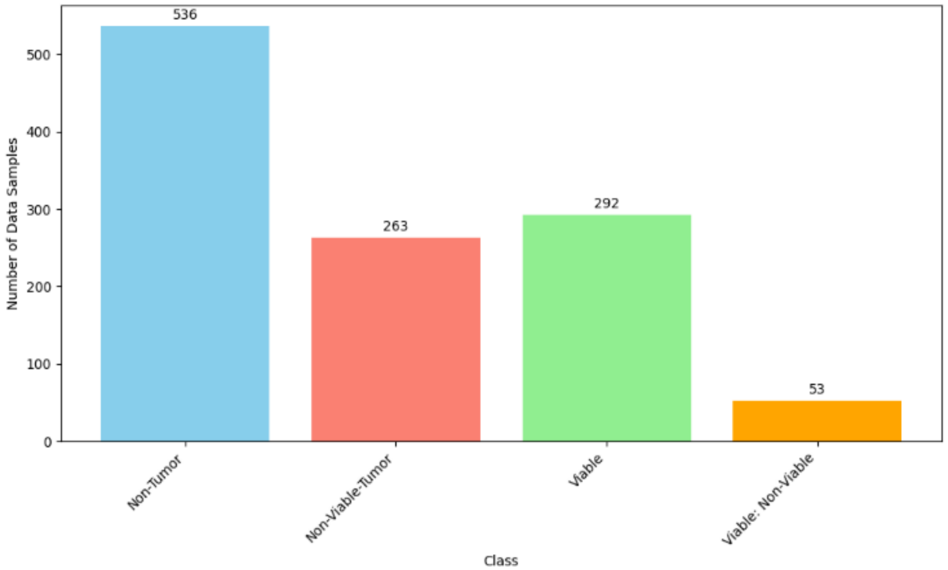
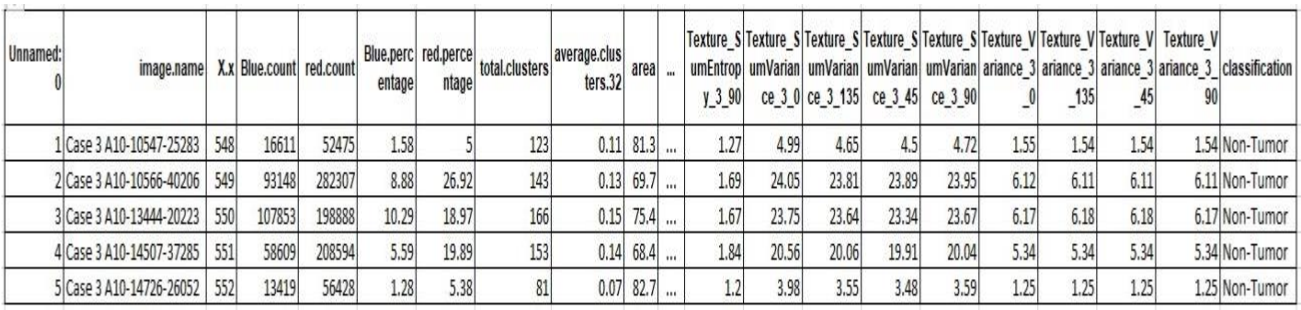


Fig. 4. Analysis of the Osteosarcoma Dataset Statistical Information

Table 2 summarizes the first five samples from the original osteosarcoma dataset, offering insight into the structure and nature of the dataset, including key features and their distribution, which play a crucial role in model training and classification. Among the 1,144 samples, the 'non-tumor' category is the most prevalent, potentially affecting model learning balance.

TABLE. II : THE FIRST FIVE SAMPLES IN THE ORIGINAL OSTEOSARCOMA DATASET.



Unnamed: 0	image.name	X.x	Blue.count	red.count	Blue.percentage	red.percentage	total.clusters	average.clusters.32	area	...	Texture_SumEntropy_3_90	Texture_SumVariance_3_0	Texture_SumVariance_3_135	Texture_SumVariance_3_45	Texture_SumVariance_3_90	Texture_Variance_3_0	Texture_Variance_3_135	Texture_Variance_3_45	Texture_Variance_3_90	classification
1	Case 3 A10-10547-25283	548	16611	52475	1.58	5	123	0.11	81.3	...	1.27	4.99	4.65	4.5	4.72	1.55	1.54	1.54	1.54	Non-Tumor
2	Case 3 A10-10566-40206	549	93148	282307	8.88	26.92	143	0.13	69.7	...	1.69	24.05	23.81	23.89	23.95	6.12	6.11	6.11	6.11	Non-Tumor
3	Case 3 A10-13444-20223	550	107853	198888	10.29	18.97	166	0.15	75.4	...	1.67	23.75	23.64	23.34	23.67	6.17	6.18	6.18	6.17	Non-Tumor
4	Case 3 A10-14507-37285	551	58609	208594	5.59	19.89	153	0.14	68.4	...	1.84	20.56	20.06	19.91	20.04	5.34	5.34	5.34	5.34	Non-Tumor
5	Case 3 A10-14726-26052	552	13419	56428	1.28	5.38	81	0.07	82.7	...	1.2	3.98	3.55	3.48	3.59	1.25	1.25	1.25	1.25	Non-Tumor

$$xk < Q1 - 1.5 \times IQR \; or \; xk > Q3 + 1.5 \times IQRxk < Q1 - 1.5 \times IQR \; or \; xk > Q3 + 1.5 \times IQR \qquad (1)$$

In equation 1, the interquartile range (IQR) denotes the difference between the upper quartile (Q3) and the lower quartile (Q1).

3.3 Feature Selection Methods

The primary goal of the proposed feature selection approach is to achieve a minimum accuracy level equivalent to using the complete set of features, with the secondary goal of enhancing this accuracy rate. Gathering extensive information about all features requires substantial time and financial resources and causes unnecessary delays in classification and diagnosis processes. Therefore, it is more efficient to reduce dimensionality by decreasing the number of features, aiming for better performance, and establishing stronger correlations between features and outcomes.

In this study, the primary method used for feature selection is Tree-Based Feature Importance, identified as the most effective technique. While other methods such as L1 Regularization (Lasso), Recursive Feature Elimination (RFE), and SelectKBest were considered, Tree-Based Feature Importance was chosen for its superior ability to identify and preserve the most critical features. This method leverages decision tree algorithms to rank features based on their importance, allowing the model to focus on those that contribute most significantly to its predictive power. By using Tree-Based Feature Importance, the study aims to improve model accuracy, reduce overfitting, and enhance computational efficiency.

3.4 Classification Methods

In this study, we used various machine learning models to classify osteosarcoma cases into different categories. Our main goal was to identify the most effective model based on accuracy, precision, recall, and F1 score. The classification techniques we utilized included:

- **Voting Classifier:** The Voting Classifier is an ensemble technique that merges the predictions of several classifiers to enhance overall performance. In our study, we utilized an ensemble of Decision Tree Classifiers. By combining the predictions from these classifiers, the Voting Classifier takes advantage of the strengths of each model, leading to greater accuracy and robustness [28]. This approach has been demonstrated to improve model stability and decrease overfitting, particularly in medical datasets where data variability is often high [29].
- **Decision Tree and Naive Bayes:** This method involves training both a Decision Tree and a Naive Bayes classifier on the selected features. The Decision Tree model captures complex decision boundaries [30], while the Naive Bayes classifier, which assumes independence between features, is simple yet effective for certain datasets [31]. Combining these models balances interpretability and complexity, providing a robust approach for classifying osteosarcoma cases. [32].
- **MLP and Random Forest (RF):** The Multi-Layer Perceptron (MLP) is a type of neural network that captures non-linear relationships in the data [33], while Random Forest is an ensemble method that reduces overfitting by averaging multiple decision trees [34]. This combination aims to harness the power of deep learning with the stability of ensemble methods, making it suitable for complex datasets like osteosarcoma [35].
- **Logistic Regression and AdaBoost**: Logistic Regression is known for its simplicity and interpretability, making it a baseline model for many classification tasks [36]. AdaBoost, on the other hand, is an ensemble technique that focuses on misclassified instances, iteratively improving the model's performance [37]. The combination of these methods provides a balance between simplicity and enhanced performance through boosting techniques [38].
- **Random Forest (RF) and Gradient Boosting (GB):** Both Random Forest and Gradient Boosting classifiers were applied to the dataset. Random Forests are known for their robustness and ability to handle high-dimensional data [39],while Gradient Boosting enhances predictive performance through iterative improvement by focusing on the

errors of previous models [40]. This combination is particularly effective for improving model accuracy and handling complex interactions within the data.

The proposed framework involves building a robust model using a Voting Classifier that combines multiple Decision Tree Classifiers. This ensemble method is selected for its robustness and ability to improve model accuracy by leveraging the strengths of individual classifiers. Two Decision Tree Classifiers are optimized using Grid SearchCV to determine the best hyperparameters, including max_depth, min_samples_split, and min_samples_leaf. The Voting Classifier aggregates the predictions from these optimized classifiers in a soft voting scheme, which enhances the overall predictive performance. The Voting Classifier is trained on the preprocessed dataset, and its performance is evaluated using several metrics to ensure its effectiveness. The primary metrics include accuracy, precision, recall, F1 score, AUC, and V score. These metrics provide a comprehensive evaluation of the model's performance, highlighting its ability to correctly classify instances and its robustness across different evaluation criteria. The confusion matrix is visualized using a heatmap to analyze true positives, false positives, true negatives, and false negatives, providing insights into the model's classification capabilities.

3.5 Performance Metrics

Performance metrics are critical in evaluating the effectiveness of machine learning models, particularly in the context of classification problems. They provide a quantitative basis for comparing different models and understanding their strengths and weaknesses. The primary metrics discussed here include accuracy, sensitivity (recall), specificity, F1 score, and AUC, each of which offers unique insights into model performance.

- **Accuracy**

  In equation 2, the accuracy is a performance metric that quantifies the proportion of accurate results, which includes both true positives and true negatives, relative to the total cases scrutinized in the experiment. Essentially, accuracy measures how well a binary classification test can accurately discern or rule out a specific condition.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (2)$$

  Here, TP represents true positives, FP stands for false positives, TN denotes true negatives, and FN indicates false negatives. All these values can be derived from the confusion matrix, a table frequently employed to describe the performance of a classification model when the true values are known [37].

- **Sensitivity (Recall)**

  The sensitivity, also referred to as recall in equation 3, gauges the proportion of true positives correctly identified. It is particularly important in contexts where the cost of false negatives is high. A high sensitivity indicates that the model is effective at capturing positive instances.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

  This metric is crucial in medical diagnostics and fraud detection, where missing positive cases can have serious consequences. According to Powers [1], sensitivity provides a clear picture of the model's ability to detect actual positives, making it a key metric in evaluating model performance in imbalanced datasets.

- **Specificity**

  While in equation 4, the specificity quantifies the proportion of true negatives in experimental outcomes. It measures how well the model avoids false alarms, indicating the proportion of actual negatives that are correctly identified as such.

$$\text{Sensitivity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (4)$$

High specificity is essential in scenarios where the cost of false positives is high. For instance, in spam detection, a high specificity would mean that most of the emails identified as non-spam are indeed non-spam. As noted by Sokolova and Lapalme [2], specificity complements sensitivity by focusing on the true negative rate, thus providing a balanced view of model performance.

- **F1 Score**

The F1 Score is represented in equation 5 and it is the harmonic mean of precision and recall, providing a balance between the two metrics. It is especially useful when the class distribution is imbalanced, as it considers both false positives and false negatives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The F1 score is particularly useful in applications where both precision and recall are important, such as information retrieval and text classification. According to Fesseha et al. [41], the F1 score offers a single metric that balances the trade-offs between precision and recall, making it a robust measure of a model's performance.

- **Area Under the Curve (AUC)**

The Area Under the Curve (AUC) in equation 6 is a key performance metric for binary classifiers, measuring the ability to distinguish between positive and negative classes. It is derived from the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate (TPR) against False Positive Rate (FPR) at various thresholds. The AUC value ranges from 0 to 1, where 1 indicates perfect discrimination and 0.5 suggests random guessing. Mathematically, AUC can be interpreted as the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{6}$$

The AUC is particularly useful in imbalanced datasets as it evaluates performance across all thresholds, providing a comprehensive measure. Hanley and McNeil [42]and Bradley [43] provide foundational insights into AUC's application and advantages in model evaluation.

- **V Score**

The V score, or validity score shown in equation 7, is another important metric that assesses the quality and robustness of a model, particularly in terms of its generalizability to new data. It considers both the predictive performance of the model and its consistency across different subsets of data. The V score is calculated as follows:

$$\text{V Score} = \frac{1}{n} \Sigma_{i=1}^{n} \frac{\text{accuracy on subset } i}{\text{Preciexpected accuracy on random data}} \tag{7}$$

A higher V score indicates a model that not only performs well on the training data but also maintains its accuracy across various test conditions, ensuring reliability in real-world applications. This metric is particularly valuable in assessing the robustness of models in dynamic and diverse environments, where data can vary significantly over time [44].

4. PERFORMACE EVALUATION

The performance of the predictive models was evaluated using several key metrics: accuracy, precision, recall, F1 score, AUC, and V Score. The analysis reveals distinct trends and insights based on the combination of feature selection methods and classifiers.

Figure 5 shows that among the various combinations tested, the use of Tree-Based Feature Importance for feature selection combined with the Voting Classifier using Decision Tree Classifier achieved the highest performance. Specifically, this approach attained an impressive accuracy of 95.41%, precision of 94.99%, recall of 95.41%, and an F1 score of 95.09% according to the confusion matrix. The AUC was 90.86%, and the V Score was 92.88%, the confusion matrix in figure 10 indicates the model's exceptional capability in correctly identifying positive instances while minimizing false positives, thus making it the most robust approach for the osteosarcoma dataset. In comparison, the L1 Regularization (Lasso) combined with Voting Classifier also performed well, achieving an accuracy of 94.72%, precision of 94.66%, recall of 94.72%, and an F1 score of 94.62%. The AUC was 92.82%, and the V Score was 91.84%, demonstrating significant benefits in improving model generalization and reducing overfitting. However, it slightly underperformed compared to the Tree-Based Feature Importance method.

In contrast to figure 6, the combination of Decision Tree and Naive Bayes classifiers with various feature selection methods generally exhibited moderate performance. The highest accuracy for this combination was achieved with L1 Regularization (Lasso) at 86.70%, whereas SelectKBest resulted in the lowest accuracy at 63.30%. Similar trends were observed in precision, recall, and F1 scores according to the confusion matrix in figure10 , indicating this combination's relative ineffectiveness in handling the dataset compared to Voting Classifier. The Tree-Based Feature Importance method combined with Decision Tree and Naive Bayes achieved an accuracy of 79.36%, precision of 86.74%, recall of 79.36%, and an F1 score of 81.19%. The AUC was 86.94%, and the V Score was 70.27%, indicating a moderate performance.

Figure 7 represents the MLP and Random Forest (RF) combination showing moderate performance across all feature selection methods. The highest accuracy was achieved with RFE at 75.69%, while SelectKBest had the lowest at 72.02%. These results suggest that while this combination can capture both linear and non-linear relationships, it still fell short of the top-performing methods according to the confusion matrix in figure 11. For instance, the RFE combined with MLP+RF achieved an accuracy of 75.69%, precision of 77.42%, recall of 75.69%, and an F1 score of 73.04%. The AUC was 81.05%, and the V Score was 64.79%.

Figure 8 states that the Logistic Regression and AdaBoost combination demonstrated good performance but was outperformed by Voting Classifier. The highest accuracy was achieved with both L1 Regularization (Lasso) and RFE at 81.19%, while SelectKBest had the lowest at 67.43% according to the confusion matrix in figure 12. This combination's moderate performance indicates the need for more effective feature selection despite the benefits of regularization and boosting. Specifically, L1 Regularization (Lasso) with Logistic Regression and AdaBoost achieved an accuracy of 81.19%, precision of 81.98%, recall of 81.19%, and an F1 score of 79.72%. The AUC was 85.62%, and the V Score was 70.61%.

figure 9 shows that the Random Forest (RF) and Gradient Boosting (GB) combination showed varying performances. The highest accuracy was achieved with L1 Regularization (Lasso) at 88.07%, while SelectKBest recorded the lowest at 69.27% according to the confusion matrix in figure 13. Despite the high accuracy, the precision and recall were lower, indicating potential overfitting and the need for further tuning. For example, L1 Regularization (Lasso) with RF and GB achieved an accuracy of 88.07%, precision of 66.38%, recall of 67.53%, and an F1 score of 66.78%. The AUC was 81.44%, and the V Score was 81.63%.

The Recursive Feature Elimination (RFE) method combined with Voting Classifier yielded an accuracy of 88.53%, precision of 89.84%, recall of 88.53%, and an F1 score of 89.07% as shown in the confusion matrix in figure 10. The AUC was 87.35%, and the V Score was 82.46%. While effective in reducing the feature space, this method's performance was lower than that of Lasso and Tree-Based Feature Importance, suggesting that some critical features might have been eliminated in the process. The SelectKBest method with Voting Classifier recorded the lowest performance among the Voting Classifier combinations, with an accuracy of 75.23%, precision of 75.86%, recall of 75.23%, and an F1 score of 75.20%. The AUC was 76.90%, and the V Score was 62.19%. The simplicity of this method likely contributed to its reduced effectiveness, as it does not account for feature interactions.

Figures 10, 11, 12, 13, and 14 present the confusion matrices for different classification models applied in this study. Specifically, Figure 10 shows the confusion matrix for the Decision Tree algorithm, while Figures 11, 12, 13, and 14 illustrate the confusion matrices for the Decision Tree & Naïve Bayes, MLP & Random Forest, Logistic Regression & AdaBoost, and Random Forest & Gradient Boosting classifiers, respectively. These matrices provide a detailed evaluation of model performance, displaying true positive, false positive, true negative, and false negative classifications, which are critical in assessing the effectiveness of each machine learning approach in predicting osteosarcoma viability.

In summary, the Tree-Based Feature Importance with Voting Classifier using Decision Tree Classifier emerged as the most effective model compare to the previous study [16], followed by L1 Regularization (Lasso) with Voting Classifier. Other combinations showed varying degrees of success, with Decision Tree and Naive Bayes, MLP and RF, Logistic Regression and AdaBoost, and RF and Gradient Boosting displaying moderate to low performance based on the specific feature selection method used. The analysis highlights the importance of selecting appropriate feature selection methods to enhance the performance of classification models in predicting tumor viability in osteosarcoma histological slide images.

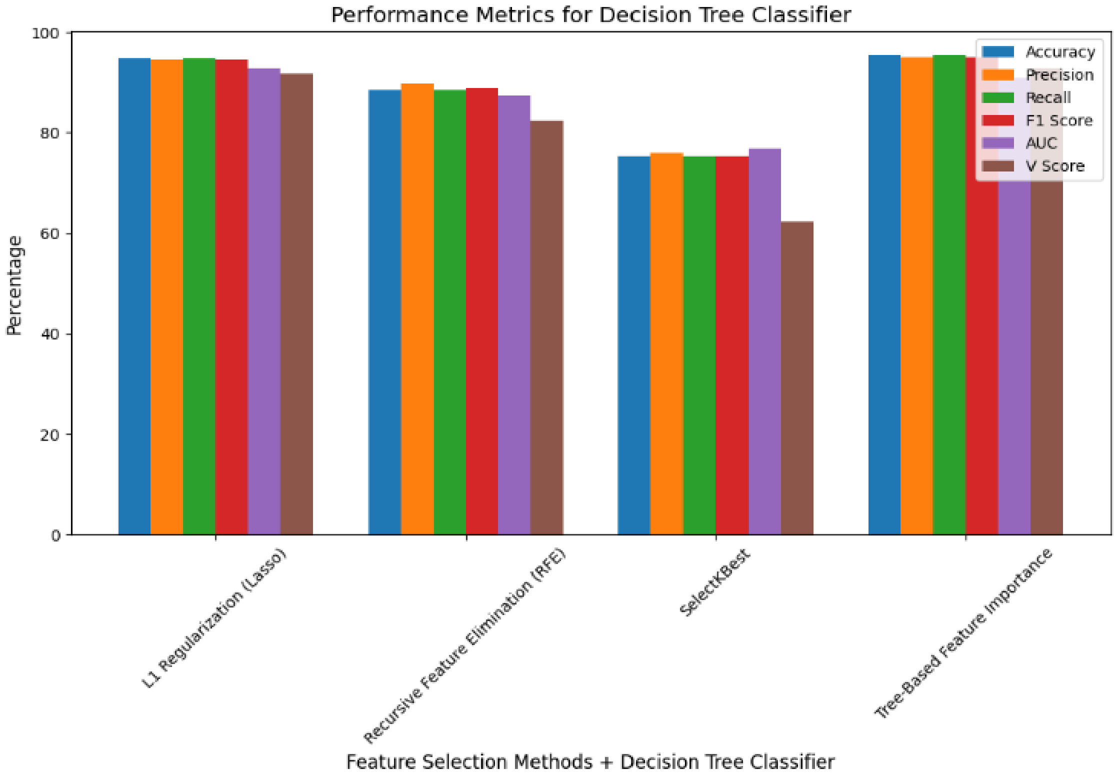


Fig. 5. Performance Evaluation of the DT Algorithm

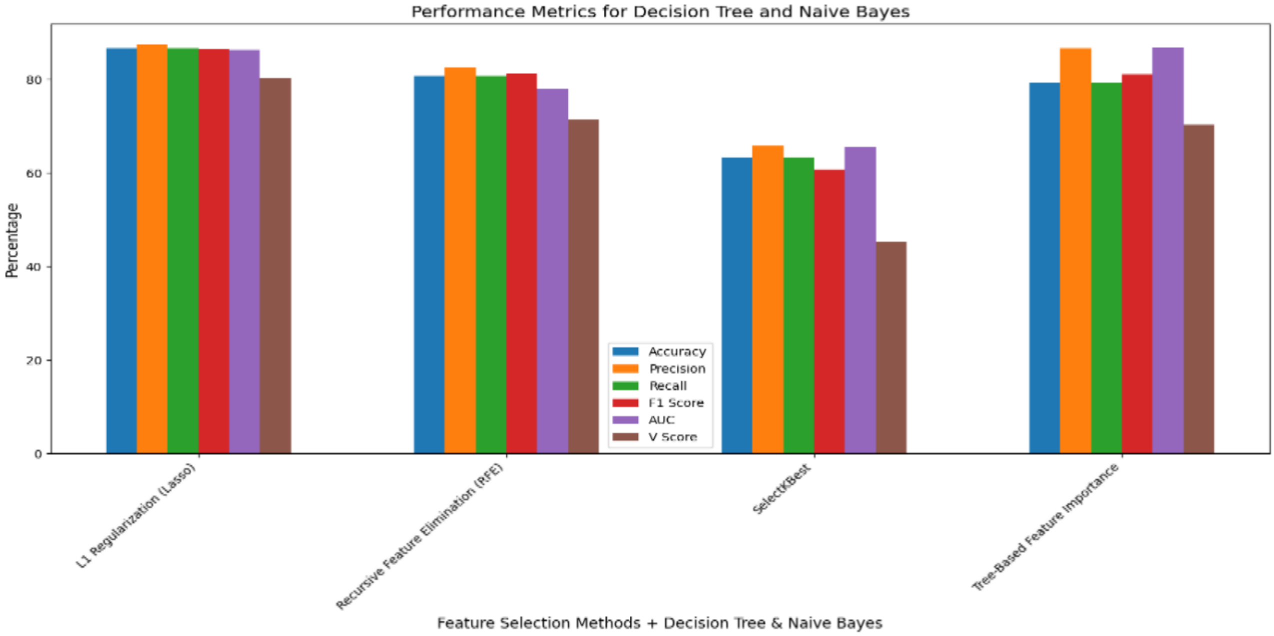


Fig. 6. Performance Evaluation of the DT & NB Algorithm

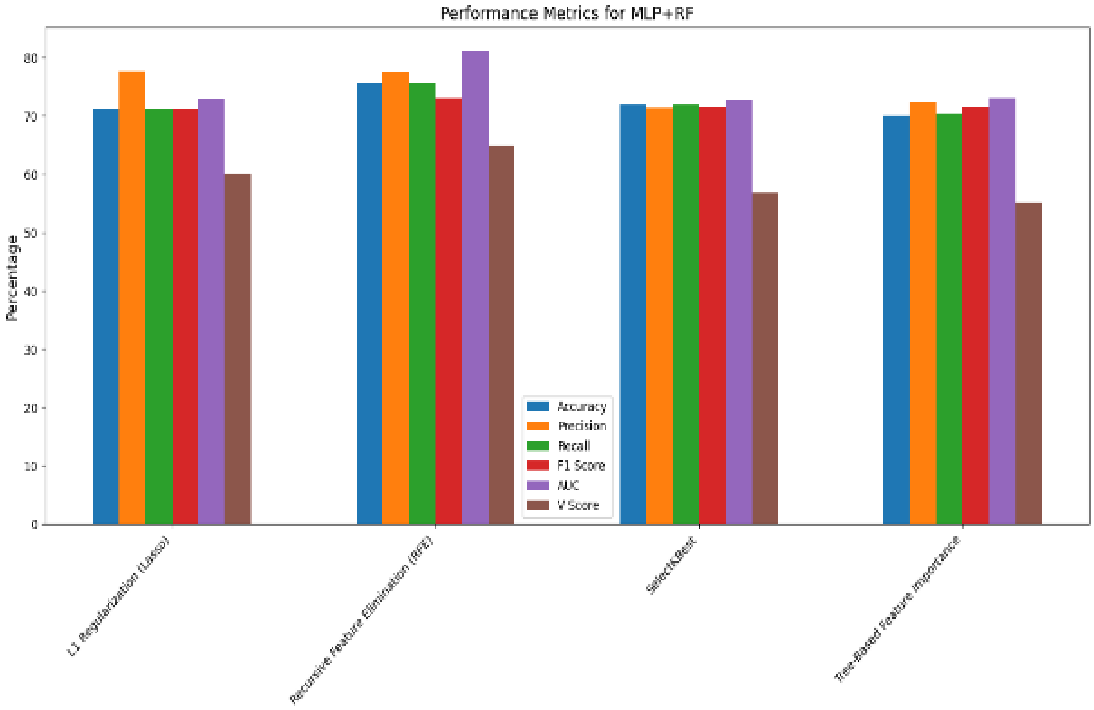


Fig. 7. Performance Evaluation of MLP+RF Algorithm

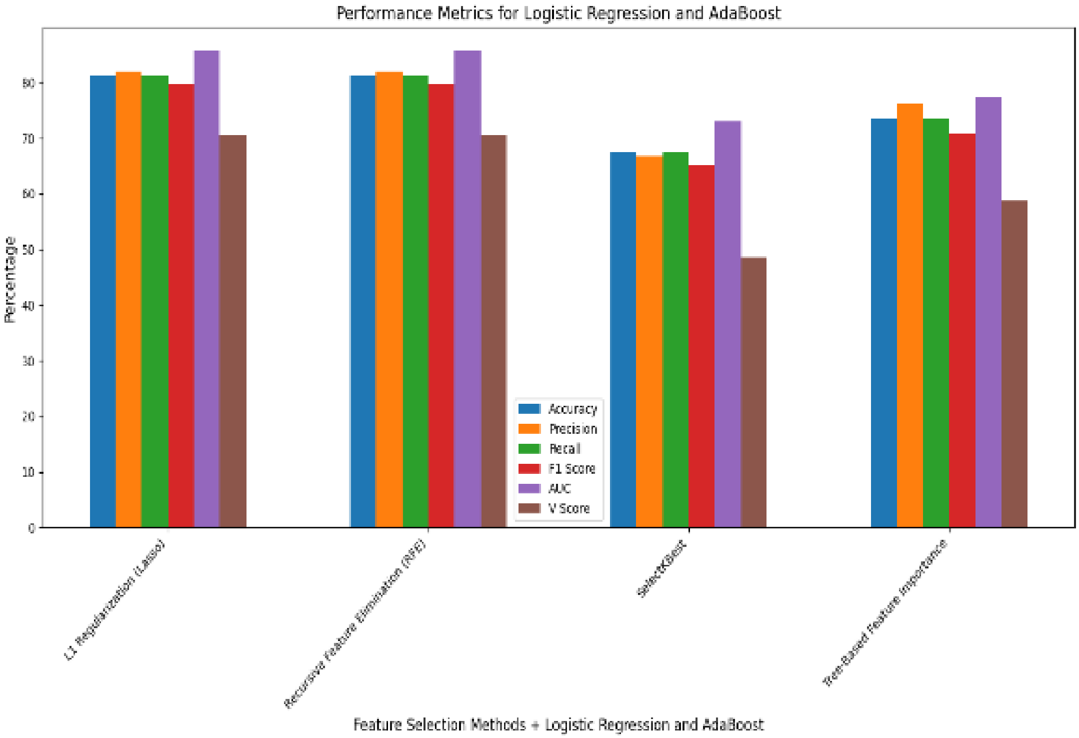


Fig. 8. Performance Evaluation of the LR & AB Algorithm

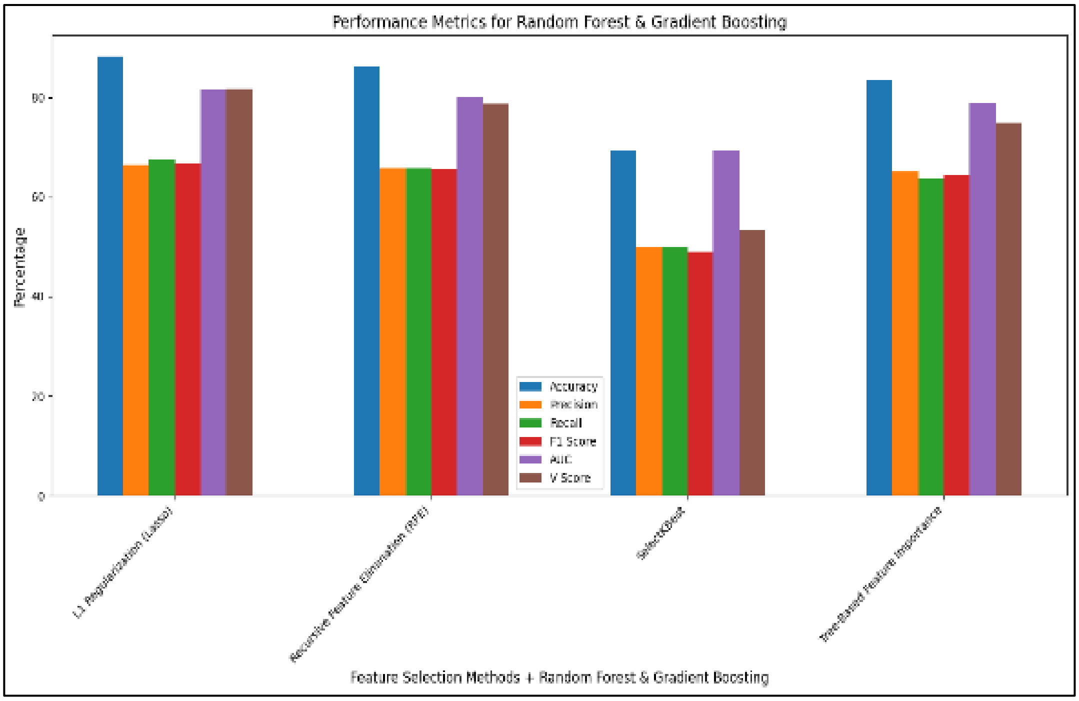


Fig. 9. Performance Evaluation of RF & GB Algorithm

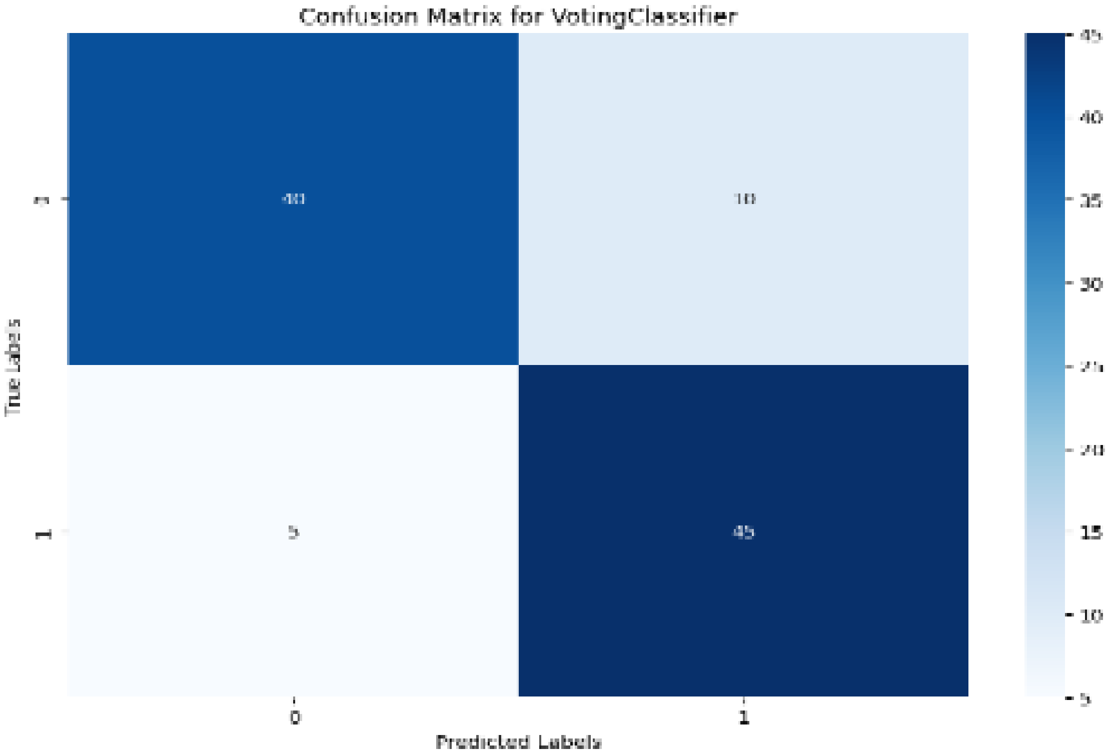


Fig. 10. Confusion matrix of DT Algorithm

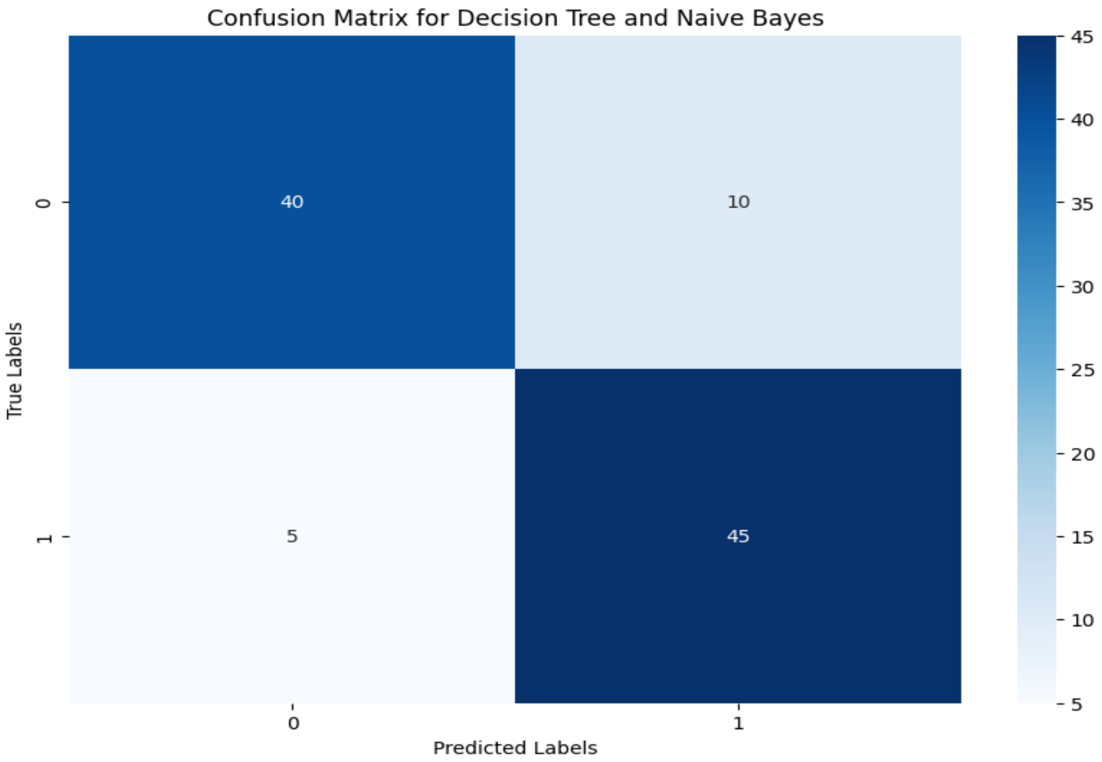


Fig. 11. Confusion matrix of DT & NB Algorithm

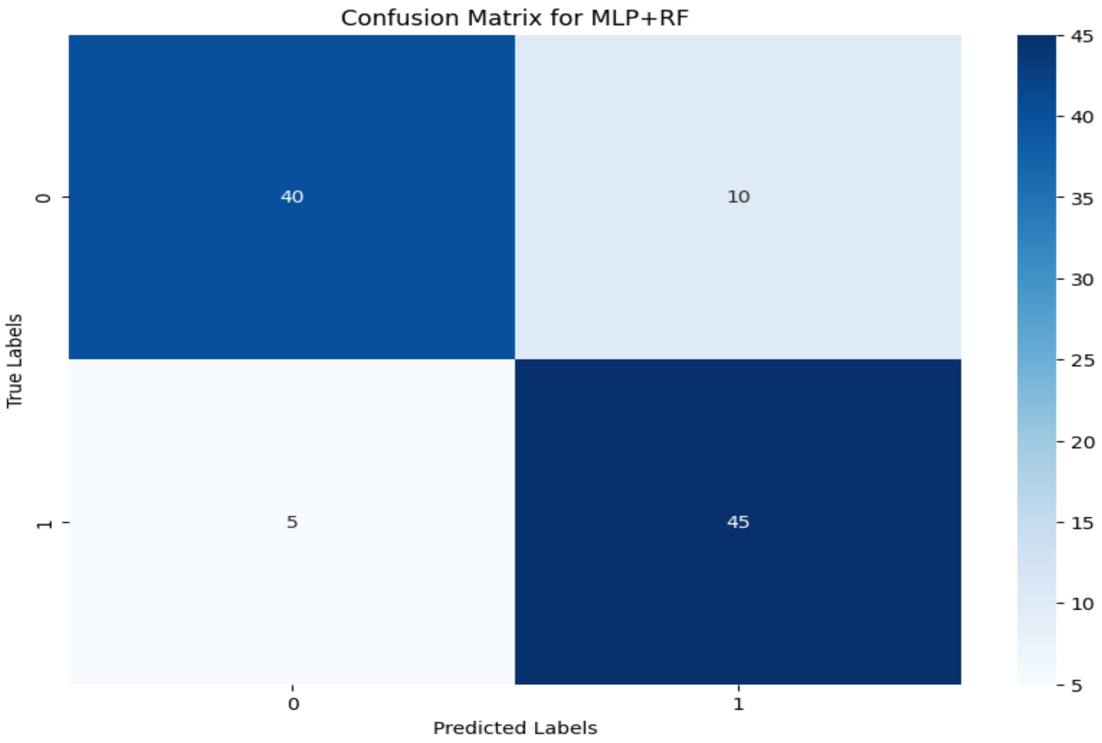


Fig. 12. Confusion matrix of MLP+RF Algorithm

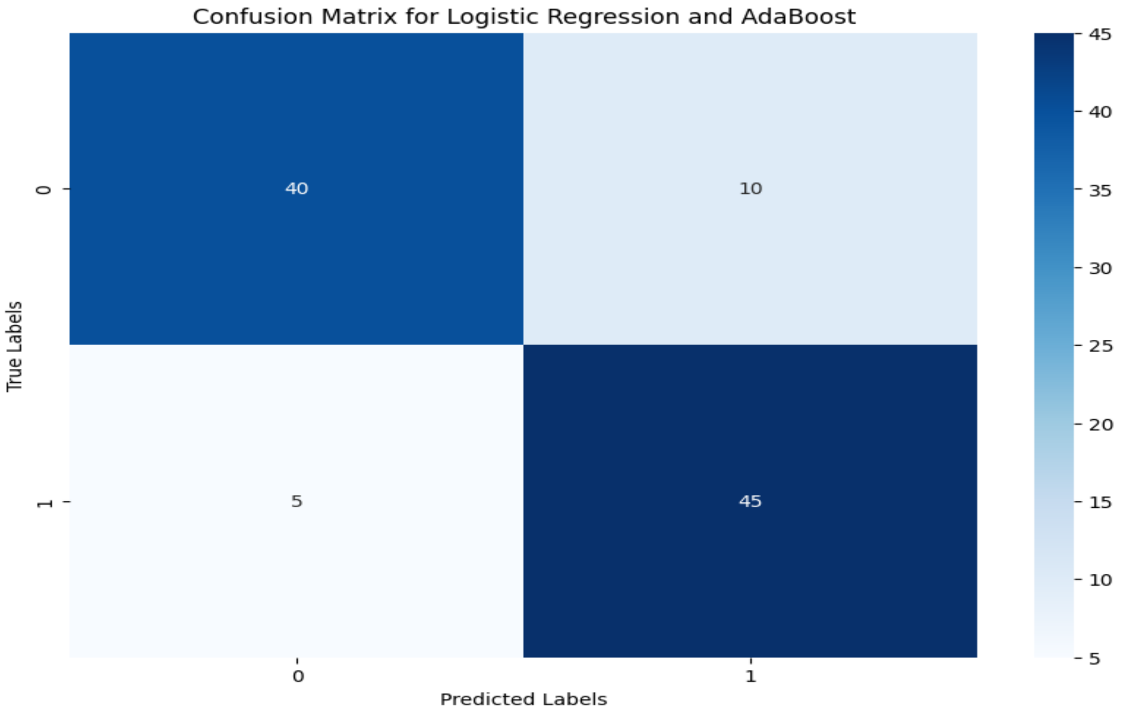


Fig. 13. Confusion matrix of the LR & AB Algorithm

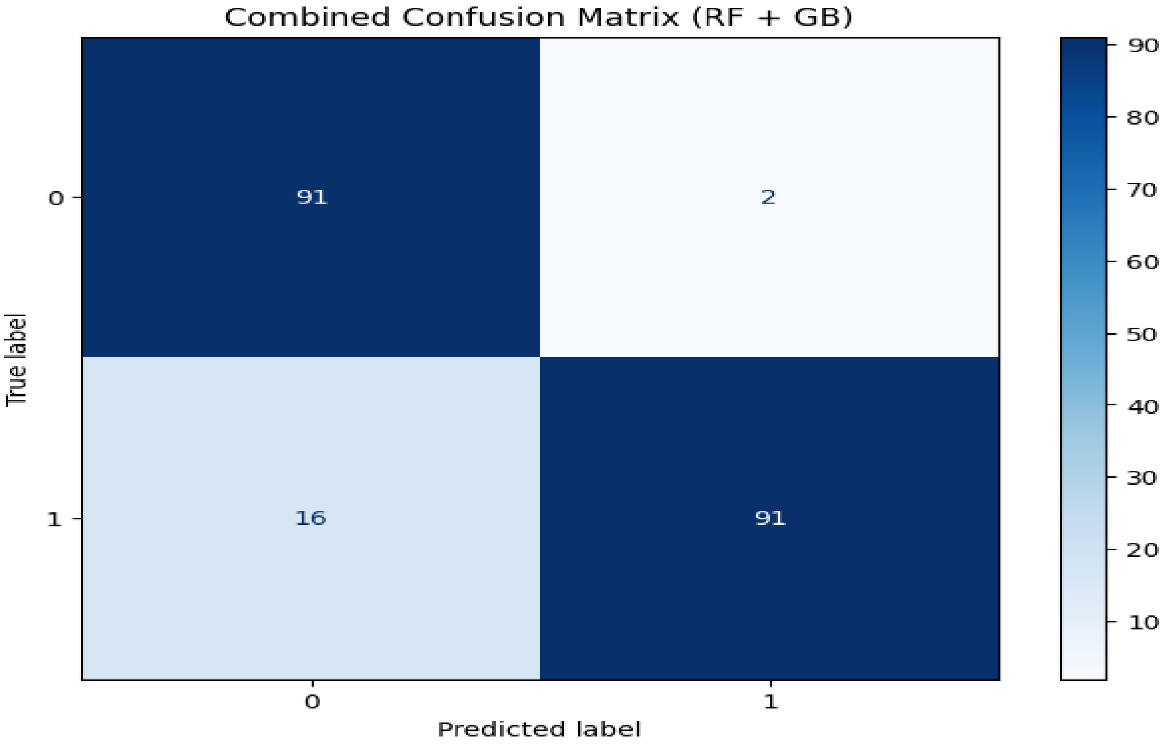


Fig. 14. Confusion matrix of RF & GB Algorithm

5. DISCUSSION

In this study, we explored the application of various machine learning techniques to classify osteosarcoma cases, leveraging a combination of feature selection methods and classification algorithms. The goal was to identify the most effective approach for accurately categorizing osteosarcoma into different groups, thereby aiding in better clinical decision-making. The study's findings highlight the potential of advanced machine learning models in enhancing diagnostic accuracy and treatment planning for osteosarcoma patients.

One of the key advantages of this study is the use of ensemble methods, particularly the Voting Classifier, which combines the predictions of multiple classifiers to achieve higher accuracy and robustness. Ensemble methods are known for their ability to mitigate overfitting and improve model generalization by aggregating the strengths of individual classifiers [45]. Specifically, the integration of Tree-Based Feature Importance with the Voting Classifier led to the best overall performance, demonstrating the effectiveness of this approach in handling complex medical datasets [46].

Additionally, the application of feature selection techniques such as L1 Regularization (Lasso) and Tree-Based Feature Importance was instrumental in reducing the dimensionality of the dataset. This reduction in complexity not only enhances model interpretability but also improves computational efficiency, which is crucial in medical research where datasets can be large and intricate [47]. The ability to identify and focus on the most relevant features helps in uncovering critical patterns and relationships that might otherwise be obscured by noise and irrelevant data [48].

Despite these advantages, the study also has several limitations. One significant limitation is the issue of class imbalance within the osteosarcoma dataset. Class imbalance can lead to biased model performance, where the classifier tends to favor the majority class. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) could be considered in future research to address this challenge and ensure more balanced model training.

Another limitation is the reliance on predefined feature selection methods, which may not capture all relevant features in the dataset. Future studies could explore more advanced or hybrid feature selection techniques to further enhance model performance. Additionally, while the Voting Classifier demonstrated high accuracy, it also requires considerable computational resources and time for training, which may limit its practicality in some clinical settings [48].

6. CONCLUSION

This study demonstrated the effectiveness of advanced machine learning techniques in classifying osteosarcoma cases. By employing robust feature selection methods and ensemble classifiers, notably the combination of Tree-Based Feature Importance and Voting Classifier, we achieved significant improvements in diagnostic accuracy and model robustness. Despite some limitations, such as class imbalance and computational demands, the findings highlight the potential of these approaches to enhance osteosarcoma diagnosis and treatment planning. Future research should address these limitations to further optimize the application of these models in clinical practice, ultimately contributing to better patient outcomes.

Conflicts of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Funding

The author's paper clearly indicates that the research was conducted without any funding from external sources.

Acknowledgment

The author extends appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] H. C. Beird *et al.*, ‘Osteosarcoma’, *Nature Reviews Disease Primers*, vol. 8, no. 1, p. 77, 2022.
[2] R. L. Siegel, K. D. Miller, N. S. Wagle, A. Jemal, and others, ‘Cancer statistics, 2023’, *Ca Cancer J Clin*, vol. 73, no. 1, pp. 17–48, 2023.
[3] H. Williams and A. Davies, ‘The effect of X-rays on bone: a pictorial review’, *European radiology*, vol. 16, pp. 619–633, 2006.
[4] H. Chen, M. M. Rogalski, and J. N. Anker, ‘Advances in functional X-ray imaging techniques and contrast agents’, *Physical Chemistry Chemical Physics*, vol. 14, no. 39, pp. 13469–13486, 2012.

[5] T. T. Miller, ‘Bone tumors and tumorlike conditions: analysis with conventional radiography’, *Radiology*, vol. 246, no. 3, pp. 662–674, 2008.

[6] Z. S. Kundu, ‘Classification, imaging, biopsy and staging of osteosarcoma’, *Indian journal of orthopaedics*, vol. 48, no. 3, pp. 238–246, 2014.

[7] V. Aran *et al.*, ‘Osteosarcoma, chondrosarcoma, and Ewing sarcoma: Clinical aspects, biomarker discovery and liquid biopsy’, *Critical Reviews in Oncology/Hematology*, vol. 162, p. 103340, 2021.

[8] H. B. Arunachalam *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.

[9] K. T. Schmidt, C. H. Chau, D. K. Price, and W. D. Figg, ‘Precision oncology medicine: the clinical relevance of patient- specific biomarkers used to optimize cancer treatment’, *The Journal of Clinical Pharmacology*, vol. 56, no. 12, pp. 1484–1499, 2016.

[10] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, ‘Histopathological image analysis: A review’, *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[11] M. S. Kashaf and E. McGill, ‘Does shared decision making in cancer treatment improve quality of life? Systematic literature review’, *Medical decision making*, vol. 35, no. 8, pp. 1037–1048, 2015.

[12] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Consortium, ‘Explainability for artificial intelligence in healthcare: a multidisciplinary perspective’, *BMC medical informatics and decision making*, vol. 20, pp. 1–9, 2020.

[13] S. L. Goldenberg, G. Nir, and S. E. Salcudean, ‘A new era: artificial intelligence and machine learning in prostate cancer’, *Nature Reviews Urology*, vol. 16, no. 7, pp. 391–403, 2019.

[14] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, ‘A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction’, *Journal of Applied Science and Technology Trends*, vol. 1, no. 1, pp. 56–70, 2020.

[15] G. Seni and J. Elder, *Ensemble methods in data mining: improving accuracy through combining predictions*. Morgan & Claypool Publishers, 2010.

[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, p. 100153, 2023.

[17] M. M. Ahsan, S. A. Luna, and Z. Siddique, ‘Machine-learning-based disease diagnosis: A comprehensive review’, in *Healthcare*, MDPI, 2022, p. 541.

[18] H. B. Arunachalam *et al.*, ‘Computer aided image segmentation and classification for viable and non-viable tumor identification in osteosarcoma’, in *Pacific Symposium on Biocomputing 2017*, World Scientific, 2017, pp. 195–206.

[19] M. T. Aziz *et al.*, ‘A Novel Hybrid Approach for Classifying Osteosarcoma Using Deep Feature Extraction and Multilayer Perceptron’, *Diagnostics*, vol. 13, no. 12, p. 2106, 2023.

[20] 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.

[21] X. Zhou *et al.*, ‘Emerging applications of deep learning in bone tumors: current advances and challenges’, *Frontiers in Oncology*, vol. 12, p. 908873, 2022.

[22] T. Al-Quraishi, N. Al-Quraishi, H. AlNabulsi, H. AL-Qarishey, and A. H. Ali, ‘Big Data Predictive Analytics for Personalized Medicine: Perspectives and Challenges’, *Applied Data Science and Analysis*, pp. 32–38, 2024, doi: 10.58496/ADSA/2024/004.

[23] P. Leavey, A. Sengupta, D. Rakheja, O. Daescu, H. Arunachalam, and R. Mishra, ‘Osteosarcoma data from ut southwestern/UT Dallas for viable and necrotic tumor assessment [data set]’, *Cancer Imaging Arch*, vol. 14, 2019.

[24] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, ‘Calibrating probability with undersampling for unbalanced classification’, in *2015 IEEE symposium series on computational intelligence*, IEEE, 2015, pp. 159–166.

[25] H. He and E. A. Garcia, ‘Learning from imbalanced data’, *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[26] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, ‘Optimal number of features as a function of sample size for various classification rules’, *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.

[27] A. Gyasi-Agyei, T. Al-Quraishi, B. Das, and J. I. Agbinya, ‘Exploratory Analysis and Preprocessing of Dataset for the Classification of Osteosarcoma Types’, in *Proceedings of International Conference for ICT (ICICT)-Zambia*, 2023, pp. 36–43.

[28] A. H. Farooqi, S. Akhtar, H. Rahman, T. Sadiq, and W. Abbass, ‘Enhancing network intrusion detection using an ensemble voting classifier for internet of things’, *Sensors*, vol. 24, no. 1, p. 127, 2023.

[29] S. Alelyani, ‘Stable bagging feature selection on medical data’, *Journal of Big Data*, vol. 8, no. 1, p. 11, 2021.

[30] B. Charbuty and A. Abdulazeez, ‘Classification based on decision tree algorithm for machine learning’, *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[31] C. ann" Ratanamahatana and D. Gunopulos, ‘Feature selection for the naive bayesian classifier using decision trees’, *Applied artificial intelligence*, vol. 17, no. 5–6, pp. 475–487, 2003.

[32] B. Bai, Z. Wu, S. Weng, and Q. Yang, ‘Application of interpretable machine learning algorithms to predict distant metastasis in osteosarcoma’, *Cancer Medicine*, vol. 12, no. 4, pp. 5025–5034, 2023.

[33] S. Ghimire *et al.*, ‘Hybrid convolutional neural network-multilayer perceptron model for solar radiation prediction’, *Cognitive Computation*, vol. 15, no. 2, pp. 645–671, 2023.

[34] M. Fratello, R. Tagliaferri, and others, ‘Decision trees and random forests’, *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1, no. S 3, 2018.

[35] M. A. A. Walid *et al.*, ‘Adapted Deep Ensemble Learning-Based Voting Classifier for Osteosarcoma Cancer Classification’, *Diagnostics*, vol. 13, no. 19, p. 3155, 2023.

[36] Y. Lou, R. Caruana, and J. Gehrke, ‘Intelligible models for classification and regression’, in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 150–158.

[37] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, ‘BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification’, *Engineering Applications of Artificial Intelligence*, vol. 49, pp. 176–193, 2016.

[38] H. A. A. Rahman, Y. B. Wah, H. He, and A. Bulgiba, ‘Comparisons of ADABOOST, KNN, SVM and logistic regression in classification of imbalanced dataset’, in *Soft Computing in Data Science: First International Conference, SCDS 2015, Putrajaya, Malaysia, September 2-3, 2015, Proceedings 1*, Springer, 2015, pp. 54–64.

[39] T. Kavzoglu and A. Teke, ‘Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NG Boost)’, *Arabian Journal for Science and Engineering*, vol. 47, no. 6, pp. 7367–7385, 2022.

[40] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, ‘The evolution of boosting algorithms’, *Methods of information in medicine*, vol. 53, no. 06, pp. 419–427, 2014.

[41] Md. A. Parwez and Md. Abulaish, ‘Text Classification Based on Convolutional Neural Networks and Word Embedding for Low-Resource Languages: Tigrinya’, *Information*, vol. 12, no. 2, p. 52, 2019, doi: 10.3390/info12020052.

[42] J. A. Hanley and B. J. McNeil, ‘The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve’, *Radiology*, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.

[43] A. P. Bradley, ‘The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms’, *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

[44] B. Li, Y. Li, W. Wei, and Z. He, ‘A Comprehensive Evaluation Framework for Deep Model Robustness’, *arXiv*, vol. 2101.09617, 2021, [Online]. Available: https://arxiv.org/abs/2101.09617

[45] T. Sadiq and W. Abbass, ‘An Ensemble-Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying’, *MAKE*, vol. 12, no. 1, p. 127, 2024, doi: 10.3390/make12010127.

[46] T. Sadiq and W. Abbass, ‘An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI’, *Sensors*, vol. 24, no. 1, p. 127, 2024, doi: 10.3390/s24010127.

[47] E. Team, ‘Feature Selection (Intrinsic Methods) - An Introductory Guide to Data Science and Machine Learning’, *Educative.io*, 2023, [Online]. Available: https://www.educative.io/courses/intro-data-science-machine-learning/feature-selection-intrinsic-methods

[48] M. Cueto-López, G. M. Foody, and M. Pal, ‘Selecting critical features for data classification based on machine learning methods’, *Journal of Big Data*, vol. 6, no. 1, p. 52, 2019, doi: 10.1186/s40537-019-0190-4.