Schrödinger:Journal of Physics Education
Vol. 7, No. 3, June 2026, pp. 138~157
ISSN: 2716-1587, DOI: 10.37251/sjpe.v7i3.3385

Machine Learning-Based Classification of Bone Tumor Severity: A Comparative Study of Classical Algorithms

Listia Rizka Triaswati[1], Dang Thi Thanh Thuy[2]
[1]Department of Physics, Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta, Jakarta, Indonesia
[2]Faculty of Physics, VNU University of Science, Hanoi, Vietnam

Article Info

Article history:

Received Mar 30, 2026
Revised Apr 27, 2026
Accepted Jun 9, 2026
Online First Jun 13, 2026

Keywords:

Bone Tumor
Clinical Decision Support System
Logistic Regression
Machine Learning Algorithms
Orthopedic Oncology

ABSTRACT

Purpose of the study: This study aims to evaluate and compare the performance of four machine learning algorithms Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest for bone tumor grade classification using structured clinical data and to identify the most effective algorithm for supporting diagnostic decision-making in orthopedic oncology.

Methodology: An experimental quantitative research design was employed using a publicly available Bone Tumor dataset from Kaggle containing 500 records. Model development was conducted in Google Colaboratory using Python and Scikit-learn. The evaluated algorithms included Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest. Data preprocessing, feature selection, and train-test splitting (80:20) were performed. Model performance was assessed using accuracy, precision, recall, and F1-score metrics.

Main Findings: The results demonstrated that all machine learning models were capable of classifying bone tumor grades with satisfactory performance. Logistic Regression achieved the best overall performance, obtaining 81% accuracy, precision values of 79–82%, recall values of 73–86%, and F1-scores of 76–84%. Decision Tree and Naïve Bayes showed moderate performance, while Random Forest exhibited reduced testing performance despite strong training results, indicating overfitting and lower generalization capability.

Novelty/Originality of this study: This study contributes a comprehensive comparison of classical machine learning algorithms for bone tumor grade classification using structured clinicopathological data rather than imaging data. The findings demonstrate that interpretable models such as Logistic Regression can achieve reliable predictive performance, providing an accessible and computationally efficient alternative for clinical decision-support systems in resource-limited healthcare.

Corresponding Author:
Listia Rizka Triaswati,
Department of Physics, Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta,
Ir. H. Juanda Street No. 95, Ciputat, South Tangerang, Banten, 15412, Indonesia
Email: listiarizaka@gmail.com

1. INTRODUCTION

Bone tumors represent a relatively uncommon group of musculoskeletal neoplasms; however, they remain clinically important because malignant forms often result in substantial morbidity, functional impairment, and premature mortality among affected patients. Osteosarcoma, chondrosarcoma, and Ewing sarcoma are

among the most frequently reported malignant bone tumors, particularly affecting children, adolescents, and young adults during critical stages of physical development [1]-[3]. Recent epidemiological evidence indicates that primary bone tumors account for only a small proportion of all malignancies, yet they contribute disproportionately to cancer-related disability and mortality in younger populations [4]-[6]. The rarity and biological heterogeneity of bone tumors frequently complicate diagnosis, treatment planning, and long-term prognosis assessment, thereby increasing the demand for more reliable and efficient diagnostic approaches [7]-[9]. Consequently, improving diagnostic accuracy and facilitating earlier identification of malignant bone tumors remain important priorities in contemporary orthopedic oncology research.

Accurate diagnosis of bone tumors remains challenging because clinical manifestations, radiological appearances, and histopathological characteristics frequently overlap between benign and malignant lesions. Physicians often rely on multiple diagnostic modalities, including radiography, computed tomography, magnetic resonance imaging, and biopsy examinations, to establish an appropriate diagnosis and determine treatment strategies [10]-[12]. Several studies have reported that substantial inter-observer variability may occur during the interpretation of bone tumor characteristics, particularly when clinical experience is limited or lesion morphology is highly heterogeneous [13]-[15]. These diagnostic complexities may contribute to delayed treatment initiation, inappropriate therapeutic decisions, and unfavorable patient outcomes when malignant lesions are not identified promptly [16]-[18]. Therefore, advanced computational techniques capable of assisting clinicians in analyzing complex clinical information are increasingly being explored to support diagnostic decision-making processes.

The rapid growth of artificial intelligence technologies has transformed many areas of healthcare by enabling automated analysis of large-scale clinical and biomedical datasets. Among these technologies, machine learning has demonstrated significant potential for extracting hidden patterns from complex data and generating predictive models that support medical decision-making [19]-[21]. Recent reviews have highlighted the increasing adoption of machine learning methods in cancer diagnosis, prognosis prediction, treatment planning, and personalized medicine applications [22]-[24]. Compared with traditional statistical approaches, machine learning algorithms can handle nonlinear relationships, high-dimensional datasets, and complex interactions among clinical variables more effectively [25]-[27]. Consequently, machine learning has become one of the most promising computational approaches for improving diagnostic performance across diverse oncology domains.

Several studies have investigated the application of artificial intelligence for bone tumor detection and classification using radiographic, computed tomography, and magnetic resonance imaging data. Deep learning architectures, particularly convolutional neural networks and DenseNet-based models, have demonstrated promising diagnostic performance and, in some cases, achieved accuracy comparable to or exceeding that of human experts [28]-[30]. Meta-analytic evidence further suggests that machine learning techniques possess considerable diagnostic value for differentiating malignant bone tumors from other musculoskeletal conditions [31]-[33]. Nevertheless, the majority of published studies primarily emphasize image-based deep learning approaches rather than structured clinical datasets commonly encountered in routine healthcare settings [7], [34], [35].

Although previous investigations have reported encouraging results from artificial intelligence applications in bone tumor diagnosis, several important limitations remain unresolved. First, most studies focus on deep learning models requiring large imaging datasets, extensive computational resources, and complex model architectures that may not be readily applicable in resource-limited healthcare environments [36], [37], [38]. Second, comparative analyses involving classical machine learning algorithms operating on structured clinical attributes remain relatively scarce within the bone tumor research domain [4]. Furthermore, limited evidence is available regarding which conventional machine learning algorithm provides the most balanced predictive performance for bone tumor classification when evaluated using multiple performance metrics simultaneously. This knowledge gap indicates the necessity for additional studies examining simpler yet effective machine learning techniques using clinically relevant datasets.

To address these limitations, the present study evaluates and compares four widely used machine learning algorithms, namely Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest, for bone tumor classification. Unlike many previous studies that emphasize image-based deep learning frameworks, this research utilizes structured clinical information derived from publicly available bone tumor datasets. The selected algorithms represent different learning paradigms, allowing a comprehensive assessment of probabilistic, statistical, tree-based, and ensemble approaches within a single experimental framework. Additionally, model performance is examined using multiple evaluation metrics, including accuracy, precision, recall, and F1-score, to provide a more robust comparison of classification effectiveness. The findings are expected to contribute practical insights regarding the applicability of classical machine learning methods in clinical decision-support systems for orthopedic oncology.

Based on the identified research gap, this study aims to determine the most effective machine learning algorithm for classifying bone tumor cases using structured clinical data. The investigation specifically compares the predictive performance of Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest models

under identical experimental conditions. Evaluation results are analyzed using multiple classification metrics to ensure a comprehensive assessment of model reliability and predictive capability. Through this comparative approach, the study seeks to provide empirical evidence supporting the development of efficient and accessible machine learning-based diagnostic tools for bone tumor management. Ultimately, the proposed framework is expected to assist future research efforts and contribute to improving clinical decision-making processes in orthopedic oncology.

2. RESEARCH METHOD

2.1. Research Design

This study is classified as experimental research employing a quantitative approach [39]. The objective of this research is to develop machine learning models using the Naive Bayes, Logistic Regression, Random Forest, and Decision Tree algorithms for predicting bone tumor data, where these algorithms are widely used in classification problems due to their effectiveness and interpretability [40], [41].

2.2. Research Instruments

The instruments and tools utilized in this study consisted of both hardware and software components to support the development and evaluation of the Machine Learning models. The hardware used was a Lenovo IdeaPad C340 laptop equipped with a Windows 11 operating system, an AMD Ryzen 5 3500U processor, 8 GB of installed RAM, and a 64-bit operating system with an x64-based processor architecture. For model development and implementation, Google Colaboratory was employed as the primary cloud-based programming environment, enabling the execution of Python code and Machine Learning algorithms efficiently.

The dataset used in this study was obtained from Kaggle.com and consisted of 500 records related to bone tumor cases. Five predictor variables were selected for model development, namely Sex, Grade, Histological Type, Site of Primary STS, and Treatment. To evaluate the performance of the proposed models, the dataset was divided into training and testing sets using two different data partitioning schemes. The first scheme allocated 80% of the data for training and 20% for testing, while the second scheme used 90% of the data for training and 10% for testing. These data splitting strategies were implemented to compare model performance under different training and testing proportions.

2.3. Research Stages

The research was conducted through several sequential stages, as illustrated in the research flowchart.

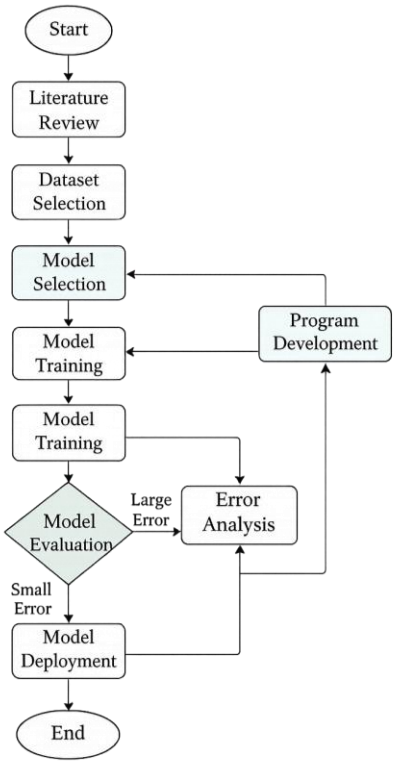

Figure 1. Research Stages Flowchart

2.4. Research Process

The research process generally consists of problem identification, literature review, program development, dataset collection, model training, and model evaluation.

Program Development

The program was developed using Python in Google Colaboratory, a cloud-based Jupyter Notebook environment provided by Google. The Machine Learning models were implemented using the Naive Bayes, Logistic Regression, Random Forest, and Decision Tree algorithms. The developed program was designed to generate performance evaluation metrics, including accuracy, precision, recall, and F1-score.

Dataset Collection

Kaggle provides numerous data science datasets that can be downloaded for research purposes. The dataset used in this study was provided by Prathap Kumar under the title "Bone Tumor." The dataset can be accessed through the following link: Bone Tumor G.

Program Testing

The data processing technique employed in this study utilizes Machine Learning methods. The process begins by downloading the dataset from the Kaggle website in Comma-Separated Values (CSV) format. The downloaded dataset is then stored in the same working directory and imported using the Pandas library through Python code executed in the development environment.

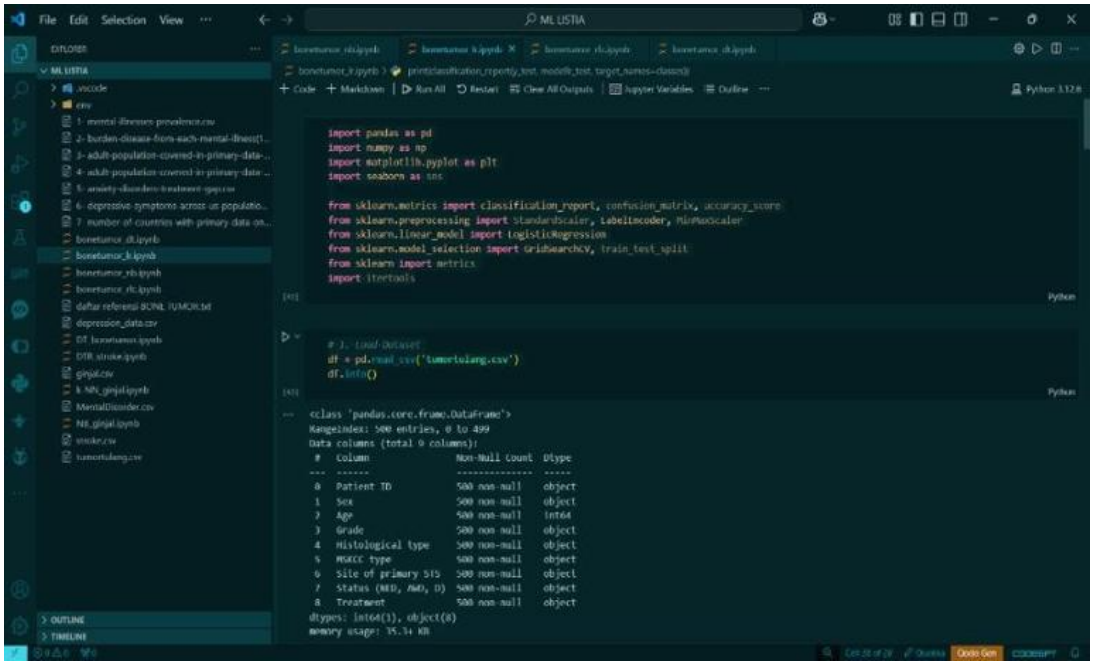

Figure 2. Input program

After loading the dataset from the CSV file, the next step was data preprocessing, in which the data were cleaned and examined to ensure their quality and suitability for analysis. Duplicate records were identified and removed using the df.duplicated().sum() function. Subsequently, feature selection was performed by defining X as the feature variables and y as the target variable. Following the separation of features and target data, the dataset was partitioned into training and testing sets using the sklearn.model_selection library, with 80% of the data allocated for training and 20% for testing. The Machine Learning models were then trained and evaluated on the prepared dataset. After model initialization and prediction, performance metrics including accuracy, precision, recall, and F1-score were calculated based on the confusion matrix. The overall workflow of the program development process is illustrated in Figure 3.

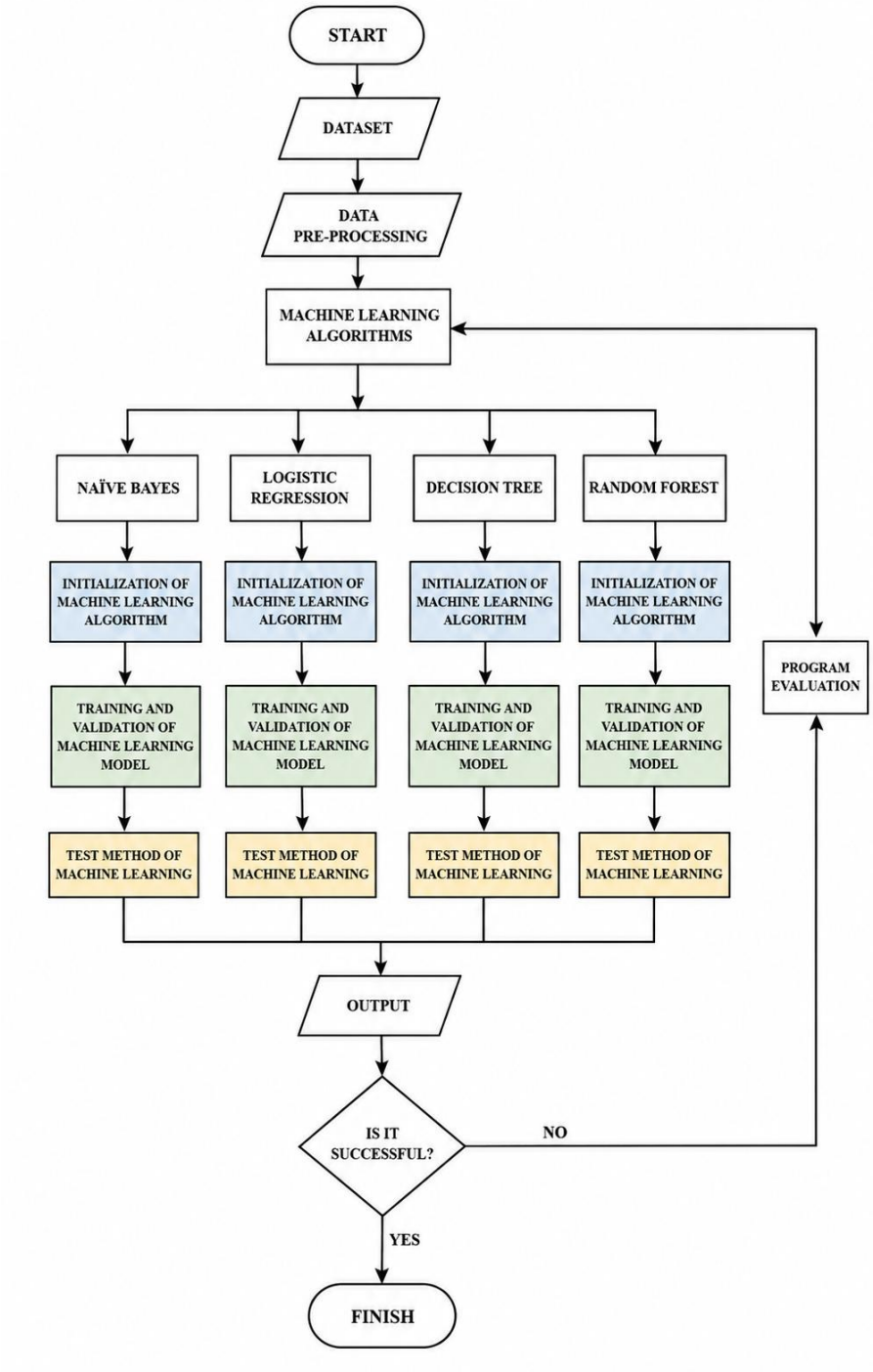

Figure 3. Program Testing Flowchart

3. RESULTS AND DISCUSSION

3.1. Characteristics of Machine Learning Model Accuracy Measurement

This study evaluated the performance characteristics of Machine Learning models using four classification algorithms: Naive Bayes, Logistic Regression, Random Forest, and Decision Tree. The dataset was divided using a data-splitting approach, with 80% allocated for training and 20% for testing. The performance of each algorithm was assessed using several evaluation metrics, including accuracy, precision, recall, and F1-score.

Accuracy represents the proportion of correctly classified instances, both positive and negative, relative to the total number of observations. A balanced classification performance is generally reflected by relatively similar numbers of false positives (FP) and false negatives (FN), indicating that the model does not favor one

class excessively over another. In this study, the final performance evaluation of each algorithm was based on the testing dataset.

3.1.1. Naive Bayes Accuracy

The accuracy results of the Naive Bayes classifier were obtained from both the training and testing datasets. The model achieved a training accuracy of 74%. For the testing dataset, the Naive Bayes algorithm for bone tumor grade classification correctly identified 48 samples as High Grade and 11 samples as Intermediate Grade, representing the true positive predictions for each category.

In contrast, the model produced 28 false positive predictions, where samples were classified as High Grade when they actually belonged to the Intermediate Grade category. Additionally, 13 false negative predictions were observed, where samples were classified as Intermediate Grade despite actually belonging to the High Grade category. The numbers of false positives and false negatives were relatively comparable, indicating a reasonably balanced classification performance.

Based on the testing dataset, the Naive Bayes classifier achieved an overall accuracy of 76%. The corresponding confusion matrix for the testing dataset is presented below.

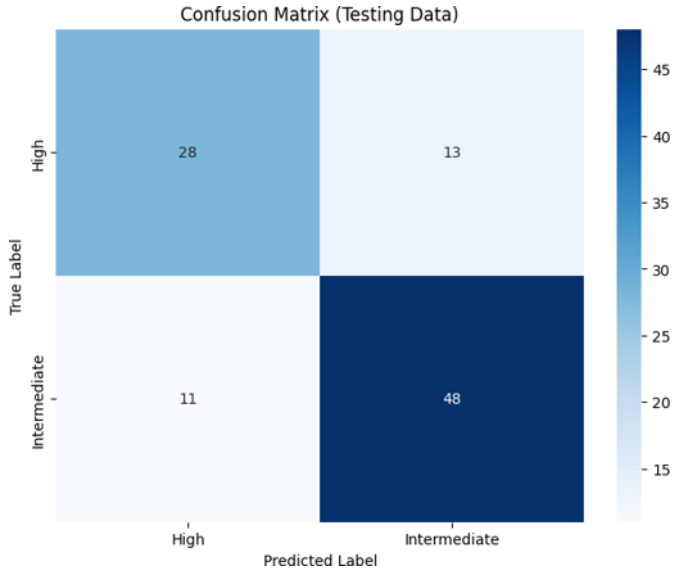

Figure 4. Confusion Matrix of the Naive Bayes Classifier

3.1.2 Logistic Regression Accuracy

The accuracy results of the Logistic Regression classifier were obtained from both the training and testing datasets. The model achieved a training accuracy of 78%, while the testing dataset yielded an accuracy of 81%. For bone tumor grade classification, the Logistic Regression algorithm correctly classified 171 samples as High Grade and 52 samples as Intermediate Grade, representing the true positive predictions for each category.

The model also generated 124 false positive predictions, where samples were classified as High Grade despite actually belonging to the Intermediate Grade category. In addition, 53 false negative predictions were observed, where samples were classified as Intermediate Grade when they actually belonged to the High Grade category. The difference between the numbers of false positives and false negatives indicates a relatively imbalanced classification performance, as the model tended to produce considerably more false positive predictions than false negative predictions.

Based on the testing dataset, the Logistic Regression classifier achieved an overall accuracy of 81%. The confusion matrix corresponding to the testing dataset is presented below.

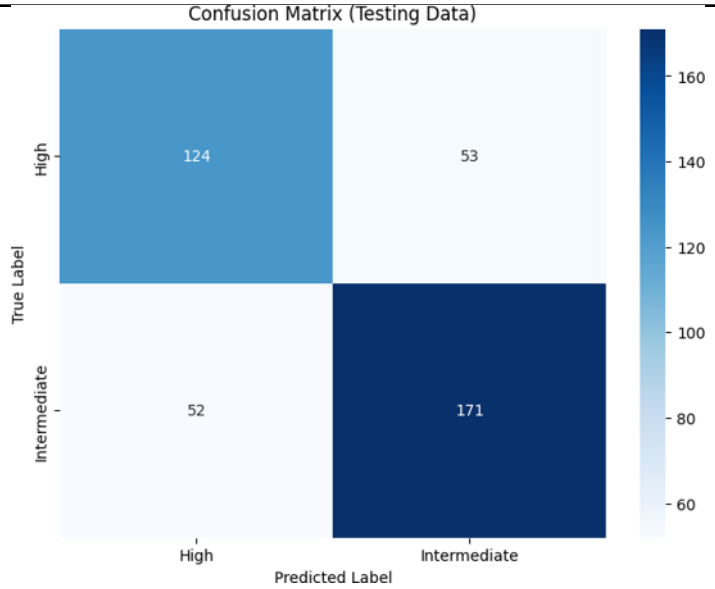

Figure 5. Confusion Matrix of the Logistic Regression Classifier

3.1.3 Decision Tree Accuracy

The accuracy results of the Decision Tree classifier were obtained from both the training and testing datasets. The model achieved a training accuracy of 80%. For the testing dataset, the Decision Tree algorithm for bone tumor grade classification correctly identified 171 samples as High Grade and 52 samples as Intermediate Grade, representing the true positive predictions for each category.

The model produced 124 false positive predictions, where samples were classified as High Grade when they actually belonged to the Intermediate Grade category. Additionally, 53 false negative predictions were recorded, where samples were classified as Intermediate Grade despite actually belonging to the High Grade category. The disparity between the numbers of false positives and false negatives indicates an imbalanced classification performance, as the false positive predictions substantially exceeded the false negative predictions.

Based on the testing dataset, the Decision Tree classifier achieved an overall accuracy of 77%. The corresponding confusion matrix for the testing dataset is presented below.

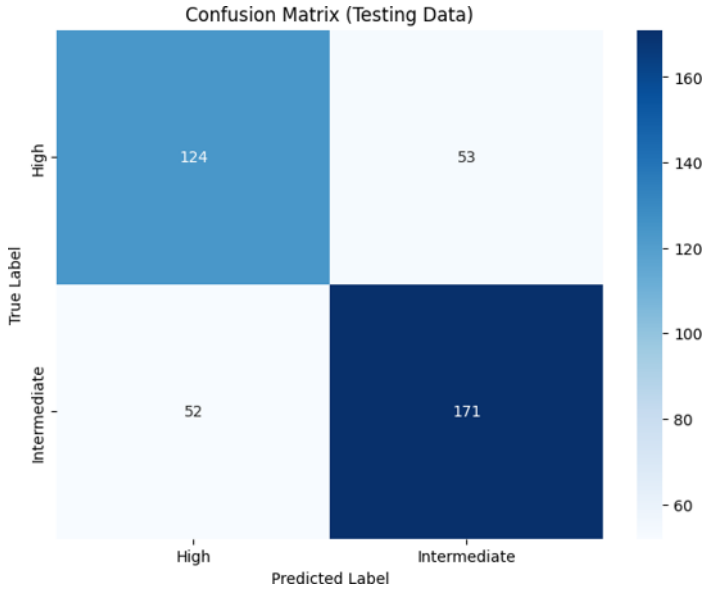

Figure 6. Confussion matrix decision tree

3.1.4 Random Forest Accuracy

The accuracy results of the Random Forest classifier were obtained from both the training and testing datasets. The model achieved a training accuracy of 87%, while the testing dataset yielded an accuracy of 71%. For bone tumor grade classification, the Random Forest algorithm correctly classified 26 samples as High Grade and 45 samples as Intermediate Grade, representing the true positive predictions for each category.

The model generated 11 false positive predictions, where samples were classified as High Grade when they actually belonged to the Intermediate Grade category. In addition, 18 false negative predictions were observed, where samples were classified as Intermediate Grade despite actually belonging to the High Grade category. The difference between the numbers of false positives and false negatives indicates a degree of imbalance in the classification results, although the disparity is less pronounced compared to some of the other models evaluated in this study.

Based on the testing dataset, the Random Forest classifier achieved an overall accuracy of 71%. The corresponding confusion matrix for the testing dataset is presented below.

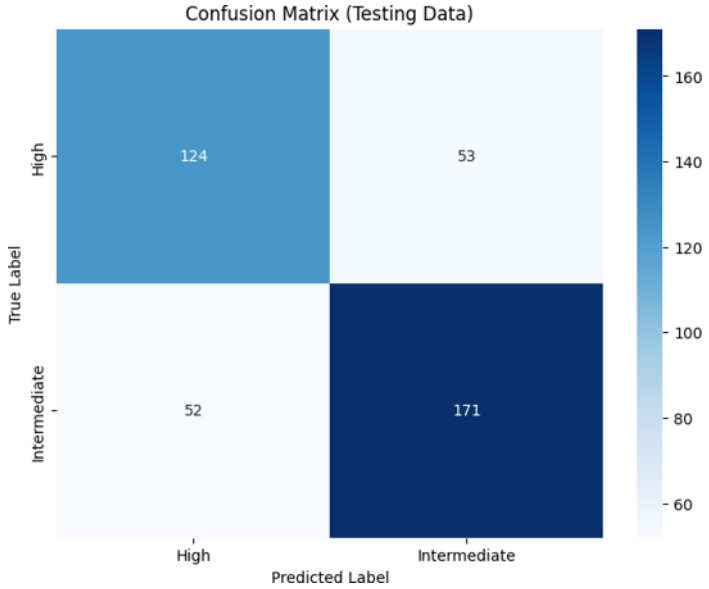

Figure 7. Confussion matrix random forest

3.2. Characteristics of Machine Learning Model Precision Measurement

In the context of bone tumor severity classification, precision represents the ratio between the number of correctly predicted positive cases and the total number of positive predictions generated by the model. This metric indicates the reliability of the model's predictions by measuring how accurately the predicted tumor severity levels correspond to the actual conditions.

3.2.1 Naive Bayes Precision

The precision results of the Naive Bayes classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a precision score of 0.70 for the Intermediate Grade class, indicating that 70% of the samples predicted as intermediate severity were correctly classified. For the High Grade class, the precision score reached 0.76, meaning that 76% of the samples predicted as high severity were correctly identified.

On the testing dataset, the precision score for the Intermediate Grade class was 0.72, indicating that 72% of the samples predicted as intermediate severity were correctly classified. Meanwhile, the High Grade class achieved a precision score of 0.79, suggesting that 79% of the samples predicted as high severity corresponded to the actual high-severity category. These results demonstrate that the Naive Bayes classifier showed slightly better precision in identifying high-grade bone tumors than intermediate-grade tumors.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.70	0.70	0.70	177
1	0.76	0.77	0.77	223
accuracy			0.74	400
macro avg	0.73	0.73	0.73	400
weighted avg	0.74	0.74	0.74	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.72	0.68	0.70	41
1	0.79	0.81	0.80	59
accuracy			0.76	100
macro avg	0.75	0.75	0.75	100
weighted avg	0.76	0.76	0.76	100

b

Figure 8. Precision Values of the Naive Bayes Classifier: (a) 80% Training Data and (b) 20% Testing Data

3.2.2 Logistic Regression Precision

The precision results of the Logistic Regression classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a precision score of 0.78 for the Intermediate Grade class, indicating that 78% of the samples predicted as intermediate severity were correctly classified. Similarly, the High Grade class obtained a precision score of 0.78, meaning that 78% of the samples predicted as high severity corresponded to the actual high-grade category.

On the testing dataset, the precision score for the Intermediate Grade class increased slightly to 0.79, indicating that 79% of the samples predicted as intermediate severity were correctly identified. For the High Grade class, the model achieved a precision score of 0.82, demonstrating that 82% of the samples predicted as high severity were correctly classified. These findings indicate that the Logistic Regression classifier exhibited strong predictive reliability, particularly in identifying high-grade bone tumor cases, as reflected by its higher precision score for the High Grade class.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.78	0.70	0.74	177
1	0.78	0.85	0.81	223
accuracy			0.78	400
macro avg	0.78	0.77	0.78	400
weighted avg	0.78	0.78	0.78	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.79	0.73	0.76	41
1	0.82	0.86	0.84	59
accuracy			0.81	100
macro avg	0.81	0.80	0.80	100
weighted avg	0.81	0.81	0.81	100

b

Figure 9. Precision Values of the Logistic Regression Classifier: (a) 80% Training Data and (b) 20% Testing Data

3.2.3 Decision Tree Precision

The precision results of the Decision Tree classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a precision score of 0.81 for the Intermediate Grade class, indicating that 81% of the samples predicted as intermediate severity were correctly classified. For the High Grade class, the precision score was 0.80, meaning that 80% of the samples predicted as high severity corresponded to the actual high-grade category.

On the testing dataset, the precision score for the Intermediate Grade class was 0.71, indicating that 71% of the samples predicted as intermediate severity were correctly identified. Meanwhile, the High Grade class achieved a precision score of 0.81, suggesting that 81% of the samples predicted as high severity were correctly classified. These results indicate that the Decision Tree classifier demonstrated higher precision in

identifying high-grade bone tumors than intermediate-grade tumors on the testing dataset, although its performance for the Intermediate Grade class decreased compared with the training dataset.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.81	0.72	0.76	177
1	0.80	0.86	0.83	223
accuracy			0.80	400
macro avg	0.80	0.79	0.79	400
weighted avg	0.80	0.80	0.80	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.71	0.73	0.72	41
1	0.81	0.80	0.80	59
accuracy			0.77	100
macro avg	0.76	0.76	0.76	100
weighted avg	0.77	0.77	0.77	100

b

Figure 10. Precision Values of the Decision Tree Classifier: (a) 80% Training Data and (b) 20% Testing Data

3.2.4 Random Forest Precision

The precision results of the Random Forest classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a precision score of 0.89 for the Intermediate Grade class, indicating that 89% of the samples predicted as intermediate severity were correctly classified. For the High Grade class, the precision score reached 0.85, meaning that 85% of the samples predicted as high severity corresponded to the actual high-grade category.

On the testing dataset, the precision score for the Intermediate Grade class was 0.68, indicating that 68% of the samples predicted as intermediate severity were correctly identified. Meanwhile, the High Grade class achieved a precision score of 0.76, suggesting that 76% of the samples predicted as high severity were correctly classified. Although the Random Forest classifier demonstrated strong precision performance on the training dataset, a decline in precision was observed on the testing dataset for both classes. This result may indicate reduced generalization performance when the model was applied to previously unseen data.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.89	0.80	0.84	177
1	0.85	0.92	0.89	223
accuracy			0.87	400
macro avg	0.87	0.86	0.86	400
weighted avg	0.87	0.87	0.87	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.68	0.63	0.66	41
1	0.76	0.80	0.78	59
accuracy			0.73	100
macro avg	0.72	0.72	0.72	100
weighted avg	0.73	0.73	0.73	100

b

Figure 11. Precision Values of the Random Forest Classifier: (a) 80% Training Data and (b) 20% Testing Data

3.3. Characteristics of Machine Learning Model Recall Measurement

In the context of bone tumor severity classification, recall measures the proportion of actual positive cases that are correctly identified by the model. This metric reflects the model's ability to detect patients who truly belong to a particular severity category. A higher recall value indicates that the model is more effective in identifying actual cases and reducing the number of false negatives.

3.3.1. Naive Bayes Recall

The recall results of the Naive Bayes classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a recall score of 0.70 for the Intermediate Grade class, indicating that 70% of the actual intermediate-grade cases were correctly identified by the model. For the High Grade class, the recall score was 0.76, meaning that 76% of the actual high-grade cases were successfully classified as high grade.

These results suggest that the Naive Bayes classifier demonstrated a slightly better ability to identify high-grade bone tumor cases than intermediate-grade cases in the training dataset. A higher recall value for the High Grade class indicates that the model was more effective in minimizing false negative predictions for severe tumor cases.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.70	0.70	0.70	177
1	0.76	0.77	0.77	223
accuracy			0.74	400
macro avg	0.73	0.73	0.73	400
weighted avg	0.74	0.74	0.74	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.72	0.68	0.70	41
1	0.79	0.81	0.80	59
accuracy			0.76	100
macro avg	0.75	0.75	0.75	100
weighted avg	0.76	0.76	0.76	100

b

Figure 12. Recall Values of the Naive Bayes Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Naive Bayes classifier achieved a recall score of 0.68 for the Intermediate Grade class, indicating that 68% of the actual intermediate-grade cases were correctly identified by the model. Meanwhile, the High Grade class obtained a recall score of 0.81, meaning that 81% of the actual high-grade cases were successfully classified as high grade. These findings indicate that the model demonstrated a stronger ability to detect high-grade tumor cases than intermediate-grade cases in the testing dataset.

3.3.2. Logistic Regression Recall

The recall results of the Logistic Regression classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a recall score of 0.70 for the Intermediate Grade class, indicating that 70% of the actual intermediate-grade cases were correctly identified. For the High Grade class, the recall score reached 0.85, meaning that 85% of the actual high-grade cases were successfully classified by the model.

These results suggest that the Logistic Regression classifier demonstrated a stronger capability in detecting high-grade bone tumor cases than intermediate-grade cases during the training phase. The higher recall value for the High Grade class indicates that the model was more effective in minimizing false negative predictions for severe tumor cases.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.78	0.70	0.74	177
1	0.78	0.85	0.81	223
accuracy			0.78	400
macro avg	0.78	0.77	0.78	400
weighted avg	0.78	0.78	0.78	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.79	0.73	0.76	41
1	0.82	0.86	0.84	59
accuracy			0.81	100
macro avg	0.81	0.80	0.80	100
weighted avg	0.81	0.81	0.81	100

b

Figure 13. Recall Values of the Logistic Regression Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Logistic Regression classifier achieved a recall score of 0.73 for the Intermediate Grade class, indicating that 73% of the actual intermediate-grade cases were correctly identified by the model. Meanwhile, the High Grade class obtained a recall score of 0.86, meaning that 86% of the actual high-grade cases were successfully classified as high grade.

These results demonstrate that the Logistic Regression classifier exhibited a stronger capability in detecting high-grade bone tumor cases than intermediate-grade cases. The higher recall score for the High Grade class suggests that the model was effective in reducing false negative predictions, thereby improving its ability to identify patients with severe tumor conditions. Overall, the testing results indicate that Logistic Regression provided robust recall performance, particularly for the detection of high-grade bone tumors.

3.3.3. Decision Tree Recall

The recall results of the Decision Tree classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a recall score of 0.72 for the Intermediate Grade class, indicating that 72% of the actual intermediate-grade cases were correctly identified by the model. For the High Grade class, the recall score reached 0.86, meaning that 86% of the actual high-grade cases were successfully classified as high grade.

These results indicate that the Decision Tree classifier demonstrated a stronger ability to detect high-grade bone tumor cases than intermediate-grade cases during the training phase. The higher recall value for the High Grade class suggests that the model was effective in identifying severe tumor cases while minimizing false negative predictions. Consequently, the classifier showed promising performance in recognizing patients with high-grade bone tumors.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.81	0.72	0.76	177
1	0.80	0.86	0.83	223
accuracy			0.80	400
macro avg	0.80	0.79	0.79	400
weighted avg	0.80	0.80	0.80	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.71	0.73	0.72	41
1	0.81	0.80	0.80	59
accuracy			0.77	100
macro avg	0.76	0.76	0.76	100
weighted avg	0.77	0.77	0.77	100

b

Figure 14. Recall Values of the Decision Tree Classifier: (a) Training Data and (b) Testing Data

For the testing dataset, the Decision Tree classifier achieved a recall score of 0.73 for the Intermediate Grade class, indicating that 73% of the actual intermediate-grade cases were correctly identified by the model. Meanwhile, the High Grade class obtained a recall score of 0.80, meaning that 80% of the actual high-grade cases were successfully classified as high grade.

These results suggest that the Decision Tree classifier demonstrated better performance in identifying high-grade bone tumor cases than intermediate-grade cases. However, compared with the training results, a slight decrease in recall was observed for the High Grade class, indicating a reduction in the model's ability to generalize to unseen data.

3.3.4. Random Forest Recall

The recall results of the Random Forest classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved a recall score of 0.80 for the Intermediate Grade class, indicating that 80% of the actual intermediate-grade cases were correctly identified. For the High Grade class, the recall score reached 0.92, meaning that 92% of the actual high-grade cases were successfully classified by the model.

These findings indicate that the Random Forest classifier exhibited excellent performance in detecting high-grade bone tumor cases during the training phase. The high recall score for the High Grade class suggests that the model was highly effective in minimizing false negative predictions and identifying patients with severe tumor conditions.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.89	0.80	0.84	177
1	0.85	0.92	0.89	223
accuracy			0.87	400
macro avg	0.87	0.86	0.86	400
weighted avg	0.87	0.87	0.87	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.68	0.63	0.66	41
1	0.76	0.80	0.78	59
accuracy			0.73	100
macro avg	0.72	0.72	0.72	100
weighted avg	0.73	0.73	0.73	100

b

Figure 15. Recall Values of the Random Forest Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Random Forest classifier achieved a recall score of 0.63 for the Intermediate Grade class, indicating that 63% of the actual intermediate-grade cases were correctly identified by the model. Meanwhile, the High Grade class obtained a recall score of 0.80, meaning that 80% of the actual high-grade cases were successfully classified as high grade.

These results demonstrate that the Random Forest classifier was more effective in detecting High Grade bone tumor cases than Intermediate Grade cases. Although the model achieved excellent recall performance

during the training phase, a noticeable decline was observed in the testing dataset, particularly for the Intermediate Grade class. This reduction suggests that the model's ability to generalize to unseen data was lower than its performance on the training data. Nevertheless, the recall score of 0.80 for the High Grade class indicates that the classifier maintained a relatively strong capability to identify severe bone tumor cases while minimizing false negative predictions.

3.4. Characteristics of Machine Learning Model F1-Score Measurement

In the context of bone tumor severity classification, the F1-score measures the overall effectiveness of a model in correctly identifying patients with Intermediate Grade or High Grade bone tumors while simultaneously evaluating its ability to accurately classify patients into the appropriate severity category. The F1-score provides a balanced assessment of model performance by considering both precision and recall.

When evaluating classification models, a trade-off often exists between precision and recall. Improving precision may lead to a decrease in recall, while increasing recall may reduce precision. Therefore, relying on a single metric may not provide a comprehensive evaluation of model performance.

The F1-score is defined as the harmonic mean of precision and recall, combining both metrics into a single performance measure. Because it is based on the harmonic mean, the F1-score becomes substantially lower when either precision or recall is low. Consequently, the F1-score is widely used to determine the balance between precision and recall and to assess the overall effectiveness of a classification model. In this study, the F1-score was employed to evaluate the ability of the Naive Bayes, Logistic Regression, Decision Tree, and Random Forest classifiers to predict bone tumor severity levels accurately and consistently.

3.4.1. Naive Bayes F1-Score

The F1-score results of the Naive Bayes classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved an F1-score of 0.70 for the Intermediate Grade class, indicating a balanced performance between precision and recall in identifying cases with intermediate tumor severity. For the High Grade class, the model obtained an F1-score of 0.77, demonstrating a stronger overall classification performance for patients with high-severity bone tumors.

These results suggest that the Naive Bayes classifier was more effective in classifying High Grade cases than Intermediate Grade cases during the training phase. The higher F1-score for the High Grade class indicates that the model achieved a better balance between precision and recall, resulting in more reliable predictions for severe bone tumor cases.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.70	0.70	0.70	177
1	0.76	0.77	0.77	223
accuracy			0.74	400
macro avg	0.73	0.73	0.73	400
weighted avg	0.74	0.74	0.74	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.72	0.68	0.70	41
1	0.79	0.81	0.80	59
accuracy			0.76	100
macro avg	0.75	0.75	0.75	100
weighted avg	0.76	0.76	0.76	100

b

Figure 16. F1-Score Values of the Naive Bayes Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Naive Bayes classifier achieved an F1-score of 0.70 for the Intermediate Grade class, indicating a balanced classification performance in terms of both precision and recall for intermediate-severity bone tumor cases. Meanwhile, the High Grade class obtained an F1-score of 0.80, demonstrating a stronger overall performance in identifying and classifying high-severity tumor cases.

The higher F1-score achieved for the High Grade class suggests that the model was more successful in maintaining a balance between precision and recall when predicting severe bone tumor conditions. These results further confirm that the Naive Bayes classifier performed better in classifying high-grade tumors than intermediate-grade tumors.

3.4.2 Logistic Regression F1-Score

The F1-score results of the Logistic Regression classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved an F1-score of 0.74 for the Intermediate Grade class, indicating a satisfactory balance between precision and recall in classifying intermediate-severity bone tumor cases. For the High Grade class, the model obtained an F1-score of 0.81, demonstrating stronger overall classification performance and a better balance between precision and recall for high-severity tumor cases.

These results indicate that the Logistic Regression classifier was more effective in identifying and classifying High Grade bone tumors than Intermediate Grade tumors during the training phase. The higher F1-score for the High Grade class suggests that the model achieved greater consistency and reliability in predicting severe tumor conditions.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.78	0.70	0.74	177
1	0.78	0.85	0.81	223
accuracy			0.78	400
macro avg	0.78	0.77	0.78	400
weighted avg	0.78	0.78	0.78	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.79	0.73	0.76	41
1	0.82	0.86	0.84	59
accuracy			0.81	100
macro avg	0.81	0.80	0.80	100
weighted avg	0.81	0.81	0.81	100

b

Figure 17. F1-Score Values of the Logistic Regression Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Logistic Regression classifier achieved an F1-score of 0.76 for the Intermediate Grade class, indicating a balanced performance between precision and recall in identifying intermediate-severity bone tumor cases. Meanwhile, the High Grade class obtained an F1-score of 0.84, demonstrating a stronger overall classification performance for high-severity tumor cases.

These results indicate that the Logistic Regression classifier maintained a favorable balance between precision and recall across both classes, with superior performance observed for the High Grade category. The higher F1-score for the High Grade class suggests that the model was more reliable and consistent in detecting and classifying severe bone tumor cases. Overall, the Logistic Regression classifier showed strong generalization performance and achieved one of the highest F1-scores among the evaluated models.

3.4.3 Decision Tree F1-Score

The F1-score results of the Decision Tree classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved an F1-score of 0.76 for the Intermediate Grade class, indicating a balanced classification performance between precision and recall for intermediate-severity bone tumor cases. For the High Grade class, the model obtained an F1-score of 0.83, demonstrating stronger overall classification performance and a better balance between precision and recall for high-severity tumor cases.

These findings suggest that the Decision Tree classifier performed more effectively in classifying High Grade tumors than Intermediate Grade tumors during the training phase. The higher F1-score for the High Grade class indicates that the model was able to maintain greater consistency in identifying severe tumor cases while balancing both precision and recall.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.81	0.72	0.76	177
1	0.80	0.86	0.83	223
accuracy			0.80	400
macro avg	0.80	0.79	0.79	400
weighted avg	0.80	0.80	0.80	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.71	0.73	0.72	41
1	0.81	0.80	0.80	59
accuracy			0.77	100
macro avg	0.76	0.76	0.76	100
weighted avg	0.77	0.77	0.77	100

b

Figure 18. F1-Score Values of the Decision Tree Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Decision Tree classifier achieved an F1-score of 0.72 for the Intermediate Grade class, indicating a balanced performance between precision and recall in classifying intermediate-severity bone tumor cases. Meanwhile, the High Grade class obtained an F1-score of 0.80, demonstrating stronger overall classification performance for high-severity tumor cases.

These results indicate that the Decision Tree classifier was more effective in identifying and classifying High Grade bone tumors than Intermediate Grade tumors. Although the model achieved relatively strong performance for both classes, the higher F1-score for the High Grade category suggests that it maintained a better balance between precision and recall when predicting severe tumor cases. Compared with the training results, a slight decrease in F1-score was observed on the testing dataset, indicating some reduction in generalization performance.

3.4.4. Random Forest F1-Score

The F1-score results of the Random Forest classifier were obtained from both the training and testing datasets. On the training dataset, the model achieved an F1-score of 0.84 for the Intermediate Grade class, indicating a strong balance between precision and recall in classifying intermediate-severity bone tumor cases. For the High Grade class, the model obtained an F1-score of 0.89, demonstrating excellent overall classification performance and a highly balanced relationship between precision and recall for high-severity tumor cases.

These results suggest that the Random Forest classifier performed exceptionally well during the training phase, particularly in identifying and classifying high-grade bone tumors. The higher F1-score achieved for the High Grade class indicates that the model was highly effective in maintaining both accuracy and consistency when predicting severe tumor cases.

Training Data Classification Report:

	precision	recall	f1-score	support
0	0.89	0.80	0.84	177
1	0.85	0.92	0.89	223
accuracy			0.87	400
macro avg	0.87	0.86	0.86	400
weighted avg	0.87	0.87	0.87	400

a

Testing Data Classification Report:

	precision	recall	f1-score	support
0	0.68	0.63	0.66	41
1	0.76	0.80	0.78	59
accuracy			0.73	100
macro avg	0.72	0.72	0.72	100
weighted avg	0.73	0.73	0.73	100

b

Figure 19. F1-Score Values of the Random Forest Classifier: (a) 80% Training Data and (b) 20% Testing Data

For the testing dataset, the Random Forest classifier achieved an F1-score of 0.66 for the Intermediate Grade class, indicating a moderate balance between precision and recall in identifying intermediate-severity bone tumor cases. Meanwhile, the High Grade class obtained an F1-score of 0.78, demonstrating stronger overall classification performance for high-severity tumor cases.

Although the Random Forest classifier achieved excellent F1-scores on the training dataset, a noticeable decrease was observed on the testing dataset for both classes. This decline suggests that the model's performance was less consistent when applied to unseen data. Nevertheless, the F1-score of 0.78 for the High Grade class indicates that the classifier maintained a relatively strong ability to balance precision and recall in detecting severe bone tumor cases.

3.5. Method Comparison

The performance comparison of the Naive Bayes, Logistic Regression, Decision Tree, and Random Forest classifiers was conducted using the testing dataset with an 80% training and 20% testing data split. The comparison was based on three evaluation metrics: accuracy, recall, and F1-score. These metrics were used to assess the effectiveness of each classification method in predicting bone tumor severity levels.

The comparative results of the performance parameters for each classification method are presented in Table 1.

Table 1. Comparison of Testing Dataset Results for All Methods

Method	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-Score (0)	F1-Score (1)
Naive Bayes	76%	72%	79%	68%	81%	70%	80%
Logistic Regression	81%	79%	82%	73%	86%	76%	84%
Decision Tree	77%	71%	81%	73%	80%	72%	80%
Random Forest	73%	68%	76%	63%	80%	66%	78%

.

The findings indicate that among the four machine learning algorithms evaluated—Naive Bayes, Logistic Regression, Decision Tree, and Random Forest—Logistic Regression demonstrated the best overall performance in classifying bone tumor grades into Intermediate Grade and High Grade categories. Logistic Regression achieved the highest testing accuracy (81%), along with strong precision (0.79–0.82), recall (0.73–0.86), and F1-score (0.76–0.84). Although Random Forest exhibited superior performance during training (87% accuracy and F1-score up to 0.89), its testing performance decreased substantially (71% accuracy), suggesting overfitting and weaker generalization capability. Meanwhile, Naive Bayes and Decision Tree produced moderate results, with testing accuracies of 76% and 77%, respectively. Across all models, classification performance was consistently better for the High Grade category than for the Intermediate Grade category, indicating that severe bone tumor cases possess more distinguishable patterns within the selected clinical features.

The results of this study indicate that Logistic Regression achieved the best overall performance among the evaluated models, with an accuracy of 81% and the highest F1-score for the High Grade class (84%). These findings are consistent with previous studies demonstrating that machine learning algorithms can effectively support bone tumor classification and grading. Li et al. [7] reported that machine learning models provide high diagnostic performance for malignant bone tumor classification, emphasizing the importance of model generalizability for clinical application. Similarly, Cuzzubbo & Carpentier [42] found that machine learning techniques utilizing radiographic features can accurately differentiate bone tumor characteristics and improve diagnostic decision-making. Furthermore, Gitto et al. [43] demonstrated that machine-learning-based radiomics models achieved reliable classification performance in bone tumor assessment using MRI data. In the present study, Logistic Regression showed better generalization performance than Random Forest and Decision Tree, as

evidenced by its superior testing accuracy, recall, and F1-score. Although Random Forest achieved excellent performance on the training dataset, its lower testing results suggest a tendency toward overfitting. Therefore, the current findings support previous evidence that simpler and more interpretable machine learning models, such as Logistic Regression, may provide more stable and reliable performance for bone tumor severity classification when applied to unseen clinical data.

The novelty of this study lies in the development and comparison of multiple machine learning approaches for bone tumor grade classification using clinicopathological variables, including sex, grade-related characteristics, histological type, primary soft tissue sarcoma site, and treatment information. While previous studies have predominantly focused on medical imaging modalities such as MRI, CT scans, and histopathological image analysis, this study demonstrates that satisfactory classification performance can be achieved using structured clinical data alone. Additionally, the study provides a comprehensive comparison based on multiple evaluation metrics (accuracy, precision, recall, and F1-score), allowing a more holistic assessment of model effectiveness and generalization performance for bone tumor severity prediction.

The findings have important implications for both clinical decision support and medical data analytics. From a clinical perspective, the Logistic Regression model could serve as a practical and interpretable decision-support tool to assist physicians in identifying high-grade bone tumors, enabling earlier intervention and treatment planning. From a methodological perspective, the results emphasize that model selection should not rely solely on training performance but should prioritize testing performance and generalization capability. The study also highlights the potential of machine learning techniques to support precision medicine by extracting predictive insights from routinely collected clinical data without requiring expensive imaging-based approaches.

Despite the promising results, several limitations should be acknowledged. First, the study relied on a single dataset with a fixed 80:20 train-test split, which may limit the generalizability of the findings to other populations and clinical settings. Second, the models were developed using a limited number of clinical variables, potentially excluding other important predictors such as genetic markers, laboratory findings, and imaging characteristics. Third, no external validation dataset or cross-validation procedure was reported, which may affect the robustness of the performance estimates. Therefore, future studies should incorporate larger multicenter datasets, additional predictive features, and external validation strategies to improve model generalizability and confirm the applicability of the proposed classification framework across diverse patient populations.

4. CONCLUSION

Based on the data processing, analysis, and testing results, it can be concluded that all evaluated machine learning algorithms were capable of classifying bone tumor severity with satisfactory performance using an 80:20 train-test data split. On the testing dataset, the Naive Bayes classifier achieved an accuracy of 76%, precision of 79%, recall of 81%, and F1-score of 80%; Logistic Regression achieved an accuracy of 81%, precision of 82%, recall of 86%, and F1-score of 84%; Decision Tree achieved an accuracy of 77%, precision of 81%, recall of 80%, and F1-score of 80%; while Random Forest achieved an accuracy of 73%, precision of 76%, recall of 80%, and F1-score of 78%. Among the evaluated models, Logistic Regression demonstrated the best overall performance, achieving the highest accuracy, precision, recall, and F1-score on the testing dataset. These findings indicate that Logistic Regression is the most effective and reliable method for classifying bone tumor severity within the dataset used in this study. For future research, it is recommended to compare the performance of Logistic Regression with advanced deep learning algorithms, such as Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Unit (GRU), and Recurrent Neural Network (RNN), as well as to utilize larger datasets and incorporate additional features to improve model accuracy and generalization performance.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to all individuals and institutions who contributed to the completion of this research. Special thanks are extended to the faculty members, colleagues, and data providers whose support and valuable insights facilitated the successful completion of this study. The authors also appreciate the constructive feedback provided throughout the research process. No specific funding was received for this research.

REFERENCES

[1] K. Zarghooni, G. Bratke, P. Landgraf, T. Simon, D. Maintz, and P. Eysel, “The diagnosis and treatment of osteosarcoma and Ewing’s sarcoma in children and adolescents,” *Dtsch. Arztebl. Int.*, vol. 120, no. 24, pp. 405–412, Jun. 2023, doi: 10.3238/arztebl.m2023.0079.

[2] E. Young, B. Kelly, and J. E. Cain, “Targeting developmental vulnerabilities in childhood sarcomas,” *Cancer Metastasis Rev.*, vol. 44, no. 4, p. 72, 2025, doi: 10.1007/s10555-025-10286-y.

[3] G. Iacobellis, A. Leggio, C. Salzillo, S. Lucà, R. Ortega-Ruiz, and A. Marzullo, "Analysis and historical evolution of paediatric bone tumours: The importance of early diagnosis in the detection of childhood skeletal malignancies," *Cancers (Basel).*, vol. 17, no. 3, p. 451, 2025, doi: 10.3390/cancers17030451.
[4] Y. Qiao, C. Eisfeld, R. von Eisenhart-Rothe, and F. Hinterwimmer, "Performance comparison and future perspectives of deep learning and classical machine learning in bone tumor applications: a systematic review (2019–2025)," *BMC Med. Inform. Decis. Mak.*, vol. 26, no. 1, p. 117, 2026, doi: 10.1186/s12911-026-03401-8.
[5] W. Li *et al.*, "Global cancer statistics for adolescents and young adults: population based study," *J. Hematol. Oncol.*, vol. 17, no. 1, p. 99, 2024, doi: 10.1186/s13045-024-01623-9.
[6] Y. Chen, Q. Lian, W. Sun, and X. Lian, "A relief or new challenges? Global, regional, and national burden of malignant neoplasm of bone and articular cartilage in children and adolescents from 1990 to 2021 and its prediction to 2035," *Adv. Orthop.*, vol. 2, pp. 117–131, 2025, doi: 10.1016/j.advop.2025.09.001.
[7] Y. Li, B. Dong, and P. Yuan, "The diagnostic value of machine learning for the classification of malignant bone tumor: A systematic evaluation and meta-analysis," *Front. Oncol.*, vol. Volume 13, 2023, doi: 10.3389/fonc.2023.1207175.
[8] Y. Guan, W. Zhang, Y. Mao, and S. Li, "Nanoparticles and bone microenvironment: a comprehensive review for malignant bone tumor diagnosis and treatment," *Mol. Cancer*, vol. 23, no. 1, p. 246, 2024, doi: 10.1186/s12943-024-02161-1.
[9] A. E. Bădilă, D. M. Rădulescu, A.-G. Niculescu, A. M. Grumezescu, M. Rădulescu, and A. R. Rădulescu, "Recent advances in the treatment of bone metastases and primary bone tumors: an up-to-date review," 2021. doi: 10.3390/cancers13164229.
[10] B. Abhisheka, S. K. Biswas, B. Purkayastha, D. Das, and A. Escargueil, "Recent trend in medical imaging modalities and their applications in disease diagnosis: a review," *Multimed. Tools Appl.*, vol. 83, no. 14, pp. 43035–43070, 2024, doi: 10.1007/s11042-023-17326-1.
[11] S. Hussain *et al.*, "Modern diagnostic imaging technique applications and risk factors in the medical field: A review," *Biomed Res. Int.*, vol. 2022, no. 1, p. 5164970, Jan. 2022, doi: 10.1155/2022/5164970.
[12] D. P. Frush, M. J. Callahan, B. D. Coley, H. R. Nadel, and R. Paul Guillerman, "Comparison of the different imaging modalities used to image pediatric oncology patients: A COG diagnostic imaging committee/SPR oncology committee white paper," *Pediatr. Blood Cancer*, vol. 70, no. S4, p. e30298, 2023, doi: 10.1002/pbc.30298.
[13] P. Pricolo *et al.*, "Whole-body magnetic resonance imaging (WB-MRI) reporting with the METastasis Reporting and Data System for Prostate Cancer (MET-RADS-P): inter-observer agreement between readers of different expertise levels," *Cancer Imaging*, vol. 20, no. 1, p. 77, 2020, doi: 10.1186/s40644-020-00350-x.
[14] Z. A. Ramadan, A. H. Elmorsy, S. E. Taman, and F. A. Denewar, "Inter-observer and intra-observer agreement of bone reporting and data system (Bone-RADS) in the interpretation of bone tumors on computed tomography," *Clin. Imaging*, vol. 117, p. 110367, 2025, doi: 10.1016/j.clinimag.2024.110367.
[15] L. Quinn *et al.*, "Interobserver variability studies in diagnostic imaging: a methodological systematic review," *Br. J. Radiol.*, vol. 96, no. 1148, p. 20220972, Aug. 2023, doi: 10.1259/bjr.20220972.
[16] C. J. R. Mullen, R. D. Barr, and E. L. Franco, "Timeliness of diagnosis and treatment: The challenge of childhood cancers," *Br. J. Cancer*, vol. 125, no. 12, pp. 1612–1620, 2021, doi: 10.1038/s41416-021-01533-4.
[17] I. Ahmad *et al.*, "New paradigms to break barriers in early cancer detection for improved prognosis and treatment outcomes," *J. Gene Med.*, vol. 26, no. 8, p. e3730, 2024, doi: 10.1002/jgm.3730.
[18] A. Rao, N. E. Rich, J. A. Marrero, A. C. Yopp, and A. G. Singal, "Diagnostic and therapeutic delays in patients with hepatocellular carcinoma," *J. Natl. Compr. Cancer Netw.*, vol. 19, no. 9, pp. 1063–1071, 2021, doi: 10.6004/jnccn.2020.7689.
[19] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, "Machine learning in clinical decision making," *Med*, vol. 2, no. 6, pp. 642–665, Jun. 2021, doi: 10.1016/j.medj.2021.04.006.
[20] S. Eloranta and M. Boman, "Predictive models for clinical decision making: Deep dives in practical machine learning," *J. Intern. Med.*, vol. 292, no. 2, pp. 278–295, 2022, doi: 10.1111/joim.13483.
[21] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of Machine Learning Tools in Healthcare Decision Making," *J. Healthc. Eng.*, vol. 2021, no. 1, p. 6679512, 2021, doi: 10.1155/2021/6679512.
[22] E. V Varlamova *et al.*, "Machine learning meets cancer," *Cancers (Basel).*, vol. 16, no. 6, p. 1100, 2024, doi: 10.3390/cancers16061100.
[23] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis, "Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 5546–5555, 2021, doi: 10.1016/j.csbj.2021.10.006.
[24] A. Sharma and R. Rani, "A systematic review of applications of machine learning in cancer prediction and diagnosis," *Arch. Comput. Methods Eng.*, vol. 28, no. 7, pp. 4875–4896, 2021, doi: 10.1007/s11831-021-09556-z.
[25] H. Wu *et al.*, "Predicting chronic pain and treatment outcomes using machine learning models based on high-dimensional clinical data from a large retrospective cohort," *Clin. Ther.*, vol. 46, no. 6, pp. 490–498, 2024, doi: 10.1016/j.clinthera.2024.04.012.
[26] M. F. Kabir, T. Chen, and S. A. Ludwig, "A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction," *Healthc. Anal.*, vol. 3, p. 100125, 2023, doi: 10.1016/j.health.2022.100125.
[27] J. Rahnenführer *et al.*, "Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges," *BMC Med.*, vol. 21, no. 1, p. 182, 2023, doi: 10.1186/s12916-023-02858-y.
[28] C. Pan, L. Lian, J. Chen, and R. Huang, "FemurTumorNet: Bone tumor classification in the proximal femur using DenseNet model based on radiographs," *J. Bone Oncol.*, vol. 42, p. 100504, 2023, doi: 10.1016/j.jbo.2023.100504.
[29] N. N. Prakash, V. Rajesh, D. L. Namakhwa, S. Dwarkanath Pande, and S. H. Ahammad, "A DenseNet CNN-based liver lesion prediction and classification for future medical diagnosis," *Sci. African*, vol. 20, p. e01629, 2023, doi:

10.1016/j.sciaf.2023.e01629.
[30] N. Hasan, Y. Bao, A. Shawon, and Y. Huang, "A DenseNet CNN-based liver lesion prediction and classification for future medical diagnosis," *SN Comput. Sci.*, vol. 2, no. 5, p. 389, 2021, doi: 10.1007/s42979-021-00782-7.
[31] P. S. Papageorgiou *et al.*, "Artificial intelligence in primary malignant bone tumor imaging: a narrative review," 2025. doi: 10.3390/diagnostics15131714.
[32] H. Hosseini, S. Heydari, K. Hushmandi, S. Daneshi, and R. Raesi, "Bone tumors: A systematic review of prevalence, risk determinants, and survival patterns," *BMC Cancer*, vol. 25, no. 1, p. 321, 2025, doi: 10.1186/s12885-025-13720-0.
[33] Y. Shen *et al.*, "The association between circulating 25-hydroxyvitamin D and pancreatic cancer: A systematic review and meta-analysis of observational studies," *Eur. J. Nutr.*, vol. 63, no. 3, pp. 653–672, 2024, doi: 10.1007/s00394-023-03302-w.
[34] G. K. Thakur, A. Thakur, S. Kulkarni, N. Khan, and S. Khan, "Deep learning approaches for medical image analysis and diagnosis," *Cureus*, vol. 16, no. 5, 2024, doi: 10.7759/cureus.59507.
[35] H. A. Helaly, M. Badawy, and A. Y. Haikal, "A review of deep learning approaches in clinical and healthcare systems based on medical image analysis," *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 36039–36080, 2024, doi: 10.1007/s11042-023-16605-1.
[36] X. Chen *et al.*, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, p. 102444, 2022, doi: 10.1016/j.media.2022.102444.
[37] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, p. 101693, 2020, doi: 10.1016/j.media.2020.101693.
[38] Y. Xu, T. M. Khan, Y. Song, and E. Meijering, "Edge deep learning in computer vision and medical diagnostics: A comprehensive survey," *Artif. Intell. Rev.*, vol. 58, no. 3, p. 93, 2025, doi: 10.1007/s10462-024-11033-5.
[39] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. Thousand Oaks, CA: SAGE Publications, 2018.
[40] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.
[41] G. James, "An introduction to statistical learning with applications in R," 2013, *Springer, New York*. doi: 10.1007/978-1-4614-7138-7.
[42] S. Cuzzubbo and A. F. Carpentier, "Applications of melanin and melanin-like nanoparticles in cancer therapy: A review of recent advances," 2021. doi: 10.3390/cancers13061463.
[43] S. Gitto *et al.*, "MRI radiomics-based machine-learning classification of bone chondrosarcoma," *Eur. J. Radiol.*, vol. 128, Jul. 2020, doi: 10.1016/j.ejrad.2020.109043.