INTERNATIONAL JOURNAL OF APPLIED METHODS IN ELECTRONICS AND COMPUTERS

www.ijamec.org

International Open Access

Volume 14
Issue 01
March, 2026

Research Article

Classification of Orthopedic Diseases Using Biomechanical Properties with Machine Learning Methods

Mehmet Cüneyt ÖZBALCI [a,*], ***Turgay Tugay BİLGİN*** [b]

[a] *Bursa Technical University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Bursa, Türkiye.*
[b] *Bursa Technical University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Bursa, Türkiye.*

ARTICLE INFO

Article history:
Received 3 February 2026
Accepted 30 March 2026

Keywords:
Biomechanical Features,
Machine Learning,
Orthopedic Disease Classification,
SMOTE.

ABSTRACT

Orthopedic diseases significantly affect the musculoskeletal system, reducing patients' functional capacity and quality of life. Accurate and early classification of such conditions is therefore critical for effective clinical decision-making. This study proposes a machine learning-based framework for orthopedic disease classification using biomechanical features, with a particular emphasis on handling class imbalance. A set of classification algorithms, including Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, XGBoost, LightGBM, and a Soft Voting Ensemble (XGBoost + LightGBM), were evaluated on a dataset of 310 patients using 10-fold cross-validation. The impact of the Synthetic Minority Over-sampling Technique (SMOTE) was systematically analyzed by comparing model performance with and without its application. Evaluation metrics included Accuracy, Precision, Recall, F1-score, and macro-average ROC-AUC. Results indicate that addressing class imbalance significantly improves model performance, particularly in terms of ROC-AUC. Among the tested methods, Logistic Regression demonstrated the most stable and competitive results. The best performance was achieved by Logistic Regression with SMOTE, yielding 87% accuracy and a macro-average ROC-AUC of 0.96. These findings highlight the importance of imbalance-aware modeling strategies in orthopedic disease classification.

1. Introduction

Orthopedics is a branch of medical science that studies the musculoskeletal system. It has achieved rapid growth with scientific advances. It examines problems related to bones, muscles, joints, and nerves. Thanks to the rapid development of computer science and technology, significant progress has been made in observing tissue behavior, predicting diseases, and conducting treatment studies. Orthopedic diseases are conditions that affect a large number of people today. As with many diseases, early diagnosis is considered crucial for starting treatment early and achieving a positive response to treatment [1].

Today, a vast amount of data from many different fields is stored and preserved. Existing data is extremely valuable in terms of generating meaningful information, providing interpretations for various purposes, and offering solutions. Data in the medical world can include patient data, disease data, tissue and organ data, and data related to chemicals in the body. In the field of orthopedics, data related to a person's muscle and bone tissues are also prominent. Based on the data of individuals with diseases, it will be possible to make an early diagnosis for people who may potentially become ill and to start treatment in a timely manner. Thanks to various orthopedic imaging techniques, it is possible to collect clinical data and then make the relevant diagnoses [2]. Biomedical imaging allows for the examination of the inside of the human body and the taking of photographs of bones, muscles, and joints. Diagnostic imaging in orthopedics involves the widespread use of ultrasound, X-

* Corresponding author. E-mail address: *mehmet.ozbalci@btu.edu.tr*
DOI: 10.58190/ijamec.2026.163

ray, CT (computed tomography) scans, and MRI [3].

Recent studies have increasingly focused on improving the performance of machine learning models for medical and orthopedic disease classification. The problem of class imbalance has also been widely investigated in recent literature. Comparative analyses of machine learning algorithms in medical diagnosis have shown that model performance is highly dependent on both data characteristics and preprocessing strategies.

Elshewey and Osman [1] proposed an optimization-based approach for orthopedic disease classification and achieved high accuracy by integrating feature selection strategies. Similarly, Zhang et al. [2] developed a hybrid SMOTE-RFE-XGBoost model for spinal disease classification, demonstrating that combining feature selection with imbalance handling techniques can significantly improve predictive performance. Kıvrak et al. [4] evaluated multiple machine learning algorithms using SMOTE and reported notable improvements in recall and F1-score for minority classes. Yang et al. [5] proposed a hybrid framework combining SMOTE with ensemble learning methods such as LightGBM and Random Forest, achieving improved classification performance in healthcare applications. Gyasi-Agyei [6] demonstrated that no single model consistently outperforms others across all scenarios, highlighting the importance of dataset-specific evaluation. In orthopedic-related applications, Rezapour et al. [7] applied machine learning techniques with SMOTE to gait analysis data and showed that imbalance-aware learning improves classification robustness.

In the study by Hafsa Binte Kibria and Abdul Matin, fusion models were created to diagnose and evaluate the severity of CVDs (cardiovascular diseases). Artificial neural networks, SVM, logistic regression, decision trees, Random Forest, and AdaBoost algorithms were applied to the heart disease dataset to perform disease prediction. To balance the imbalanced dataset, Randomoversampler was used only for multi-class classification, and a weighted score fusion approach was used to improve classification performance. In the proposed approach, the highest accuracy for multi-class classification was found to be 75.41% with logistic regression and artificial neural networks, while for binary classification, it was 95.08% with AdaBoost and decision trees [8]. In the study conducted by Shenglong Li and Xiaojing Zhang, an orthopedic assistant classification prediction model developed with the XGBoost algorithm is proposed. The model was created using the Random Forest algorithm, XGBoost, and the associated classification algorithm, respectively. The results obtained show that the XGBoost algorithm has a higher accuracy with a value of approximately 95% [9]. The study conducted by Cai-Jin Ling and his team investigated the effectiveness of biomedical sensors and medical image segmentation algorithms in the diagnosis and treatment of orthopedic diseases. This study aims to develop image segmentation methods by collecting data from MR and scan images and then using machine learning techniques. The data was divided into two groups. The first group is a control group that attempts to diagnose orthopedic diseases using X-ray images with traditional methods, while the second group is an experimental group that uses machine learning techniques with MR or CT scanning methods to diagnose orthopedic diseases. The proposed method was evaluated using standard performance metrics such as VAS score, Ramsay score, disease classification, treatment effect, and observation index. In the study, CNN, KNN, Fast KNN, and Mask R-CNN machine learning methods were applied and their performance was compared. Looking at all the results obtained, Mask R-CNN stands out as the method that achieved the highest performance [10].

Nadia Rubaiyat and colleagues aims to predict orthopedic diseases early. To this end, three different machine learning algorithms—logistic regression, random forest, and KNN—were applied to a dataset of 310 patients containing six biomechanical features describing the pelvic and lumbar conditions of the patients. When the results were compared, the random forest algorithm yielded the best result with an accuracy rate of 89%. KNN achieved 85% accuracy, while logistic regression achieved 81% accuracy [11]. In the study conducted by Nasrin Jahan and colleagues, a dataset containing 310 examples was used to classify patients using various machine learning algorithms. Classification was performed in two stages. In the two-stage classification, patients were classified as normal or abnormal in the first stage, and then classified according to their physical condition in the second stage. Performance metrics were calculated along with the performance rates. The KNN algorithm yielded the highest accuracy rate. Most other classification algorithms also achieved successful results with an accuracy above 80% [12]. In the study conducted by Kamrul Hasan and colleagues, an attempt was made to predict the presence or absence of orthopedic disease. The classification performance of various classification algorithms was measured. Each patient in the dataset is represented by six biomechanical features derived from the shape and orientation of the pelvis and lumbar spine. While most of the applied algorithms achieved over 90% accuracy, the decision tree (DT) algorithm achieved 99% accuracy, showing a significant difference [13]. In the study conducted by Qiuchong Chen and colleagues, AKI risk prediction in elderly patients after orthopedic surgery is performed based on machine learning algorithm models. Nine different machine learning algorithms were applied to an experimental group of 1000 postoperative AKI patients who underwent orthopedic surgery. Intraoperative information and preoperative clinical characteristics were used in the prediction phase. The ideal model was

evaluated by calculating AUC, sensitivity, specificity, and accuracy. In this way, the most suitable model was selected. Logistic regression yielded the best results [14].

In the study conducted by Mantzaris and colleagues, various algorithms were evaluated in terms of classification performance for osteoporosis, an orthopedic disease. The aim of the study was to assist experts in predicting osteoporosis by avoiding unnecessary further tests using bone densitometry. Multi-layer perceptrons (MLP) and probabilistic neural networks (PNN), which are feedforward networks, were used to predict osteoporosis risk factors. PNNs were applied with propagation values ranging from 0.1 to 50 and with 4 or 2 neurons in the output layer, depending on the coding of the desired outcome of osteoporosis. Looking at the experimental results, it was observed that PNNs performed better. In addition, as the spread values increased, the overfitting problem was more frequently observed in MLPs [15]. In the study conducted by Seok Won Chung and colleagues, a study was performed to detect and classify proximal humerus fractures using CNN with plain anterior-posterior shoulder radiographs. The results were evaluated by measuring AUC, sensitivity, specificity, and Youden index. To distinguish normal shoulders from proximal humerus fractures, 96% accuracy, 1.00 AUC, 0.99 sensitivity, 0.97 specificity, and 0.97 Youden index were obtained. To classify the fracture type, 65-86% accuracy, 0.90-0.98 AUC, 0.88-0.97 sensitivity, 0.83-0.94 specificity, and 0.71-0.90 Youden index were obtained [16]. Jakub OLCZAK and colleagues classified ankle fractures. A neural network based on the ResNet architecture was trained using 4,941 radiographic ankle examination results. All images were classified according to the AO/OTA (AO Foundation/Orthopedic Trauma Association) 2018 classification, and the network performance was evaluated against a test set of 400 patients independently reviewed by two expert observers. The results showed that the AUC value reached up to 0.93 for type B fractures. Type A fractures showed the worst performance [17]. This study investigated the efficient detection of regions of interest in hip MRI and the accurate identification of the femoral head and necrotic region. The study, conducted with 484 patients, utilized 3,937 labeled hip MRI images. Lesion detection achieved a mAP value of up to 94.7%. Cao and colleagues aimed to efficiently identify the region of interest in hip MRI and accurately define the femoral head and necrotic region. Comparative analyses were performed using Swin-Unet, U-net, and YOLO models. The study, conducted with 484 patients, utilized 3,937 labeled hip MRI images. The lesion detection mAP value reached up to 94.7% [18]. A study on the classification of orthopedic diseases was conducted by Elshevey and Osman. For feature selection in the study, binary particle swarm optimization (BPSO), binary gray wolf optimization (BGWO), and binary whale optimization algorithm (BWAO) were used. An average error rate reduction of 47.29% was observed in the breadth-first search method. The highest accuracy rate was achieved with BFS-Random Forest, with 99.41% accuracy [19].

Although there are numerous studies in the literature using the same dataset, studies that systematically compare the effect of SMOTE on different classifiers in terms of macro-average metrics are limited. This study aims to fill this gap.

1.1. Key Contributions of This Study

A comprehensive machine learning framework is proposed for the classification of orthopedic diseases using biomechanical features, addressing a clinically relevant multi-class problem. The effect of class imbalance is systematically investigated by integrating the Synthetic Minority Over-sampling Technique (SMOTE) into the modeling pipeline and comparing results with and without its application. A wide range of machine learning algorithms, including Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, XGBoost, LightGBM, and a Soft Voting Ensemble model, are evaluated under a unified experimental setup using stratified 10-fold cross-validation.

Model performance is assessed using macro-average evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC), providing a balanced and reliable comparison across imbalanced class distributions. Experimental results demonstrate that Logistic Regression, particularly when combined with SMOTE, achieves the most stable and competitive performance, offering practical insights for clinical decision support systems.

The study provides a systematic and comparative analysis of imbalance-aware learning strategies, contributing to the limited literature on SMOTE-based evaluation using macro-average metrics in orthopedic disease classification.

2. Material and Methods

The steps of the study are shown step by step in Figure 1 below.

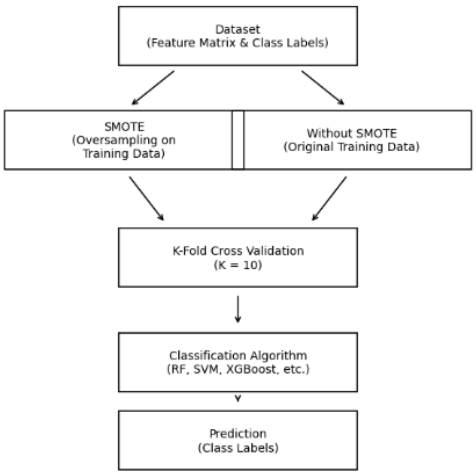

Figure 1. Phases of Study

Results were obtained and compared when SMOTE was applied and when it was not applied during this process. SMOTE is applied only to the training data, followed by classification using 10-fold cross-validation and multiple machine learning algorithms. The hyperparameters used in all models were selected from the default values commonly used in the literature. Complex hyperparameter optimization was not applied in order to prevent overfitting. Finally, performance metrics are obtained.

2.1. Dataset

The dataset [20] used contains the biomechanical characteristics that define the pelvic and lumbar conditions of orthopedic patients and, if a disease condition is present, the name of that disease. These biomechanical characteristics consist of six features: pelvic incidence, pelvic tilt numeric, lumbar lordosis angle, sacral slope, pelvic radius, and grade of spondyloisthesis. The condition of each of these characteristics triggering the disease has been examined. It investigates whether 310 different individuals have one of two different diseases or neither of these two diseases, i.e., whether they are healthy. All six different characteristics of each individual are present in the dataset. The diseases are spondylolisthesis and hernia. In addition to these two diseases, there is also a class designated as healthy, or normal. Thus, there are three different classes in total. Of these three classes, 60 belong to the hernia class, 150 to the spondylolisthesis class, and 100 to the normal class. The data is the Biomechanical features of orthopedic patients dataset obtained from the Kaggle website. The dataset used in this study is publicly available and does not contain any personally identifiable information; therefore, no ethics committee approval is required. The dataset and its characteristics are provided in Table 1 below.

Table 1. Description of Biomechanical Features Used in the Dataset

Number	Feature	Definition
1	Pelvic incidence	It refers to the angle between the line perpendicular to the sacral plate at its midpoint and the line connecting this point to the femoral head axis.
2	Pelvic tilt	It refers to the angle formed by the vertical line passing through the femoral head and the line connecting the center of the sacral plane.
3	Lumbar lordosis angle	It refers to the angle formed by the inward curvature of the lumbar spine region located just above the hips.
4	Sacral slope	It refers to the angle between the line tangent to the sacral plateau and the vertical line passing through the exact center point of the sacral plateau.
5	Pelvic radius	It refers to the distance from the hip axis to the posterior-superior corner of the S1 endplate.
6	Grade of spondylolisthesis	It indicates the degree of spinal displacement.

To ensure that model performance can be evaluated in a reliable and generalizable manner, the Stratified K-Fold Cross Validation method was used in this study. The dataset was divided into 10 sub-sets (folds) while preserving the class distribution; in each iteration, one sub-set was used as test data while the remaining nine sub-sets were evaluated during the training phase. This minimized bias in random data division. Due to the limited size of the dataset, 10 folds were deemed appropriate.

2.2. Synthetic Minority over Sampling Technique (SMOTE)

The primary goal of SMOTE is to increase the number of examples belonging to minority classes while preserving the interpolation of these examples in the feature space. This approach aims to improve the model's generalization ability by reducing class imbalance [21]. In the SMOTE method, synthetic samples are generated using linear interpolation points between minority class samples. This makes the dataset relatively balanced. As a result, it provides balanced and reliable classification performance for minority classes [22].

The Synthetic Minority Over-sampling Technique (SMOTE) was used to reduce the potential negative effects of class imbalance on classification performance. SMOTE generates synthetic data by creating linear combinations of existing samples for minority classes [23]. This relatively balances the number of samples between classes. This method particularly improves macro-average performance metrics. Furthermore, it has a positive effect on the model's generalization ability on minority classes.

The SMOTE process was applied only on the training subset in each cross-validation iteration. Test data were not

included in any synthetic sampling process.

2.3. Random Forest

The Random Forest algorithm is a flexible machine learning algorithm that can be applied for various purposes. Random Forest can provide solutions for classification and regression problems. It appears as an advanced derivative of decision trees [24]. In this algorithm, partitioning is applied randomly. The goal is to find the optimal branch. Sub-branches are then created based on this [25].

Random Forest is an ensemble learning algorithm. The main goal in ensemble learning methods is to improve performance by combining models. In Random Forest, continuous and non-continuous variables can be used together in large datasets [26].

2.4. Support Vector Machines

SVM is an effective method that produces successful results and is commonly used for classification. It enables the prediction of labels using one or more feature vectors [27]. The classification process is performed by creating a decision boundary between classes [28]. There are three commonly used kernel functions for classifiers: the radial basis kernel, linear kernel, and polynomial kernel function. In addition, the sigmoid kernel function can also be used in some cases. While data preprocessing and similar operations performed beforehand are of great importance, the kernel function applied also has a significant impact on performance [29]. The RBF kernel function, which yielded the most successful results in this study, was used. The equation of RBF given below.

$$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|)^2 \quad (1)$$

In the above equation, x_i,x_j represents the data points, while $\|x_i-x_j\|^2$ is the square of the Euclidean distance. Here γ is the spread parameter of the kernel.

Data classification is performed using hyperplanes. The best classifier is the most ideal hyperplane. The distance between hyperplanes is called the margin [30]. The small number of operations during the learning phase in the SVM algorithm positively affects performance.

2.5. Naive Bayes Classifier

Naive Bayes is a powerful classification algorithm widely used in classification problems and based on Bayes' theorem. In this method, which is shaped on the basis of conditional probabilities, the effect and dominance of all attributes are investigated [31]. This method is based on selecting the option with the highest probability [32]. The classification process is performed by combining the effects of different features on the result [33]. To achieve the optimal result, the final calculation is made based on the combined probabilities of different situations. The formula of Bayes' theorem used in calculating the combined probability is as follows.

$$P(C_k|x)=\frac{P(x|C_k)P(C_k)}{P(x)} \quad (2)$$

In the above equation C_k represents class, while x represents feature vector.

2.6. Logistic Regression

Logistic regression is a statistical method widely used in predictive analysis. It was developed to model the relationship between one or more independent variables and a dependent variable, and it is particularly effective in linear classification problems [34].

Logistic regression is preferred because it offers a more interpretable structure compared to other multivariate analysis methods. Thanks to this flexibility, it is an effective method that can provide reliable and successful classification performance.

2.7. XGBoost

Extreme Gradient Boosting (XGBoost) is a high-performance machine learning algorithm based on the gradient boosting approach. It works by sequentially training decision trees to minimize the errors of previous models. XGBoost also demonstrates successful performance in parallel processing [35].

XGBoost relatively prevents overfitting thanks to its regularization mechanism. Performance is also improved through parallel computing and efficient memory usage. This makes it a classification method that stands out for its successful performance, especially on imbalanced datasets.

2.8. LightGBM

Light Gradient Boosting Machine (LightGBM) has a structure that uses gradient boosting-based decision trees. It is a machine learning algorithm that stands out for its high speed and efficient memory consumption [36]. LightGBM is successful in learning complex situations with its leaf-wise tree growth structure. The histogram-based learning approach reduces computational costs. LightGBM has become a popular method, especially for high-dimensional datasets.

2.9. XGBoost + LightGBM Soft Voting Ensemble

XGBoost + LightGBM Soft Voting Ensemble is an ensemble method based on combining the outputs of gradient boosting-based XGBoost and LightGBM algorithms. In the soft voting strategy, classification is performed using probabilities obtained by weighted averaging and combining the class probabilities produced by each base model. Combining the strengths of both algorithms, this method leverages XGBoost's regularization capabilities and LightGBM's fast learning ability to improve the model's classification performance.

2.10. Statistical Analysis

In this study, accuracy, recall, precision, specificity, and

F1-Score [37] were computed for a three-class classification problem using confusion matrices.

TP: Correct prediction of the positive class.

TN: Correct prediction of the negative class.

FP: Incorrect prediction of the positive class.

FN: Incorrect prediction of the negative class.

2.10.1. Accuracy

Accuracy represents the overall accuracy of the model by measuring the proportion of correctly classified examples out of all examples.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

2.10.2. Recall

Recall, known as sensitivity or true positive rate, measures a model's ability to correctly identify positive examples. It is the ratio of true positives to the total of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

2.10.3. Precision

Precision measures the proportion of correctly predicted positive examples in all positively predicted examples.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

2.10.4. Specificity

Specificity, known as the true negative rate, measures the proportion of correctly identified negative examples.

$$\text{Specificity}= \frac{TN}{TN+FP} \quad (6)$$

2.10.5. F1 Score

The F1 score is calculated as the harmonic mean of precision and recall rates, providing a balanced measure of both metrics.

$$\text{F1 Score}= \frac{2*(\text{Precision}*\text{Recall})}{\text{Precision}+\text{Recall}} \quad (7)$$

3. Results and Discussion

The performance of the Random Forest classifier was evaluated using macro-average Receiver Operating Characteristic (ROC) analysis, confusion matrices, and cumulative Area Under the Curve (AUC) scores.

Macro-average ROC-AUC was preferred in this study due to the imbalanced class distribution in the dataset. Micro-averaging aggregates the contributions of all samples and is therefore dominated by majority classes, potentially masking the performance on minority classes. In contrast, macro-averaging computes the ROC-AUC independently for each class and then averages the results, ensuring equal importance is given to all classes. This makes it a more appropriate and reliable metric for evaluating model performance in imbalanced multi-class classification problems.

3.1. Results obtained with Random Forest

An average AUC value of 0.949 ± 0.029 was obtained from the macro-average ROC curve. The average confusion matrix showed balanced classification performance between classes. Feature importance analysis based on impurity reduction revealed that the degree of spondylolisthesis was the most influential feature on classification.

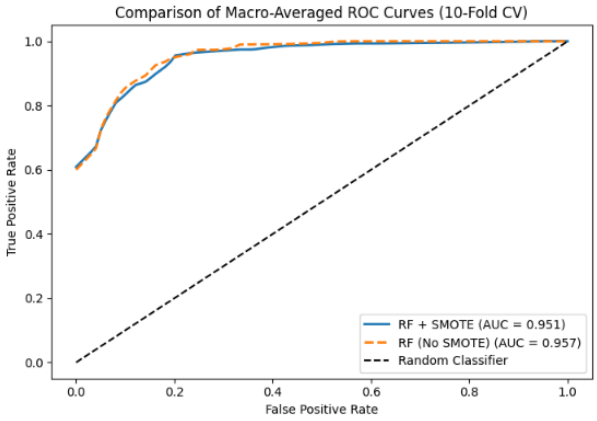

Figure 2. Macro-Averaged ROC Curves of Random Forest: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in the Figure 3, the Random Forest classifier has high discriminative power in both cases. The fact that the AUC value obtained without applying SMOTE (0.957) is only slightly higher than the value obtained with SMOTE (0.951) indicates that class imbalance in this dataset did not significantly negatively affect RF performance.

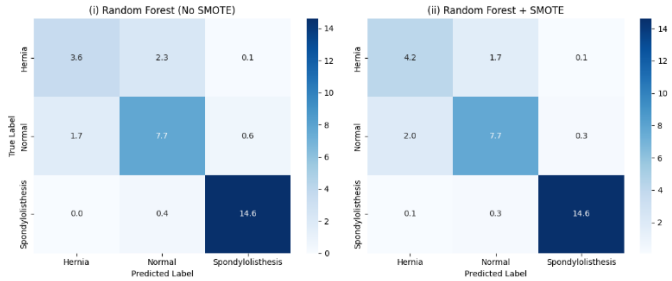

Figure 3. Average confusion matrices of Random Forest: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 4, the application of SMOTE has had a slightly positive effect on the Hernia class. No significant difference was observed for the other two classes. Looking at the overall situation in the confusion matrix, a very slight improvement was observed with SMOTE.

Table 2. Metrics obtained with Random Forest (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.85	0.06	0.84	0.06
2	Precision (Macro)	0.83	0.08	0.81	0.08
3	Recall (Macro)	0.81	0.07	0.78	0.06
4	F1-Score (Macro)	0.81	0.07	0.77	0.07
5	ROC-AUC (Macro, OvR)	0.95	0.04	0.95	0.03

As shown in Table 2, the SMOTE application provides consistent improvements in macro-average Precision, Recall, and F1 score, showing an increase of approximately 2-4% compared to the setting without SMOTE. The macro-average Recall increased from 0.78 to 0.81, indicating increased sensitivity towards minority classes. While the overall Accuracy remained similar in both settings (0.85 and 0.84), the SMOTE-based model showed higher Precision, Recall, and F1 scores. The macro-average ROC-AUC was high and close in both cases (0.95). This indicates that the nature of class separability is inherently strong.

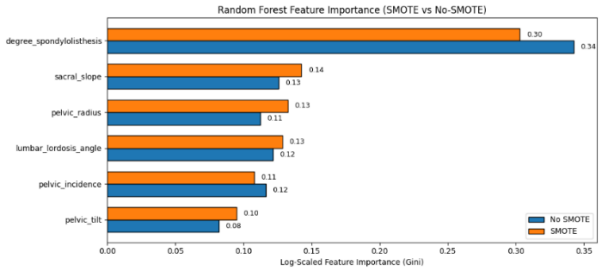

Figure 4. Comparison of Random Forest feature importance scores with and without SMOTE

In Figure 5, the feature importance analysis performed using the Random Forest model shows the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis emerges as the most dominant feature. This is followed by sacral tilt and lumbar lordosis angle. Although there are minor differences in the relative importance values between the two configurations, the overall feature hierarchy is similar. This indicates that the application of SMOTE does not alter the fundamental decision structure of the model.

3.2. Results obtained with SVM

The macro average ROC curve yielded an average AUC value of approximately 0.95 both when SMOTE was used and when SMOTE was not used. This value indicates high discrimination power between classes. The average confusion matrices showed stable and balanced classification performance with limited confusion between classes. The results in SVM demonstrate that the SVM model provides robust and consistent performance both with and without SMOTE.

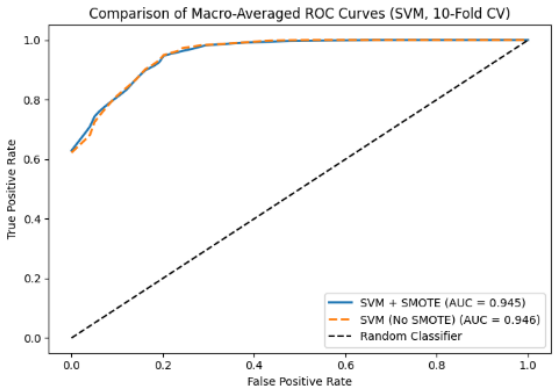

Figure 5. Macro-Averaged ROC Curves of SVM: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 6, the SVM model exhibits very similar macro-average ROC-AUC values when SMOTE is applied and when it is not applied (AUC ≈ 0.945–0.946). This indicates that SVM is relatively robust to class imbalance in the dataset. Both ROC curves remain well above the random baseline at all false positive rates.

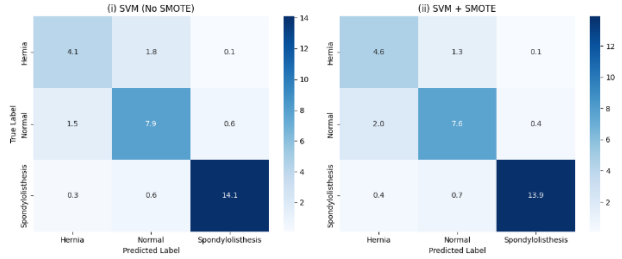

Figure 6. Average confusion matrices of SVM: **(i)** no SMOTE, **(ii)** SMOTE (10-fold CV)

As shown in Figure 7, the confusion matrices obtained from the SVM classifier indicate that the model achieves high true positive rates for the Spondylolisthesis class under both conditions. With the inclusion of SMOTE, a slight improvement is achieved in the correct classification of the Hernia class, while there is a very slight deterioration for the Normal and Spondylolisthesis classes. These results show that the results in SVM are close to each other when SMOTE is included and when it is not included.

Table 3. Metrics obtained with SVM (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.84	0.09	0.84	0.08
2	Precision (Macro)	0.82	0.10	0.82	0.10
3	Recall (Macro)	0.82	0.09	0.80	0.09
4	F1-Score (Macro)	0.81	0.10	0.81	0.10
5	ROC-AUC (Macro, OvR)	0.95	0.03	0.95	0.03

As shown in Table 3, the SVM model performs well both with and without SMOTE. While accuracy, precision, F1 score, and ROC-AUC values remain nearly identical in both settings, a slight improvement in the macro-average recall value is observed when SMOTE is applied.

3.3. Results obtained with Naive Bayes

The macro average ROC analysis yielded an average AUC value of approximately 0.95. These values indicate that the model performs well in distinguishing between classes. The average confusion matrices showed consistent classification performance across all classes. They yielded successful classification rates. Additionally, the results obtained without using SMOTE showed only a limited improvement in the performance of the Naive Bayes classifier.

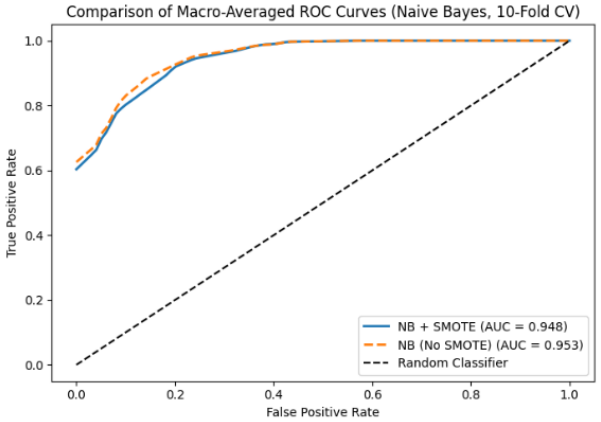

Figure 7. Macro-Averaged ROC Curves of Naive Bayes: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 8, the Naive Bayes model achieves a slightly higher value than the macro-average ROC-AUC value obtained when SMOTE is not used. When SMOTE was not used, the macro average AUC value was slightly higher. Overall, both ROC curves yielded successful results and remained well above the random baseline for all false positive rates.

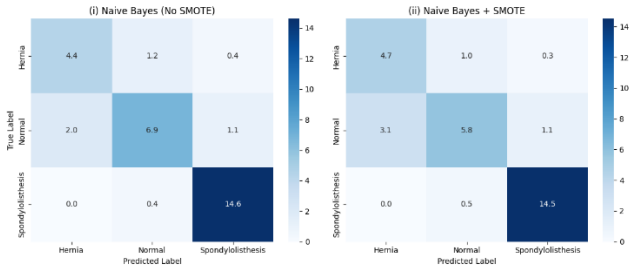

Figure 8. Average confusion matrices of Naive Bayes: **(i)** no SMOTE, **(ii)** SMOTE (10-fold CV)

As shown in Figure 9, the confusion matrices obtained from the Naive Bayes classifier indicate that the model achieves high true positive rates for the Spondylolisthesis class under both conditions. When SMOTE is applied, the classification performance for the Hernia and Normal classes is higher compared to when SMOTE is not applied. In particular, a significant difference of 1.1 points is observed for the Normal class. Looking at the overall results, it is seen that when SMOTE is not used, the Naive Bayes classifier successfully distinguishes all classes.

Table 4. Metrics obtained with Naive Bayes (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.81	0.05	0.84	0.07
2	Precision (Macro)	0.79	0.07	0.82	0.09
3	Recall (Macro)	0.78	0.06	0.80	0.08
4	F1-Score (Macro)	0.77	0.07	0.80	0.09
5	ROC-AUC (Macro, OvR)	0.95	0.03	0.95	0.03

As shown in Table 4, the Naive Bayes classifier performs better when SMOTE is not applied. Higher values were obtained without SMOTE for all metrics.

3.4. Results obtained with Logistic Regression

The macro average ROC analysis yielded an average AUC value of approximately 0.96. These values indicate that the model performed well in distinguishing between classes. The average confusion matrices showed good classification performance across all classes. Looking at the overall situation, the results obtained when SMOTE was used revealed that the Logistic Regression classifier performed more successfully.

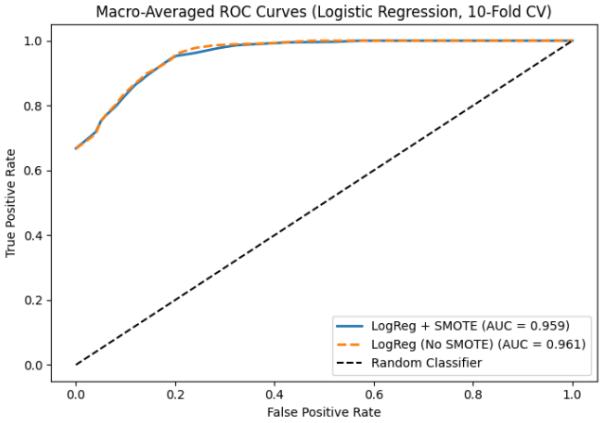

Figure 9. Macro-Averaged ROC Curves of Logistic Regression: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in the Figure 10, the macro-average AUC values are quite close to each other when SMOTE is applied and when it is not applied. Logistic regression performed quite well with an AUC value of approximately 0.96. These results show that logistic regression produced quite successful results in both SMOTE and non-SMOTE conditions.

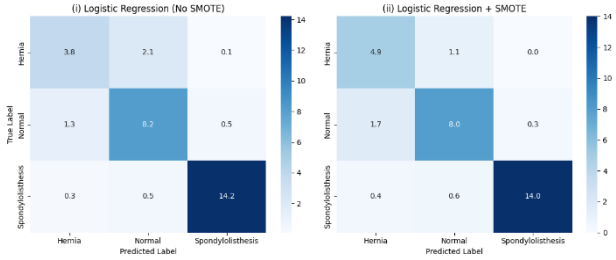

Figure 10. Average confusion matrices of Logistic Regression: **(i)** no SMOTE, **(ii)** SMOTE (10-fold CV)

As shown in Figure 11, when applied with the SMOTE logistic regression model, a small increase is observed in the accuracy of the Hernia class, while a very slight decrease is observed in the Spondylolisthesis and Normal classes.

Table 5. Metrics obtained with Logistic Regression (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.87	0.07	0.85	0.07
2	Precision (Macro)	0.85	0.08	0.83	0.10
3	Recall (Macro)	0.85	0.08	0.80	0.08
4	F1-Score (Macro)	0.84	0.08	0.80	0.09
5	ROC-AUC (Macro, OvR)	0.96	0.03	0.96	0.03

Table 5 shows the performance metrics obtained with and without SMOTE applied to the logistic regression model, indicating that SMOTE yielded more successful results in all metrics except ROC-AUC. With the use of SMOTE, a slight increase was observed in accuracy, macro-average precision, recall, and F1-score values. The ROC-AUC value was 0.96 in both cases. This indicates that the model has a high ability to distinguish between classes. It appears that SMOTE improves the balance between classes rather than its separation power.

3.5. Results obtained with XGBoost

An average AUC value of approximately 0.95 was obtained from the ROC curve when XGBoost was used with and without SMOTE. The average confusion matrix showed balanced classification performance between classes. Positive effects of applying SMOTE were observed in the performance metrics.

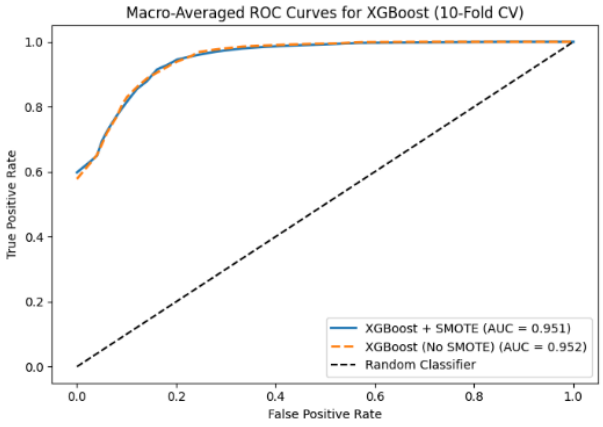

Figure 11. Macro-Averaged ROC Curves of XGBoost: (i) no SMOTE, (ii) SMOTE (10-fold CV)

The macro-average ROC curves presented in Figure 12 show that the XGBoost model achieves high AUC values both with and without SMOTE. In both cases, the AUC is approximately 0.95, and the fact that the curves are significantly above the random classifier line indicates that the model has a high level of ability to distinguish between classes.

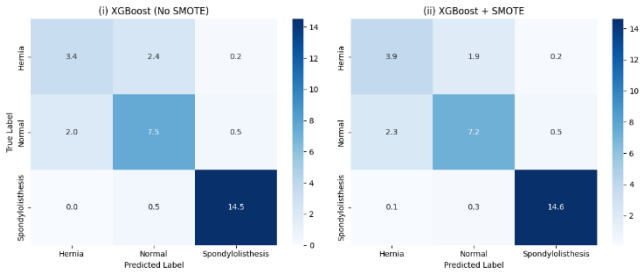

Figure 12. Average confusion matrices of XGBoost: **(i)** no SMOTE, **(ii)** SMOTE (10-fold CV)

As seen in the Figure 13, the classification performance of the XGBoost model before and after applying SMOTE is similar. Before SMOTE, the model was more successful in distinguishing the Hernia class, but relatively less successful in the Spondylolisthesis and Normal classes.

Table 6. Metrics obtained with XGBoost (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.83	0.05	0.82	0.06
2	Precision (Macro)	0.80	0.07	0.78	0.07
3	Recall (Macro)	0.78	0.06	0.76	0.06
4	F1-Score (Macro)	0.78	0.06	0.76	0.07
5	ROC-AUC (Macro, OvR)	0.95	0.03	0.95	0.03

Table 6 shows that the XGBoost model using SMOTE performed slightly better than the ROC-AUC for all metrics. The AUC value remaining stable at approximately 0.95 in both cases indicates that the model successfully distinguishes between classes.

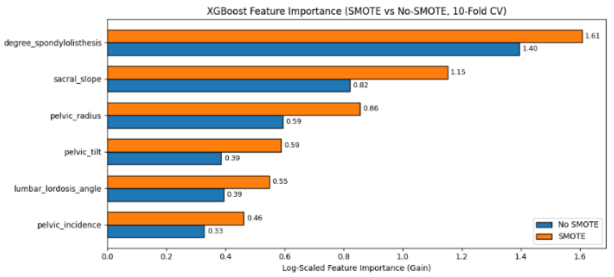

Figure 13. Comparison of XGBoost feature importance scores with and without SMOTE

Figure 14 shows that the feature importance analysis performed using the XGBoost model indicates the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis was observed to be the most dominant feature. This was followed by sacral tilt and lumbar lordosis angle. It was observed that the application of SMOTE did not change the basic decision structure of the model.

3.6. Results obtained with LightGBM

When LightGBM was used with and without SMOTE, an average AUC value of approximately 0.94 was obtained from the ROC curve. The average confusion matrix

showed similar performance in both cases. A slight positive effect of applying SMOTE was observed in the performance metrics.

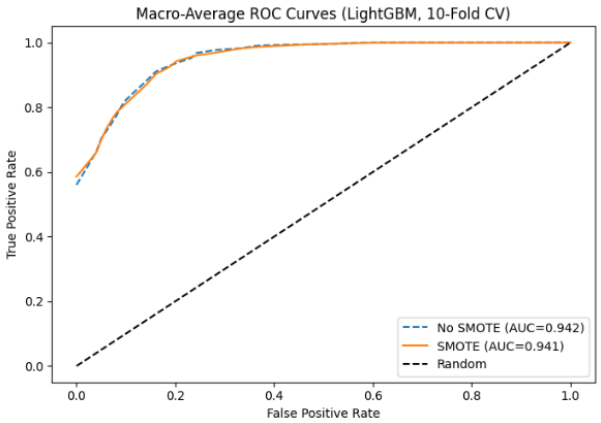

Figure 14. Macro-Averaged ROC Curves of LightGBM: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 15, the LightGBM model achieves an AUC value of approximately 0.94 with and without SMOTE applied, indicating nearly identical performance. Thus, the model demonstrates successful classification performance in both scenarios.

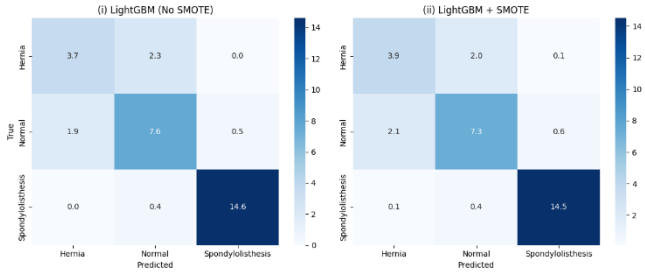

Figure 15. Average confusion matrices of LightGBM: **(i)** no SMOTE, **(ii)** SMOTE (10-fold CV)

As seen in the Figure 16, the confusion matrices of the LightGBM model before and after applying SMOTE show similar values for all classes. After applying SMOTE, a slight improvement is observed in the Hernia class. Before applying SMOTE, the performance of the Normal and Spondylolisthesis classes was slightly higher.

Table 7. Metrics obtained with LightGBM (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.84	0.04	0.83	0.05
2	Precision (Macro)	0.81	0.06	0.80	0.08
3	Recall (Macro)	0.78	0.05	0.78	0.07
4	F1-Score (Macro)	0.78	0.06	0.78	0.07
5	ROC-AUC (Macro, OvR)	0.94	0.02	0.94	0.03

As shown in Table 7, applying SMOTE results in a slight increase in the Accuracy and Precision (Macro) metrics, while the Recall, F1-Score, and ROC-AUC values remain nearly the same.

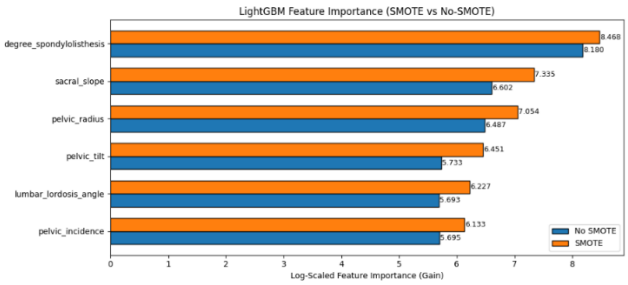

Figure 16. Comparison of LightGBM feature importance scores with and without SMOTE

Figure 17 shows the feature importance analysis performed using the LightGBM model, indicating the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis was observed to be the most dominant feature. This was followed by sacral tilt and lumbar lordosis angle. The other features were observed to have similar values.

3.7. Results obtained with XGBoost + LightGBM Soft Voting Ensemble

When XGBoost + LightGBM Soft Voting Ensemble was used with and without SMOTE, an average AUC value of approximately 0.94 was obtained from the ROC curve. The average confusion matrix showed similar performance in both cases. No significant positive effect has been observed in the performance metrics of the SMOTE application.

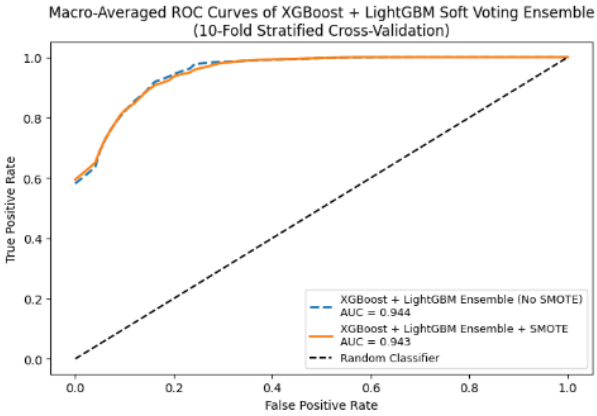

Figure 17. Macro-Averaged ROC Curves of XGBoost + LightGBM Soft Voting Ensemble: (i) no SMOTE, (ii) SMOTE (10-fold CV)

As shown in Figure 18, when SMOTE is applied with the XGBoost + LightGBM soft voting ensemble model, the AUC values obtained are approximately 0.94, which is almost the same as when SMOTE is not applied. It can be seen that the AUC value produced by the ensemble structure is close to the average of both algorithms.

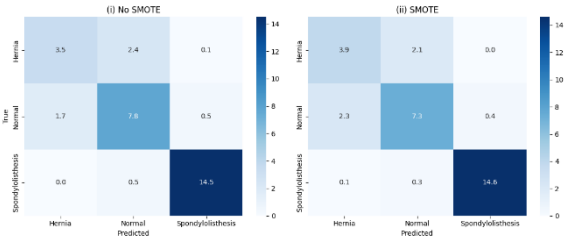

Figure 18. Average confusion matrices of XGBoost + LightGBM Soft Voting Ensemble: **(i)** no SMOTE, **(ii)** SMOTE (10-fold CV)

As shown in the Figure 19, the confusion matrices of the XGBoost + LightGBM Soft Voting Ensemble model before and after applying SMOTE show close values for all classes. After applying SMOTE, a slight improvement is observed in the Hernia class. Before applying SMOTE, the performance of the Normal and Spondylolisthesis classes was slightly higher.

Table 8. Metrics obtained with XGBoost + LightGBM Soft Voting Ensemble (Stratified 10-Fold CV)

number	metric	mean	std	mean (No-SMOTE)	std (No-SMOTE)
1	Accuracy	0.83	0.06	0.83	0.05
2	Precision (Macro)	0.80	0.08	0.79	0.07
3	Recall (Macro)	0.78	0.07	0.78	0.06
4	F1-Score (Macro)	0.78	0.08	0.78	0.06
5	ROC-AUC (Macro, OvR)	0.94	0.03	0.94	0.03

As shown in Table 8, the XGBoost + LightGBM Soft Voting Ensemble model produced very similar values both when SMOTE was applied and when it was not applied. Only a slight improvement in Precision (Macro) was achieved with SMOTE.

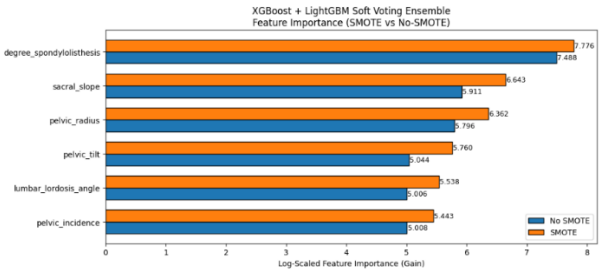

Figure 19. Comparison of XGBoost + LightGBM feature importance scores with and without SMOTE

Figure 20 shows the feature importance analysis performed using the XGBoost + LightGBM model, indicating the importance of features when SMOTE is used and when it is not used. In both cases, the degree of spondylolisthesis was again observed as the most dominant feature. This was followed by sacral tilt and lumbar lordosis angle. Other features also showed similar values.

3.8. Overall Results

The results obtained for all classifiers using both SMOTE and without SMOTE in the experiments conducted are presented in Table 9 below.

Table 9. Performance comparison of classifiers with and without SMOTE (Stratified 10-Fold CV)

Classifier	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC-AUC
Random Forest+SMOTE	0.85±0.06	0.83±0.08	0.81±0.07	0.81±0.07	0.95±0.04
Random Forest	0.84±0.06	0.81±0.08	0.78±0.06	0.77±0.07	0.95±0.03
SVM+SMOTE	0.84±0.09	0.82±0.10	0.82±0.09	0.81±0.10	0.95±0.03
SVM	0.84±0.08	0.82±0.10	0.80±0.09	0.81±0.10	0.95±0.03
Naive Bayes+SMOTE	0.81±0.05	0.79±0.07	0.78±0.06	0.77±0.07	0.95±0.03
Naive Bayes	0.84±0.07	0.82±0.09	0.80±0.08	0.80±0.09	0.95±0.03
Logistic Regression+SMOTE	0.87±0.07	0.85±0.08	0.85±0.08	0.84±0.08	0.96±0.03
Logistic Regression	0.85±0.07	0.83±0.10	0.80±0.08	0.80±0.09	0.96±0.03
XGBoost+SMOTE	0.83±0.05	0.80±0.07	0.78±0.06	0.78±0.06	0.95±0.03
XGBoost	0.82±0.06	0.78±0.07	0.76±0.06	0.76±0.07	0.95±0.03
LightGBM+SMOTE	0.84±0.04	0.81±0.06	0.78±0.05	0.78±0.06	0.94±0.02
LightGBM	0.83±0.05	0.80±0.08	0.78±0.07	0.78±0.07	0.94±0.03
XGBoost+LightGBM+SMOTE	0.83±0.06	0.80±0.08	0.78±0.07	0.78±0.08	0.94±0.03
XGBoost+LGBM	0.83±0.05	0.79±0.07	0.78±0.06	0.78±0.06	0.94±0.03

As seen in the experimental results, the Logistic Regression classifier achieved the highest performance across all metrics. In addition, it was observed that applying SMOTE had a positive effect on all classifiers except Naive Bayes.

Table 10. Wilcoxon Signed-Rank Test Results for SMOTE vs. No-SMOTE Models

Classifier	Accuracy (p-value)	F1-score (p-value)	ROC-AUC (p-value)	Interpretation
Random Forest	0.063	0.063	0.492	Not Significant
SVM	1.000	0.578	0.734	Not Significant
Naive Bayes	0.031	0.055	0.275	Only Accuracy Significant
Logistic Regression	0.313	0.078	0.678	Not Significant
XGBoost	1.000	0.688	0.432	Not Significant
LightGBM	0.750	1.000	0.820	Not Significant
XGBoost+LGBM	0.250	0.844	0.492	Not Significant

The Wilcoxon signed-rank test was applied to compare the performance of models with and without SMOTE across cross-validation folds. The reported p-values indicate whether the differences in performance metrics are statistically significant at the 0.05 significance level.

The results of the Wilcoxon Signed-Rank test indicate that the application of SMOTE does not lead to statistically significant improvements across most classifiers and evaluation metrics ($p > 0.05$). Only the Naive Bayes model demonstrates a statistically significant improvement in

terms of accuracy ($p < 0.05$). These findings suggest that the impact of SMOTE on model performance is generally limited and model-dependent, providing measurable benefit only in specific cases rather than consistently enhancing performance across all classifiers.

4. Discussion

The results obtained from comparative analyses show that the Logistic Regression method demonstrated the highest performance in almost every metric, both in SMOTE and non-SMOTE settings. This indicates that it provides high performance in class separation.

Feature importance rankings in SMOTE and no-SMOTE configurations showed similar characteristics across all applied models. Features related to spinal alignment, particularly the dominance of spondylolisthesis degree, sacral tilt, and lumbar lordosis angle, are in good agreement with established clinical knowledge regarding spinal pathologies. This consistency demonstrates that the oversampling strategy improves prediction balance without altering the fundamental biomechanical relationships learned by the model.

The limited sample size of the dataset used is the most significant limitation of this study. Since the dataset belongs only to a specific group of patients, the model's generalizability to different populations remains limited. Furthermore, it is believed that including clinical, demographic, or laboratory data in addition to biomechanical properties would yield more decisive and reliable results. The study compared three classical machine learning algorithms. The methods applied could be supported by deep learning-based models or hybrid methods if the dataset could be further expanded. Finally, since the dataset used was not fully balanced, class distribution differences may have caused performance drops in some algorithms.

In comparative analyses, it is generally observed that class separation is successfully achieved in both SMOTE and non-SMOTE settings. The improvements observed in macro-average accuracy, precision, recall, and F1 scores after applying SMOTE demonstrate its effectiveness. SMOTE has been shown to improve macro-average sensitivity and F1 scores, particularly in underrepresented classes, by synthetically balancing minority classes; however, these improvements remained within the standard deviation range in some models.

Despite promising results, this study has several limitations that should be considered. First, the experiments were conducted on a limited dataset, which may limit the generalizability of the findings to broader clinical populations. Second, although SMOTE effectively addresses class imbalance, it relies on synthetic sample generation and may not fully capture the complex distribution of minority class samples in the real world. Enhancing the dataset by increasing the number of real samples is a factor that could improve consistency.

The similarity in feature importance rankings across SMOTE and no-SMOTE configurations indicates the consistency of the Random Forest, XGBoost, and LightGBM models. Among the features related to spinal alignment, the degree of spondylolisthesis is particularly prominent in terms of feature importance. The consistent results obtained show that the model maintains a consistent prediction balance without altering the fundamental biomechanical relationships. This consistency is of significant clinical importance. The proposed approaches produce meaningful output not only in terms of classification performance but also in decision support systems for diagnosing spinal disorders.

Among all the machine learning models evaluated in this study, Logistic Regression consistently demonstrated the best overall performance across all evaluation metrics, particularly when combined with SMOTE, and achieved superior accuracy, recall, and F1 score values. Additionally, Random Forest and Support Vector Machine also demonstrated strong and stable performance with high ROC-AUC values and balanced classification results across classes. On the other hand, ensemble and boosting-based methods such as XGBoost, LightGBM, and their Soft Voting combination yielded competitive but not superior results compared to simpler models. These findings suggest that for this dataset, simpler and more interpretable models such as Logistic Regression may be preferred, especially when combined with class imbalance handling techniques like SMOTE.

The impact of SMOTE on ensemble-based models such as XGBoost and LightGBM has been widely reported in the literature as generally improving classification performance, particularly for imbalanced medical datasets. For instance, previous studies have shown that combining SMOTE with boosting-based and ensemble learning methods can enhance minority class detection and improve overall predictive performance.

Previous studies, such as the integration of SVM-SMOTE with ensemble learning for cerebrovascular disease prediction [38], have reported significant improvements in minority class detection and overall model performance, highlighting the effectiveness of combining data balancing techniques with ensemble methods. In the study Karamti et al. [39], SMOTE, when used in conjunction with ensemble learning and advanced preprocessing techniques, leads to improved performance. The highest accuracy of 99.9% was achieved when XGB, RF, and ETC were used together.

However, the findings of this study indicate a different trend. Although SMOTE provided slight improvements in certain metrics such as Precision and Recall, it did not result in a significant increase in overall performance for XGBoost, LightGBM, or their Soft Voting combination.

In particular, the macro-average ROC-AUC values remained largely unchanged after the application of SMOTE.

5. Conclusion

This study has demonstrated the effectiveness of machine learning methods in the early diagnosis and classification of orthopedic diseases. In the study, which was conducted using 7 different methods, accuracy, precision, recall, F1-Score, and ROC-AUC values were calculated. The results obtained indicate that biomechanical properties may be decisive in the diagnosis of orthopedic diseases. The most successful results were obtained using SMOTE with logistic regression, achieving an accuracy of 0.87, precision of 0.85, recall of 0.85, F1-Score of 0.84, and ROC-AUC of 0.96.

In addition to successful classification results, this study has some limitations that should be considered. The dataset is publicly available and limited. This may limit the generalizability of the findings to broader clinical populations. Furthermore, although SMOTE effectively addresses class imbalance, this method, which relies on synthetic sample generation, may not always fully capture the complex distribution of class samples in the real world.

It is anticipated that expanding the dataset, increasing the number of samples, and enriching the variety of biomechanical parameters will enhance the model's generalization capability. Furthermore, applying more complex deep learning models could further improve classification performance. It is recommended that future studies be expanded to include demographic factors such as different age groups and gender distributions.

Based on the data obtained from the evaluation of various criteria of orthopedic patients, this study can serve as a guide for disease detection and early diagnosis. It can also serve as a guide for similar studies to be conducted in the future. It is believed that more accurate results could be obtained if the content of the dataset could be expanded and the determining factors could be identified more clearly. The study makes an important contribution to the literature in terms of its concrete findings regarding disease detection.

To enhance the robustness and applicability of the proposed framework, future studies may evolve in various directions. The use of larger and more diverse datasets, including data from different populations and clinical settings, will improve the generalizability of the models. Incorporating clinical, demographic, and imaging data alongside biomechanical features could enable a more comprehensive representation of patient conditions and help achieve more consistent results. Furthermore, the application of advanced deep learning and hybrid model such as neural networks combined with ensemble learning strategies could further improve classification performance. Another key aspect is the integration of explainable artificial intelligence (XAI) techniques to interpret model decisions and enhance clinical confidence.

Declaration of Ethical Standards

The authors declare that this study was conducted in accordance with accepted ethical standards in research and publication. All authors have contributed honestly to the work, appropriate citations have been provided, and the manuscript represents original research that has not been published or submitted elsewhere.

Credit Authorship Contribution Statement

All authors contributed to the conception and design of the study. Data analysis and model development were performed collaboratively. All authors participated in the interpretation of results, drafting, and critical revision of the manuscript, and approved the final version for submission.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding / Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data used in this study are derived from the publicly available Biomechanical Features of Orthopedic Patients dataset.

References

[1] A. M. Elshewey and A. M. Osman, "Orthopedic disease classification based on breadth-first search algorithm," Scientific Reports, vol. 14, no. 1, Art. no. 23368, 2024.

[2] Y. Zhang, X. Liu, and H. Wang, "Spinal disease classification using SMOTE-RFE-XGBoost model," PeerJ Computer Science, vol. 9, e1280, 2023.

[3] L. Frizziero et al., "New methodology for diagnosis of orthopedic diseases through additive manufacturing models," Symmetry, vol. 11, no. 4, Art. no. 542, 2019, doi: 10.3390/sym11040542.

[4] S. Kıvrak, M. A. Aydın, and H. Polat, "Evaluation of machine learning models with SMOTE for imbalanced medical datasets," Diagnostics, vol. 14, no. 23, Art. no. 2634, 2024.

[5] Z. Yang et al., "A hybrid machine learning approach using SMOTE and ensemble methods for healthcare prediction," Scientific Reports, vol. 15, Art. no. 92722, 2025.

[6] Y. Gyasi-Agyei, "Comparative analysis of machine learning algorithms for medical diagnosis," Results in Engineering, vol. 21, Art. no. 101234, 2025.

[7] S. Rezapour, M. H. Mahoor, and R. L. Figueroa, "Machine learning-based gait analysis for orthopedic disorder prediction using SMOTE," arXiv preprint arXiv:2309.15990, 2023.

[8] H. B. Kibria and A. Matin, "The severity prediction of binary and multi-class cardiovascular disease—a machine learning-based fusion approach," Comput. Biol. Chem., vol. 98, Art. no. 107672, 2022, doi: 10.1016/j.compbiolchem.2022.107672.

[9] S. Li and X. Zhang, "Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm," Neural Comput. Appl., vol. 32, pp. 1971–1979, 2020, doi: 10.1007/s00521-019-04378-4.

[10] C. J. Ling et al., "Machine learning-based segmentation of images to diagnose orthopedic diseases and to guide orthopedic surgeries," Soft Comput., 2023, doi: 10.1007/s00500-023-08503-3.

[11] N. Rubaiyat et al., "Classification and prediction of orthopedic disease based on lumbar and pelvic state of patients," in Proc. IEEE Int. Conf. Electr., Comput. Commun. Technol. (ICECCT), pp. 1–4, 2019, doi: 10.1109/ICECCT.2019.8869540.

[12] N. Jahan et al., "Classification of orthopedic patients using supervised machine learning techniques," in Intelligent Computing and Optimization, Springer, pp. 659–669, 2021, doi: 10.1007/978-3-030-68154-8_61.

[13] K. Hasan et al., "A machine learning approach on classifying orthopedic patients based on their biomechanical features," in Proc. 7th Int. Conf. Informatics, Electron. Vis. (ICIEV) and 2nd Int. Conf. Imaging, Vis. Pattern Recognit. (icIVPR), pp. 289–294, 2018, doi: 10.1109/ICIEV.2018.8641004.

[14] Q. Chen, Y. Zhang, M. Zhang, Z. Li, and J. Liu, "Application of machine learning algorithms to predict acute kidney injury in elderly orthopedic postoperative patients," Clin. Interv. Aging, pp. 317–330, 2023.

[15] D. H. Mantzaris, G. C. Anastassopoulos, and D. K. Lymberopoulos, "Medical disease prediction using artificial neural networks," in Proc. IEEE Int. Conf. Bioinformatics and Bioengineering, pp. 1–6, 2008.

[16] S. W. Chung et al., "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm," Acta Orthop., vol. 89, no. 4, pp. 468–473, 2018, doi: 10.1080/17453674.2018.1453714.

[17] J. Olczak et al., "Ankle fracture classification using deep learning: Automating detailed AO/OTA 2018 malleolar fracture identification," Acta Orthop., vol. 92, no. 1, pp. 102–108, 2020, doi: 10.1080/17453674.2020.1837420.

[18] L. Cao, R. Li, D. Zhou, M. Zhao, and W. Huang, "Deep learning-based diagnosis and classification of femoral head necrosis," in Proc. 5th Int. Conf. Artif. Intell. Ind. Technol. Appl. (AIITA), pp. 1225–1228, 2025.

[19] A. M. Elshewey and A. M. Osman, "Orthopedic disease classification based on breadth-first search algorithm," Sci. Rep., vol. 14, no. 1, Art. no. 23368, 2024.

[20] UCI Machine Learning Repository, "Biomechanical features of orthopedic patients," 2023. [Online]. Available: Kaggle dataset.

[21] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," Inf. Sci., vol. 505, pp. 32–64, 2019.

[22] P. Dutta, S. Paul, and M. Majumder, "An efficient SMOTE based machine learning classification for prediction and detection of PCOS," 2021.

[23] Abacı, İ., & Yıldız, K. SMOTE vs. KNNOR: An evaluation of oversampling techniques in machine learning. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 13(3), 767-779, 2023.

[24] Z. Sun et al., "An improved random forest based on the classification accuracy and correlation measurement of decision trees," Expert Syst. Appl., vol. 237, Art. no. 121549, 2024.

[25] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," Comput. Stat. Data Anal., vol. 52, no. 4, pp. 2249–2260, 2008, doi: 10.1016/j.csda.2007.08.015.

[26] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[27] Y. Arora and S. K. Gupta, "Brain tumor classification using weighted least square twin support vector machine with fuzzy hyperplane," Eng. Appl. Artif. Intell., vol. 138, Art. no. 109450, 2024.

[28] S. Huang et al., "Applications of support vector machine learning in cancer genomics," Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[29] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," in Proc. Int. Conf. Adv. Technol. Eng. (ICATE), pp. 1–9, 2013, doi: 10.1109/ICAdTE.2013.6524737.

[30] O. N. Manjrekar and M. P. Dudukovic, "Identification of flow regime in a bubble column reactor with a combination of optical probe data and machine learning technique," Chem. Eng. Sci.: X, vol. 2, Art. no. 100023, 2019, doi: 10.1016/j.cesx.2019.100023.

[31] A. G. Abdulameer, A. S. Hammood, F. M. Abdulwahed, and A. A. Ayyash, "Naïve Bayes algorithm for timely fault diagnosis in helical gear transmissions using vibration signal analysis," Int. J. Interact. Des. Manuf., vol. 19, no. 5, pp. 3695–3706, 2025.

[32] Ö. Tonkal and H. Polat, "Traffic classification and comparative analysis with machine learning algorithms in software defined networks," Gazi Univ. J. Sci. C: Des. Technol., vol. 9, no. 1, pp. 71–83, 2021, doi: 10.29109/gujsc.869418.

[33] B. Cömert, "Alın bölgesinden alınan elektrookülogram (EOG) işaretleri için ölçüm devresi tasarımı ve sınıflandırılması," M.S. thesis, Balıkesir Univ., Inst. Sci., Balıkesir, Türkiye, 2016.

[34] H. Li et al., "Prediction of urban forest aboveground carbon using machine learning based on Landsat 8 and Sentinel-2," Remote Sens., vol. 15, no. 1, Art. no. 284, 2023, doi: 10.3390/rs15010284.

[35] A. Tasic et al., "Towards sustainable societies: Convolutional neural networks optimized by modified crayfish optimization algorithm aided by AdaBoost and XGBoost for waste classification tasks," Appl. Soft Comput., vol. 175, Art. no. 113086, 2025.

[36] S. Kakkar et al., "Enhancing energy efficiency and classification modeling through a combined approach of LightGBM and stratified k-fold cross-validation," Electr. Power Compon. Syst., pp. 1–19, 2024.

[37] S. Dörterler, "Hybridization of k-means and meta-heuristics algorithms for heart disease diagnosis," New Trends Eng. Appl. Nat. Sci., p. 55, 2022.

[38] Nithya, R., Kokilavani, T., & Beena, T. L. A. Balancing cerebrovascular disease data with integrated ensemble learning and SVM-SMOTE. Network Modeling Analysis in Health Informatics and Bioinformatics, 13(1), 12, 2024.

[39] Karamti, H., Alharthi, R., Anizi, A. A., Alhebshi, R. M., Eshmawi, A. A., Alsubai, S., & Umer, M. Improving prediction of cervical cancer using KNN imputed SMOTE features and multi-model ensemble learning approach. *Cancers*, *15*(17), 4412, 2023.