Machine Learning for Osteoporosis Detection from Medical Images

Saud Maher Abdul Razzaq and Baydaa Ibraheem Khaleel
Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, 41002 Mosul, Iraq
saud.23csp48@student.uomosul.edu.iq, baydaaibraheem@uomosul.edu.iq

Keywords: Machine Learning, Deep Learning, Classification, Osteoporosis, Knee, X-Ray Images.

Abstract: Osteoporosis is a disease characterized by decreased bone density resulting from a higher rate of bone loss compared to bone formation, which increases the risk of fractures. Therefore, early detection of osteoporosis helps reduce the risk of fractures. Traditional methods for diagnosing knee osteoporosis are time-consuming and require specialized personnel. Machine learning has emerged as an effective tool for image analysis, we proposed in research the development of an intelligent system for detecting knee osteoporosis using X-ray images. The research included both binary classification and multi-classification to differentiate images into five main categories: normal, Doubtful, mild, moderate, and severe osteoporosis, using two separate datasets. We proposed a new feature extraction technique called Hybrid Multi-Modal Feature Fusion (HMMFF), which integrates traditional feature extraction techniques, such as GLCM, HOG, and ORB, with deep learning models, including YOLO11, VGG-16, and ResNet50. Feature selection was performed using the harmony search optimization algorithm, and the selected features were then fed into several machine learning classifiers for training, evaluation, and classification. Experiments showed that the XGBoost model achieved the best performance compared to the other models, with an accuracy rate of 94.5% in binary classification and 93.3% in multi-class classification. The results indicate that the application of the HMMFF technique contributed to the extraction of comprehensive and diverse features. Furthermore, the feature selection using harmony search optimization helped identify important and effective features, thereby increasing the accuracy of classification and diagnosis of osteoporosis and consequently reducing the risk of fractures.

1 INTRODUCTION

Osteoporosis is a bone disorder that causes a progressive loss of bone mass and bone structure and makes the person susceptible to fractures. Early onset of osteoporosis is also troublesome because the disease develops with time. This highlights the significance of early diagnosis and successful control. The healthcare industry has been under notable technological developments such as the adoption of cloud computing services in the treatment of their patients, provision of big data analytics, strong data storage solutions, and safe and speedy access to patient records. These have led to the reason why there is a higher life expectancy and lower mortality rates resulting into a higher percentage of the population being composed of the elderly demographic which is more susceptible to osteoporosis than any other age category [1]-[3]. Several elements of the skeletal system are prone to osteoporosis which is a disease that is silent since the disease is usually manifested without obvious symptoms. This is a great challenge to early diagnosis and predisposes to bone fracture. The condition is caused by the lack of balance between the bone remodeling and bone resorption process, along with other risk factors, which include age, genetic, lifestyle, dietary, and gender factors, thus, it is a significant health burden [4], [5].

Conventional options of osteoporosis diagnosis are subject to various difficulties as the changes in the bone density are subtle and require special knowledge, which may result in different views and human error. These problems have led to more use of artificial intelligence, including machine learning and deep learning that have helped in better diagnostic accuracy and efficiency, and early detection, thereby lowering costs and offering better care solutions to people. Our design in this study is a proposed intelligent system to diagnose knee osteoporosis in the X-ray images in order to enhance the accuracy and efficiency of osteoporosis diagnosis. Knee bone analysis based on a set of machine learning algorithms, including Random Forest, ExtraTrees, Gradient Boosting, Bagging, XGBoost, and Decision Trees, will help to offer a new scheme within the frames of which the clinical diagnosis will be enhanced, and the economic, social, and psychological costs of this chronic disease will be minimized.

2 RELATED WORKS

A number of studies have investigated the application of intelligent techniques for osteoporosis detection and bone condition assessment, as the disease affects various parts of the skeletal system. The main findings of these studies are summarized below.

In 2020, Yousfi et al. [6] proposed a method for distinguishing osteoporosis patients from healthy individuals using texture analysis, genetic algorithms, and artificial intelligence applied to two-dimensional bone X-ray images. Genetic algorithms were employed to optimize the parameters of the gray-level co-occurrence matrix (GLCM), including the pixel distance coefficient and the number of gray levels used during preprocessing. In addition, genetic algorithms were utilized to select the optimal feature subset derived from run-length matrix (RLM) and GLCM features. The results demonstrated that combining genetic algorithms with GLCM and BSIF features improved the classification accuracy to 87.50%, compared with 77.8% when using GLCM features alone.

Lee et al. [7] developed a CNN-based framework for feature extraction from spinal X-ray images. The extracted features were subsequently classified using machine learning algorithms. Their best-performing configuration employed VGG for feature extraction and a Random Forest classifier, achieving a classification accuracy of 71% and an AUC of 0.858.

In 2022, Bello et al. [8] evaluated two transfer learning models, VGG16 and GoogLeNet, for osteoporosis classification using knee X-ray images. The dataset was analyzed using Python Scikit-learn and divided into grayscale and RGB image categories. GoogLeNet required 42 minutes for training on grayscale images and 50 minutes on RGB images, whereas VGG16 required 37 and 44 minutes, respectively. Performance evaluation showed that GoogLeNet achieved an accuracy of 90%, slightly outperforming VGG16, which achieved an accuracy of 87%.

Yang [9] investigated the use of deep learning techniques, including CNNs, VGG16, and a Late-Fusion approach, for the classification of knee X-ray images. The study demonstrated the effectiveness of deep learning methods in identifying osteoporosis-related changes in knee radiographs. Experimental results indicated that the VGG16 model achieved an accuracy of 82% on the osteoporosis dataset, while the Late-Fusion model obtained an accuracy of 77% on the osteoarthritis dataset.

Nakamoto et al. [10] proposed a computer-aided screening system for osteoporosis prediction using panoramic radiographs of women aged 50 years and older. Three deep CNN architectures—AlexNet, VGG16, and GoogLeNet—were evaluated. All models demonstrated strong agreement with expert radiologist assessments, with concordance rates ranging from 86.0% to 90.7%. For lumbar spine osteoporosis prediction, the models achieved sensitivities between 78.3% and 82.6%, specificities between 71.4% and 79.2%, and accuracies between 74.0% and 79.0%. For femoral neck osteoporosis prediction, sensitivities ranged from 80.0% to 86.7%, specificities from 67.1% to 74.1%, and accuracies from 70.0% to 75.0%.

Abubakar et al. [11] evaluated transfer learning techniques based on deep convolutional neural networks for osteoporosis classification in knee X-ray images. The study compared a standard VGG16 model with a fine-tuned VGG16 model. Performance was assessed using sensitivity, specificity, and accuracy metrics. The results showed that fine-tuning significantly improved model performance, increasing classification accuracy from 80% to 88%.

Sollmann et al. [12] investigated the use of convolutional neural networks for estimating volumetric bone mineral density from spinal X-ray images. The predicted values were compared with measurements obtained from conventional CT scans. The proposed CNN-based approach demonstrated promising diagnostic performance, achieving an AUC of 0.862.

In 2023, Kumar et al. [13] compared several pre-trained deep learning models for osteoporosis detection from knee X-ray images, including VGG16, VGG19, InceptionV3, Xception, DenseNet169, DenseNet201, ResNet101, and ResNet152. Using a dataset of 372 X-ray images divided into training and testing sets, the models were evaluated using accuracy, precision, recall, and F1-score metrics. Among all evaluated architectures, VGG16 achieved the highest accuracy of 86.36%, whereas ResNet101 demonstrated the lowest performance with an accuracy of 81.82%.

Naguib et al. [14] proposed a deep learning framework for classifying knee X-ray images into normal, osteoporosis, and osteopenia categories. The study introduced a Superfluity DL architecture and compared its performance with transfer learning models based on ResNet50 and AlexNet. Experiments conducted on two knee X-ray datasets demonstrated that the proposed architecture outperformed the benchmark models, achieving accuracies of 85.42% and 79.39% on the two datasets, respectively.

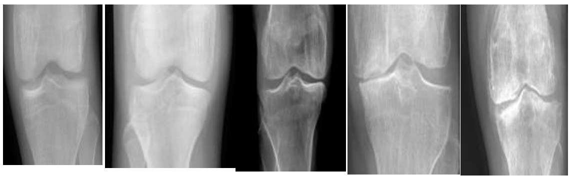
In 2024, Shen [15] presented a deep learning approach for osteoporosis detection using knee X-ray images. The study employed the VGG19 architecture to extract detailed features from two-dimensional radiographs. Using datasets obtained from Mendeley and Kaggle, the proposed model achieved an accuracy of 89%, demonstrating the potential of deep learning techniques for efficient and non-invasive osteoporosis screening.

3 METHODOLOGY

In this research, we attempt to analyze X-ray images of the knee by developing a proposed intelligent system to assess the condition of the bones and detect the presence of the disease. The proposed system is implemented in two stages, utilizing a database comprising two separate sets. The first is to implement binary classification to distinguish between normal images and images affected by osteoporosis, and the second is to implement multi-classification into cases that represent degrees of osteoporosis, such as normal knee bone condition, Doubtful condition, mild, moderate, and severe osteoporosis, using multiple machine learning models to perform the classification process and compare the results of these models.

3.1 Data Set

The X-ray image database comprised two separate sets of common cases of knee osteoporosis, obtained from the Kaggle website. Each set contained 1650 X-ray images for binary and multiclass classification. The database was divided into 75% training data and 25% testing data. Figure 1 illustrates the classification of X-ray images of knee osteoporosis. Table 1. shows the number of samples in each category for binary classification, while Table 2. shows the number of samples in each category of knee osteoporosis in the multiclass classification dataset.



Normal **Doubtful** **Mild** **Moderate**

Figure 1: Categories of X-ray images of knee osteoporosis cases.

Table 1: Number of samples for each category in binary classification.

Class	Total
Normal	503
disease	1147

Table 2: Number of samples for each category of knee osteoporosis in the multiple classification.

Class	Total
Normal	503
Doubtful	488
Mild	232
Moderate	221
Severe	206

3.2 Pre-images Processing

One of the major challenges facing medical data processing in general, and medical images in particular, is the problem of distinguishing between medical images captured of different types of the human body, this is essential for detecting the presence of disease by detecting, identifying, and analyzing abnormalities in the data [16]. This is accomplished by performing preprocessing operations on the input images to identify and distinguish them, this is accomplished by improving the image by removing noise, increasing contrast, and smoothing the details, image preprocessing is a fundamental and important process for feature extraction [17].

In our method, image processing is performed through a series of steps when processing X-ray images to facilitate computational complexity, the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique is applied to improve contrast, address inhomogeneous lighting in the image, and also improve the quality of images with unclear details, the Gaussian Blur technique is then applied to reduce noise in the image, improve image clarity, and highlight important details, during the implementation of the Sharpening technique, the Morphological Gradient was also applied to identify and highlight the edges in the images. In order to help increase the efficiency of the calculation process and reduce the complexity, the resulting image will be clearer and more prepared to extract important features from the input image, Figure 2 shows the pre-processing process for the knee X-ray images in our study.

3.3 Features Extracting

Algorithms used for object discrimination, classification, or diagnosis rely heavily on the extracted features that are entered into these algorithms to implement what is required of them; in the event of the presence of unimportant and unnecessary features, this may cause problems in the operation of the algorithms and also affect their efficiency in implementation due to the increase in the search space for the problem to be solved and thus the increase in the size of data and the time required to perform the task, whether for classification, detection or discrimination [18]. Therefore, the process of extracting features is considered a basic and important element in the image processing system, as the feature is represented through a set of values that represent a description of the details and contents of the image to be used according to the goal of the system, such as diagnosis, discrimination, classification or grouping [19].

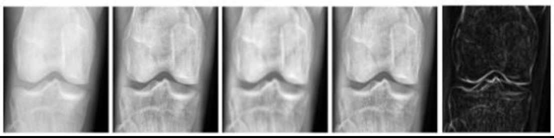


Figure 2: Pre-processing process of knee X-ray images.

Our paper proposes a new feature extraction method called Hybrid Multi-Modal Feature Fusion (HMMFF), the proposed feature extraction and fusion method relies on a comprehensive, multi-level approach, this fusion aims to leverage the complementary capabilities of each set of features, techniques such as HOG, GLCM, and ORB provide features based on local patterns, textures, and geometry, respectively. VGG16, YOLOv11, and ResNet50 provide more abstract hierarchical representations learned from large amounts of data, the features extracted from each method are then combined and unified to obtain an effective representation of the image, which is then selected and fed into machine learning models for classification.

The Gray Level Co-occurrence Matrix (GLCM) is used to analyze images through texture analysis, which shows how grayscale values are distributed in an image by counting the number of times a pair of grayscale values co-occurs at a given location [20]. GLCM features are extracted at a distance of 1 and for angles (0, 45, 90, 135) degrees with 256 gray levels, five texture attributes are then calculated, including contrast, dissimilarity, homogeneity, energy, and correlation, to characterize the spatial relationships between pixel densities, the Gradient Oriented Gamma Gradient (HOG) features are then calculated, this is an algorithm in computer vision used to extract features that match images by detecting objects within them and identifying their corresponding features [20]. It is done using cells with dimensions of 16 × 16 pixels and a block size of 1 × 1, resulting in a feature vector containing 1,177 features, the Oriented Fast and Rotated Brief (ORB) algorithm is also employed, which is a commonly used algorithm in computer vision for identifying, recognizing, and describing the characteristic points of an image to extract features, its properties are rotational and invariant to variations in illumination [21]. ORB descriptions are created with a maximum of 100 key points, to ensure consistent dimensions, and a flat vector of 500 dimensions representing local invariant features is obtained.

Deep learning models such as YOLO11, VGG-16, and ResNet50 were also used to extract only very fine-grained features without performing classification, this was done by disabling the fully connected layer in these models, which is responsible for implementing classification in the deep learning model, deep convolutional features are extracted from the VGG-16 model, which is pre-trained on ImageNet, resulting in a vector containing 512 dimensions that represent high-level features, the features were also extracted using the YOLOv11 model and the ResNet50 model, both of which were pre-trained, resulting in feature vectors containing 128 and 2,048 features, respectively, the final feature representation was then constructed by combining all the extracted features, resulting in a comprehensive vector containing 4,385 features for each image that represent low-level texture information, geometric patterns, local keypoint descriptors, and high-level semantic features, this created a robust and discriminative feature space suitable for subsequent classification tasks, this combined vector, after performing feature selection, was then applied as input to subsequent classification models, Figure 3 shows the steps of the HMMFF method for feature extraction.

3.4 Feature Selection

The feature selection procedure will help to reduce the feature space by taking a subset of features extracted which are more significant and affecting as compared to the initial set of features. They are then used to create and attach machine learning models, and on this basis, improve the performance of the models [22], [23].

The feature selection application of the harmony search optimization algorithm was used in our research. This approach allows the solution space of the feature search algorithms to be explored more widely, thus improving the generalizability of the model, its resistance to variations in the data distribution. The proposed feature selection method proved its efficiency in implementing feature selection, where 2214 features were selected for each image in binary classification and 2195 features were selected for each image in multi-class classification from a total of 4385 features for each image. The approach used in the study helps in discovering an optimal subset of features by reducing the complexity of the resulting feature set after merging, which will be used in machine learning models for classification.

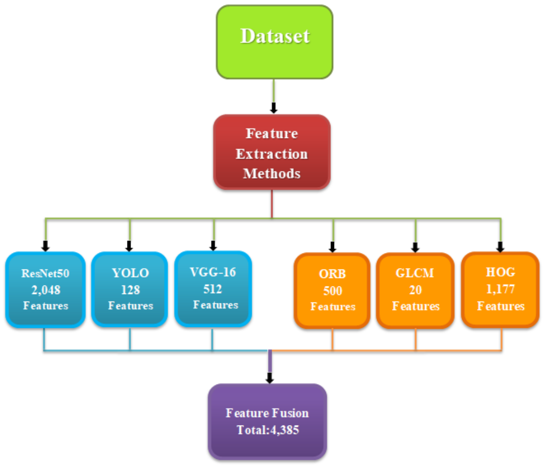


Figure 3: HMMFF method for feature extraction.

3.5 Harmony Search Algorithm

The Harmony search algorithm is a new computational method that was first proposed in 2001 by the Korean scientist Zong Woo Geem, it is considered one of the metaheuristic algorithms inspired by the musical improvisation process, which aims to enhance musical harmony observed in musicians while improvising a new piece of music [24]. It is based on the idea of searching for optimal harmony, much like musicians do when composing music, it relies on collaboration and collective search to efficiently explore the solution space, each musical note represents a variable in the problem, while the complete musical harmony represents a potential solution to the problem at hand The foundation of HAS consists of a harmony memory update equation, a harmony memory acceptance rate update equation, and the use of a pitch modulation rate to determine the optimal solution, [25], [26]. Figure 4 illustrates the flowchart of the harmony search algorithm [27].

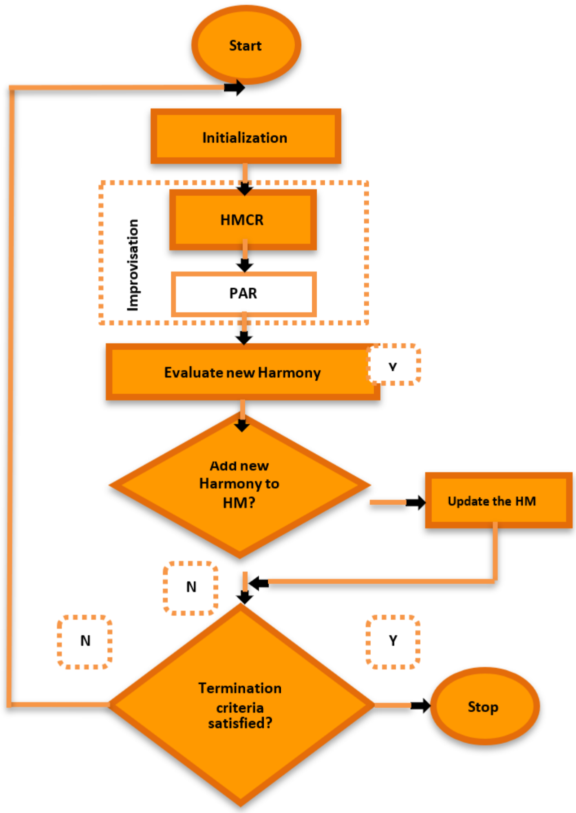


Figure 4: Flowchart of the harmony search algorithm.

Main equations of the Harmony search algorithm.

The following equations are used to express the Harmony search algorithm through the following steps [28], [29]:

1) Harmony Memory Initialization.

An array representing the HM of a specific size is created, containing random initial solutions, where each row represents a possible solution:

$$HM = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{HMS,1} & x_{HMS,2} & \cdots & x_{HMS,D} \end{bmatrix}, \qquad (1)$$

where HMS is the size of the compatibility memory and D is the number of variables in each solution.

2) Generating a New Harmony.

For each variable in the new solution (x_j') The value is chosen in two ways:

- Either from the harmony memory with probability HMCR.
- Or it is generated randomly with probability 1 - HMCR.

If the value is chosen from memory, it can be modified with probability PAR.

$$x_j' = \begin{cases} x_{i,j} \in HM & HMCR \\ x_j^{rand} \in [x_j^{min}, x_j^{max}] & (1 - HMCR) \end{cases}. \quad (2)$$

If the tone is modified:

$$x_j' = x_j' \pm rand() \times bw. \quad (3)$$

Where bw is the modulation range.

3) Evaluate New Harmony.

Calculate the objective function $f(X')$ for the new solution and find how good it is compared to other solutions.

4) Update Harmony Memory.

If the new solution is better than the current worst solution in HM, it is replaced.

If: $f(X') < f(\text{worst in HM}) \Rightarrow$
Replace worst with X'

5) Check Stopping Criterion.

The search terminates when a specified number of iterations is completed or a target value is reached.

3.6 Performance Metrics

To evaluate the proposed system for detecting knee osteoporosis based on knee X-ray images, we relied in this research on several standard criteria such as Precision, Recall, and F1-score in valuation the effectiveness of the system by relying on the import library (confusion Matrix) in the Python language to calculate the system performance evaluation metrics specified below [30]:

1) Accuracy. It shows what proportion of the system's predictions are accurate:

Accuracy = TP /TP + FP *100. (4)

2) Precision. The ratio of correct positive forecasts to all positive predictions is known as accuracy:

Precision=TP/ TP + FP. (5)

3) Recall. Recall expresses the ratio of correct positive predictions to the total actual positive samples:

Sensitivity-(Recall)= TP /TP + FN. (6)

4) F1-score. It is the harmony mean of precision and recall, and is used to evaluate the overall performance of the model:

F1-score=2*(Precision * Recall)/(Precision + Recall). (7)

3.7 Proposed Method

The proposed model for detecting and classifying osteoporosis was developed using several machine learning algorithms, such as Random Forest, ExtraTrees, Gradient Boosting, Bagging, XGBoost, and Decision Trees. The classification process for the input images is carried out after applying necessary steps, such as data balancing to address imbalances in the datasets and prevent bias in the algorithm predictions. Preprocessing of the input images is then applied to prepare them for feature extraction using Hybrid Multi-Modal Feature Fusion (HMMFF) technology, which is implemented using traditional techniques such as GLCM, HOG, and ORB, and deep learning models such as YOLO, VGG-16, and ResNet50, while disabling the Fully Connected layer responsible for classification in the model. The extracted features were merged into two sets to obtain a comprehensive and efficient representation of the image, resulting in a total of 4,385 extracted features for each image. Subsequently, a selection of the merged features was performed using the harmony search optimization algorithm, resulting in the selection of (2,214) features. A feature in the binary classification, as well as the selection of (2195) features for the multi-classification, will be entered into the proposed machine learning models to perform the binary and multi-classification processes for the condition of the bones in the knee, and then to compare the results of the different machine learning models that were used. Figure 5 below shows the proposed method of work.

4 RESULTS AND DISCUSSION

This section presents an evaluation of the results of the proposed system based on a number of basic metrics in evaluating the effectiveness of selected machine learning models such as Precision, Recall, F1-score and Accuracy.Binary, and multi-classification was made in the Python language and on two separate datasets with the aim of comparing and determining the effectiveness and efficiency in the detection of osteoporosis in knee x-ray images.

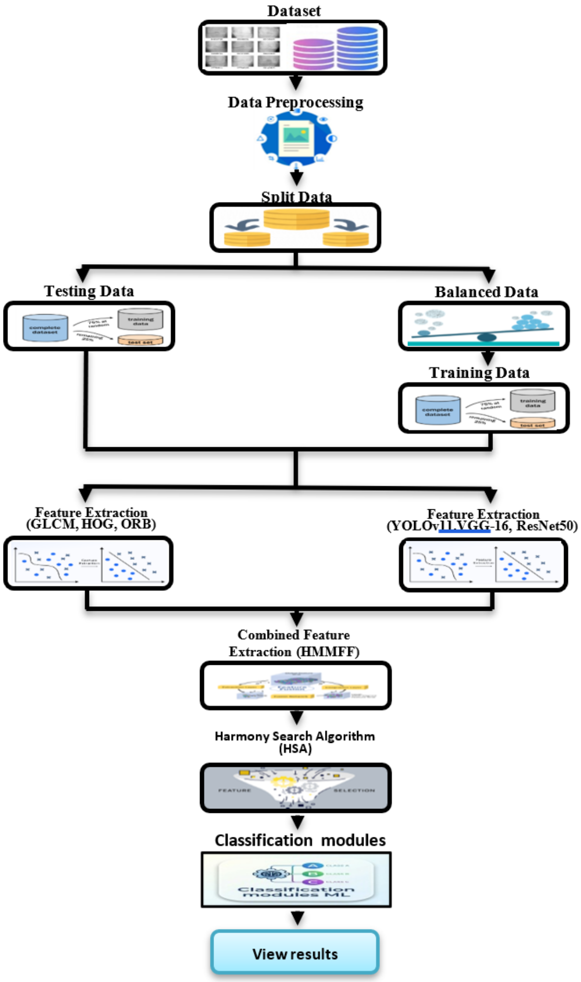


Figure 5: Flowchart of the proposed method of work.

4.1 Binary Classification Results

Table 1. presents the binary classification data to be used in the training and testing of the proposed machine learning models. The proposed system has been applied in the binary and multi-factor classification based on a comparative methodology, where the effectiveness of each of the models is gauged by its effectiveness in classifying the data. The experiments showed that the various machine learning models had varying performance in the binary classification process, which seeks to distinguish X-ray images to obtain normal bone images and osteoporotic bone images. Table 3. presents the results that were achieved and shows the performance of the proposed machine learning models during binary classification. The XGBoost model performed the most, compared to the other machine learning models and this is because it has 94.5 accuracy rate. It also got the highest scores in the other measures with Precision of 94% and Recall and F1-score of 93, respectively. This shows that the model effectively draws the line between the two categories and contributes to minimizing the number of false positives and false negatives. In the meantime, the Gradient Boosting model performed the second best with the accuracy rate of 94.2. The Bagging and Random Forest models had the accuracy rates of 93.3 and 93%. The ExtraTrees model has the highest level of accuracy at 92.7% with the Decision Tree being the worst of the suggested models with the lowest level of accuracy at 90.8%. The results of these models are given in a breakdown by the category in Table 4. We observe that the XGBoost model had an equal performance with F1-score of 0.91 and 0.96 with both normal and disease groups and precision value of 0.95 and a recall value of 0.97 in the classification of the fragility condition. These results are also illustrated graphically in the confusion matrices provided in Figure 6. Figure 6e of the confusion matrix of the XGBoost model indicates that the misclassified sample is not very high, it is lying outside the main diagonal, this indicates that the data balancing process was successful and the model was able to learn the two categories.

4.2 Multi- Classification Results

The multi-class model of classification of five categories used to determine the level of osteoporosis in the knee including the following categories, namely, Normal, Doubtful, Mild, Moderate, and Severe, was implemented based on the data presented in Table 2. The proposed machine learning models were also trained and tested and the experiment showed that the XGBoost model performed best compared to other machine learning models and the rate of classification accuracy is 93.3 percent. It comes after the Gradient Boosting model, as seen in Table 5. which also had outstanding performance with an accuracy rate of 93%. An evaluation of the performance of the other classifiers, the Random Forest and Extra Trees models, showed a very good performance and accuracy of 91.18 and 90.88 respectively. In addition, Table 6 presents a closer insight into the results of all the five categories, which is valuable when it comes to identifying the strong points of each model in this classification. The XGBoost model showed very fine results in predicting severe cases with the recall rate being 100 per cent. This implies that all the severe cases were identified by the model and this is important in medical needs because cases with severe cases may be overlooked. Moderate and mild cases also had very high F1 values of 0.97 and 0.93 respectively by the model. The model however showed a certain weakness in the ability to differentiate the doubtful cases with a recall rate of 70 only indicating that this category is the hardest to categorize. Gradient Boosting model was strong in moderate and severe categories with the F1 score of 0.97 and 0.98, respectively. Nevertheless, the model was not able to differentiate the Doubtful category, as it could be seen in the confusion matrices in Figure. 7. There is a majority of errors in the classification made among the similar categories, e.g., between Doubtful and Mild cases, as demonstrated by these matrices. This is because of the visual resemblance of these initial stages of the disease. Moreover, the nature of medical imaging data to extract the characteristic features of knee X-rays resulted in different performances of the proposed machine learning models especially at the initial stages of the disease where alterations are subtle and hardly noticeable on bone shape and structure. Conversely, in moderate and severe cases, one can observe evident structural alterations on the X-ray images and therefore can be easily differentiated.

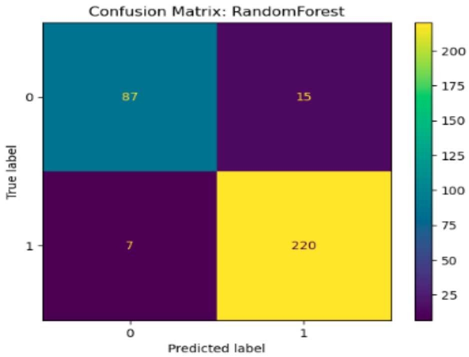
The results obtained indicate the potential of using the proposed system as a diagnostic aid, particularly in medical settings that lack specialized expertise in interpreting bone X-ray images. This system provides a reliable initial assessment of obvious cases, as well as an initial evaluation of cases that require more detailed, specialized review. Figure 8 illustrates a comparison of the performance of machine learning models in binary and multi-class classification.

Table 3: Performance results of the proposed machine learning models in binary classification.

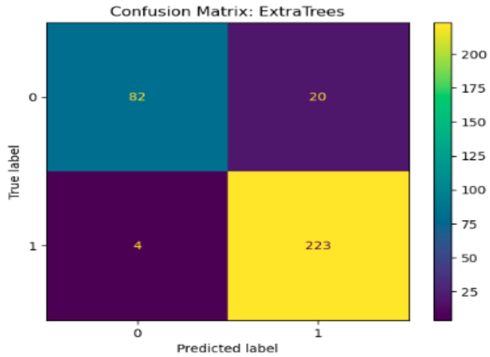
Classifier	Accuracy	Precision	Recall	F1
Random Forest	0.933131	0.930851	0.911052	0.920068
ExtraTrees	0.927052	0.935592	0.893150	0.910638
Gradient Boosting	0.942249	0.938055	0.925758	0.931564
Bagging	0.930091	0.919229	0.916947	0.918075
XGBoost	0.945289	0.940476	0.930660	0.935349
Decision Tree	0.908815	0.891966	0.896130	0.893999

Table 4: Detailed performance results of machine learning models in binary classification.

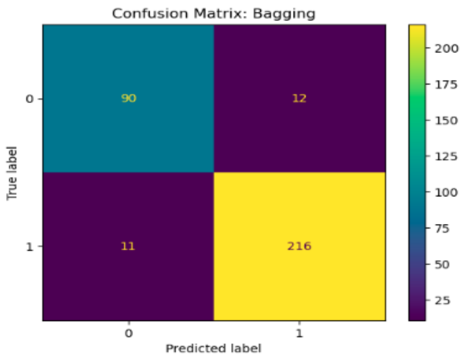
Proposed model	Disease Class	Precision	Recall	F1-score
Random Forest	normal	0.93	0.85	0.89
	disease	0.94	0.97	0.95
ExtraTrees	normal	0.95	0.80	0.87
	disease	0.92	0.98	0.95
Gradient Boosting	normal	0.93	0.88	0.90
	disease	0.95	0.97	0.96
Bagging	normal	0.89	0.88	0.89
	disease	0.95	0.95	0.95
XGBoost	normal	0.93	0.89	0.91
	disease	0.95	0.97	0.96
Decision Trees	normal	0.85	0.86	0.85
	disease	0.94	0.93	0.93



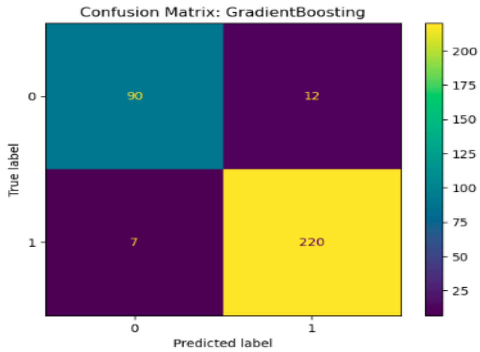
(a)



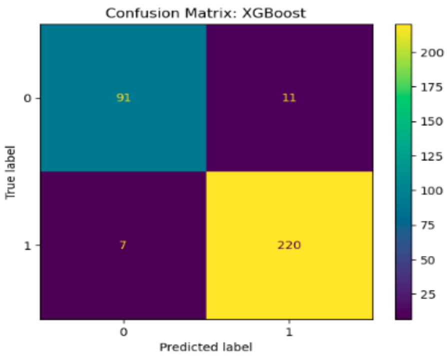
(b)



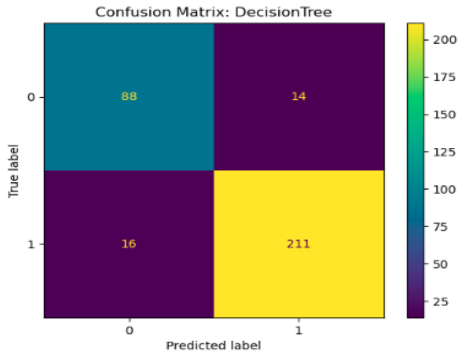
(c)



(d)



(e)



(f)

Figure 6: Confusion matrices of machine learning models in binary classification: a) Random Forest, b) ExtraTrees, c) Gradient Boosting, d) Bagging, e) XGBoost, f) Decision Trees.

Table 5: Performance results of the proposed machine learning models in multi-classification.

Classifier	Accuracy	Precision	Recall	F1
Random Forest	0.911854	0.932369	0.896268	0.909163
Extra Trees	0.908815	0.927959	0.895364	0.905295
Gradient Boosting	0.930091	0.938286	0.915973	0.924630
Bagging	0.908815	0.913551	0.900279	0.905935
XGBoost	0.933131	0.937307	0.918992	0.922268
Decision Tree	0.844985	0.830824	0.833609	0.831078

Table 6: Detailed performance results of machine learning models in multi-classification according to each disease category.

Proposed model	Disease Class	Precision	Recall	F1-score
Random Forest	normal	0.84	0.94	0.89
	Doubtful	1.00	0.72	0.84
	Mild	0.89	0.91	0.90
	Moderate	0.93	0.96	0.95
	Severe	1.00	0.95	0.97
ExtraTrees	normal	0.83	0.97	0.89
	Doubtful	1.00	0.70	0.82
	Mild	0.91	0.91	0.91
	Moderate	0.95	0.92	0.94
	Severe	0.95	0.98	0.96
Gradient Boosting	normal	0.87	0.97	0.92
	Doubtful	0.95	0.77	0.85
	Mild	0.93	0.91	0.92
	Moderate	0.97	0.96	0.97
	Severe	0.98	0.98	0.98
Bagging	normal	0.87	0.92	0.89
	Doubtful	0.93	0.81	0.86
	Mild	0.89	0.91	0.90
	Moderate	0.93	0.94	0.94
	Severe	0.95	0.93	0.94
XGBoost	normal	0.91	0.96	0.93
	Doubtful	1.00	0.70	0.82
	Mild	0.91	0.95	0.93
	Moderate	0.95	0.98	0.97
	Severe	0.91	1.00	0.95
Decision Trees	normal	0.85	0.86	0.86
	Doubtful	0.76	0.68	0.72
	Mild	0.76	0.86	0.81
	Moderate	0.90	0.88	0.89
	Severe	0.88	0.88	0.88

4.3 Comparison with Related Works

The results achieved in this research demonstrate the efficiency of the proposed system in classifying knee X-ray images with high accuracy compared to previous work mentioned in Table 7. The proposed model achieved an accuracy rate of 93.3% in multi-class classification and 94.5% in binary classification, surpassing the accuracy rates reported in all the aforementioned previous studies. Furthermore, it was evaluated on two database sets containing 1650 images for both binary and multi-class classification, the largest number of images used in previous studies. This increases the reliability and generalizability of the results compared to some previous studies that relied on smaller datasets. Additionally, the superior performance of the proposed model is attributed to the use of the HMMFF method, which combines features extracted using traditional techniques with higher-level features such as YOLOv11, VGG-16, and ResNet50. This resulted in a comprehensive and informative feature vector comprising 4385 features, representing diverse aspects of the image, from low-level fine details to high-level features. This provides an effective representation of the pathological condition. The use of harmony search optimization in the feature selection process further facilitated precise exploration of the feature space. And efficiency in selecting an ideal subset of the most influential and effective features. The number of features was reduced to 2,214 features in the binary classification and 2,195 features in the multi-class classification. This contributed to providing a comprehensive and effective representation of bone features in X-ray images, followed by their classification and achieving the best results, which demonstrates the efficiency of integrating traditional and advanced techniques in improving the automated diagnosis of osteoporosis.

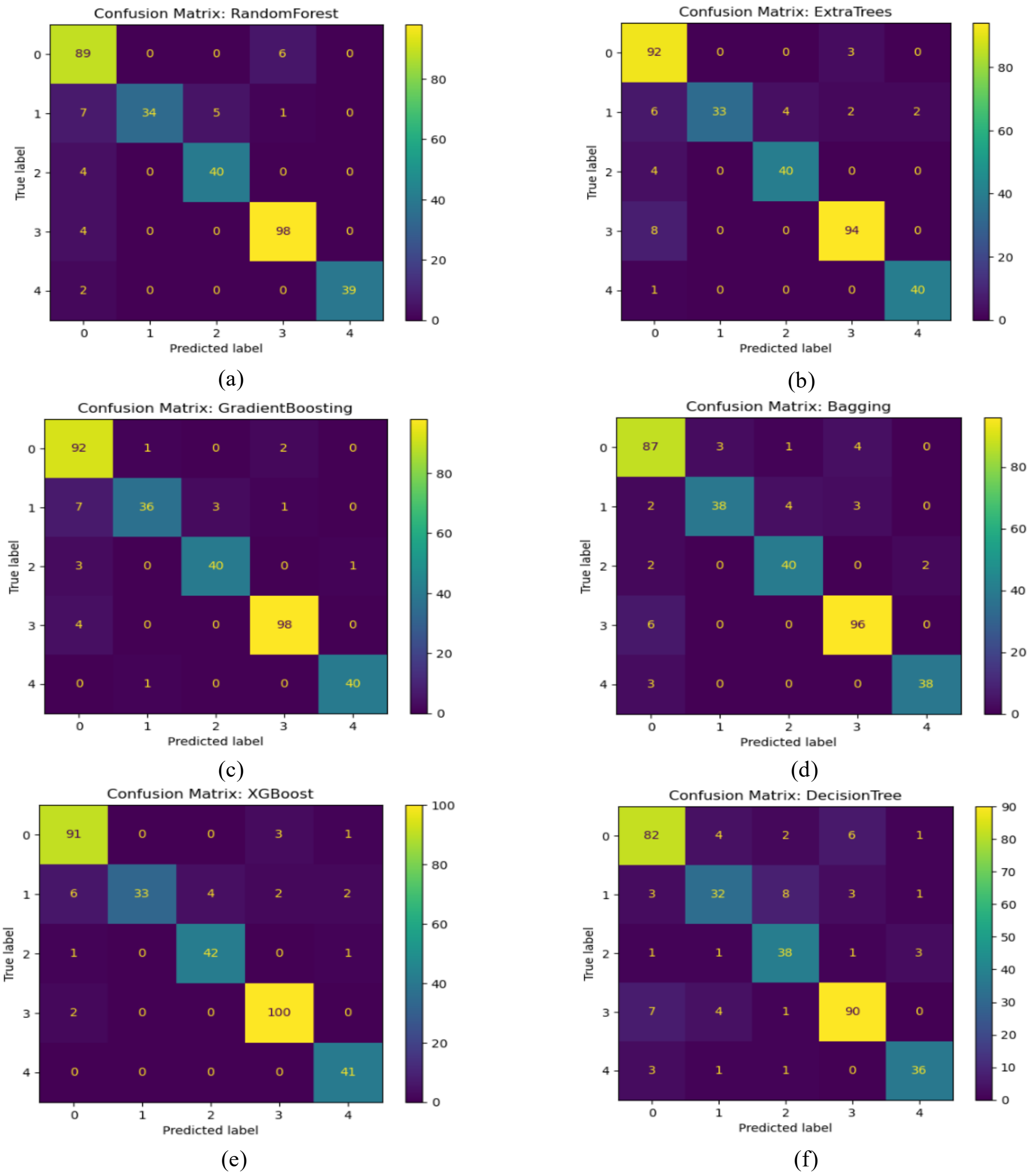


Figure 7: Matrices for machine learning models in multi-classification: a) Random Forest, b) ExtraTrees, c) Gradient Boosting, d) Bagging, e) XGBoost, f) Decision Trees.

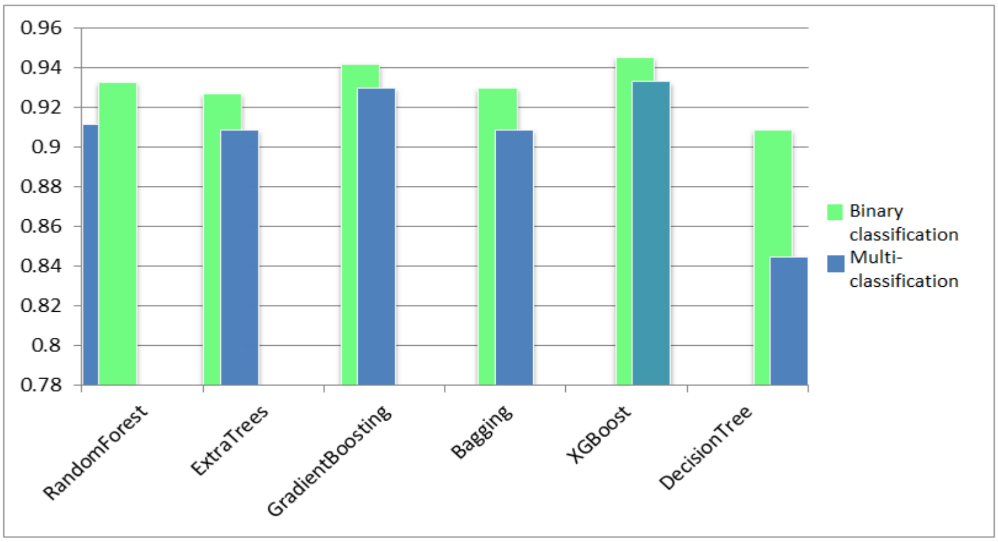


Figure 8: Comparison of the performance of machine learning models in multi-classification the y-axis represents a measure of the accuracy rate of the proposed models.

Table 7: Comparison between the results of the proposed method and the results of related works.

No.	Authors	Year	Imaging Modality	Methods	Dataset	Result
1	Yousfi et al.	2020	X-RAY	GLCM method. ,RLM method. BSIF method. Genetic Algorithms for feature selection	174 images.	ACC = 87.50%
2	Lee et al.	2020	X-RAY	Vgg-16, Random forest	334 Images	AUC: 0.858
3	Bello et al.	2022	X-RAY	GoogLeNet, VGG-16	646 images	GoogLeNet ACC: 90% VGG-16 ACC: 87%
4	Yang,Tsai Shih	2022	X-RAY	VGG16, Late-Fusion	1650 images	VGG16 ACC: 82% Late-Fusion ACC: 77%
5	Nakamoto et al.	2022	X-RAY	Alexnet, GoogLeNet, and VGG-16	1500 images	spine osteoporosis ACC: 79.0% .neck osteoporosis ACC: 75.0%.
6	Abubakar et al.	2022	X-RAY	fine-tuning, VGG16	646 images	VGG16 ACC: 88%
7	Sollmann et al.	2022	X-RAY	DCNN	144 patients	AUC: 0.862
8	Kumar et al	2023	X-RAY	VGG16,VGG19,InceptionV3,Xception, DenseNet169, DenseNet201, ResNet101, and ResNet152	372 images	86.36 ACC:
9	Naguib et al.	2024	X-RAY	superfluity and Transfer learning AlexNet and ResNet50	240 images	Superfluity DL model ACC: 85.42% for dataset1 and ACC: 79.39% for dataset2
10	Shen et al.	2024	X-RAY	VGG19	1400 images	ACC: 89%
11	In this research	2025	X-RAY	VGG-16,YOLOv11, ResNet-50 GLCM,HOG,ORB, harmony search, and using ML methods: Random Forest, Extra Trees,GradientBoosting, Bagging XGBoost, Decision Trees	1650 images	ACC: 94.5% for dataset1(binary classification) and ACC: 93.3% for dataset2(multi-classification)

5 CONCLUSIONS

The implementation of the proposed system and the results obtained demonstrated the feasibility of designing an intelligent system based on machine learning models to accurately classify knee X-ray images, this paper demonstrates that using the HMMFF feature extraction method, a combination of traditional feature extraction techniques and deep learning models, which exploit the diversity of patterns and structural information in images, and then combining them, along with feature selection using the harmony search algorithm, helped improve the system's efficiency in accurately distinguishing different knee conditions, this enhances diagnostic efficiency and contributes to improving the quality of healthcare, by combining traditional feature extraction techniques with deep learning models, a more comprehensive and accurate set of features was extracted, feature selection using the harmony search algorithm also helped improve classification accuracy, the proposed system provides an effective tool that supports clinicians in making more accurate and rapid decisions, contributing to improved patient outcomes and reduced treatment costs, the results obtained support the research hypothesis that combining traditional feature extraction techniques with deep learning models, along with feature selection applications, represents a promising approach that opens new horizons in the fields of computer vision and medical diagnosis, it could also contribute to the development of automated diagnosis in medicine, such as applications that require the analysis of X-ray image data or medical images in general.

It is also recommended that future studies include clinical validation using real clinical data, given its crucial importance in validating the system's results.

REFERENCES

[1] M. M. Sobh et al., “Secondary osteoporosis and metabolic bone diseases,” J. Clin. Med., vol. 11, no. 9, p. 2382, 2022, [Online]. Available: https://doi.org/10.3390/jcm11092382.

[2] A. Aibar-Almazán, A. Voltes-Martínez, Y. Castellote-Caballero, D. F. Afanador-Restrepo, M. del C. Carcelén-Fraile, and E. López-Ruiz, “Current status of the diagnosis and management of osteoporosis,” Int. J. Mol. Sci., vol. 23, no. 16, p. 9465, 2022, [Online]. Available: https://doi.org/10.3390/ijms23169465.

[3] Z. N. Al-Kateeb and D. B. Abdullah, “A Smart Architecture Leveraging Fog Computing Fusion and Ensemble Learning for Prediction of Gestational Diabetes,” Fusion: Practice & Applications, vol. 12, no. 2, 2023, [Online]. Available: https://doi.org/10.54216/fpa.120206.

[4] C.-H. Cheng, L.-R. Chen, and K.-H. Chen, “Osteoporosis due to hormone imbalance: an overview of the effects of estrogen deficiency and glucocorticoid overuse on bone turnover,” Int. J. Mol. Sci., vol. 23, no. 3, p. 1376, 2022, [Online]. Available: https://doi.org/10.3390/ijms23031376.

[5] N. Salari et al., “The global prevalence of osteoporosis in the world: a comprehensive systematic review and meta-analysis,” J. Orthop. Surg. Res., vol. 16, pp. 1-20, 2021, [Online]. Available: https://doi.org/10.1186/s13018-021-02772-0.

[6] L. Yousfi, L. Houam, A. Boukrouche, E. Lespessailles, F. Ros, and R. Jennane, “Texture analysis and genetic algorithms for osteoporosis diagnosis,” Int. J. Pattern Recognit. Artif. Intell., vol. 34, no. 05, p. 2057002, 2020, [Online]. Available: https://doi.org/10.1142/S0218001420570025.

[7] S. Lee, E. K. Choe, H. Y. Kang, J. W. Yoon, and H. S. Kim, “The exploration of feature extraction and machine learning for predicting bone density from simple spine X-ray images in a Korean population,” Skeletal Radiol., vol. 49, pp. 613-618, 2020, [Online]. Available: https://doi.org/10.1007/s00256-019-03342-6.

[8] U. B. Abubakar, M. M. Boukar, S. Adeshina, and S. Dane, “Transfer learning model training time comparison for osteoporosis classification on knee radiograph of RGB and grayscale images,” WSEAS Trans. Electron., vol. 13, pp. 45-51, 2022, [Online]. Available: https://doi.org/10.37394/232017.2022.13.7.

[9] T. S. Yang, “Recognition and classification of knee osteoporosis and osteoarthritis severity using deep learning techniques,” Dublin, National College of Ireland, 2022, [Online]. Available: https://doi.org/10.1302/3114-220769.

[10] T. Nakamoto, A. Taguchi, and N. Kakimoto, “Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network,” Dentomaxillofacial Radiol., vol. 51, no. 6, p. 20220135, 2022, [Online]. Available: https://doi.org/10.1259/dmfr.20220135.

[11] U. B. Abubakar, M. M. Boukar, and S. Adeshina, “Evaluation of parameter fine-tuning with transfer learning for osteoporosis classification in knee radiograph,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 8, 2022, [Online]. Available: https://doi.org/10.14569/IJACSA.2022.0130829.

[12] N. Sollmann et al., “Automated opportunistic osteoporosis screening in routine computed tomography of the spine: comparison with dedicated quantitative CT,” J. Bone Miner. Res., vol. 37, no. 7, pp. 1287-1296, 2020, [Online]. Available: https://doi.org/10.1002/jbmr.4575.

[13] S. Kumar, P. Goswami, and S. Batra, “Enriched diagnosis of osteoporosis using deep learning models,” Int. J. Performability Eng., vol. 19, no. 12, p. 824, 2023, [Online]. Available: https://doi.org/10.23940/IJPE.23.12.P7.824833.

[14] S. M. Naguib, M. K. Saleh, H. M. Hamza, K. M. Hosny, and M. A. Kassem, “A new superfluity deep learning model for detecting knee osteoporosis and osteopenia in X-ray images,” Sci. Rep., vol. 14, no. 1, p. 25434, 2024, [Online]. Available: https://doi.org/10.1038/s41598-024-75549-0.

[15] M. Shen, “Utilizing Deep Learning for Osteoporosis Diagnosis through Knee X-Ray Analysis,” in 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024), Atlantis Press, 2024, pp. 553-560, [Online]. Available: https://doi.org/10.2991/978-94-6463-512-6_58.

[16] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, “Deep learning for medical anomaly detection-a survey,” ACM Comput. Surv., vol. 54, no. 7, pp. 1-37, 2021, [Online]. Available: https://doi.org/10.1145/3464423.

[17] A. A. Torres-García, O. Mendoza-Montoya, M. Molinas, J. M. Antelis, L. A. Moctezuma, and T. Hernández-Del-Toro, “Pre-processing and feature extraction,” in Biosignal Processing and Classification Using Computational Learning and Intelligence, Elsevier, 2022, pp. 59-91, [Online]. Available: https://doi.org/10.1016/B978-0-12-820125-1.00014-2.

[18] L. K. Singh and K. Shrivastava, “An enhanced and efficient approach for feature selection for chronic human disease prediction: a breast cancer study,” Heliyon, vol. 10, no. 5, 2024, [Online]. Available: https://doi.org/10.1016/j.heliyon.2024.e26799.

[19] S. Aouat, I. Ait-Hammi, and I. Hamouchene, “A new approach for texture segmentation based on the Gray Level Co-occurrence Matrix,” Multimed. Tools Appl., vol. 80, no. 16, pp. 24027-24052, 2021, [Online]. Available: https://doi.org/10.1007/s11042-021-10634-4.

[20] I. Zine-dine, J. Riffi, K. El Fazazy, I. El Batteoui, M. A. Mahraz, and H. Tairi, “A review: Machine learning techniques of brain tumor classification and segmentation,” Mach. Graph. Vis., vol. 34, no. 3, pp. 31-55, 2025, [Online]. Available: https://doi.org/10.22630/mgv.2025.34.3.2.

[21] F. Siddiqui, S. Zafar, S. Khan, and N. Iftekhar, “Computer vision analysis of BRIEF and ORB feature detection algorithms,” in International Conference on Computing in Engineering & Technology, Springer, 2022, pp. 425-433, [Online]. Available: https://doi.org/10.1007/978-981-19-2719-5_40.

[22] D. Theng and K. K. Bhoyar, “Feature selection techniques for machine learning: a survey of more than two decades of research,” Knowl. Inf. Syst., vol. 66, no. 3, pp. 1575-1637, 2024, [Online]. Available: https://doi.org/10.1007/s10115-023-02010-5.

[23] D. Effrosynidis and A. Arampatzis, “An evaluation of feature selection methods for environmental data,” Ecol. Inform., vol. 61, p. 101224, 2021, [Online]. Available: https://doi.org/10.1016/j.ecoinf.2021.101224.

[24] C. Peraza, O. Castillo, P. Melin, J. R. Castro, J. H. Yoon, and Z. W. Geem, “A type-3 fuzzy parameter adjustment in harmony search for the parameterization of fuzzy controllers,” Int. J. Fuzzy Syst., vol. 25, no. 6, pp. 2281-2294, 2023, [Online]. Available: https://doi.org/10.1007/s40815-023-01499-w.

[25] O. Lopez-Rincon, O. Starostenko, and A. Lopez-Rincon, “Algorithmic music generation by harmony recombination with genetic algorithm,” J. Intell. Fuzzy Syst., vol. 42, no. 5, pp. 4411-4423, 2022, [Online]. Available: https://doi.org/10.3233/JIFS-219231.

[26] F. Makhmudov, D. Kilichev, and Y. Im Cho, “An application for solving minimization problems using the Harmony search algorithm,” SoftwareX, vol. 27, p. 101783, 2024, [Online]. Available: https://doi.org/10.1016/j.softx.2024.101783.

[27] A. E. Yıldırım, “Optimization of fuel cost in electric power systems using harmony search algorithm,” Int. J. Eng. Res. Dev., vol. 13, no. 2, pp. 531-544, 2021, [Online]. Available: https://doi.org/10.29137/umagd.814025.

[28] E. Uray, S. Carbas, Z. W. Geem, and S. Kim, “Parameters optimization of Taguchi method integrated hybrid harmony search algorithm for engineering design problems,” Mathematics, vol. 10, no. 3, p. 327, 2022, [Online]. Available: https://doi.org/10.3390/math10030327.

[29] J. Gholami, K. K. A. Ghany, and H. M. Zawbaa, “A novel global harmony search algorithm for solving numerical optimizations,” Soft Comput., vol. 25, no. 4, pp. 2837-2849, 2021, [Online]. Available: https://doi.org/10.1007/s00500-020-05341-5.

[30] R. Diallo, C. Edalo, and O. O. Awe, “Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score,” in Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA, Springer, 2024, pp. 283-312, [Online]. Available: https://doi.org/10.1007/978-3-031-72215-8_12.